DCU

A Perceptually Based Computational Framework
for the Interpretation of Spatial Language

by

John D. Kelleher

A dissertation submitted in partial fulfilment
of the requirements for the award of
Doctor of Philosophy (Ph.D.)

Supervised by
Prof. Josef van Genabith
Dr. Fintan Costello
Dr. Mark Humphrys

School of Computing
Dublin City University
September 2003

I hereby certify that this material, which I now submit for assessment on the programme of study leading to the award of Doctor of Philosophy, is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the text of my work.

Signed: *John Kelleher*
John Kelleher
97970948

Date: 16/9/03

"Surely a major source of widespread scepticism about 'machine understanding' of natural language is that such systems almost never avail themselves of anything like a visual workspace in which to parse or analyse the input If they did, the sense that they were actually understanding what they processed would be greatly heightened (whether or not it would still be, as some insist, an illusion) As it is, if a computer says, 'I see what you mean' in response to input, there is a strong temptation to dismiss the assertion as an obvious fraud " (Dennett 1991)

"A semantic theory having no contact with the world, a mere translation of one set of words into another, is a ladder without rungs " (Miller and Johnson-Laird 1976)

Acknowledgements

Despite what quantum mechanics says is possible, a dissertation does not appear out of thin air. Indeed, the writing of this dissertation has been a long and difficult process. It would not have been possible without the help of many people who I am delighted to have this opportunity to acknowledge.

My supervisors Josef van Genabith, Fintan Costello and Mark Humphrys have improved this work through their insight and ideas. Josef's inexhaustible patience, constant encouragement and friendship have contributed to this work on a daily basis and helped to develop my abilities as a researcher. For this, I owe him a special debt of gratitude.

Many other people at DCU also deserve to be recognised: Monica Ward, Barry McCaul, Noel O'Hara, Tom Soedring, Ger Hayes, Ray Walsh and Darragh O'Brien have, through their friendship and support, helped me to keep a grasp on reality during my time at DCU.

My good friends (in alphabetical order) Alan, Brian, Cathal, Conor, Gar, John, Lorraine, Neal, Neil, Niall, Orlaith, Shirley and Suzanne have in so many different ways and at so many different times helped me. I have many wonderful things to say about them, but this page is too small to contain it all.

The members of the St. Vincents Basketball Club also deserve to be mentioned. In particular, Joey Boylan and the Donnelly family. Joey's dedication as a coach not only helped me develop as a basketball player but also gave me the self-discipline to complete this work. The friendship that the Donnelly family have shown me, in particular Dave, has been a tremendous support to me over the years. Thanks for all the tea and hob-nobs!

I would also like to express my deepest affection and appreciation to my family. The love and support shown by my parents, John and Betty, my sisters, Liz and Marianne, my grandparents, Desmond and Bridie, Mary and my niece Kerry throughout my life have made me into the man I am today. Thank you all, I love you very much. Mum you can go to Mc Elhinneys for my graduation.

Finally, I want to thank Aphra for the immense patience, understanding, and love that she has shown me. I also want to acknowledge that her input to this work extends

beyond that of support I am indeed fortunate that my girlfriend is an accomplished and intelligent academic It was Aphra who first showed me what a Ph D entails Without her I would never have really started this work let alone have completed it Thank you, we made it, I love you very much and I can't wait for our holiday

Table of Contents

ii

Abstract

The goal of this work is to develop a semantic framework to underpin the development of natural language (NL) interfaces for 3 Dimensional (3-D) simulated environments The thesis of this work is that the computational interpretation of language in such environments should be based on a framework that integrates a model of visual perception with a model of discourse

When interacting with a 3-D environment, users have two main goals the first is to move around in the simulated environment and the second is to manipulate objects in the environment In order to interact with an object through language, users need to be able to refer to the object There are many different types of referring expressions including definite descriptions, pronominals, demonstratives, one-anaphora, other-expressions, and locative-expressions Some of these expressions are anaphoric (e g , pronominals, one-anaphora, other-expressions) In order to computationally interpret these, it is necessary to develop, and implement, a discourse model Interpreting locative expressions requires a semantic model for prepositions and a mechanism for selecting the user's intended frame of reference Finally, many of these expressions presuppose a visual context In order to interpret them this context must be modelled and utilised

This thesis develops a perceptually grounded discourse-based computational model of reference resolution capable of handling anaphoric and locative expressions There are three novel contributions in this framework a visual saliency algorithm, a semantic model for locative expressions containing projective prepositions, and a discourse model

The visual saliency algorithm grades the prominence of the objects in the user's view volume at each frame This algorithm is based on the assumption that objects which are larger and more central to the user's view are more prominent than objects which are smaller or on the periphery of their view The resulting saliency ratings for each frame are stored in a data structure linked to the NL system's context model This approach gives the system a visual memory that may be drawn upon in order to resolve references

The semantic model for locative expressions defines a computational algorithm for interpreting locatives that contain a projective preposition Specifically, the prepositions

in front of, behind, to the right of, and *to the left of* There are several novel components within this model First, there is a procedure for handling the issue of frame of reference selection Second, there is an algorithm for modelling the spatial templates of projective prepositions This algorithm integrates a topological model with visual perceptual cues This approach allows us to correctly define the regions described by projective preposition in the viewer-centred frame of reference, in situations that previous models (Yamada 1993, Gapp 1994a, Olivier *et al* 1994, Fuhr *et al* 1998) have found problematic Thirdly, the abstraction used to represent the candidate trajectors of a locative expression ensures that each candidate is ascribed the highest rating possible This approach guarantees that the candidate trajector that occupies the location with the highest applicability in the prepositions spatial template is selected as the locative's referent

The context model extends the work of Salmon-Alt and Romary (2001) by integrating the perceptual information created by the visual saliency algorithm with a model of discourse Moreover, the context model defines an interpretation process that provides an explicit account of how the visual and linguistic information sources are utilised when attributing a referent to a nominal expression It is important to note that the context model provides the set of candidate referents and candidate trajectors for the locative expression interpretation algorithm These are restricted to those objects that the user has seen

The thesis shows that visual salience provides a qualitative control in NL interpretation for 3-D simulated environments and captures interesting and significant effects such as graded judgments Moreover, it provides an account for how object occlusion impacts on the semantics of projective prepositions that are canonically aligned with the front-back axis in the viewer-centred frame of reference

List of Figures

List of Algorithms

List of Equations

List of Tables

List of Acronyms

3-D – 3 Dimensional

2-D – 2 Dimensional

ANOVA – Analysis of Variance

BRP – Bounding Right Parallelepiped (Gapp 1994a)

CSR-3D – Computation of Spatial Relations in 3D-Space (Gapp 1994a)

DRS – Discourse Representation Structure (Kamp and Reyle 1993)

DRT – Discourse Representation Theory (Kamp and Reyle 1993)

ERP/ERPs – Event-Related Potential(s)

FSM – Finite State Machines

HCI – Human Computer Interaction

LD/LDs – Linguistic Domain(s)

LDL – Linguistic Domains List

LM – Landmark

LTM – Long Term Memory

ModSAF – Modular, Semi-Automated Forces (Stent *et al* 1999)

NL – Natural Language

NLP – Natural Language Processing

NLVR – Natural Language Virtual Reality System

VPD/VPDs – Visual Perceived Domain(s)

VDL – Visual Domains List

SLI – Situated Language Interpreter

SPRINT – Spatial Representation INTerpreter (Yamada 1993)

STM – Short-Term Memory

STSS – Short-Term Sensory Storage

TR – Trajector

WIP – Words In Pictures (Olivier *et al* 1994, Olivier and Tsuji 1994)

Mathematical Notation

The dot operator ( ) is used to refer to an entity that is a subset or attribute of another entity For example, $x\ y$ denotes the entity y that is a subset or attribute of the entity x

The notation $x[]$ is used to denote that the entity x contains a group of other entities, i e , x is a list, an array, or a set Furthermore, the notation $x[y]$ denotes the element of $x[]$ that is indexed by y

Following McCawley (1993), the notation in Table 1 is used to denote logical and relational operations and set definitions

Symbol	Informal Description
$\exists$	there exists
$\forall$	for all
$\wedge$	and
$\vee$	or
$\sim$	not
$\wedge (p_1, p_2,\quad p_n)$	$p1, \wedge\ p2, \wedge\quad, \wedge\ p_n$
$\vee (p_1, p_2,\quad p_n)$	$p1, \vee\ p2, \vee\quad, \vee\ p_n$
$=$	assignment
$\neq$	not equals
$==$	equality
$\{\,fx\quad gx\,\}$	set having as members all items fx for which x meets the condition gx
$\cup$	union of two sets
$\cap$	intersection of two sets
$A - B$	the members of set A not in set B

$\in$	is an element of
$\notin$	is not an element of
$\vert A \vert$	the number of members of set A
$(\forall / \exists \ Fx)_x(Gx)$	All/There Exists Fs are G

Table 1: The logical, relational, and set theory symbols used in the definition of the algorithms in this thesis.

Typographic Conventions

Bold	the definition of a technical term that is used throughout the thesis (an index of these definitions is given at the end of the thesis)
Italics	word used as an example
Bold and Italics	mathematical variable
<u>Underline</u>	emphasis
*	a semantically malformed phrase

Table 2   The typographic conventions used in this thesis

1 Introduction

This thesis presents the design, implementation, and evaluation of a semantic framework to underpin the development of natural language (NL) interfaces, allowing a user to navigate through and interact with a rendered 3-D environment in real-time The framework integrates visual perceptual, linguistic, and conceptual information, and provides a unified model of reference resolution capable of handling both anaphoric and deictic references expressions

While Natural Language Processing (NLP) computer systems are reasonably adept at handling vocabulary, syntax, and grammar, they have difficulty with ambiguity, polysemy, vagueness, and deixis Humans often use perceptual cues to resolve these issues Indeed, psycholinguistic studies (Spivey-Knowlton *et al* 1998) have demonstrated the dependence of spatial language on a visual context

> "Given these results, approaches to language comprehension that assign a central role to encapsulating linguistic subsystems are unlikely to prove fruitful More promising are theories in which grammatical constraints are integrated into processing systems that coordinate linguistic and non-linguistic information as the linguistic input is processed " (Spivey-Knowlton *et al* 1998 pp 211-212)

Computational models interpret spatial language better if they utilise perceptual information from a visual context shared with the user While a visual information source is not feasible for all computer systems, there are a growing number of applications that incorporate a 3-D graphical element where a NL interface is advantageous with respect to cost, user comfort, and ease of use To test the feasibility of this approach a **natural language virtual reality**[1] (NLVR) system – the Situated Language Interpreter (SLI) – was developed, containing a rendered 3-D environment and NL user interface The SLI's

---

[1] A natural language virtual reality system is a computer system that allows a user to interact with a simulated environment through a NL interface

1

NL interface is grounded in a visual context by integrating visual salience into the discourse model

The central tenet of the framework is the grounding of spatial semantics in visual perception There are three main components within the framework a model of synthetic vision, a discourse model, and a semantic model for locative expressions containing the projective prepositions *in front of*, *behind*, *to the right of*, and *to the left of*

1.1 Synthetic Vision

A synthetic model of vision is a computational framework that attempts to capture the visual information within an **avatar's[2] view volume[3]** in a manner analogous to human vision Given a geometric description of an environment and the avatar's viewpoint, a synthetic vision system computes what the avatar sees

Renault *et al* (1990) describes a synthetic model of vision that was used as an aid to virtual character animation This model was later adapted as a navigation module to guide autonomous animated characters through changing virtual environments (Noser *et al* 1995, Kuffner and Latombe 1999) More recently, the model has been integrated as part of a goal-driven memory model, directing the gaze of autonomous virtual humans (Peters and O'Sullivan 2002)

Here, this model of vision is used as the main information channel between the simulated environment and the language-interpretation module The central idea is to model the knowledge of the environment the user has gained through their visual experiences in it and use this information as the basis for the interpretation process Not only is this a novel use of the synthetic visual model, it is also a novel approach to supplying knowledge of the environment to a language interpreter Previous systems, (SHRDLU (Winograd 1973), CITYTOUR (Andre *et al* 1986, Andre *et al* 1987), CSR-3-D (Gapp 1994a), SPRINT (Yamada 1993), WIP (Olivier *et al* 1994, Olivier and Tsuji

---

[2] The term avatar denotes the data structure that represents the embodiment of the user in a rendered environment

[3] The view volume defines the region of the simulated environment that is visible to the user

1994), Situated Artificial Communicator (Socher and Naeve 1996, Socher *et al* 1996, Vorwerg *et al* 1997, Fuhr *et al* 1998), Virtual Director (Mukerjee *et al* 2000), CommandTalk (Dowding *et al* 1999, Stent *et al* 1999, Goldwater *et al* 2000)) which have attempted to interpret language in a visual domain, have given their interpretive module complete access to all the objects in the environment This, however, is phenomenologically unrealistic, and it is impractical for large environments which contain many objects In these environments, it is likely that more than one world object fulfils the linguistic descriptions of a given referring expression If the interpretive process does not have a mechanism such as visual salience, that allows it to create a perceptual local context and rate the objects within this and new contexts, it will be unable to uniquely resolve these references Moreover, such a system will not be able to handle references that suppose a visual context, for example, a pronominal or other-anaphoric reference to an object which has been seen but not been previously mentioned in the linguistic dialogue

In this thesis, the synthetic vision model is extended to ascribe a saliency to each object it observes, based on its size and centrality in the user's view at the time it is observed The saliency model is based on the assumption that objects which are larger and closer to the centre of a user's view are more prominent than objects which are smaller or on the periphery of their view The resulting saliencies for each frame are stored in the system's NL **context model**[4] This approach gives the system a visual memory that may be used to resolve references as the perceived context evolves

1.2 Semantic Model of Projective Prepositions

Semantically modelling locative expressions is a complex task The main issues are how to resolve the landmark reference, frame of reference selection, the location of the frame of reference's origin, the dependency and scalability of the spatial template associated with a given preposition on the extension of the landmark, the gradation of

---

[4] A context model is a data structure that attempts to model how the context of a discourse changes as a dialogue evolves

3

applicability across a preposition's spatial template, and recognising and handling the occlusion of trajectors as well as rating and selecting trajectors Generally, the proposed models fall into two categories neat and scruffy (Mukerjee 1998) Neat models (Cooper 1968, Bennett 1975, Miller and Johnson-Laird 1976) propose definitions of spatial prepositions that parse space into discrete regions There are many problems with such strict definitions of semantics and not surprisingly many counter examples can be found for every proposed definition Herskovits's (1986) work attempts to extend these definitions by treating them as ideal meanings (prototypes of a category) from which use types can be derived based on sense shifts and tolerance functions However, even this loosening of definitions cannot save predicate-based models Several scruffy or continuum models have also been proposed (Yamada 1993, Gapp 1994a, Olivier et al 1994, Fuhr et al 1998, Mukerjee et al 2000) The advantage of these models is their ability to distinguish between different locations within a spatial template by assigning each point an applicability rating This simplifies the trajector rating and selection process However, some of these models only work in 2-D (Yamada 1993, Olivier et al 1994, Mukerjee et al 2000), one (Fuhr et al 1998) has problems distinguishing between the position of trajectors that are fully enclosed within a region, most rely on problematic bounding box representations of objects (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994, Fuhr et al 1998) and those that do not (Mukerjee et al 2000) are dependent on locating the local minimum within the continuum field of a preposition Furthermore, all of these models abstract to a purely topological analysis, ignoring perceptual information, in particular the issue of object occlusion Moreover, none of these models attempt to handle the issue of reference frame selection or propose mechanisms for handling anaphoric references

In this thesis, a novel semantic model of projective prepositions is developed The two main components are

    1   A computational approach based on linguistic and psycholinguistic work (Carlson-Radvansky and Irwin 1993, Carlson-Radvansky and Irwin 1994, Carlson-Radvansky 1996, Levelt 1996, Levinson 1996, Logan and Sadler 1996) that attempts to select the user's intended frame of reference This

procedure does not claim to represent the cognitive processes used by a human in selecting a frame of reference, but aims to model general preferences

2   A semantic model for the projective prepositions *in front of, behind, to the right of,* and *to the left of* that defines prepositions in terms of both perceptual and topological axioms The basis of this model is a parameterised continuum function that works in 3 dimensions One of the most important aspects of this model is the shifting of the reference frame's origin based on the user's view of the landmark This dynamic location of the reference frame's origin avoids many of the problematic situations with models that default to using the landmark's bounding box centre as the origin (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994, Fuhr *et al* 1998) Furthermore, following the theoretical work of Clark (1973), Vandeloise (1991), and Jackendoff and Landau (1992) object occlusion is integrated into the definition of prepositions along the viewer-centred front-back axis This approach defines the regions surrounding landmarks with complex geometries in a consistent manner Finally, the integration of perceptual information into the trajector selection process manages the issue of object occlusion

1.3   Integrating Linguistic Discourse and Visual Information

When developing a NL interface for a computational system that is to interpret language at anything but the shallowest level or to interact in a mode natural to a user it is impossible to consider a user's commands in isolation Often a user's commands can only be understood by considering them as part of an ongoing dialogue Consequently, the main issue in developing a natural language system is how to incrementally build up and use a model of the dialogue

The requirement for a discourse model is never more apparent than in the analysis of referring expressions People use referring expressions to introduce entities into the

5

discourse and then to re-mention them later There are many different types of referring expression, for example, definite descriptions, pronominals, demonstratives, one-anaphora, other-expressions, locative-expressions, etc

While several discourse models have been proposed – some of the best known being (Kamp and Reyle 1993, Grosz *et al* 1995) – they concentrate on how to update, manage, and extract information from a purely linguistic context model However, none of these models handle the semantics of locative expressions at anything but the most abstract level Since people often use locative expressions when navigating and interacting with spatial environments, these models are problematic from the perspective of this work Furthermore, the majority of discourse models (Kamp and Reyle 1993, Grosz *et al* 1995) neglect the impact of perceptual context on discourse Those that do, however, (Salmon-Alt and Romary 2001) give no description of how the perceptual domain is to be modelled or integrated with the linguistic information when resolving references

In this thesis, a discourse framework that adapts and extends the model proposed by Salmon-Alt and Romary (2001) is developed, which includes a novel method for integrating perceptual information into the context model and an explicit description of how this perceptual information is combined with the linguistic information to resolve references

1.4 Summary of the Thesis

The semantic computational framework developed in this thesis is based on an approach which grounds the semantics of spatial language in visual perception There are three major components within the framework a novel application and extension of a synthetic model of vision that uses a graphics technique called false colouring, a discourse model that adapts and extends the model proposed by Salmon-Alt and Romary (2001), an innovative algorithm for interpreting locative expressions The tripartite nature of this thesis is evident in the structure of its chapters, with each chapter devoting one section to each component in turn

Chapter 2 introduces the concepts, terminology, and problems related to each of the framework components Section 2 2 introduces the link between language and visual perception and highlights some of the problems for computational systems that model human visual perception Section 2 3 focuses on a particular type of referring expression – locative expressions – describing what locative expressions are and their importance in spatial language A general outline of the steps required to interpret locative expressions is given Each of the steps in the interpretive algorithm is examined in detail and the problems in computationally modelling this process are described Section 2 4 describes the computational problems in interpreting NL due to the contextual nature of language, introduces the concept of a discourse model, and concludes by noting that all the information required to compute a unique interpretation of an utterance at the time it occurs in the discourse is not always available from the linguistic context provided by the discourse Consequently, reference resolution is a canonical artificial intelligence problem, requiring the combination of information from multiple sources linguistic, conceptual, and perceptual

Chapter 3 introduces the linguistic approach adopted by this thesis Langacker's (1987, 1991b, 1994) cognitive grammar The motivation for adopting Langacker's linguistic model is its emphasis on situating language within wider general cognitive faculties

Chapter 4 continues a review of linguistic models Here, the focus is on previous models of discourse In particular DRT (Kamp and Reyle 1993), Centering Theory (Grosz *et al* 1995), and a reference resolution framework proposed by (Salmon-Alt and Romary 2001) are critically reviewed

Having reviewed the linguistic theory and frameworks relevant to this dissertation in Chapters 3 and 4, Chapter 5 critically reviews related computational research Section 5 2 reviews previous models of visual attention, Section 5 3 reviews previous work relevant to the interpretation of locative expressions, Section 5 4 reviews previous systems that have integrated language and vision

In Chapter 6, the Situated Language Interpreter (SLI) system is introduced providing a high-level overview of the system's architecture and an example user-system

interaction scenario that illustrates some of the system's functionality The SLI system implements the interpretive framework developed in this thesis

Chapters 7, 8, and 9 describe the SLI framework in detail The basic premise underlying the SLI framework is that computational systems that attempt to interpret NL input must model the user's perceptual context at the time of the utterance In Chapter 7, the SLI computational model of visual perception is developed, using the false colouring technique The SLI model of visual perception is a novel design which is suitable as an interface between a rendered environment and a linguistic interpretive module In order to adapt the false colouring model of synthetic vision to this design it was necessary to extend the model to rate the observed objects based on their saliency within the viewed scene

In Chapter 8, the SLI algorithm for the interpretation of projective locative expressions is described In particular, in Section 8 3, a novel algorithm based on psycholinguistic work (Carlson-Radvansky and Irwin 1993, Carlson-Radvansky and Irwin 1994, Carlson-Radvansky 1996, Taylor *et al* 2000), which attempts to predict a user's intended frame of reference, is developed In Section 8 4, a semantic model of projective prepositions that defines prepositions in terms of perceptual and topological axioms is developed One of the important aspects of this model is the dynamic location of the spatial template's origin in the viewer-centred frame of reference based on the user's location relative to the landmark Another key element is the integration of perceptual factors as parameters within the spatial templates of prepositions in the viewer-centred frame of reference By combining the SLI algorithm for locating the spatial template's origin with the SLI semantic model for projective prepositions, which integrates topological and perceptual factors, the SLI algorithm is able to define the regions surrounding landmarks with complex geometries in a consistent manner

In Chapter 9, the SLI discourse model is developed which adapts and extends Salmon-Alt and Romary's (2001) reference resolution framework by integrating the SLI model of visual perception and the SLI algorithm for interpretive projective locative expressions into the discourse model The novelty of the SLI discourse model rests on its integration of perceptual information into its context model and an explicit description of

8

how this perceptual information is combined with the linguistic information to resolve references

Chapter 10 describes a set of psycholinguistic experiments designed to examine different aspects of the SLI framework The results of these experiments indicate that

1   The assumption that an object's absolute size affects its visual salience, and consequently the probability of it being interpreted as the referent for an expression, is valid

2   The process of selecting a frame of reference impacts on the shape of the spatial template associated with the prepositions *in front of* and *behind*

3   There is a bias towards the use of the intrinsic frame of reference for the prepositions *in front of* and *behind*

4   The SLI reference resolution algorithms, which integrate both perceptual and linguistic information, are cognitively plausible

Finally, this thesis draws upon several disciplines artificial intelligence, linguistics (including computational linguistics, semantics, psycholinguistics), graphics, cognitive psychology, and software engineering As such, depth is sacrificed in order to gain breadth, which may fail to fulfil the expectations of researchers from any one of these fields Also, the framework developed in this thesis does not purport to be a model of cognitive or linguistic processes in the brain Rather it seeks to inform the development of interfaces to facilitate human computer interaction (HCI)

2 Modelling Visual Salience, Discourse, and Locative Expressions: Theory and Problems.

2.1 Introduction

This chapter has three goals the first is to make the reader familiar with the concepts, terminology, and approaches used in the different disciplines that this work draws upon, the second is to highlight some of the problems relating to the development of an NL interface for spatial language, and the third is to give some initial indication of how these problems are tackled in the framework developed here In Chapter 1 it was noted that there are three major components in the framework a model of synthetic vision, a semantic model for locative expressions containing the projective prepositions *in front of*, *behind*, *to the right of*, and *to the left of*, and a discourse model Accordingly, the main body of this chapter describes the terminology, background, and issues that are relevant to each of these components Section 2 2 deals with the field of synthetic vision, Section 2 3 focuses on locative expressions, and Section 2 4 describes discourse models

The keynote of Section 2 2 is that modelling human visual perception is difficult because there is a myriad of conflicting factors that impact on this cognitive process This section begins by describing Herb Clark's (1973) analysis of the correlation hypothesis Clark's work illustrates the connection between language and perception and by so doing provides a theoretical justification for basing a model of language on a model of perception Next, the importance of visual attention as a selective process in human perception and the difficulties in modelling the complexity of this process are examined This introduction to visual attention finishes by concluding that by abstracting visual attention to its most general and basic determiner, location in the scene, the complexity of the model is reduced and the genericness of the model is increased

Section 2 3 describes the problems affecting computational systems that attempt to interpret locative expressions

How to computationally select a user's intended frame of reference?

How to model the semantics of a preposition?

How to represent and adjudicate between the candidate referents of the expression?

Section 2 4 introduces one of the main problems in developing a computational natural language system how to model the contextual nature of language? Furthermore, it highlights the need to extend the purview of computational discourse models to include information from the visual context as well as the linguistic context

2.2 Perception and Language

The information required to compute a unique interpretation of an utterance is not always available at the time the utterance occurs in the discourse Indeed, people often draw on perceptual information in order to understand language One of the primary perceptual sources of information is the visual context of the discourse Following this, it is natural for an NLVR system to draw on the visual context when interpreting user input However, the modelling of visual perception and the integration of this with a linguistic framework has historically proven to be problematic

2 2 1 The Correlation Hypothesis

This section reviews Clark's (1973) analysis of the correlation hypothesis There are two factors motivating this review (1) it will illustrate the connection between language and perception and by doing so support the premise that in order to interpret language, the perceptual context of the utterance must be modelled, (2) several of the experiential concepts highlighted in this section impact on the issue of frames of reference which are discussed later in this chapter

11

In his influential paper "Space, Time, Semantics, and the Child" (1973), Clark's basic premise is the **correlation hypothesis**, which is that there is a close correlation between the structure of the human perceptual domain (**P-space**) and the semantic structures of human spatial terms. The hypothesis is based on the premise that, as children have knowledge of space and time before they learn the terms for space and time, the acquisition of these expressions is achieved by applying these expressions to their prior knowledge; i.e., the linguistic spatial domain (**L-space**) is based on the P-space. Accordingly, any model of the L-space must be cognisant of the pertinent characteristics of P-space.

2.2.1.1 P-space Properties

Clark notes that humans are inhabitants of an environment containing objects, people, space, and time and that one's perception of these entities and their interrelations is affected by one's biological makeup. "Clearly man's physical and biological environment itself places a large number of constraints – a priori constraints – on how he can describe the location of objects" (Clark 1973 pg. 30). Indeed, the nature of the physical environment requires that the description of the location of an object in space must always be relative to other positions in that space, or points of reference. In a 3-D space an object's location is optimally specified by directed distances away from three reference planes[5]. There are two invariant aspects in man's P-space: gravity and the terrestrial plane. They define two natural planes of references: verticality and ground level. For these reference planes to be optimally used in specifying a location, they require a positive and negative directionality to be imposed on them. This directionality can be accounted for in a natural manner through the asymmetries inherent in the P-space.

Man's primary perceptual apparatus (eyes, ears, nose, mouth, etc.) are most sensitive to stimulation from the front of the body and least sensitive to stimulation from behind the body. This defines a front-back plane of perceptual sensitivity. This perceptual

---

[5] The advantages of such an approach can be seen in geometry, where the Euclidean space is used.

configuration also defines a second plane at the base of the feet "Objects above ground level are characteristically visible, audible, smellable, tasteable, and touchable, whereas objects below ground level are not" (Clark 1973 pg 33) These perceptual asymmetries suggest a positive (i e more sensitive) and negative (i e less sensitive) directionality for the two environmentally defined reference planes Aligning the positive directionality of the ground level reference plane with the perceptually sensitive regions around the body results in the forward direction being labelled positive and the backward direction being labelled negative Applying this perceptual sensitivity criterion to the vertical reference plane aligns upward with the positive directionality and downward with the negative directionality

At this point, only two reference planes have been accounted for The third reference plane is based on the bilateral symmetry of the human body, which defines a vertical plane defining left and right However, the symmetry inherent in the definition of this plane makes it inappropriate to apply positive or negative values

Another characteristic of P-space is the human bipedal stance In general, humans are normally upright, this is the optimal perceptual position – Clark calls it the **canonical position** Extending this notion of canonical position to the human social environment, Clark defines the **canonical encounter**, which describes "the characteristics of the most usual interaction between two people" (Clark 1973 pg 34) In a one-to-one conversation, people usually face each other a short distance apart In Figure 2-1, the labelling of the horizontal axis around B are from A's perspective during a canonical encounter If A and B were two people in a canonical encounter, the axis defined around B from A's perspective would be B's canonical front and back axis, which are the opposite of A's, and A's left and right axis

Left

Back

Front

Right

A

B
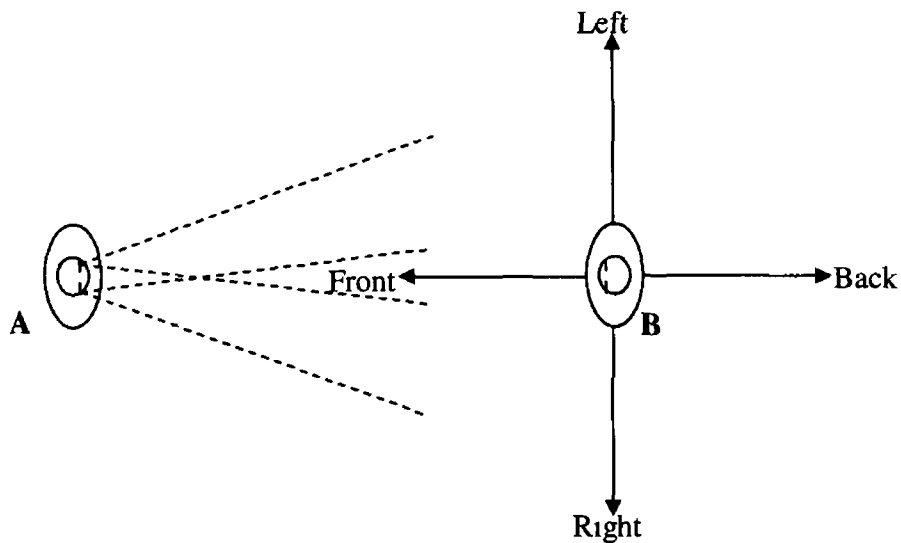
Figure 2-1: Bird's Eye view of a canonical encounter between two people

While the transposing of the front-back axis from the observer's point of view in a canonical encounter can be explained through the notion of canonical position, the failure to reverse the left–right axis is less clear Clark has suggested that this failure to reverse can be attributed to the symmetry across the left-right plane

"The reason for this failure to reverse is not clear Perhaps it is because the *left* and *right* directions in space are symmetrical, so the terms are difficult to apply to objects in a canonical encounter We have no trouble with the asymmetrical pairs *top-bottom* and *front-back* in this situation, because their criteria for application are intrinsic to the asymmetries in the situation But *left* and *right*, even in their normal use, are applied under fairly arbitrary criteria, the reversal of this application in a canonical encounter would seem unnecessarily complex " (Clark 1973 pp 46-47)

14

This notion of canonical encounter has an enormous impact on spatial language in the area of frames of reference In summary, human P-space

- constrains the description of the location of an object in space by requiring that it be relative to other positions in that space Usually this means with respect to other objects in the space

- contains three natural reference planes with associated directionality based on perceptual asymmetries By aligning heightened perceptual sensitivity with a positive direction the ground level reference plane can be described with upward as positive, the vertical left-to-right reference plane with forward as positive, the vertical front-to-back plane with both left and right as positive

- introduces the notions of canonical position and canonical encounter

2 2 1 2 L-Space Properties

There are two general properties of the L-space that are required by English spatial terms point of reference and direction The point of reference concept follows exactly from the definition of point of reference discussed in Section 2 2 1 1 above For locative expressions the object of the preposition serves as the point of reference Indeed, this holds for all English prepositions In example (1), *the house* is the object of the preposition and serves as the point of reference used in specifying the location of the tree

(1) *The tree to the right of the house*

This requirement for point of reference also applies to spatial adjectives Spatial adjectives have two points of reference One is a zero point, or point of reference, from which measurements are taken – Clark (1973) calls this the primary point of reference The secondary point of reference is the implied standard measurement For example, consider the adjectives *high*

15

"*High* has two implicit reference points ground level (the primary one) and some standard height (the secondary one) *The balloon is high* may therefore be paraphrased as 'The balloon is above some standard height from the ground level' " (Clark 1973 pg 37)

The notion of direction in L-space is apparent in both the positivity/negativity implicit in the scales associated with adjectives and prepositions, and in the use of English relational prepositions that locate an object by specifying its direction from a point of reference

Clark (1973) uses the linguistic notion of **markedness** in his analysis of spatial terms Markedness is a structural linguistic concept that ranks the complexity of a linguistic term relative to its morphological or formally related complementary "In its most general sense, this distinction refers to the presence versus the absence of a particular linguistic feature" (Crystal 1985 pg 188) For example, *countess* would be said to be marked with respect to *count* since *countess* contains the extra suffix *-ess* (Clark 1973) The range of structural indications of markedness are broad but in all cases the more complex term is said to be marked with respect to the less complex term In general, the marked member of a complementary pair is restricted in the range of contexts in which it occurs relative to its unmarked partner Taking the pair of spatial adjectives *tall/short* as an example, Clark (1973) illustrates the restriction of the marked case by examining the sentences

(2a) *How tall is Harry?*

(2b) *How short is Harry?*

Question (2a) is a neutral question about Harry's height, question (2b), however, has an additional presupposition that Harry is short This additional presupposition increases the number of conditions that must be met before the question can be used felicitously and, therefore, *short* is said to be marked with respect to *tall*

The unmarked member of a complementary adjectival pair is used as the basis for the scale name associated with the pair One can always define a positive direction along

16

the scale extending infinitely away from the primary point of reference in the direction associated with the unmarked member The marked member usually defines what has been referred to as a defective scale extending only from the secondary reference point in a negative direction towards the primary reference point

Apart from these general properties, "the use of spatial terms in English can be divided generally into two categories those that demand reference to the ego as either a primary or secondary point of reference, and those that do not" (Clark 1973 pg 37) The introduction of the **ego**, Clark's term for speaker, into the specification of English spatial terms complicates their structure, consequently Clark first examines the simpler non-egocentric form before analysing the egocentric L-space

Clark conjectures[6] that English spatial adjective pairs can be grouped into three categories based on the presupposed dimensionality of the objects they describe one-dimensional, two-dimensional, or three-dimensional In additions, "they can also be classified as to whether they specify the extent of an object or the position of an object" (Clark 1973 pg 38)

Taking as a specimen the adjectival pair *high/low*, it is evident that their primary point of reference is a plane, usually the ground level, unless some other reference plane is specified In Clark's analysis, to say something is high or low is really to say something is high or low of the ground High/low can be categorised as positional adjectives This suggests that English L-space contains a "(1) ground level plane of reference, and (2) verticality, the direction perpendicular to ground level, as a reference direction" (Clark 1973 pg 39)

Frequently, the scale defined by the adjectival pair *deep/shallow*, takes the same reference plane as *high/low* as its primary point of reference, i e , ground level However, the implied direction of the depth scale is downwards away from the ground level

---

[6] Clark's (Clark 1973) categorisation of English spatial adjectival pairs is based on the assumption of a schematization process (see Section 2 3 4 1 2) underlying spatial language which results in objects being abstracted to simple geometric forms While this approach has been adopted by other researchers (Talmy 1983, Herskovits 1986, Fillmore 1997, Herskovits 1998), it has also been criticised In particular Vandeloise has argued that the role of dimensionality in language is "indirect and secondary" (1991 pg 6), see Section 2 3 4 1 3

Clark's comparison of the *deep/shallow* pair against the *high/low* pair reveals that depth is marked with respect to height and following the above discussion on the defectiveness of a marked scale concludes "that distance up from ground level is positive, and distance down is negative" (Clark 1973 pg 39)

According to Clark, in English *at, on,* and *in* constitute the set of fundamental prepositions All three assert the location of an object X at some point of reference Y, e g , X is *at/on/in* Y While their meanings can overlap, the dimensionality they ascribe to their landmark, the term used to describe the object of a locative expression, is quite different *At* assigns no particular dimensionality to its landmark, while *on* imputes it as being a line or a plane (i e , two-dimensional), and *in* requires its referent to be a bounded two-dimensional or three dimensional space (Clark 1973, Fillmore 1997, pp 28-29), see Section 2 3 4 1 2 and contrast with Section 2 3 4 1 3

Relational prepositions, e g , *front, back, left, right,* etc, also indicate location "but they do so by specifying a direction from a point of reference in which the object is located" (Clark 1973 pg 42) *Above-below, over-under, on top of-beneath* all require a vertical direction, "but this vertical could be defined (1) by direct reference to gravitational vertical or (2) with reference to the top and bottom sides of the landmark, which are in turn defined (canonically) with respect to gravitational vertical" (Clark 1973 pg 42) Although more complex, the second definition is preferable as it allows a non-canonical definition of verticality not coincident with gravity and accounts for explicit references to the top and bottom sides of an object Furthermore, this definition dovetails with the use of front-back terms that also refer to intrinsic properties of referent objects

"To use these terms, one must define the front and back of the point of reference – say the front and back of a car – and then refer to the space adjacent to the front and back sides as *in front of* and *in back of*, respectively "

(Clark 1973 pg 42)

18

"For animate beings having a certain degree of complexity, the front is that portion of it which contains its main organs of perception and which arrives first whenever it moves in its most characteristic manner of movement"

(Fillmore 1997 pg 33)

In considering the definition of front and back, Clark concludes, "it is the front that is always defined in a positive way" (1973 pg 43), and this indicates that "L-space has a *front-back* dimension that coincides exactly in its asymmetry properties with P-space" (1973 pg 43)

The two candidate definitions for verticality given above highlight a possible ambiguity in the interpretation of relational prepositions This ambiguity arises because English recognises two kinds of verticality gravitational and intrinsic Some objects are considered to have intrinsic tops and bottoms that are not defined relative to gravity "Indeed, tops and bottoms in these cases appear to be defined relative to a canonical position, the upright position" (Clark 1973 pg 43) This is particularly apparent in the convention for describing the head-to-toe measurement of people in English Babies whose canonical position is horizontal are described as long, while adults who are usually encountered in an upright position are described as tall

With the introduction of the ego into L-space the canonical encounter and other P-space properties also appear in L-space Once the ego has been introduced into the domain it may now serve as a point of reference Taking distance as an adjectival example, it is evident that the ego is the point of reference in the unmarked case, see (3 a, b, c)

(3a) *It is far to San Francisco*

(3b) *San Francisco is far away*

(3c) *San Francisco is 30 miles away*

(Clark 1973 pg 44)

The introduction of the ego is also significant for relational prepositions, in particular to words referring to front and back Similar to the distance example, these prepositions use the ego as the unmarked point of reference (4)

(4) San Francisco is ahead

(Clark 1973 pg 45)

Example (4) means *San Francisco is ahead of me* (Clark 1973 pg 45) In the discussion of front and back above, Fillmore's definition for the intrinsic front of an object was given However, these terms are often applied to objects which do not have specifiable fronts and backs These usages are only explicable through the use of the ego as a point of reference and its function in a position in a canonical encounter If a speaker is looking at a ball and a tree, they may say *the ball is in front of the tree* By this, they mean that the ball is between them and the tree As a tree[7] has no intrinsic front or back it is evident that the speaker has anthropomorphized the tree as the other 'person in a canonical encounter and labelled the directions around the tree accordingly

A consequence of this is that there are two forms of front and back in English one is based on an intrinsic front dependent on an object's characteristics and the other is based on an egocentric front that is defined by the canonical encounter Indeed, it is through this process that the different frames of reference discussed in Section 2 3 3 arise

---

[7] Although trees are often cited exemplars as objects which do not have an intrinsic front, it should be noted that this is in fact a convention of English and other European languages Bowerman notes that "for speakers of the African language Chamus, they do¹ – the front of a tree is the side toward which it leans, or, if it does not lean, the side on which it has its longest branches" (1996 pg 400) Furthermore, Levinson (1996) notes that some Nilotic cultures make the assumption that a tree has a front, away from the way it leans

20

2 2 1 3   Correlation Hypothesis Summary

In summary, L-space has properties that are identical to P-space These are

- L-space shows the universal use of points, lines, and planes of reference in describing the location of an object

- L-space has three specific primary planes of reference Figure 2-2(a) ground level, with upward positive and downward negative, Figure 2-2(b) vertical left-right plane through the body, with forward positive and backward negative, Figure 2-2(c) vertical front-back plane of symmetry through the body, with right and left both equally positive

- L-space requires the use of canonical position to define uses of vertical expressions for dimensions that do not coincide with the gravitational vertical

- L-space requires the notion of canonical encounter to account for the speaker-centric uses of terms like front and back
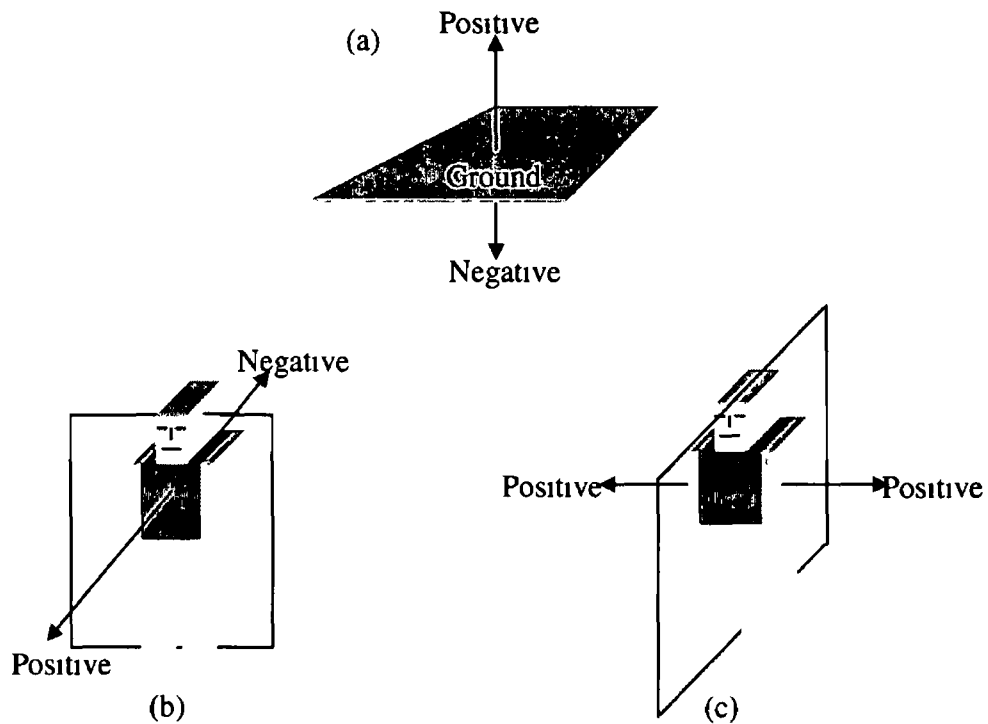
Figure 2-2 The three primary planes of reference in L-space (a) ground level, with upward positive and downward negative; (b) vertical left-right plane through the body, with forward positive and backward negative; (c) vertical front-back plane of symmetry through the body, with right and left both equally positive.

2 2 2 Perception and Attention

Although visual perception seems effortless, "psychophysical experiments show that the brain is severely limited in the amount of visual information it can process at any moment in time" (Reynolds 2001 pg 1) In effect, there is more information perceived than can be processed

The human faculty of attention is the "selective aspect of processing" (Kosslyn 1994 pg 84) Attention regulates the processing of perceived visual stimuli by selecting a region within the visual buffer for detailed processing Our knowledge of the human

attention process is not complete, "but it appears to consist of a set of mechanisms that exhibit different, sometimes opposing effects" (Hewett 2001 pg 9)

In computational systems, a saliency map is used to estimate the regions within an image that receive visual attention (Yee *et al* 2001) These saliency maps are built upon a model of user visual perception As noted above there are many different and sometimes competing factors that affect the location of the region a perceiver attends to For example, Landragin *et al* (2001) list visual familiarity, intentionality, an object's physical characteristics, and the structure of the scene This multiplicity makes the modelling of visual perception and consequently the creation of a saliency map extremely difficult

Several of these factors are so dependent on subjective considerations that they are impossible to generically model in a computational form Visual familiarity is dependent on a person's prior learning For example, if a footballer is walking through a park, a set of football goals in the distance might be more salient than the trees nearby The opposite might be the case for a botanist Intentionality is dependent on a viewer's task or goal If you invite colleagues into your offices, you will normally search the visual scene for chairs to offer them During this search, chairs are more salient than other pieces of furniture (Landragin *et al* 2001)

The dependency of an object's saliency on its physical characteristics, although less subjective, is no less difficult to model Gestalt theory (Ungerer and Schmid 1996, Landragin *et al* 2001) is one approach to this issue The term **gestalt** describes the concept of a perceived whole or unity This theory focuses on the perceptual grouping of stimuli into object or stable forms The fundamental claim of this theory is that visual perception is organised along a set of principles Moreover, the closer a configuration of elements adheres to these principles, the greater the tendency for them to be perceived as a unity called **Pragnanz** by gestaltists Configurations which exhibit a high Pragnanz are called good forms and are more salient The most important of these organisational principles are

- principle of proximity: elements with a small distance between them will be perceived as being related
- principle of similarity: elements which appear similar tend to be perceived as a common segment
- principle of closure: perceptual organisation tends towards closed figures
- principle of continuation. elements with few interruptions will be perceived as unities

(Ungerer and Schmid 1996)

There are, however, many difficulties with modelling the gestalt principles Firstly, each principle requires a different algorithm, this increases the complexity of the implementation Secondly, these principles are not always congruent, this may result in several different predictions for the organisation of the scene These situations are problematic because it is not known when it is better to use one gestalt principle instead of another and consequently there is often no way of adjudicating between conflicting principles (Landragin *et al* 2001)

The final set of saliency criteria reviewed and indeed the most promising from a computational perspective is the dependency of attention to a region on the structure of the scene "The strong points are classically the intersection of the horizontal and vertical lines at the 1/3-2/3 of the rectangular frame" (Landragin *et al* 2001 pg 2) However, these are not the only locations where scene structure directs attention, Landragin *et al* (2001) list several others As a result, attempting to computationally model structural salience suffers from many of the difficulties as gestalt-based approaches However, unlike the gestalt principles, the determinants of structural visual salience can be hierarchically described with some criteria being categorised as more fundamental or basic (Gapp 1995c, Landragin *et al* 2001)

Research has shown that "humans cannot attend to more that one region of space (i e , one set of contiguous locations) at a single time" (Kosslyn 1994 pg 90) Furthermore, although not invariant, "normally, the eye fixation and attentional locus are highly correlated" (Ericksen 1990 pg 3)

A priori, one of the major functions of visual attention is object identification With this in mind, an important factor when considering modelling visual attention is the difference between foveal and peripheral vision The fovea is a shallow pit in the retina which is located directly opposite the pupil, consisting of cones and is the site of highest visual acuity, the ability to recognise detail It "drops 50 percent when an object is located only 1° from the centre of the fovea and an additional 35 percent when it is 8° from the centre" (Forgus and Melamed 1976 pg 228) Identifying an object requires the use of foveal vision, occurring when a person looks directly at the object, causing the image of the object falling on the retina to be centred on the fovea The dependence of object identification on foveal vision implies a relationship between foveal vision and attention Moreover, this gradation across visual acuity is congruent with the gradation of attention theory This theory posits that "attention is greatest at a single point, and drops of gradually from that point" (Kosslyn 1994 pg 90)

Following this, the more central a location is with respect to the centre of an eye fixation the higher the location's salience Indeed, the most common computational mechanism for modelling visual attention is a filtering of visual data by removing portions of the input located outside a spatial focus of attention (Hewett 2001) By abstracting visual salience to this most general and basic factor, the input to the proposed language interpreting module discourse model is restricted, the complexity of the model is reduced, and the genericness of the model is increased

2.3 Locative Expressions

This section introduces the concept of a locative expression and describes a general outline of the steps required to interpret locative expressions with a detailed description of the background and issues associated with each of these stages

2 3.1 What are Locative Expressions?

The term **locative expression** is used to describe "an expression involving a locative prepositional phrase together with whatever the phrase modifies (noun, clause, etc)" (Herskovits 1986 pg 7) In the simplest form of locative expression, a prepositional phrase has an adjectival role modifying a noun phrase and locates an object Example (5) shows a simple locative expression

(5) *The book* [subject] *on the table* [object]

Following Herskovits' terminology, *The book* is the subject of the preposition in example (5) and *the table* is the object of the preposition in example (5) There is a wealth of terms used in the literature analysing simple locative expressions The terms local object, figure object, trajector, or target are used to describe the subject of a locative expression while the terms reference object, ground, landmark, or relatum are used to describe the object of a locative expression This work adopts the terminology of Landmark (LM) and Trajector (TR) (Langacker 1987)

The English linguistic conception of space is basically relativistic (Miller and Johnson-Laird 1976) the location of the trajector is specified relative to the landmark whose location is usually assumed by the speaker to be known by the hearer Examples illustrating the landmark-trajector distinction were presented by Jackendoff and Landau (1992)

(6a) *The book* [trajector] *is lying on the table* [landmark]

(6b) *The train* [trajector] *reached the station* [landmark]

(6c) *The star* [trajector] *is inside the circle* [landmark]

Figure 2-3 illustrates the syntactic structure of simple locative expressions and the position of the subject and object noun phrases
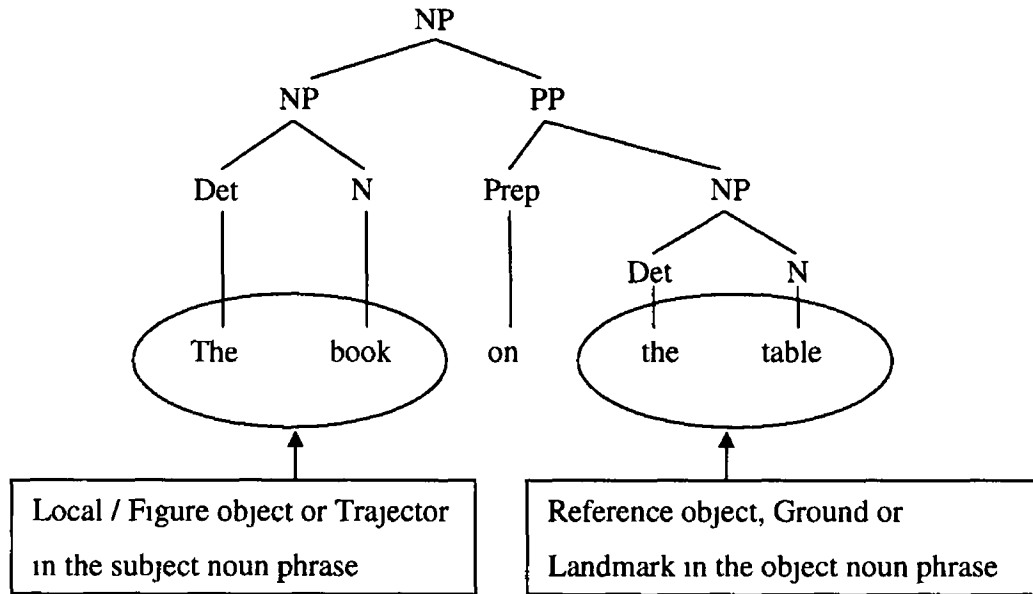
Figure 2-3· A syntax tree for a simple locative expression

Understanding a spatial locative involves coordination between a perceptual event and a linguistic utterance. The steps involved in this coordination include "identifying the components of the linguistic event, identifying the components of the perceptual event, and mapping the sets of components together within a mental representation of space such as a spatial mental model" (Carlson-Radvansky and Irwin 1994 pg 646) The process of coordinating the spatial relation between objects in the perceptual event and the linguistic spatial expression is referred to as **spatial term assignment.** Following (Carlson-Radvansky 1996), the four basic stages to this process are

1   Identify the landmark

2   Select a frame of reference and superimpose it on the landmark

3   Define the area of search for the trajector as defined by the spatial template associated with the preposition

4   Identify the primary trajector within the search area

The following sections describe the main issues involved in each of these stages

2.3 2   Identifying the Landmark

The syntactic structure of simple locatives aids the selection of the landmark by trivialising the extraction of the linguistic description of the landmark from the phrase Once the description has been extracted, it can be used in the selection of the landmark However, the content of the landmark's description can vary immensely, ranging from a definite description containing a noun and adjectives to a pronominal reference The inclusion criteria on the set of candidate landmarks vary with this content range, i e , the greater the amount of content in the description, the stricter the criteria Ambiguity can range across the whole spectrum of landmark description, even where a detailed definite description is given To resolve this ambiguity, the resolution process should integrate the linguistic description with both perceptual and previous linguistic information

When trying to define an algorithm that identifies the intended landmark of a spatial locative, it is important to note the asymmetry inherent in the linguistic parsing of space because this asymmetry defines the general characteristics associated with a landmark As described in Section 2 3 1, the conceptual mechanism underlying spatial locatives is to characterise the spatial location of the trajector by describing its location relative to the landmark Implicit in such characterisations is the assumption that the object functioning as the landmark is suitable for this role because its own spatial disposition is known This is to say that its prominence within a scene makes the extraction of its location from the visual context possible Talmy lists the characteristics which generally make an object easy to locate These are that the object is more

28

permanently located, larger, taken to have greater geometric complexity than other objects in the scene (Talmy 1983 pp 230-231) From this, it is evident that in general the more salient an object is, the more suitable it is to be used as a landmark Based on this observation, here the process of landmark selection is defined as extracting the most salient object from the visual context that matches the linguistic description, exactly what the general model of reference resolution proposed here attempts to achieve Given this, the process for interpreting a locative expression developed in this thesis treats this stage as a general case of reference resolution, with the issues attending to that process (see Section 2 4) being the main points of concern

2.3 3   Superimpose a Frame of Reference on the Landmark

In English, there are three different types of frames of reference absolute, intrinsic, and relative or viewer-centred (Rets-Schmidt 1988, Gapp 1995c, Hernandez and Mukerjee 1995, Carlson-Radvansky 1996, Levelt 1996, Levinson 1996, Taylor et al 2000) As a result, the appropriate (intended by the speaker) frame of reference must be selected before it may be superimposed on the landmark, this is a non-trivial task[8] This section begins with a description of what a frame of reference is and definitions for each type Next, an explanation of why the absolute and viewer-centred frames of reference are assumed to be collinear in this work is given This reduces the set of possible frames of reference to the intrinsic and viewer-centred Following this, the cognitive basis for these frames of reference is delineated and some of the issues specific to each frame of reference are described Having defined the different frames of reference, the focus shifts to how they interact, how a frame of reference may be explicitly marked within an utterance and the difficulties in selecting a frame of reference in the absence of a linguistic cue

---

[8] It should be noted that the issue of frames of reference is only applicable to locative expressions containing a projective preposition, e g , *in front of, behind* These prepositions have a canonical direction associated with them within a given frame of reference

2 3 3 1   *What are Frames of Reference ?*

The concept of frame of reference has a long history in the study of spatial cognition Levinson (1996) traces it back to Aristotle However, the modern linguistic interpretation of the phrase is that a **frame of reference** consists of six half-line axes with origin at the landmark These axes are sometimes referred to as the **base axes** (Herskovits 1986) In English, these axes are usually labelled *front, back, right, left, up,* and *down* Significantly, a frame of reference's base axes are not fixed in space, but may be rotated dependent on the perspective used Consequently, many frames of reference are possible While there is a consensus that English uses a tripartite system of frames of reference (absolute, intrinsic and viewer-centred), the distinctions between the different frames of referenceare not always clearly defined

Levinson (1996) describes the main points in this debate and points out that the confusion between researchers is "not merely terminological but results from the failure in the literature to distinguish coordinate systems from their origins or centres " The crux of this debate centres on the deictic versus intrinsic contrast, of which there has been at least three different interpretations

1   Speaker-centric versus non-speaker centric
2   Centred on any of the speech participants versus not so centred
3   Ternary versus binary spatial relations

(Levinson 1996)

Using an analysis of a set of sentences to illustrate the inconsistencies that arise through distinctions based on a reference frame's characteristic but variable origin (points (1) and (2) above), Levinson argues that frames of reference must be distinguished based on the cardinality of the relation they utilise and defined qua coordinate systems Moreover, he describes the term deictic frame of reference as a malapropism, preferring instead to use the term relative to describe the viewer-centred or ego-based frame of reference

30

To illustrate the cardinality based distinctions Levinson uses the spatial scenario of a man located in front of a house and describes this within the three different systems In all three cases, the trajector is the man and his position is described relative to the landmark – the house In Levinson's analysis, the intrinsic and absolute frames are binary requiring only two terms to locate an object the trajector and the landmark Intrinsically the man's position can be described as *the man is in front of the house*, meaning close to the house's intrinsic front Using an absolute frame of reference, the situation may be described as *the man is north of the house* The relative system adds the location of the viewer, making relative relations ternary If the viewer is located away from the house's intrinsic left they could describe the location of the man as *the man is to the left of the house* i e , the man is to the left of the house with respect to the speaker's left from their current location and orientation

Following Levinson, this thesis distinguishes between the frames of reference based on the cardinality of their relations However, the term viewer-centred is preferred over the classical linguistic term deictic or Levinson's term relative to describe the ego based coordinate system This preference is based on the desire to highlight the location this frame of reference's origin at the viewer

- **Absolute (extrinsic, environmental, world-based) frame** this is a binary reference frame that locates a trajector relative to a landmark The labelling of the landmark axes is dependent on salient environment features, e g , gravity, magnetic poles, etc

- **Intrinsic (object-centred, landmark-based) frame** involves binary relations that locate a trajector relative to a landmark The axes of the coordinate system are oriented around the landmark based on its canonical position

- **Viewer-centred (egocentric, relative, deictic) frame.** presupposes a viewpoint with ternary relations that locate an object relative to a landmark The axes of the landmark are oriented based on a canonical encounter between an observer and the landmark

31

Figure 2-4 and Figure 2-5 illustrate the intrinsic and viewer-centred frames of reference.
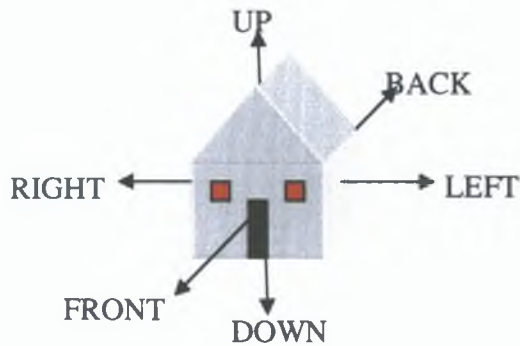
Figure 2-4: A house's intrinsic frame of reference.
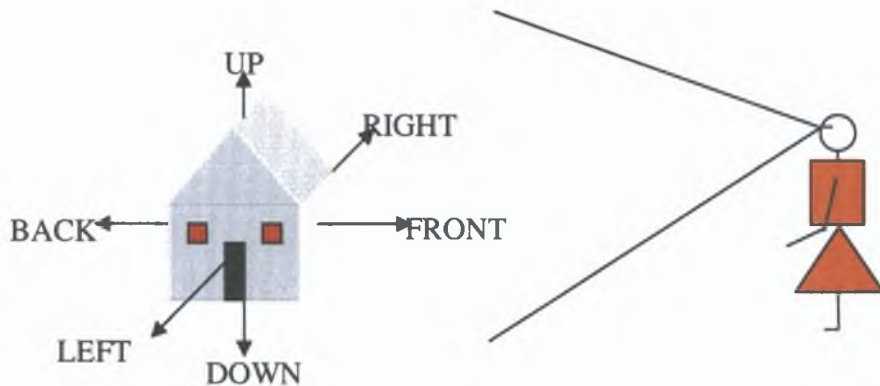
Figure 2-5: A viewer's viewer-centred frame of reference of a house.

When using an absolute frame of reference, the interlocutors must agree on absolute orientation, for instance on what is *north*. However, in a virtual environment these orientations are difficult to define except by stipulation.

Typically a computer user is in an upright position. In this stance, the viewer-centred and absolute frames of reference are aligned; i.e., they assign the same prototypical directions to the projective prepositions (Carlson-Radvansky and Irwin 1993). Figure 2-6, based on an illustration in (Levelt 1996 pg. 90), formally depicts

scenes used by Carlson-Radvansky and Irwin (1993) to investigate the appropriateness of saying *the ball is above the chair*. The position taken by the speaker is depicted in these examples. The appropriateness of the description within a reference frame is shown by a + (appropriate) or a − (not appropriate). The numbers below each scene show the percentage of subjects's *above* responses for each configuration.

	(a)		(b)		(c)		(d)	
Absolute	+		+		+		+	
Viewer-centred	+	(.95)	+	(.63)	-	(.93)	-	(.76)
Intrinsic	+		-		+		-	

	(e)		(f)		(g)		(h)	
Absolute	-		-		-		-	
Viewer-centred	+	(.28)	+	(.01)	-	(.30)	-	(.00)
Intrinsic	+		-		+		-	
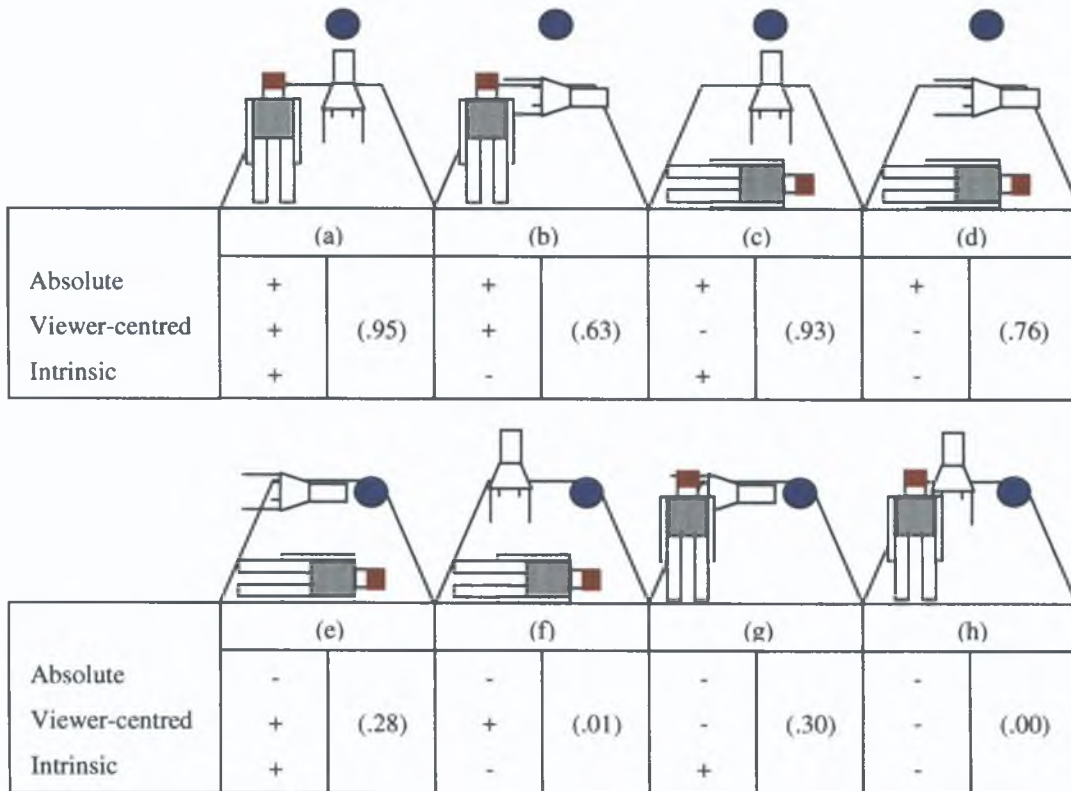
Figure 2-6: Formal representations based on an image from (Levelt 1996) of scenes used by (Carlson-Radvansky and Irwin 1993) to analyse "*the ball is above the chair*". The + and − signs indicate for each scene which perspective this description is appropriate for. The numbers below each scene show the percentage of *above* responses for each configuration.

The results of the viewer-centred and absolute frames of reference are not identical across the range of examples, in (c, d, e, and f) they differ. This variance between the perspective systems is restricted to situations where the observer is in a non-canonical position. In examples (a, b, g, and h), the observer/speaker is in a canonical/upright

position, the absolute and viewer-centred frames of references produce identical results This correlation between the viewer-centred and absolute frames of reference occurs in all psycholinguistic experiments where the subjects are in their canonical upright position For example, Carlson-Radvansky and Logan note that "in these experiments, the viewer-centred reference frame was always aligned with the environment-centred[9] reference frame (e g , subjects were upright)" (1997 pg 412) Furthermore, Miller and Johnson-Laird's (1976) semantically motivated analysis of spatial relations takes a similar approach concluding that the core of English spatial conception is a relativistic 3-D universe of locations within which an object can be located through two strategies "Ego's location and orientation can define the space deictically, or some other object can provide the point of origin, in which case its intrinsic parts orient the coordinates" (1976 pg 405) This view is echoed by Levinson who states that "by and large psychologists have considered notions of 'absolute' space irrelevant to theories of the naive spatial reasoning underlying language" (1996 pg 128) From Carlson-Radvansky and Irwin (1993) and supported by the work of (Miller and Johnson-Laird 1976, Levinson 1996, Carlson-Radvansky and Logan 1997) the thesis proposes that the scope of reference available to a user is not restricted by computational systems that treat the viewer-centred and absolute frames of reference as collinear Building on the concepts of canonical position and canonical encounter (c f Section 2 2 1 1), how the intrinsic and viewer-centred frames of reference arise is described and some of the issues specific to the intrinsic frame of reference are examined below

2 3 3 2   Intrinsic Frame of Reference

The correlation hypothesis posits that the structure of L-space is based on human perceptual experience or P-space The analysis of P-space (see Section 2 2 1 1) revealed three primary planes of reference The result of combining these planes of reference and their positive and negative directionality is similar to the classical Euclidean space axes (see Figure 2-7).

---

[9] Carlson-Radvansky and Logan use the term environmental-centred frame of reference to describe the frame of reference denoted in this thesis by the term absolute frame of reference
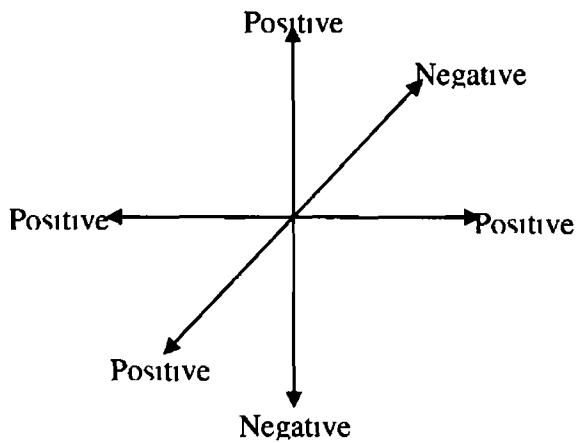
Figure 2-7. The resulting axes after combining the three primary planes of reference

Labelling the axes of the primary planes of references based on the canonical position of an object at the origin, in this case a humanoid figure called Jim, results in Figure 2-8 This is a representation of an intrinsic frame of reference
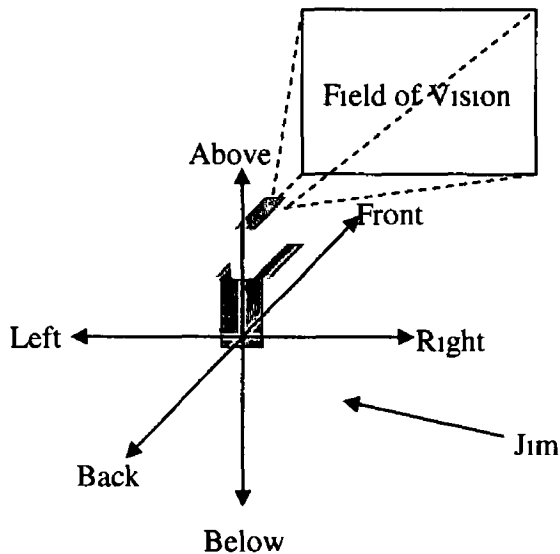
Figure 2-8: The labelling of the base axes based on Jim's experience of canonical position – illustrating the intrinsic frame of reference for a human.

35

The dependency of the intrinsic frame on the canonical position can be illustrated by rotating Jim and re-labelling the axis accordingly (as in Figure 2-9)
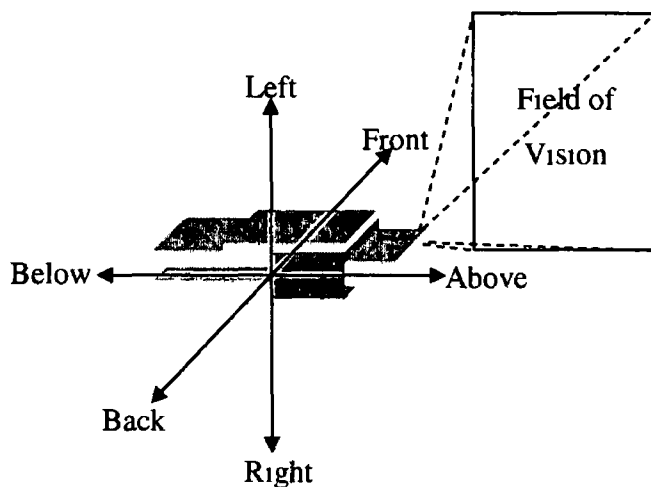
Figure 2-9: The labelling of the axes in an intrinsic frame of reference when Jim is not in his canonical position

A corollary of this dependency is that the intrinsic system requires that the interlocutors be aware of the landmark's orientation The utterance *the chair is to Jim's left* can only effectively localise the chair for the hearer if not only Jim's position, but also his orientation is known

2 3 3 2 1 Strategies for Defining an Object's Intrinsic Horizontal Axes

The felicity of speaker/hearer coordination in the intrinsic system is crucially dependent on a shared image of the landmark's location and orientation (Levelt 1996) An implicit sine qua non of this condition is that use of the intrinsic system is only possible if the landmark is oriented "The question of whether an object is considered to have an intrinsic top is relatively straightforward, it depends on whether it has a characteristic

orientation to the vertical" (Miller and Johnson-Laird 1976 pg. 400). However, defining an object's intrinsic horizontal axes is more difficult.

Both animate and inanimate objects can have intrinsic fronts. However, "frontness is an interpretative category, not a strictly visual one" (Levelt 1996 pg. 87). There is no visual feature that characterises the front of an entity. Fillmore lists two criteria for defining the front of an animate being: "the front is that portion of it which contains its main organs of perception and which arrives first whenever it moves in its most characteristic manner of movement" (1997 pg. 33). However, these two criteria are not invariantly aligned. Crabs are the exemplar for animals in which these criteria differ. Noting that these are described as moving sideways and not having heads on the sides of their bodies, Fillmore concludes, "the location of the main organs of perception outweighs the direction of movement criterion" (1997 pg. 33). Inanimate objects may also have intrinsic fronts and backs defined for them through visual analogy with living beings or functional use by living beings. Levelt (1996) gives the examples of a desk, a chair, and a church as inanimate objects with fronts. Figure 2-10 based on an illustration in Levelt (1996) shows these objects with their fronts labelled.
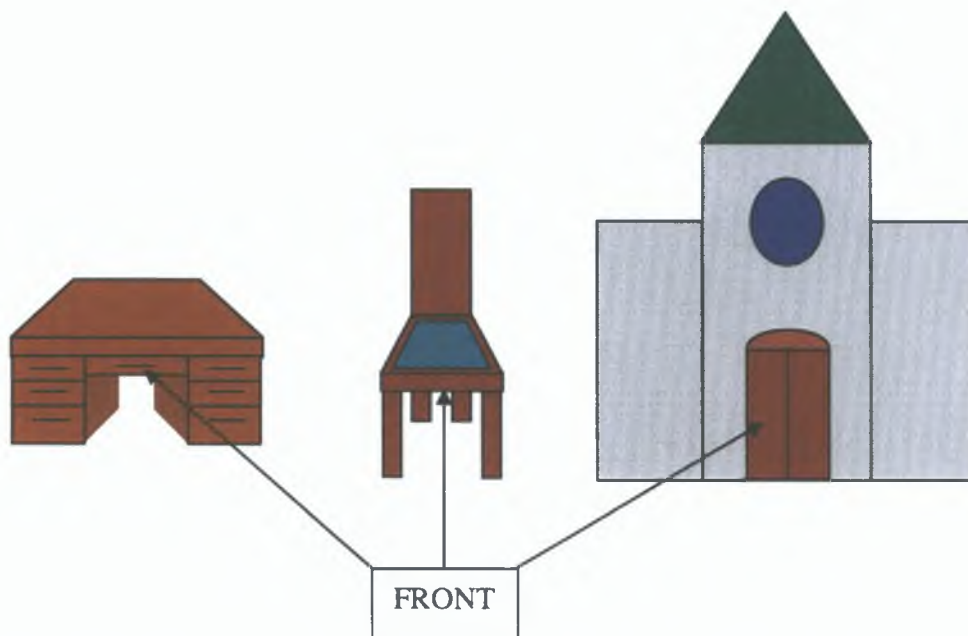
Figure 2-10: A desk, a chair, and a church. The definition of the front of these objects is dependent on their functional properties.

Fillmore (1997) describes four possible processes through which a front can be defined for an inanimate object

- Analogy If an object has some surface similarity to a front-back oriented animal, the portion of the object designated as its front is so designated on analogy with the model
- Motion Objects which have a fixed orientation when they are in motion have that part which arrives earlier designated as the front
- Function The part of an object that is oriented towards a user when they are using the object in its usual manner may be designated as the front
- Access The part of an object which a user typically, or symbolically, has access to, may be designated as its front

The last two processes above are based on a user's experience of an object Similar to the double criteria for defining the front of an animate being, these experiential criteria are not invariantly aligned and may lead to some uncertainties Churches are the exemplar for objects with functionally defined and access defined fronts that are different "One end of the church is thought of as its front on the inside, the opposite end on the outside" (Fillmore 1997 pg 33) The functionally defined front of a church is the end containing the altar, the access-defined front of a church is the end containing the door

Resolving references to the functionally and access defined fronts of churches is dependent on the location of the conversation and/or on the form of complex spatial preposition used If the conversation takes place outside the church, the access criteria for defining the front is dominant, this is evident in the labelling of the front of the church in Figure 2-10 as the end containing the door Fillmore's (1997) analysis of complex spatial prepositions highlights how the locative used can inform the resolution process If an object is outside the landmark along the front-back axis and close to its front it is described as being *in front of* the landmark If, on the other hand, the object is located inside the landmark the expression used to indicate it is close to the front extremity is *in*

38

the front of (Fillmore 1997) Following this analysis, if someone arranges to meet you _in front of the church_ they are referring to the outside of the church and therefore to the access defined front, the end containing the door If, on the other hand, someone arranges to meet you _in the front of the church_ they are referring to the inside of the church and therefore to the functionally defined front, the end containing the altar

"If an object has both an intrinsic top and bottom, and an intrinsic front and back, the remaining two sides are intrinsically left and right" (Miller and Johnson-Laird 1976 pg 401) However, the alignment of an object's left and right with respect to the front-back axis is not fixed, but is dependent on its characteristic use (Levelt 1996) Although the front-back axes of the chair and desk in Figure 2-10 are parallel, the right-left axes are reversed Miller and Johnson-Laird (1976) explain this phenomenon by distinguishing between two kinds of characteristic use inside and outside If the characteristic use of an object involves the user being inside the object _car, chair, clothing, etc_ the part of the object adjacent to their right hand will become the object's intrinsic right side through analogy with the body If during the characteristic use of an object a user is positioned outside the object, the part of the object adjacent to their right hand will become the object's intrinsic right side

2 3 3 2 2 Intrinsic Frame of Reference Summary

In summary, the experience of objects in their canonical position is at the core of the intrinsic frame of reference A priori, the use of the intrinsic frame of reference requires that the interlocutors have a shared conception of the location and orientation of the landmark Inherent in this is the requirement that the landmark be oriented Within this frame of reference, an object may have top/bottom axes defined for it without any of the horizontal axis aligned, by virtue of its canonical position relative to the vertical Defining the front-back axes is more problematic, however For animate beings, this axis is dependent on the location of the main organs of perception For inanimate objects, the front-back axes can be aligned through analogy with living beings, fixed orientation when in motion, functional use, or access considerations The axes defined through these

39

processes may not always be parallel, as in the church example above However, ambiguities arising between conflicting criteria can often be resolved through analysis of the spatial preposition used Aligning the right-left axis presupposes that the vertical and front-back axes have been defined, left and right being assigned to the remaining two sides For animate beings, they are usually defined through analogy with human beings For inanimate objects, the relationship between the left-right axes and the front back axes is not fixed, being dependent on the position taken by the user during the characteristic use of the object inside versus outside

2 3 3 3   Viewer-Centred Frame of Reference

The experience of canonical encounter impacts on the construction of frames of reference Figure 2-11 depicts a canonical encounter The axes around the encountered object are labelled from the observer's viewpoint This egocentric labelling of the axis by an observer is known as a viewer-centred frame of reference

The strategy for labelling the axis around the landmark in this figure is based on the rules of a canonical encounter, which follow what Clifford Hill (1982) described as a **mirror imagery strategy.** This involves the axes of the speaker being translated to the landmark and then the front back axes being rotated While this is the strategy employed by European languages, it is not universal Hill (1982) describes an in-tandem imagery strategy where the axes are only translated Speakers of the West African language Hausa among others use this strategy Using this strategy, the sentence *the lion in front of the tree* in Hausa describes a situation which an English speaker would characterise as *the lion behind the tree*
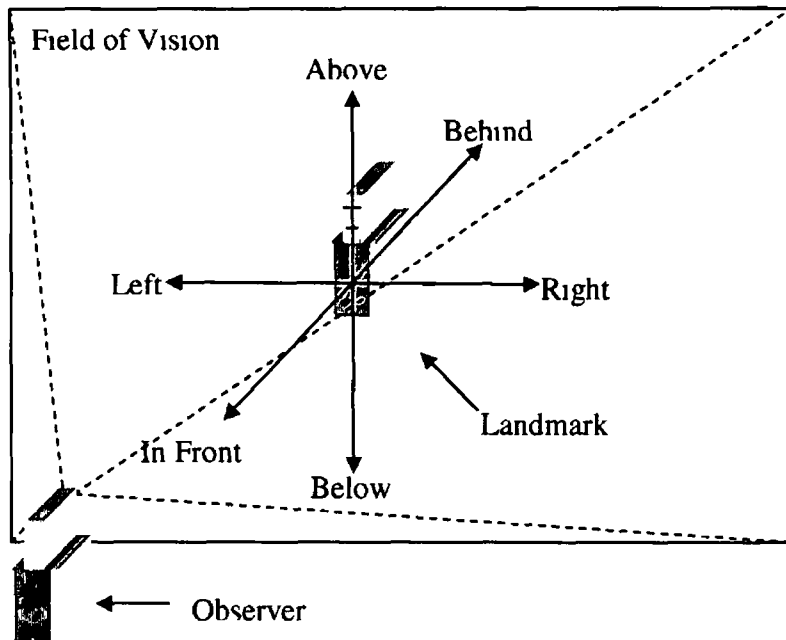
Figure 2-11: The labelling of the axis around an object based on a canonical encounter. The labelling of the axis is done from the observer's point of view – demonstrating a viewer-centred frame of reference

Comparing Figure 2-8 and Figure 2-11 highlights the transposition of the labels on the front-back axis when switching between intrinsic and viewer-centred frames of reference

2 3 3 4 Interaction between Frames of Reference

If the linguistic spatial locative contains a projective preposition, the vertical and horizontal base axes of the perceptual event must be oriented with respect to one of the above frames of reference so that the spatial terms can be assigned a direction (Miller and Johnson-Laird 1976, Carlson-Radvansky and Irwin 1993) However, many spatial terms are common between intrinsic and viewer-centred systems, along the horizontal plane, they both evince the same opposition pairs *left* versus *right* and *front* versus *back*, while

41

on the vertical axis *above* and *below* are also common to both systems Coupled with the fact that the frame of reference is usually implicit within the prepositional phrase, the task of coordinating between the speaker and the hearer's perspective is more difficult Miller and Johnson-Laird's (1976) analysis of the imperatives (7a) and (7b) highlights this difficulty

(7a) *Put it in front of the chair*

(7b) *Put it in front of the rock*

In (7a), the landmark, the chair, has an intrinsic front, so the imperative "is ordinarily understood to mean that 'it' should be put in a location determined by the orientation of the chair" (Miller and Johnson-Laird 1976 pg 396) In contrast, the landmark in (7b), the rock, does not have an intrinsic front Here, the coordinate system around the rock must be aligned using a canonical encounter strategy, i e , using the viewer-centred reference system Consequently, this imperative in English is understood to mean that it should be placed between the rock and the viewer[10] Note that this strategy may also be applied to (7a) However, when viewers are reclined or objects appear in non-canonical orientations, the reference frames are dissociated, and thus assign conflicting directions to spatial terms (Carlson-Radvansky 1996) In these situations, a misinterpretation based on frame of reference ambiguity may occur Levelt (1996) uses the term **coordination failure** to describe such misinterpretations In (7a), the object's intrinsic reference frame may be aligned differently to the viewer-centred frame of reference If this is the case, the different reference systems will assign conflicting directions to the spatial term *front* Figure 2-12 illustrates situations where conflicts between reference system alignments arise In (a) and (b), the chair's intrinsic front and the reader's viewer-centred front are aligned However, in (c), there is a conflict between the reference systems, the intrinsic front is aligned with the viewer-centred right, while

---

[10] In a real world environment the viewer-centred interpretation may be ambiguous as to whether the axes of the reference system should be aligned from the speaker's or the hearer's viewpoint However, in a rendered environment this ambiguity is avoided as the speaker (the user) and the hearer (the avatar) have a common viewpoint

the viewer-centred front is aligned with the intrinsic right. Furthermore, (b) illustrates a conflict between intrinsic and viewer-centred frames for the spatial term *above*. Here, the viewer-centred *above* is aligned with the intrinsic *left* and the intrinsic *above* is aligned with the viewer-centred *left*. The ability to interpret locatives whose landmark has an intrinsic frame of reference in a viewer-centred manner can cause ambiguity.
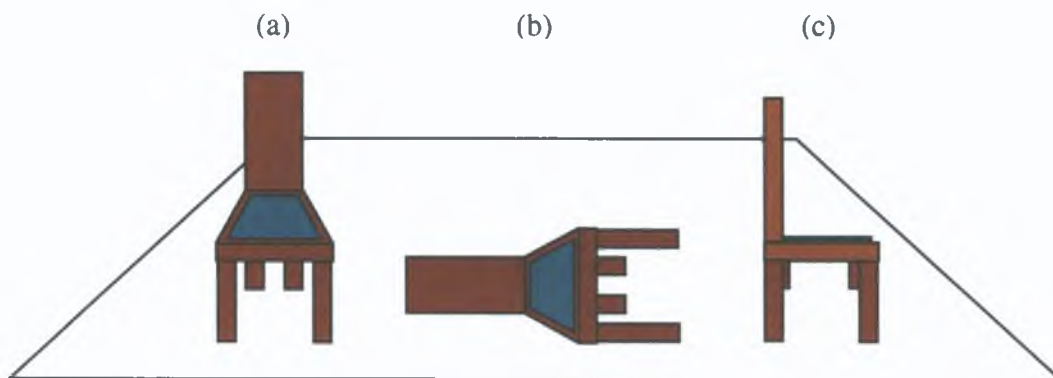
Figure 2-12: Three chairs in different positions – illustrating conflicts between frames of reference.

With such scope for ambiguity it is remarkable how infrequently coordination failures occur. Clearly, the interlocutors must in some way agree on the perspective system used in an utterance. In some instances, this may be achieved by the speaker using an explicit linguistic cue.

2.3.3.5 Linguistic Cues of Reference Frame Selection

In some cases, the speaker may explicitly mark the intended frame of reference; for example, if at a wedding someone described the groom with the phrase *The man to the bride's right*, it would be clear that they were describing man (a) in Figure 2-13. This is because the use of the genitive form of the landmark noun *bride's* indicates that its intrinsic reference frame should be used. In contrast, if the speaker described the groom

43

using *The man to the right of the bride* they could be describing either man (a) or man (b) in Figure 2-13.
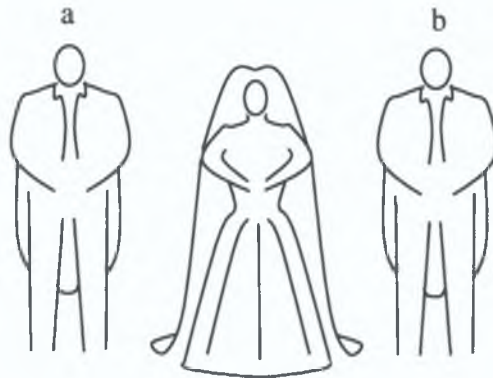
Figure 2-13: A bride and two men.

There are other linguistic cues. For example, the use of the determiner *the* in a noun phrase which describes a spatial region X, such as *the X*, implies that an intrinsic frame of reference is being used. The region that is *on top of X* could apply to any of the frames described; in contrast, the region that is *on the top of X* could only apply to intrinsic frames of reference (Landau and Munnich 1998). Another form of cue is the use of a phrase that explicitly references the point of observation being used or the observer such as *from where I stand* or *from where you stand*. For example, *The dog is to the right of the library from where I stand*. In this case, it is evident that the frame of reference being used is not the library's intrinsic frame, but the speaker's (see also Herskovits 1986). The phrases (8 a, b, c) are examples of explicit linguistic cues that indicate the intended frame of reference.

(8a)   *The chair is to Jim's left.*
         Frame of reference: intrinsic.
         Linguistic cue: genitive form of the landmark noun.

(8b)   *The chair to the left of Jim, from where you stand.*
         Frame of reference: viewer-centred.
         Linguistic cue: *from where you stand.*

44

(8c)   *The chair is to the left of Jim, from my perspective*
         Frame of reference viewer-centred
         Linguistic cue *from my perspective*

Explicit linguistic cues are exceptional in general, the intended frame of reference is tacit within the statement Apart from these cues, selecting the frame of reference intended by the speaker becomes a complex task involving many contextual factors

2 3 3 6   Computationally Selecting a Frame of Reference

In Section 2 3 3 4, the commonality of spatial terms across the frames of reference in English and how this can lead to coordination failures was noted Section 2 3 3 5 described how these may be avoided by the use of explicit linguistic cues, however, such linguistic cues are exceptional Given this, how is one to computationally select a user's intended frame of reference?

It should first be noted that few of these cases are problematic, there may, for example, be no intrinsic frame associated with the landmark, in which case a viewer-centred frame of reference is the only possibility In other cases, the landmark is in its canonical position and the frames of reference are aligned However, if the landmark is not in its canonical position, a process for selecting the frame of reference is required

A sensible way to start developing an algorithm to model this selection process would be to look for a default reference frame As has been noted above, research has pointed to individual languages favouring one perspective system over another (see Footnote **Error' Bookmark not defined**) For English, however, the experts do not agree on a dominant or default perspective Some researchers (Gapp 1995c) argue for a viewer-centred reference system based on the ease of cognitive computation the viewer-centred frame matches the speaker's body axes and is immediately available through perception, while other reference systems require cognitive mechanisms such as mental rotations to be computed However, little evidence has been found to support the view leading to theories championing the absolute or intrinsic frame (Taylor *et al* 2000)

Tversky (1996) notes this variance contrasting the view of Levelt against Miller and Johnson-Laird

"Still, it is a general finding that the dominant or default system for most speakers is deictic reference, either primary or secondary" (Levelt 1989, pg 52)

"But intrinsic interpretations usually dominate deictic ones, if a deictic interpretation is intended when an intrinsic interpretation is possible, the speaker will usually add explicitly 'from my point of view' or 'as I am looking at it'" (Miller and Johnson-Laird 1976, pg 398)

Tversky (1996) argues that for English there is no default perspective, instead different perspectives are adopted in different situations depending on a range of pragmatic considerations Herskovits (1986, see also Tversky 1996) lists some of the factors that can inform this selection "cohesion, topic, speaker's and addressee's mutual beliefs (in particular about the contents of the addressee's awareness), purpose of communication, perceptual salience, visibility of perceptual evaluation of alignment, of right angles etc " (1986 pp 172-173) Modelling such a wide range of factors is impossible However, in this thesis, a general heuristic algorithm whose results are aligned with user preferences in the majority of situations is all that is required

To date, there have to been many approaches adopted by systems confronted with this issue The first approach is to adopt a default frame of reference and force the user to adopt this for all input The second approach is to allow the user to switch between frames of reference if they use an explicit marker in the input, e g , *from where I stand* Neither of these approaches is satisfactory from a HCI perspective, as both force a user to learn how to communicate with the system This thesis proposes a procedure based on linguistic and psycholinguistic work (Carlson-Radvansky and Irwin 1993, Carlson-Radvansky and Irwin 1994, Carlson-Radvansky 1996, Levelt 1996, Levinson 1996, Logan and Sadler 1996) that attempts to select the intended frame of reference This procedure does not claim to represent the cognitive processes used by humans in selecting a frame of reference, but aims to model their general preferences Once a frame

46

of reference has been selected, it is imposed on the selected landmark This allows the canonical directions associated with each of the projective prepositions to be aligned relative to the landmark

2 3 4 Defining the Area Described by a Preposition

In the context of a NL interface to a 3-D-rendered environment, one of the main purposes of locative expressions is to narrow the domain of search for the trajector that the user is intending on The determining factor defining the area of search is the preposition used Although the class of English spatial prepositions is relatively small (approximately 80 elements not including composites (Landau 1996)), there are still far too many for an exhaustive analysis here Instead, this thesis will focus on a subset of static prepositions

Following (Herskovits 1998), a **static preposition** describes the location of a stationary trajector, while a motion preposition, such as *along, across, over*, describe the direction of the path of the trajector For example, *on* in (9a) is a static preposition that simply describes the location of the trajector *the man*, but does not constrain its path in a particular direction In contrast, the motion preposition[11] *across* (9b) describes both the location and the path of the trajector Comparing examples (9b) and (9c) highlights the effect of a motion preposition on constraining the path of the trajector In (9c), *across* has been replaced by another motion preposition *along* Figure 2-14(a) illustrates the directional freedom of the trajector complementing the preposition *on*, while Figure 2-14(b) and Figure 2-14(c) illustrate the directional constraints on the paths of the trajector imposed by motion prepositions *across* and *along*

(9a) *The man walked on the road*

---

[11] It should be noted that motion prepositions may be used to describe static as well as dynamic scenes For example in, *the tree lay across the road*, the trajector *the tree* is a stationary object However, even when describing a static scene, prepositions of this class still specify more than the location of the trajector In this example *across* specifies both the location of the trajector and the orientation of its pose relative to the landmark i e its dominant axis is orthogonal to the road rather than parallel to the road

47

(9b) *The man walked across the road*

(9c) *The man walked along the road*

(a) *on the road*

(b) *across the road*

(c) *along the road*
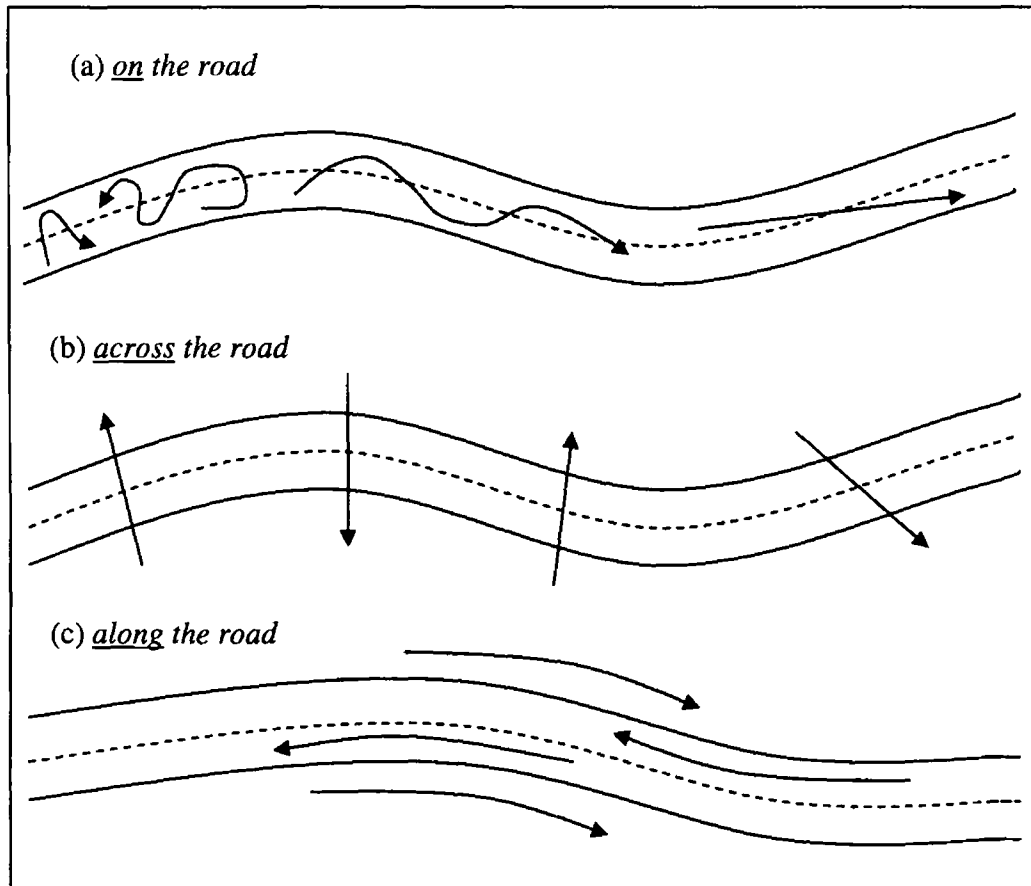
Figure 2-14 Illustration (a) depicts the freedom of direction for trajectors complementing static prepositions Illustrations (b) and (c) represent the directional constraints on the path of trajectors complementing motion prepositions

In general,[12] static spatial locatives can be grouped into two classes depending on the type of preposition used topological and projective **Topological prepositions** are the category of prepositions referring to a region that is proximal to the landmark, e g , *at, on, in,* etc **Projective preposition** describe a region in a particular direction, e g , *to the right*

---

[12] The preposition *between* takes an exceptional position among static spatial locatives because it refers to two landmarks (Gapp 1994a)

48

of, to the left of, in front of, behind, etc Moreover, the specification of the desired direction is dependent on the frame of reference being used (see Section 2 3 3) Section 2 3 4 1 examines the literature on topological prepositions Although the framework developed in this thesis does not attempt to model the semantics of topological prepositions, it is appropriate to review this literature because such a review allows us to

1   illustrate the impact of the distance between a landmark and trajector on a preposition's applicability (Section 2 3 4 1 1)

2   introduce the concept of a spatial template (Section 2 3 4 1 1)

3   introduce the process of schematisation in a full context, i e , the arguments for schematisation (Section 2 3 4 1 2) and those against it (Section 2 3 4 1 3)

4   explain why topological prepositions are not modelled in this thesis (Section 2 3 4 1 4)

Following the discussion of topological prepositions, the main factors impacting on the semantics of projective prepositions are introduced (Section 2 3 4 2) In Section 2 3 4 2 1, some of the psycholinguistic work (Gapp 1995a, Logan and Sadler 1996) that has examined the semantics of projective prepositions is described followed by a critical review of these experiments and their results In Section 2 3 4 2 2, the impact of the trajector's proximity to the landmark on the semantics of a projective preposition is highlighted and an explanation of why this factor was not found by Gapp's (1995a) experiments or Logan and Sadler's (Logan and Sadler 1996) experiments is given Section 2 3 4 2 3 criticises the methodology used in both sets of experiments (Gapp 1995a, Logan and Sadler 1996), because it excluded perceptual cues, such as object occlusion Furthermore, some of the linguistic theorists (Clark 1973, Vandeloise 1991, Jackendoff and Landau 1992) and psycholinguistic evidence (Gapp 1995a, Logan 1995) that argue for the inclusion of perceptual cues within the semantics of projective prepositions are noted Despite the methodology used in the experiments, Gapp's (1995a) results indicate that perceptual cues impact on the semantics of projective prepositions Section 2 3 4 2 4 highlights the importance of defining the origin of a spatial template

and illustrates how the incorrect location of this point can result in a paradoxical parsing of space Finally, in Section 2 3 4 2 5, the main characteristics that any computational model of a projective preposition's semantics should accommodate are defined

2 3 4 1   Topological Prepositions

The fundamental concept in the use of topological prepositions is punctual location (Clark 1973), the set of topological prepositions model this constraint *At, on, in,* and *near* are examples of topological prepositions The primary constraint on applicability of these prepositions is proximity to the landmark This constraint often results in an overlap in their range of applicability Differentiating between the applicability of these prepositions is a problematic issue requiring recourse to conceptual and/or functional information

2 3 4 1 1   Topological Prepositions Proximity Constraint

The above overview of the constraints defining the applicability of a topological preposition highlighted the requirement of the trajector's proximity to the landmark as the primary factor The main obstacle to modelling this constraint is the gradation of applicability across the region associated with a preposition

Figure 2-15 depicts two scenes, analysing the locative expression *the chair near the plant* in each of these scenes illustrates the importance of modelling the gradation in the interpretation process The chair intended by the locative in scene (a) and (b) are different However, the chair described by this locative expression in scene (b) is also present in scene (a) Clearly, it fits the conceptual and pragmatic constraints associated with the preposition *near* Yet it would still not be selected as *the chair near the plant* in scene (a) because of the presence of the extra chair in this scene This ranking of selection is due to the different applicability ratings of the chairs within the spatial template associated with the preposition *near*

50

(a)

(b)

Figure 2-15: Diagrams depicting how the gradation of applicability across the spatial template associated with a preposition affects its interpretation. The interpretation of *the chair near the plant* in scenes (a) and (b) is different because of this gradation.

How can this gradation be computationally modelled? Logan and Sadler (1996) present an analysis of the representations and processes involved in apprehending spatial relations They propose that "people decide whether a relation applies by fitting a spatial template to the object's regions of acceptability for the relation in question" (1996 pg 496) A **spatial template** is a representation of the regions of acceptability associated with a given preposition It is centred on the landmark and identifies for each point in its space the acceptability of the spatial relationship between the landmark and a trajector appearing at that point being described by the prepositions Psycholinguistic work (Logan and Sadler 1996) has shown that there are roughly three areas of applicability good, acceptable, and bad Good regions receive the highest acceptability ratings and correspond to the best uses of a spatial term Acceptable regions receive intermediate acceptability ratings, the distinction between good and acceptable regions is not sharp, rather these regions blend into one another gradually Finally, the bad regions correspond to unacceptable locations for a trajector to be located with respect to the landmark and described by the preposition, there was a sharp distinction in acceptability ratings between the bad regions and the adjacent good and acceptable regions Most importantly, the candidate trajectors can be assessed and rank-ordered by comparing their locations to the template and rating their candidacy based on the region of acceptability they are located within The candidate object with the highest acceptability rating is then selected as the trajector

To date, there have been several **continuum models** or **potential field models** proposed that aim to capture these regions These potential field models define equations that rate the inclusion of a point in space within a region based on variables such as distance from another point or angular deviation from a vector These will be reviewed in Chapter 5 Chapter 8 presents the model proposed in this thesis

Incorporating a potential field model into the interpretation of a preposition allows the ranking of candidate trajectors based on their location in the template However, as was mentioned at the start of Section 2 3 4 1, there are conceptual and functional factors that also require consideration during interpretation In the following sections, some of the approaches to these issues are introduced and reviewed

Many researchers (Clark 1973, Talmy 1983, Herskovits 1986, Fillmore 1997, Herskovits 1998) have conjectured that objects are cognitively characterised as simple geometric shapes when locating them with a preposition

> "The preposition 'at' is said to ascribe no particular dimensionality to the referent of its associated noun, the preposition 'on' is said to ascribe to the referent of its complement the property of being a line or a surface, and the preposition 'in' is said to ascribe to the referent of its complement the notion of a bounded two-dimensional or three-dimensional space " (Fillmore 1997 pp 28-29)

> "Prepositions contain certain presuppositions about their point of reference – e g whether it is one-, two-, or three-dimensional " (Clark 1973 pg 40)

The generalisation is based on the restrictions on the types of objects that may complement certain prepositions For example (10a) and (10b)

(10a) *The man stood in the yard*
(10b) **The man stood on the yard*

Note that, throughout this thesis an asterisk symbol is used to indicate semantically malformed phrases It is conjectured that the use of the preposition *in* in (10a) implies that the landmark object *yard* must be conceptualizable as a 3-D object, while in (10b) the use of *on* implies that the landmark is conceptualizable as a 2-D surface or plane Furthermore, it is posited that the unnaturalness of conceptualising a yard as a 2-D surface explains the semantic oddity of (10b) This analysis is compatible with Miller and Johnson-Laird's (1976 pg 384) paraphrases of Leech's (1969) interpretation of the prepositions *in* and *on*

"x in y x is 'enclosed' or 'contained' either in a two-dimensional or in a three-dimensional place y " (1976 pg 384)

"x on y x is contiguous with the place of y, where y is conceived of either as one-dimensional (a line) or as two-dimensional (a surface) " (1976 pg 384)

The term **schematisation** is used to describe the cognitive process that reduces a detailed scene to a sparse schematic content "schematisation – a process that involves the systematic selection of certain aspects of a referent to represent the whole, while disregarding the remaining aspects" (Talmy 1983 pg 225)

Although the existence of some form of geometric conceptualisation mediating between spatial language and perception is broadly accepted, it is not ubiquitously so Claude Vandeloise (1991) posits a functional approach to spatial language claiming that "the dimensionality of the object is often only a superficial consequence of the preposition itself, and not an essential characterisation of the use of the prepositions" (1991 pg 7)

2 3 4 1 3 Topological Prepositions Functional Constraints

Vandeloise argues for the rejection of the primary role of schematisation by positing that the proposed dimensionality required by the preposition of its complement is only a corollary of any specific definition assigned to the preposition, furthermore, for any locative expression, selecting a particular dimensional representation for the landmark is only one of several possible interpretations (Vandeloise 1991) Echoing Langacker's (1991b, 1994) concept of construal, Vandeloise maintains that "the speaker is free to consider one single object from an infinite number of perspectives, each of which may alter the importance accorded to any dimension" (1991 pg 7)

In order to illustrate the arbitrariness of a geometrically based solution to the constraint placed on a preposition's complements, Vandeloise gives examples of

acceptable sentences which illustrate the indifference of the preposition *in* to the dimensionality of its object

(11a) *The jewels are in the box*

(11b) *The cow is in the field*

(11c) *The priest is in the line*

(Vandeloise 1991 pg 6)

Although the landmarks in (11a) and (11b) may be described as a 3-D and bounded 2-D respectively, in example (11c) the landmark *the line* contravenes the accepted analysis that *in* ascribes the schematised form of a bounded two-dimensional or three-dimensional space to its complement (see (Fillmore 1997) Section 2 3 4 1 2) Vandeloise (1991) proposes a semantic model of prepositions based on functional relationships, such as bearer on burden or container in contained, between the landmark and the trajector The semantics of a spatial expression emerges through the linking of these functional relationships with "the extralinguistic knowledge of space shared by the speakers of one language" (Vandeloise 1991 pg 13)

Returning to the preposition *in*, Vandeloise explains the constraints placed on the form of object that may complement it not as a geometrical constraint but rather as a functional one "Its object must be a potential container, as long as the object satisfies this condition, the number of dimensions of the object is unimportant" (Vandeloise 1991 pg 19) For example, Vandeloise associates the functional relationship of container/contained as necessary to describe a spatial configuration using the preposition *in* Following this, in (12) the relationship between the trajector *John* and the landmark *his bed* can be described as the landmark contains the trajector or container/contained, therefore *in* is applicable to this relation and the expression is semantically well formed

(12) *John is lying in his bed*

However, there are problems with Vandeloise's functional approach According to Vandeloise an important trait of a container is the force it exerts on the objects in it "the

container controls the position of the contained object and not the reverse" (Vandeloise 1991 pg 225) This condition is a necessary part of Vandeloise's framework as it allows it to predict the semantic malformedness of expressions such as "*the bottle is in the cap" (Vandeloise 1991 pg 215) and "*the cat is in the collar" (Vandeloise 1991 pg 215) without resorting to geometric invariants such as the trajector being smaller than the landmark However, as Garrod *et al* note "it is quite natural to describe a plane as being in a cloud when there is topological enclosure (i e when it is completely surrounded by the cloud), even though we would not judge the cloud to control the location of the plane" (1999 pg 186) This demonstrates that prototypical enclosure licences the use of *in* even when the container does not control the location of the object in it

2 3 4 1 4 Topological Prepositions Summary

In summary, the main issues inherent in the semantic modelling of topological prepositions are modelling the proximity constraint and developing a mechanism to handle the geometric and/or functionally based factors that differentiate between the uses of the different topological prepositions

Unfortunately, the issue of whether it is geometric of functional constraints that determines the applicability of a topological preposition to a spatial relationship between a particular landmark and trajector is currently unresolved Furthermore, attempting to computationally model either of these approaches would require an explicit description of the geometric forms and or functional roles that each object assumes when functioning as a landmark As the set of objects that English can describe is open ended, the set of object descriptions defining the geometric forms and or functional roles would also be open ended In effect, these approaches require a computational system to semantically model each element of an open-ended class individually Clearly creating such a database would be extremely difficult Following this, the set of prepositions modelled in this thesis will be restricted to the projective prepositions *in front of, behind, to the right of, to the left of* The motivating criterion for selecting these prepositions is the minimal impact of pragmatic factors on their interpretation Consequently, they offer the best hope for a

successful treatment This said, however, interpreting even these prepositions is no easy task In the following sections the main issues that attend the semantic model of these prepositions are introduced

2 3 4 2 Projective Prepositions

The above analysis of topological prepositions describes the pragmatic issues inherent in modelling their semantics Indeed, it concluded with the proposition that the complexity of the epistemic factors involved make computational models of their semantics potentially intractable Consequently, the model of locative expressions developed in this thesis focuses on projective prepositions, more precisely the prepositions, *in front of*, *behind*, *to the left of*, and *to the right of* The principle for selecting this subset of prepositions is the relatively reduced role of pragmatics in their individual meanings The main issues that will be attended to in the modelling of these prepositions are (1) the orientation of the canonical direction of a preposition, (2) the origin of the spatial template, (3) the constituency of the spatial template, and (4) the scale dependency of the spatial template on the extension of the landmark

The pre-eminent factor in the construction of a projective preposition's spatial template is the preposition's canonical direction The **canonical direction** of a projective preposition is the usual directional relationship between a landmark and trajector that a projective preposition describes within a particular frame of reference (see Section 2 3 3) This directional constraint is represented by an axis projected from the landmark called the **search axis** It is the reliance of projective prepositions for the alignment of their search axis on the intended frame of reference that makes frame of reference selection such an integral part of any computational approach to modelling spatial locatives A resolution to this issue will be discussed in greater detail within the sections pertaining to frames of reference (Sections 5 3 1 and 8 3) However, here it will be assumed that the search axes have been aligned and the focus will be on the factors affecting the shape and regions of acceptability around a projective preposition's search axis that defines the area of search for a trajector In other words, this discussion will focus on examining the

factors that define a projective preposition's spatial template and the issues that arise in modelling it A number of psycholinguistic experiments (Gapp 1995a, Gapp 1995b, Logan and Sadler 1996) have examined the spatial templates of projective prepositions Section 2 3 4 2 1 describes these experiments (Gapp 1995a, Logan and Sadler 1996) and their results This description is followed by a critical review Section 2 3 4 2 2 highlights the impact of the trajector's proximity to the landmark on the semantics of a projective preposition and explains why this factor was not found by Gapp's (1995a), experiments or Logan and Sadler's (Logan and Sadler 1996) experiments Section 2 3 4 2 3 criticises the methodology used in both sets of experiments (Gapp 1995a, Logan and Sadler 1996), because it excluded perceptual cues, such as object occlusion Moreover, it is noted that some linguistic theorists (Clark 1973, Vandeloise 1991, Jackendoff and Landau 1992) and some of the psycholinguistic evidence (Gapp 1995a, Logan 1995) indicates that perceptual cues should be included within the semantics of projective prepositions It should be noted that, despite the methodology used in the experiments Gapp's (1995a) results indicate that perceptual cues impact on the semantics of projective prepositions Section 2 3 4 2 4 highlights the importance of defining the origin of a spatial template and illustrates how the incorrect location of this point can result in a paradoxical parsing of space Finally, in Section 2 3 4 2 5, a set of criteria for modelling the spatial templates of projective prepositions are defined

2 3 4 2 1 Projective Prepositions and Spatial Templates Psycholinguistic Evidence

Logan and Sadler describes the result of psycholinguistic work that aimed to define the constituency of spatial templates for the prepositions *above, below, over, under, left of, right of, next to, away from, near, far from, in, on* Figure 2-16 is a representation of the regions of acceptability within the spatial template of the projective preposition *above* as described in (Logan and Sadler 1996), the arrow in the figure represents the search axis for the preposition
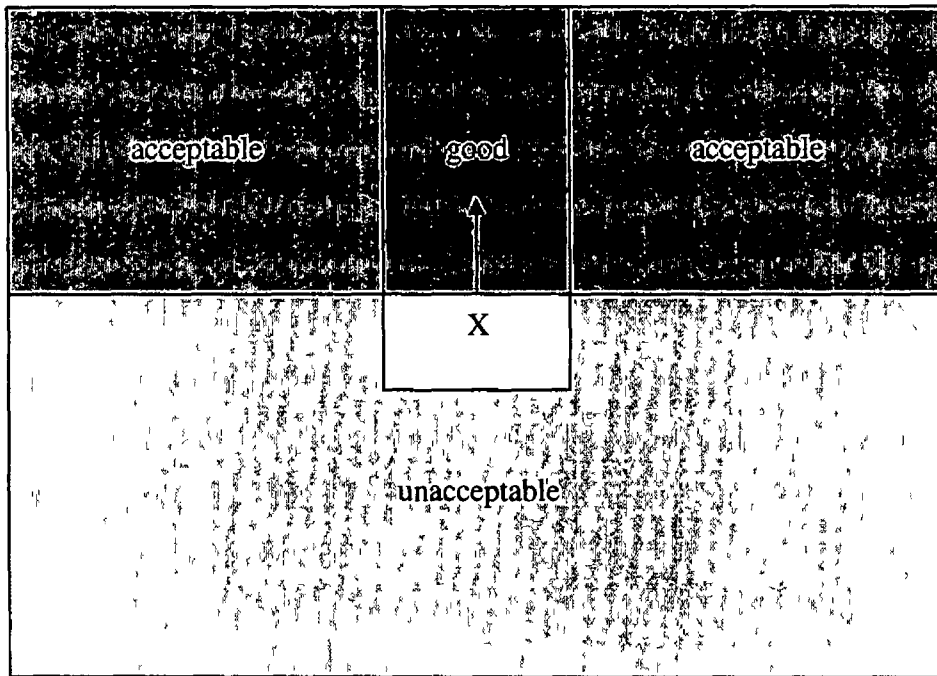
Figure 2-16. Representation of the regions of acceptability in the spatial templates for the projective preposition *above* defined in (Logan and Sadler 1996).

Logan and Sadler carried out four different types of experiments each designed to test a particular facet of spatial templates The first experiment was a production task aimed at assessing the regions of space that correspond to the greatest acceptability In the second experiment, subjects were shown sentences followed by pictures and were asked to rate how well the sentences described the pictures The purpose of this experiment was to capture the areas corresponding to good, acceptable, and bad regions The results of the first two experiments suggested similarities in spatial templates among classes of prepositions

"Templates corresponding to *above*, *below*, *over*, *under*, *left of*, and *right of* have similar shape but differ from each other in orientation and direction Templates corresponding to *next to*, *away from*, *near to*, and *far from* have different shapes from *above*, *below*, and so on, but are similar to each other except that *next to* and *near to* are reflections of *away from* and *far from* " (Logan and Sadler 1996 pg 514)

The third experiment was designed to test whether these similarities persisted when subjects were given lexicalised stimuli, i e , the subjects were shown words rather than pictures as stimuli The purpose of this experiment was to test whether there was a common knowledge structure underlying the cognitive processes in both verbal and visual spatial relations Subjects were asked to rate the similarity between pairs of relations from the set *above, below, left of, right of, over, under, next to, away from, near to, far from, in,* and *on* Correlating the results between this experiment and the previous experiment revealed a similar grouping across visual and verbal stimuli Based on this, Logan and Sadler concluded "that subjects used spatial templates to perform both tasks"(1996 pg 519) An important point, discussed later (see Section 8 4), and one that was not highlighted by Logan and Sadler, is that the results of this experiment revealed not only a grouping of spatial templates with respect to prepositional type (topological versus projective), but also within these groupings there were sub-groupings composed of prepositional pairs, i e , the spatial templates for *left* and *right* were grouped as similar

The fourth and final experiment tested the idea that spatial templates are applied in parallel This experiment used a reaction time task in which subjects were required to verify spatial relations between a landmark and a trajector The distance between these objects was varied systematically between trials The idea behind this experiment was that as spatial templates are conjectured to be applied in parallel to the whole visual field, the distance between the landmark and trajector should not matter The reaction times were not what had been expected, it was proposed ,however, that the variance in reaction times could reflect "a process of reference frame adjustment" (Logan and Sadler 1996 pg 523)

In describing their results, Logan and Sadler noted 5 main points

1   The area covered by a good acceptability in the spatial templates for projective prepositions is aligned with a parallel projection of the landmark along the search axis

2   Increasing the distance between the trajector and the landmark only has a slight impact on the acceptability rating of the trajector

3   The good and acceptable regions blended into one another

4   There was a sharp boundary between bad and acceptable regions

5   Similarities in the meanings of spatial terms can be accounted for in terms of the similarities in the spatial templates that correspond to them

The empirical data presented by Logan and Sadler (1996) indicates that the acceptable regions extend to an angular deviation of $90°$ from the search axis Gapp (1995a)[13] describes a set of experiments that refined the results of (Logan and Sadler 1996) The main areas of interest for this work that were examined by Gapp were

1   How does the angle of deviation between the vector describing the trajector position and the search axis influence the acceptability of a projective preposition describing the spatial relationship between the landmark and trajector?

2   How does the distance between the trajector and the landmark impact on the preposition's applicability?

3   How does the landmark's shape influence the prepositions applicability?

4   Are there distinctions between the applicability of regions of *in front of- behind*, *right-left*, and *above-below*?

The experiments examining the impact of the angle of deviation reveal that the acceptability of a projective preposition decreases linearly as the angular deviation increases, acceptability approached 0 as the angular deviation approached $90°$ The results of the distance experiments found no significant effect on acceptability Both these findings were expected and convergent with (Logan and Sadler 1996)

The third area studied was the influence of the landmark's shape on the spatial template of the preposition Gapp (1995a) proposes that the extension of a landmark orthogonal to the preposition's canonical direction affects the scale of the angular

---

[13] At the time Gapp wrote his 1995 paper (Logan and Sadler 1996) was in press

deviation and consequently a relation's degree of applicability. Figure 2-17[14], which is based on an figure in (Gapp 1995a), illustrates this: "even though the object L[15] is in the same absolute position compared to R for all configurations (a, b, c), the applicability of the relationship <right L R>[16] increases from (a) to (c)" (1995a pg. 2).
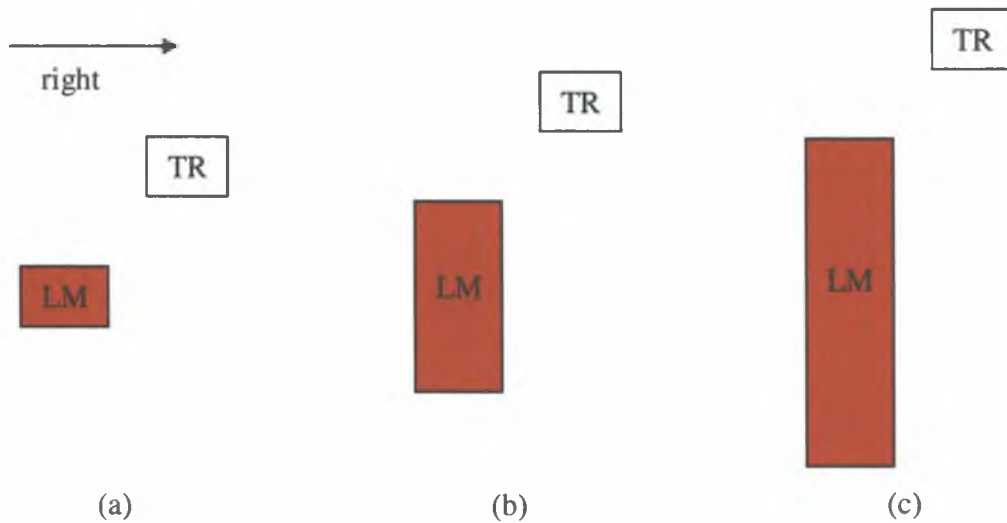
Figure 2-17: The influence of the landmark's extension on the angular deviation of the spatial template, based on a figure in (Gapp 1995a). The object labelled LM represents the landmark and the object labelled TR represents the trajector.

The study found a direct link between the extension of the landmark perpendicular to the direction of the preposition and the angular deviation encompassed

---

[14] Gapp (1995a) uses the term reference object to describe the landmark and located object to describe the trajector. For the sake of consistency in terminology across the dissertation the labels *R* and *L*, used in the figures in the original to denote the reference object and the local object respectively, have been replaced with *LM* and *TR*, which symbolise the landmark and trajector, respectively.

[15] For the sake of consistency in terminology across the thesis, in Figure 2-17 the trajector object in each diagram that Gapp (1995a) refers as L in this quotation is labelled TR and the the landmark object in each diagram that Gapp (1995a) refers as R in this quotation is labelled LM.

[16] In the terminology of this thesis, the relationship Gapp (1995a) defines by <right L R> would be defined as <right TR LM>.

by the spatial template "The larger the extension of the reference object[17] perpendicular to the canonical direction of the relation, the larger the relation's regions of applicability in this perpendicular direction" (Gapp 1995a pg 5)

As to whether there were distinctions between the applicability of regions of *in front of-behind, right-left,* and *above-below,* Gapp found that there was a slight tendency to rate *in front of-behind* and *above-below* regions higher than the *right-left* regions In a later paper, Gapp proposed that "this slightly higher rating might be due to the fact that the *in front of-behind* and the *above- below* axes are easier to perceive" (1995b pg 9)

To review, the results of (Gapp 1995a) and (Logan and Sadler 1996) reveal several factors that impact on projective spatial templates

1   Similarities in the meanings of spatial terms can be accounted for in terms of the similarities in the spatial templates that correspond to them

2   There are three areas within the spatial template good, acceptable, and bad

3   The areas within a spatial template are symmetrical around the search axis

4   The good and acceptable regions blend into one another

5   There is a sharp boundary between bad and acceptable regions

6   The acceptability of a projective preposition decreases linearly as the angular deviation increases

7   Acceptability approached 0 as the angular deviation approached 90°

8   The distance between the landmark and trajectors has no real impact on the acceptability rating of the preposition

9   The extension of the landmark orthogonal to the preposition's canonical direction effects the scale of the angular deviation and consequently a relation's degree of applicability

10 There is a distinction between the angular dependence of the three main directions *in front of-behind, right-left,* and *above-below* with the *in front of -behind* direction rated highest followed by the *above-below* direction

---

[17] As noted previously, Gapp (1995a) uses the term reference object to describe the landmark object

For this thesis, however, there are several weaknesses in the method and design of these experiments In particular, none of the experiments tested the distance factor in a situation where there was more than one trajector at a differing distance but with the same angular deviation Moreover, none of the experiments used visible cues where object occlusion could occur Furthermore, none of the experiments examined the location of the spatial template origin These weaknesses impact on the applicability of the above results to the design of NL interfaces These issues will be discussed in detail in Sections 2 3 4 2 2, 2 3 4 2 3, and 2 3 4 2 4 below

2 3 4 2 2 Projective Prepositions and Spatial Templates The Effect of Distance

Firstly, in both sets of experiments, (Gapp 1995a) and (Logan and Sadler 1996), the subjects were only asked to rate how applicable a preposition is to describe a spatial configuration between a landmark and one trajector There were, however, no experiments that tested the distance factor in situations where there was more than one trajector at differing distances but with the same angular deviation from the search axis In this thesis it is proposed that this omission resulted in the proposition that the distance between the landmark and the trajector has no real impact on the applicability rating of the trajector within a given spatial template For example, in Figure 2-18, it is clear that trajector (a) is a more likely candidate for *the X to the right of the landmark* than trajector (b), even though the angular deviation of both (a) and (b) from the search axis is identical

(a)

Landmark ϑ Search axis
            ϑ

(b)
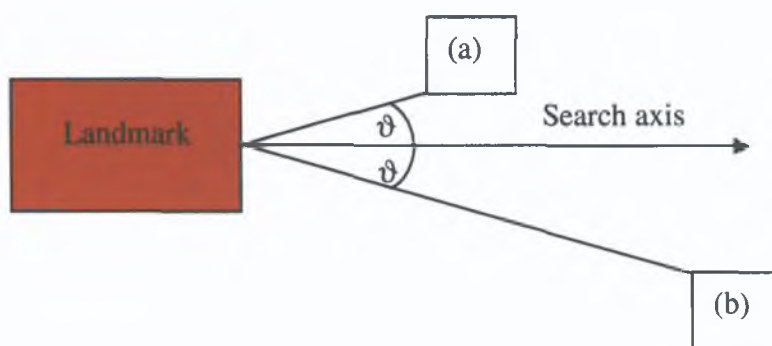
Figure 2-18: An example illustrating the effect of distance on the rating of a trajector within a spatial template.

Clearly any computational model of the semantics of projective prepositions must include distance as well as angular deviation in defining its spatial templates.

2.3.4.2.3 Projective Prepositions and Spatial Templates: Perceptually Based Differences

Secondly, it is important to note that both sets of experiments, (Gapp 1995a) and (Logan and Sadler 1996), were carried out in such a way that visible cues such as object occlusion did not occur in the tests; the stimuli given to the subject were either lexicalised or in the case where visible stimuli were given and object occlusion was relevant (i.e., in experiments testing *in front of-behind*) it was precluded by using a bird's eye view of the stimuli. Here it is proposed that this design impacts on the question as to whether all projective prepositions have a similar spatial template based on an identical set of factors or whether each preposition defines its own region based on a particular set of factors unique to itself.

Logan and Sadler's (1996) results indicate a similarity in the shape of spatial templates associated with prepositions of the same type: "Templates corresponding to *above, below, over, under, left of*, and *right of* have similar shapes" (1996 pg. 514). However, (Gapp 1995b) found that there was a difference between spatial templates of prepositions aligned with different spatial axes; regions in the spatial templates for *in front of-behind* and *above-below* were rated slightly higher than regions with the same

65

angular deviation in the *right-left* spatial template Moreover, he proposed that these higher ratings could be a result of the ease of perceiving the asymmetry along the *in front of-behind* and *above-below* axes This finding echoes the results of earlier psycholinguistic work (Logan 1995, cited in Logan and Sadler 1996) which found that "Subjects were faster with *above* and *below* than with *front* and *back*, and faster with *front* and *back* than with *left* and *right*" (Logan and Sadler 1996 pg 505) Both these findings are convergent with Clark's (1973) and Vandeloise's (1991) analysis of the impact of perception on spatial language It is worth noting that Jackendoff and Landau (1992) also posit that perceptual cues, such as object occlusion, impact on the semantics of projective prepositions

From this, in this thesis it is proposed that there are differences in the spatial templates across the set of projective prepositions Furthermore, although spatial templates exhibit a family resemblance indicating a set of core factors affecting all the projective prepositions, it is perceptual factors impacting on the definition of some prepositions, but not others, that cause the differences between the spatial templates More precisely, the perceptual phenomenon of occlusion impacts on the spatial template of prepositions along the *front-back* axis but not *right-left* axis

2 3 4 2 4 Projective Prepositions and Spatial Templates The Point of Origin

Finally, an issue that has been relatively neglected by research to date is the location of the origin of the spatial template, i e , point at which the search axes originate

In a seminal paper, Landau and Jackendoff (1993) presented evidence that object identification and localisation are performed by separate neural subsystems which they christened the "what" and "where" systems Furthermore, they claimed that the where system used a very coarse representation of the world when locating an object, while the what system used a much finer detailed description when viewing objects Based on this, many previous computational models (Gapp 1994a, Olivier and Tsuji 1994) schematised the landmark to its centre of gravity or the centre of the its bounding box The **bounding box** of an object is the minimal rectangle that encompasses the geometry of the object

66

The dashed lines in Figure 2-19(a) delineate the bounding boxes of a pyramid; in Figure 2-19(b) the green sphere represents the centre of the bounding box and Figure 2-19(c) illustrates the location of this origin relative to the object.

(a)   (b)   (c)
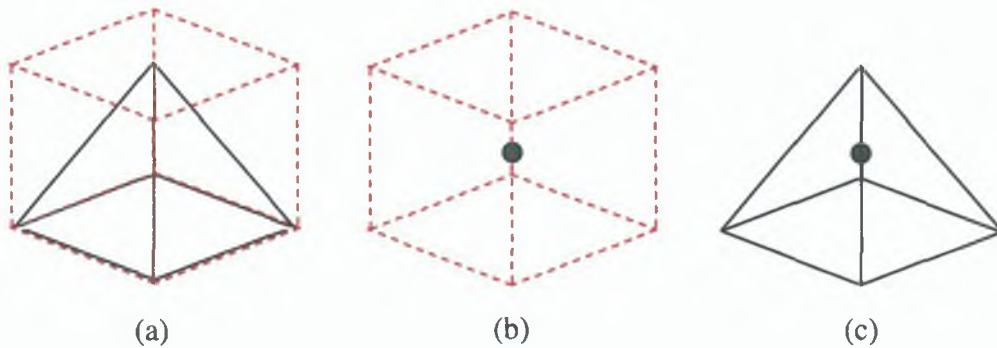
Figure 2-19: Illustrations of (a) the bounding box of a pyramid, (b) the centre of the bounding box, (c) the location of the centre of the bounding box relative to the pyramid.

While this approach works well for most simple solid objects, applying it to more complex shapes can be problematic. For example, when applied to concave or U-shaped objects the centre of the bounding box may be outside the object. This can result in paradoxical classification of regions as illustrated in Figure 2-20. Diagram (a) is a bird's eye view of a concave building – the dashed red lines delimit its bounding box, the green circle marks the centre of its bounding box, and the blue circle represents the location of a speaker. Taking the green circle as the origin for the spatial templates and aligning the search axis according to the mirror imagery strategy (see Section 2.3.3.3) employed by European languages, the area on the opposite side of the green circle to the speaker will be defined as *behind the building*, which paradoxically includes the grey area in diagram (b).

(a)   (b)
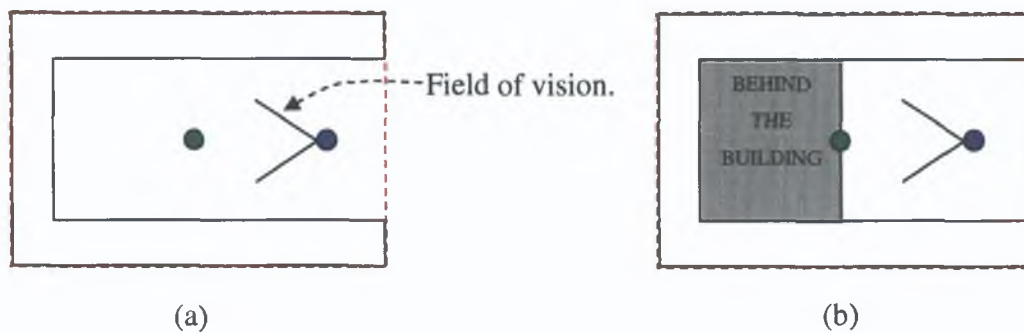
Figure 2-20: Illustrates the problem with locating the origin of the spatial templates at the centre of the landmark's bounding box: using such an approach results in the grey area in diagram (b) being classified as *behind the building* from the perspective of the viewer represented by the blue circle.

Another approach, which was adopted by Fuhr *et al.* (1998), is to represent the landmark using its bounding box. However, this bounding box representation is problematic when applied to locative expressions which refer to objects located inside the bounding box.

Clearly a new approach for locating the origin of the spatial template which avoids such paradoxes must be developed. In Chapter 8, such an algorithm is developed.

2.3.4.2.5 Projective Prepositions and Spatial Templates: Summary

In summary, there are many factors that impact of the shape and size of a projective preposition's spatial template. Some of these factors impact on the interpretation of all prepositions of this type. This set of core factors exhibits itself in the family resemblance manifest across the associated spatial templates. However, there are other factors; in particular, perceptually based cues such as object occlusion, which affect the interpretation of certain prepositions but not others. These factors result in the variance between the spatial templates aligned along different axes. Furthermore, a factor that has been relatively ignored in the literature is the location of the origin of the spatial template; successfully locating this point is a crucial step for any computational model. In

conclusion, the following list defines the characteristics of spatial templates that have emerged in the above discussion

1. There are three areas within the spatial template good, acceptable, and bad

2. These areas are symmetrical around the search axis

3. The good and acceptable regions blend into one another

4. There is a sharp boundary between bad and acceptable regions

5. The acceptability of a projective preposition decreases linearly as the angular deviation increases

6. Acceptability approached 0 as the angular deviation approaches $90°$

7. The extension of the landmark orthogonal to the preposition's canonical direction effects the scale of the angular deviation and consequently a relation's degree of applicability

8. The distance between the landmark and trajector impacts on the acceptability rating of the trajector by virtue of the fact that if there are two trajectors located at the same angular deviation from the search axis the trajector closer to the landmark will have a higher acceptability rating

9. There is a distinction between the angular dependence of the three main directions *in front of-behind*, *right-left*, and *above-below* with the *in front of-behind* direction rated highest followed by the *above-below* direction It is conjectured that perceptual cues are the basis of this variance

The ability to accommodate these characteristics will be used to assess the computational models reviewed in Chapter 5 Moreover, these characteristics will inform the design parameters of the model developed in Chapter 8

69

2 3 5   Locatmg the Trajector

The precedmg section addressed the issue of how to define the area of search for a trajector The concept of a spatial template was introduced and a set of criteria for modellmg such a template was described From a computational perspective, the main advantage of a spatial template is its ability to rate the location of different candidate trajectors and, by domg so, to rank them However, havmg developed such a template, it is still incumbent upon an interpretative computational system to define an abstraction to represent the location of the candidate trajectors within the spatial template Here, it is proposed that to select the optimal object as the intended trajector such a system must include a model of perceptual accessibility to the trajectors in conjunction with their rankmg within the spatial template This section begins with a discussion of the different abstractions that a computational system may use to represent the location of a trajector followed by an example demonstratmg the necessity of checkmg for the occlusion of a trajector

2 3 5 1   Modellmg the Trajector

There are many ways in which a computational system may represent the location of a trajector when interpreting a locative These range from a reasonably concrete description based on the vertices of its geometric mesh to an abstract characterisation of the object as a point, usually taken as the trajector's centre of mass

Although adopting the set of points defined by the mesh of the object provides the maximal information on the trajector's location, it does so at the huge computational cost Moreover, Landau and Jackendoff's (1993) research (see Section 2 3 4 2 4) indicates that such a specificity is not required Following this, most computational systems (Gapp 1994a, Olivier and Tsuji 1994) treat the trajector as a point, usually the centroid of the object While this approach has the advantage of computational efficiency, choosing the trajector's centre of mass as the representative location for the object can have undesirable results

70

Figure 2-21 illustrates how this abstraction can distort the measurement of the distance between a trajector and a landmark and the angle of deviation between the trajector and the landmark Both diagrams in Figure 2-21 use a bird's eye view of a spatial configuration The speaker is represented by the blue circle, the landmark by the red rectangle, and the trajector by the white rectangle Both diagrams contain a representation of a projective spatial template constructed around the illustrated search axis, this spatial template arbitrarily drawn could represent the area described by the preposition *to the right of the landmark* from the viewer perspective However, it illustrates the general shape of a projective spatial template, where the applicability of the preposition decreases as the angular deviation from the search axis of the trajector's location and the distance of the trajector from the landmark increases The applicability of the preposition within the spatial template is represented by colour the darker the region, the higher the applicability The green line in diagram (a) depicts the distance between the trajector's centre of mass and the origin of the spatial template It is evident that the trajector's centroid is located in the second area of applicability in the spatial template, even though large portions of the trajector are located in a higher region of applicability Diagram (b) illustrates how the angle of deviation between the trajector's location and the search axis may be exaggerated by only measuring its centre of mass, in this instance, parts of the trajector actually lie on the search axis, however, the computed angular deviation would locate the trajector outside the regions of applicability for this spatial template
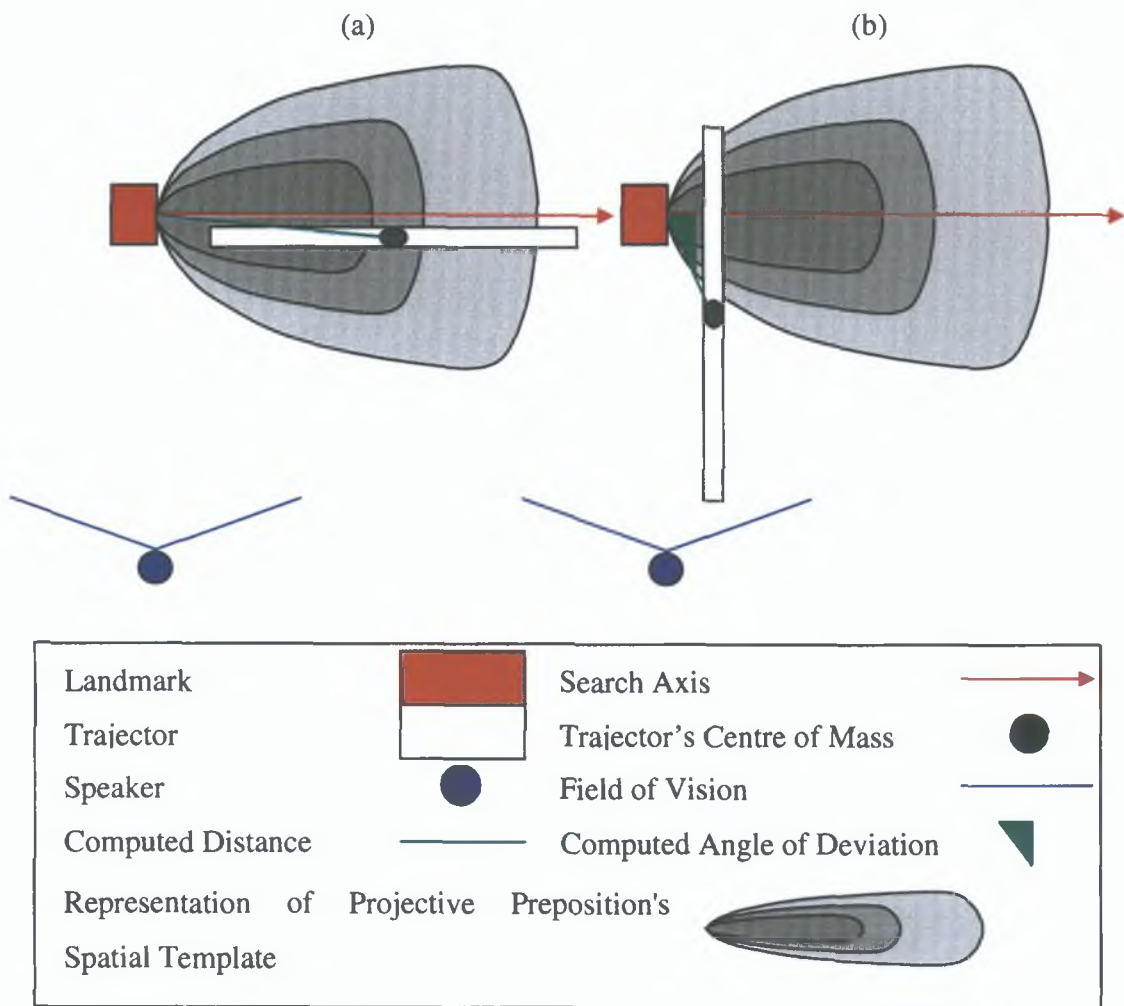
Figure 2-21: Diagram (a) illustrates how the abstraction of the trajector to its centre of mass can distort the computation of the distance between the landmark and the trajector. Diagram (b) illustrates how the abstraction of the trajector to its centre of mass can distort the computation of the angle of deviation between the trajector and the search axis.

2 3 5 2 Occluded Trajectors

Another issue that impacts on locating a trajector is the perceptual access of the user to the trajectors The majority of previous models have neglected this issue, relying merely on a topological algorithm to rank the trajectors

Figure 2-22 illustrates the fallacy of such an approach In these diagrams, there are two trajectors (assumed to be trees) and the landmark (a house) It is reasonable to predict that in diagram (a), trajector (1) is the object intended on by the phrase *the tree to the right of the house* However, in diagram (b), trajector (1) is occluded from the speaker's view, for example, by a wall In this instance, trajector (2) is the most likely candidate when interpreting the phrase It is important to note that in both diagrams, trajector (1) will have the highest rating within the spatial template Furthermore, a purely topological model that does not consider the occlusion of trajectors will select trajector (1) erroneously in situation (b)
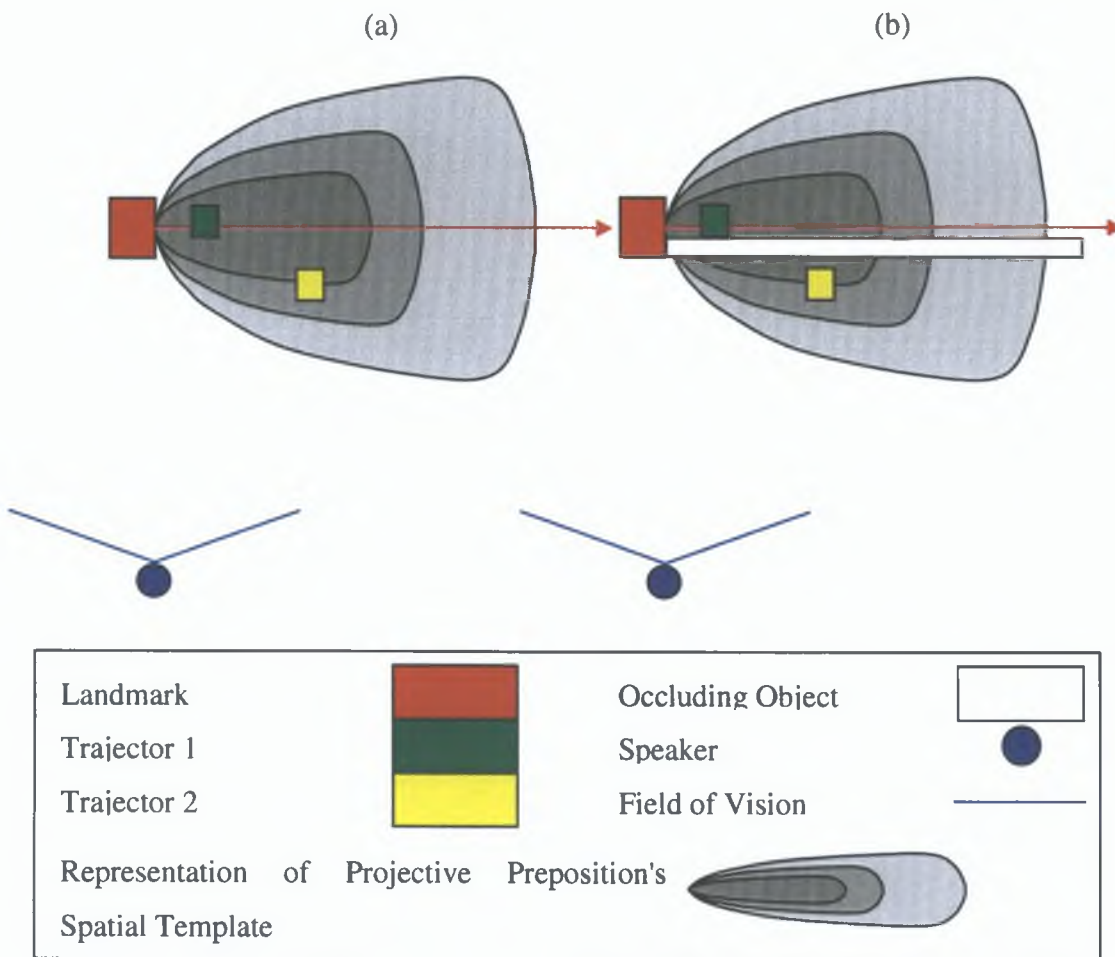
Figure 2-22: Diagrams illustrating the impact of object occlusion on the selection of a trajector. In diagram (a) trajector (1) has the highest applicability rating among the candidate trajectors due to its location in the spatial template. As such it should be selected as the primary trajector. In diagram (b), however, trajector (1) is occluded from the view of the speaker. Consequently, trajector (2) should be selected as the primary trajector.

2.4 Discourse Models

When developing an NL interface for a computational system that is to interpret language at anything but the shallowest level or to interact in a mode natural to a user, it is impossible to consider a user's commands in isolation. Often, user commands can only be understood by considering them as part of an ongoing dialogue. Consequently, a major issue in developing a NL system is how to incrementally build up, and use, a model of the dialogue.

The requirement for a discourse model is never more apparent than in the analysis of referring expressions, particularly in the case of anaphoric expressions. People use **referring expressions** to intend on objects using language. There are many different types of referring expressions, including: definite descriptions, pronominals, demonstratives, one-anaphora, other-expressions, and locative expressions. In the context of this work[18], a referring expression is a linguistic structure that intends upon an object on the screen or an entity that has previously been introduced into the discourse. The isolated entity is labelled the **referent** of the expression; the term **antecedent** may be used where the referring expression is anaphoric. Some of these expressions are anaphoric (e.g., pronominals, one-anaphora, other-expressions). The term **anaphora** describes "any abbreviated back-reference to an entity mentioned, explicitly or implicitly, earlier in a text or conversation" (Hirst 1994 pg. 3487). Pronouns are the best example of this grammatical phenomenon; however, they are not the only form of expression that utilises this mechanism.

During dialogue, an interlocutor may refer to elements of discourse through a variety of expressions, including definite descriptions, demonstratives, pronominal

---

[18] A referring expression in discourse can serve other purposes in addition to isolating an entity. For example, they may convey some additional information about the entity (referring expressions are italicised): (a) *That dog* is getting quite aggressive, (b) Yesterday, *the old flea ridden mutt* bit my hand. The referring expression in (b) describes that age and hygiene of the dog and provides information on the speaker's attitude towards the dog. However this thesis only considers their ability to intend on entities.

75

reference, one-anaphora, locatives expressions, etc Some of these mechanisms are illustrated in the following sequence of expressions (13)

(13a) Make *the red square* blue (*the red square* – definite description)

(13b) Move *the other one*₁ up (*the other one* – one-anaphora)

(13c) Make *it*₁ wider (*it* – pronominal reference coindexed with the expression *the other one* in 13b)

(13d) Put *this on it*₁ (*this* – demonstrative reference, *it* – pronominal reference coindexed with the pronominal reference *it* in 13c, *this on it* – locative expression)

Note that referring expressions are italicised and coindexing (i e , assigning identical subscripts) is used to indicate anaphoric relations This variety of referring linguistic constructs at a user's disposal increases the complexity of computationally modelling the domain Moreover, this complexity is exacerbated by the apparent variety of approaches required to deal with each of these expressions

- accommodation The entity referred to by the expression is novel to the discourse
- linking / accessing The entity referred to by the expression is established in the discourse
- bridging The entity referred to by the expression is related to an entity previously established in the discourse
- evoking / inference The entity referred to by the expression is known to the listener but new to the discourse

(Byron 1998, Salmon-Alt and Romary 2001)

The term **reference resolution** describes the process by which a listener identifies as accurately as possible the referents of a speaker's discourse During this process, a listener may utilise one or more of the above approaches At the heart of many reference resolution algorithms is the notion of a context or **discourse model** Its purpose is to save contextual information that can inform the selection of referring expressions' referents

76

Classically, there are two basic problems the discourse model must address (a) what information should the model carry forward that may be useful in resolving future references? and (b) how does the model attribute a referent to a given referring expression (Byron 1998)?

The main obstacle in addressing these issues is that, often, all the information necessary to compute a unique interpretation of an utterance, at the time it occurs in the discourse, is not available from the linguistic context of the discourse Indeed, reference resolution is a canonical artificial intelligence problem, requiring the combination of information from multiple sources linguistic, epistemic, and perceptual Although this is widely recognised (Hirst 1994, Grosz et al 1995, Byron 1998, Salmon-Alt and Romary 2001), the majority of context models proposed to date have concentrated solely on linguistic sources Moreover, the models that do admit perceptual information into their framework (Salmon-Alt and Romary 2001), give no explicit description of how (a) the model is to gather or structure the perceptual information and (b), once this information has been captured, how it combines with linguistic knowledge to resolve a reference

2.5 Visual Salience, Locative Expressions, and Reference Resolution

Section 2 2 examined the link between language and perception and the importance of attention as a regulating process in human perception and proposed an approach to creating a computational model of visual salience Section 2 3 analysed the problems inherent in the interpretation of locative expressions In Section 2 4, the concept of a referring expression was introduced Referring expressions were described as the linguistic mechanism that people use to intend on objects in the world The question addressed now is what is the relationship between these topics?

The link between locative expressions and reference resolution is that a locative expression is a complex type of referring expression, to interpret a locative is to resolve a reference A more difficult question, however, is what is the relationship between resolving a referring expression and a computational model of visual salience?

77

The answer to this question is **mutual knowledge**[19] Mutual knowledge "consists of the set of things that are taken as 'known' by the participants in a discourse" (McCawley 1993 pg 355) Following Grice's (1989) cooperative principle, a conversational implicature of a cooperative speaker using a referring expression is that the speaker will use the referring expression to denote an object they assume the hearer has knowledge of In other words, a cooperative speaker will only refer to objects they assume are in the mutual knowledge set Furthermore, when interpreting a referring expression, the hearer will select a referent from the objects they assume are in the mutual knowledge set

In the SLI context, the system manipulates objects in the simulation in response to user commands The user's primary information source about the simulation is what they see on screen Consequently, the set of objects the user has seen approximates their knowledge of the simulated world Crucially, this set of objects also approximates the set of objects the user can assume the system has knowledge of In effect, the set of objects the user has seen comprises what the user understands as the mutual knowledge in the user-system dialogue

Following this, it is proposed here that the interpretive module of an NLVR system should restrict the set of entities in the world model that it treats as possible referents for

---

[19] There have been many names used to describe the concept of mutual knowledge shared knowledge, context, common ground, pragmatic presuppositions, tacit assumptions, normal beliefs The motivation for adopting Clark and Marshall's (1981) term mutual knowledge is to highlight the interactive nature of the overall discourse context A further point of note is that there are several terms used in this thesis that are related to the term mutual knowledge discourse model, context model, cognitive domain, reference domains At this point a brief description of how these terms are related might be useful in clarifying the later discussions The discourse model developed in this thesis, has two components, an interpretive module and a context model The context model represents the system's model of the user-system discourse mutual knowledge set Consequently, in this thesis the terms mutual knowledge and context model are taken to be synonymous The term cognitive domain (Langacker 1987) describes the local context that an expression is interpreted in The mutual knowledge model or context model component of the discourse model developed in this thesis is comprised of a set of domains of reference These reference domains function as local contexts for interpreting user input Hence, each of these domains of reference model a cognitive domain In summary, the mutual knowledge model / context model component of the discourse model is composed of a set of cognitive domains / reference domains

78

user input to those objects the user has seen; i.e., the set of objects a user will treat as mutual knowledge. In this context, resolving a user's natural language reference entails the selection of a specific object from the set of objects the user has seen in the simulation. There are two processes by which the mutual knowledge of the user-system discourse can be extended:

1. A perceptual event: in this context, a new object rendered on the screen.
2. A linguistic event: the user making a reference to an object in the discourse extends the set of linguistic entities in the mutual knowledge set.

Considering this, the three different uses of reference that the SLI system is concerned with can be defined as what Clark and Marshall[20] (1981) have called:

1. **Visible situation use**. A speaker can refer to an object so long as the object is "visible to both the speaker and listener" (Clark and Marshall 1981 pg. 22).
2. **Immediate situation use**. A speaker can refer to an object even though it "is not visible so long as its existence can be inferred from the situation" (Clark and Marshall 1981 pg. 22); i.e., it has been rendered during the user-system interaction.
3. **Anaphoric use**. A speaker can refer to an object that has previously been introduced into a "shared previous discourse set" (Clark and Marshall 1981 pg. 22).

But what is the link between resolving reference and a visual saliency algorithm? Modelling visual salience allows us to capture the perceptual events that cause the entry of an object into what the user considers as mutual knowledge. This, in conjunction with a model of discourse, allows the system to incrementally build a model of what the user

---

[20] The definitions of reference use described in (Clark and Marshall 1981) were restricted to the use of definite reference. Here these definitions are broadened to include other forms of referring expressions.

considers mutual knowledge Since a cooperative user will only refer to objects they consider to be in the mutual knowledge domain, modelling this domain allows the system's interpretive module to restrict the set of objects it should consider as referents to user referring expressions As an aside, it should be noted that in this thesis the term **deictic reference** is understood to describe a visible situation or immediate situation reference use

2.6 Chapter Summary

In Chapter 1, it was noted that perceptual and linguistic input should be combined to improve on existing systems which interpret spatial language A broad outline of the interpretive framework proposed by this thesis was also presented This framework has three major components a model of synthetic vision, a semantic model for locative expressions containing the projective prepositions *in front of, behind, to the right of,* and *to the left of,* and a discourse model that integrates perceptual and linguistic information The structure of Chapter 2 reflected the tripartite architecture of the proposed framework

Section 2 2 examined the link between language and perception It began by reviewing Herb Clark's (1973) analysis of the correlation hypothesis which illustrates the connection between language and perception Following this, the importance of attention as a regulating process in human perception was examined Furthermore, the range of conflicting factors that determine visual salience was highlighted and the subsequent difficulties in modelling such a complex process was noted It was concluded that by abstracting visual attention to its most general and basic determiner, (i e , location in the scene) the complexity of the model is reduced and the genericness of the model is increased The computational model of vision proposed by this thesis is based on this abstraction of visual attention (see Chapter 7)

Section 2 3 analysed the problems inherent in the interpretation of locative expressions The initial stage of the interpretation process is the identification of the landmark It was concluded that in order to resolve the landmark reference a

80

computational model must accommodate the different mechanisms and the attending issues of referential expressions

The next stage of the interpretation process considers the speaker's intended frame of reference Section 2 3 3 began by defining the different frames of reference used in English and described the alignment of the viewer-centred and absolute frame of reference for a computer user Section 2 3 3 4 examined the common terminology shared between the reference frames and how this could result in coordination failures The final section, Section 2 3 3 6, which highlighted the need for a computational algorithm to interpret a user's intended frame of reference This section explored the problems generated by the lack of a default frame of reference in English and the lack of a suitable computational algorithm for selecting a frame of reference

Once a frame of reference has been selected, the canonical direction of the projective preposition can be orientated relative to the landmark Subsequent to this, a model of the area described by the locative and a mechanism for selecting a trajector within this region must be defined Section 2 3 4 ascribed the preposition the primary role in the definition of the search region and introduced the concept of static prepositions Next, the classification of prepositions was refined by distinguishing between topological and projective prepositions The analysis of the pragmatic factors involved in distinguishing between the semantics of topological prepositions revealed difficulties that are beyond the scope of this work For this reason, this thesis focuses on projective prepositions, in particular *in front of*, *behind*, *to the left of*, and *to the right of* Adopting the approach of (Carlson-Radvansky and Irwin 1994, Carlson-Radvansky 1996, Logan and Sadler 1996) it is proposed that people use spatial templates when interpreting a locative expression Based on an examination of psycholinguistic evidence, a set of criteria that must be considered when modelling the spatial template of projective prepositions was defined This set included the effect of distance as well as angular deviation (see Section 2 3 4 2 2), the perceptually based variance across the spatial templates of projective prepositions (see Section 2 3 4 2 3), and the issues pertaining to the location of the spatial template origin (see Section 2 3 4 2 4) The final stage in the interpretation process related to the selection of the trajector, the main problems being

(a) the problems associated with the different representations of a trajector and (b) the need to check for the occlusion of candidate trajectors

Section 2 4 introduced the process of reference resolution and the need for a discourse model to capture contextual information Although the majority of this section focused on the problems relating to modelling the linguistic context, it highlighted the need to incorporate perceptual information into the discourse model and concluded by stating two problems facing any discourse model that models perceptual information (a) how does the model gather or structure the perceptual information? (b) once this information has been captured, how does the model combine it with linguistic knowledge to resolve a reference?

The final section, Section 2 5, described the relationship between the topics introduced in the previous sections modelling perception, locative expressions, reference resolution It was noted that locative expressions are a complex form of referring expression Following this, the relationship between modelling perception and resolving a referring expression was examined The concept of mutual knowledge was introduced as the foundation of this relationship Here, it is proposed that by computationally modelling visual salience NLVR systems extend their model of mutual knowledge and, by doing so, their ability to resolve references Finally, it should be noted that the model of reference resolution developed in this thesis is an attempt to extend the current accounts of how people use words as a means to refer and, furthermore, as a step towards a theory of how people use referring expressions to refer to things in a visual environment This contrasts with linguistically inspired theories, which focus on modelling co-reference – accounts of the way people use words to refer to words

82

3 Theoretical Linguistic Foundation

3.1 Introduction

Cognitive linguistics is an approach to language that is based on people's experience of the world and the way they perceive and conceptualise it (Ungerer and Schmid 1996) Langacker's **cognitive grammar** (1987, 1991a, 1991b, 1994) is one of this field's pre-eminent theories This chapter reviews Langacker's (1987, 1991b, 1994) cognitive grammar The linguistic approach adopted by this thesis is framed within this paradigm

Cognitive grammar postulates a conceptualist, as opposed to truth conditional, approach to linguistic semantics and regards grammar as a symbolic system that structures the conceptual content Moreover, Langacker posits that language can only be understood within a cognitive context

> "It is claimed instead that semantics structures (which I call **'predications'**) are characterised relative to **'cognitive domains'**, where a domain can be any sort of conceptualisation a perceptual experience, a concept, a conceptual complex, an elaborate knowledge system, etc " (Langacker 1991b pg 3)

In this thesis Langacker's model is adopted as a linguistic basis, because he admits visual perceptual information into the framework and by doing so allows for the introduction of elements into the context domain which have not been explicitly referred to in the linguistic/textual discourse While such an approach is more conducive to dealing with issues arising through visual perceptual context, from a computational perspective its lack of formalisation can make it difficult to implement

3.2 Cognitive Grammar

The title of Langacker's book "Concept, Image and Symbol" (1991b) refers to the central claims of his linguistic theory of cognitive grammar The fundamental claim is that linguistic semantics is grounded in conceptualisations that reside in cognitive processes A linguistic expression evokes conceptual content which has one of many possible images imposed on it through the grammatical structure of an expression Furthermore, grammatical elements and constructs provide a mechanism for symbolising the construed image

"This model assumes that language is neither self-contained nor describable without essential reference to cognitive processing (regardless of whether one posits a special *faculte de language*) Grammatical structures do not constitute an autonomous formal system or level of representation they are claimed instead to be inherently symbolic, providing for the structuring and conventional symbolization of conceptual content Lexicon, morphology, and syntax form a continuum of symbolic units, divided only arbitrarily into separate components" (Langacker 1991b pg 1)

A network of interrelated senses describes the conventional meaning of most lexical items In general, these networks cannot be reduced to a single structure, e g , a prototype or highest-level schema Instead, Langacker defines semantic structures, predications, relative to cognitive domains that can be any sort of conceptualisation This position is based on the observation that some concepts presuppose others in their definition, e g , the concept of a right angle triangle is necessary to characterise a hypotenuse This observation introduces the notion of "hierarchies of conceptual complexity" (Langacker 1991b pg 3), where concepts are built up from operations performed on concepts at lower levels Moreover, "linguistic predications can occur at any level in such hierarchies" (Langacker 1991b pg 3)

"The implications of this position are apparent the full and definitive characterisation of a semantic structure must incorporate a comparable description of its domain, and ultimately of the entire hierarchy of more fundamental conceptions on which it depends " (Langacker 1991b pg 4)

At the lowest level of these hierarchies are basic domains These are cognitively irreducible representations of human experiences of the world, e g , how humans conceptualise their experience of time and space Langacker admits the possibility "that certain linguistic predications are characterised solely in relation to one or more basic domains" (1991b pg 4) Most expressions, however, reference non-basic domains

An expression's characterisation may require more than one domain In such situations, Langacker uses the term complex matrix to describe the set of domains required Similar to the hierarchy of concepts within a domain, Langacker describes a hierarchy of domains within a complex matrix For a given situation, some domains will be more likely activated as central than others Indeed, some domains may be included as components of others

Figure 3-1 illustrates a section of a complex matrix for a knife In this figure, there are three diagrams each representing a domain that the concept of a knife is incorporated in, from left to right the shape or space domain, the cutting domain, and the silverware domain The shape domain, which is represented by a black knife, is incorporated in the cutting and silverware domains The dotted lines indicate that the shape domain corresponds across the domains that it is incorporated in (i e , it is construed as identical) The arrow in the cutting domain delineates the typical direction of motion of a knife within that domain

85
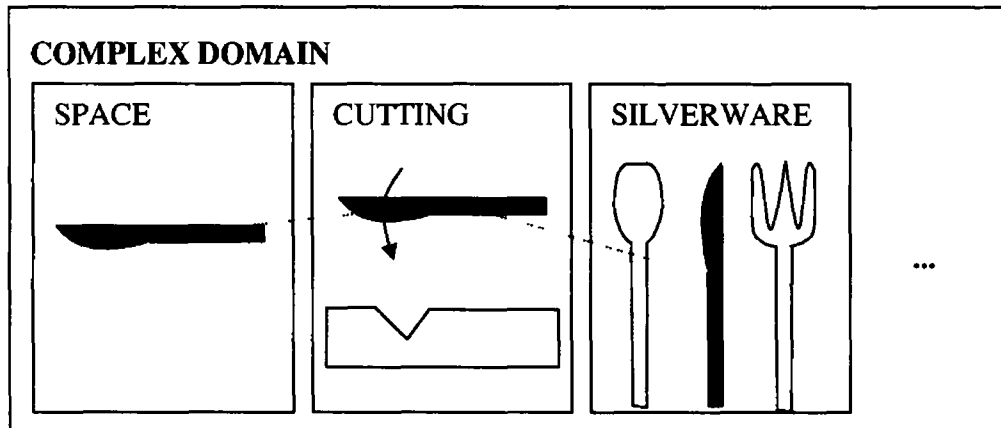
Figure 3-1: Section of a complex matrix characterising a knife. Based on a figure in (Langacker 1991b pg. 5)

The above architecture for conceptual content is not sufficient to characterise a linguistic predication A mechanism for imposing conventional imagery or **construal** onto the conceptual content is required By conventional imagery Langacker is referring to the human ability to "construe the content of a domain in alternate ways" (Langacker 1991b pg 5)

"People have the capacity to construe a scene by means of alternative images, so that semantic value is not simply received from the objective situation but is instead in large measure imposed on it " (Langacker 1991b pg 35)

Langacker describes six dimensions that can impact on how humans construe an expression

1   "The imposition of a **profile** on a base The base of a predication is its domain (or each domain in a complex matrix) Its profile is a substructure elevated to a special level of prominence within the base, namely that substructure which the expression 'designates' [ ] An expression's

semantic value does not reside in either the base or the profile individually, but rather in the relationship between the two " (Langacker 1991b pg 5)

2 Level of specificity Typically one of the component expressions in a grammatical construction elaborates a schematic substructure within the other

3 Scale and scope of a predication The scope of a predication is the extent of its coverage in relevant domains The scale of a predication is the implied size of objects it characterises

4 Prominence The relative salience of a predication's substructures One factor is the special prominence associated with profiling (see 1 above) Another is the asymmetry inherent between trajector and landmark objects in relational predications The final facet is the enhanced salience of elements that are explicitly mentioned

5 The construal of a situation relative to different background assumptions and exceptions

6 Perspective which includes factors such as frame of reference and how objectively an entity is construed "to the extent that a scene has a visual aspect it can be portrayed as if observed from different vantage points and orientations" (Langacker 1991b pg 35)

The semantic image of an expression is imposed on its conceptual content through grammar "When we use a particular construction or grammatical morpheme, we thereby select a particular image to structure the conceived situation for communicative purposes" (Langacker 1991b pg 12) For example

(14a) *The car in front of the tree*

(14b) *The tree behind the car*

(14c) *The car is in front of the tree*

Noun phrase (14a) designates *the car*, (14b) designates *the tree*, while sentence (14c) designates a locative relationship *is in front of* through a span of time Langacker's

theory explains these designations through the grammatical structure of the sentences In sentences where a head combines with a modifier, the head's conceptual substructure is profiled within the cognitive domain of the expression In (14a) and (14b), the prepositions *in front of* and *behind* are modifiers This results in the conceptual substructure of the head in each of these noun phrases being designated, that is *the car* and *the tree* respectively However, in (14c), the addition of *is* to the prepositional phrase modifier converts it from a modifier into a process predication Hence, the head is not combining with a modifier but with a process predication In such instances, it is the entity designated by the predication (i e , the extension of the locative relationship) that is imposed as a profile on the domain

Langacker's framework directly links grammatical construction and semantics He rejects the notion of abstract deep structures and any linguistic paradigm that conceives grammar as generative, constructive, or distinct from semantics A necessary complement to his definition of the role and scope of grammar is a model of grammatical organisation

In Langacker's model, grammar is organised as a structured inventory of established linguistic units Some of these units can function as components of others and some are schematic templates of conventions used in the assembly of complex symbolic structures From a user's perspective, these units are holistic, i e , they can be activated without the user attending to their internal composition A further "pivotal claim of cognitive grammar is that grammatical units are intrinsically symbolic" (Langacker 1991b pg 16) These symbolic units are bipolar, having a semantic and a phonological aspect (e g , [[SEM]/[PHON]]) that can vary in complexity and specificity The complexity of "a unit is minimal (a morpheme) if it contains no other symbolic components" (Langacker 1991b pg 16) A unit's facet of specificity can vary from highly specific to maximally schematic A symbolic unit representing a basic grammatical category (e g , noun – [[THING]/[X]]) is maximally schematic semantically and phonologically Schematic templates are complex and act as grammatical rules Figure 3-2 depicts a morphological rule given by Langacker to illustrate schematic templates

```
[[PROCESS]/[Y]] - [[ER]/[er]]]
```

Figure 3-2: Example of a schematic template in Langacker's model (1991b pg. 16).

This schematic template describes how to categorise deverbal nominalizations *teacher*, *helper*, etc. The rule shows how the verb schema [[PROCESS]/[Y]] is integrated ("-") with a grammatical morpheme [ER]/[er]]. "Its internal structure is exactly parallel to that of an instantiating expression; e.g., [[[TEACH]/[teach]]-[[ER]/[er]]], except that in lieu of a specific verb stem it contains the schema for the verb-stem category" (Langacker 1991b pg. 17). The grammatical framework treats these constructional schemas as symbolic resources; consequently, a schema may be incorporated as a component of another.

These schematic rules allow conventional units to be categorised; however, they do not show how new conventional units emerge. This is an essential process in a description of any linguistic model. Langacker describes a mechanism that allows frequently used/encountered novel expressions to be integrated into the grammar's inventory. Novel expressions are those expressions whose meaning is obvious from the context they are used in and whose constituent parts may be described within the grammar. However, the overall categorisation and/or constructional schema is not described by the conventions of its components within the grammar. If such an expression recurs frequently it may become established as a conventional unit taking a schematised form of its contextual meaning as its meaning in the grammar.

Allowing the emergence of new schemas to represent new units necessitates the grammar characterising the new unit's predications relative to a cognitive domain. For each cognitive domain, a schematic unit representing the shared content of its class members is extracted. New units are then categorised based on a judgement of whether they instantiate the category's representational schematic. In the majority of cases, this is achieved based on their intrinsic semantic and/or phonological content. "The vowel [i], for example, is classified as a high vowel by virtue of the categorising unit [[HIGH VOWEL] → [i]], where [HIGH VOWEL] is a schematic phonological structure which

neutralises the properties that distinguish one high vowel from another" (Langacker 1991b pg 19)

The main problem with this approach is that categorising new units is not always possible through their semantic or phonological properties, for example, the class of nouns that voice *f* to *v* in the plural (*leaf/leaves*, but *reef/reefs*) (Langacker 1991b) Indeed, it is the unpredictability of syntactic and morphological behaviour which forms the basis on which many linguistic models posit the independence of grammatical categories from their meaning or use Langacker refutes this distinction between grammar and semantics and accommodates the unpredictable behaviour of exceptions by allowing extra symbolic structures describing the behaviour of exceptions in the inventory of the grammar "To say that leaf (but not reef) voices f to v in the plural is simply to say that the composite symbolic structure leaves (but not reeves) is included among the conventional units of the grammar" (Langacker 1991b pg 19)

Starting with the notional descriptions of a noun as a thing, which profiles a "region in some domain" (Langacker 1991b pg 20), and a relation, which profiles the interconnections among entities, Langacker describes five basic classes of semantic structure or predications

1 Entity can be either a thing or a relation

2 Thing noun – "region in some domain", count noun – "bounded region in some domain"

3 **Simple atemporal** or stative relation profiles the interconnections between two or more conceived entities Adjectives and many prepositions have this character

4 Complex atemporal relation profiles a sequence of stative relations scanned in a summary manner In *John walked along the fence*, the preposition *along* designates a series of locative configurations defining the path of the trajector relative to the landmark

5 Processes complex temporal relations They define verbs and profile a set of relations whose trajector is always a thing Furthermore, the component

90

states of the process are conceived as distributed across time in a sequential (rather than holistic) manner

(a)	(b)	(c)	(d)	(e)
ENTITY	THING	STATIVE	COMPLEX	PROCESS
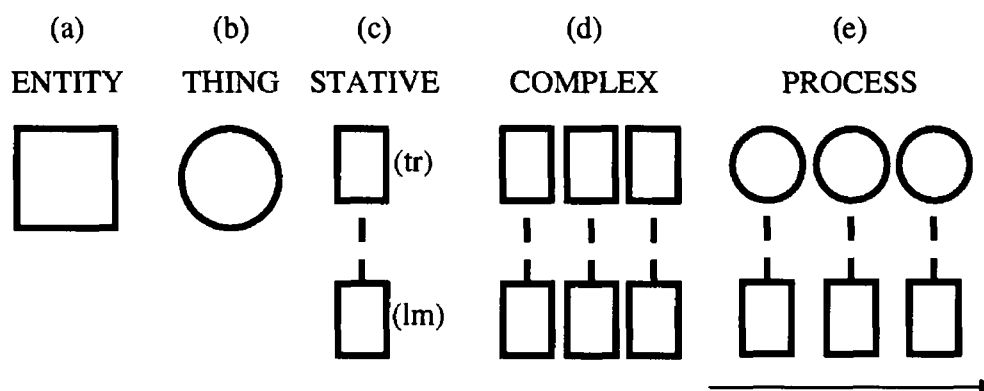
Figure 3-3 The abbreviatory notations for the basic classes of predications Based on an illustration in (Langacker 1991b pg. 23) The tr and lm symbols in diagram (c) stand for trajector and landmark respectively

Langacker's cognitive linguistic model propounds that these five classes of predication are sufficient to categorise any linguistic expression Encoding a complex expression is a process of integrating component structures to create a more elaborate composite structure which may in turn be integrated as a component of another even more elaborate composite structure This integration involves establishing correspondences between the component structures "The composite substructure is obtained by superimposing the specifications of the corresponding substructures" (Langacker 1991b pg 24) Figure 3-4 outlines the core structures and relationships in a grammatical construction Here [SEM3/PHON3] is the composite structure formed by integrating the component expressions [SEM1/PHON1] and [SEM2/PHON2] Four symbolic relationships are indicated in Figure 3-4

"S1 and S2 hold between the semantic and the phonological poles of each component expression, whereas S3 indicates that the composite phonological structure symbolizes the composite semantic structure The fourth relationship, S1,

91

reveals an important sense in which grammar is said to be inherently symbolic the integration of component structures at the phonological poles serves to symbolise the integration of the corresponding component structures at the semantic pole " (Langacker 1991b pg 23-24)
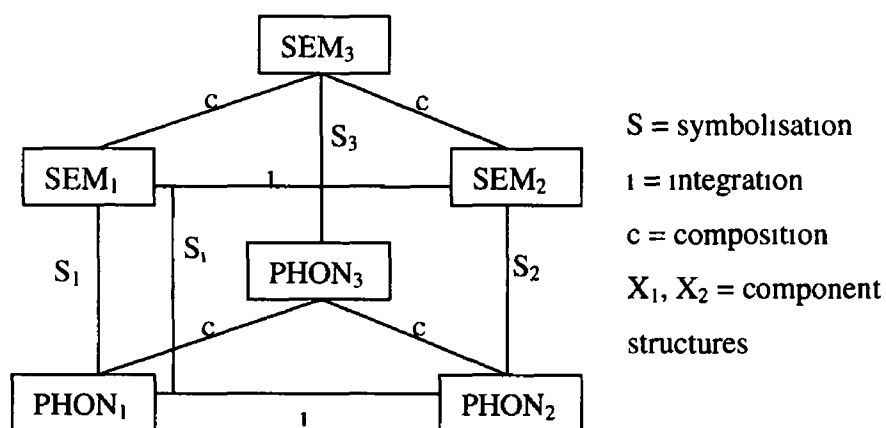
Figure 3-4 The essential structures and relationships in grammatical construction
Based on Figure 11 (Langacker 1991b pg. 24)

On this account, semantics cannot be fully compositional Often a composite substructure for a novel expression will refer to a domain or specification not discernible from the component substructures, or other conventional units Here, the contextual use of the expression will elucidate the meaning Once the expression has been assimilated into the grammar, the required contextual information will be incorporated into its conventional semantic value While allowing this proviso toward contextual semantic analysis, Langacker posits constructional schemas describing conventional patterns of composition that delineate the main aspects of a composite structure's organisation A pertinent example is the schema for prepositional phrase construction
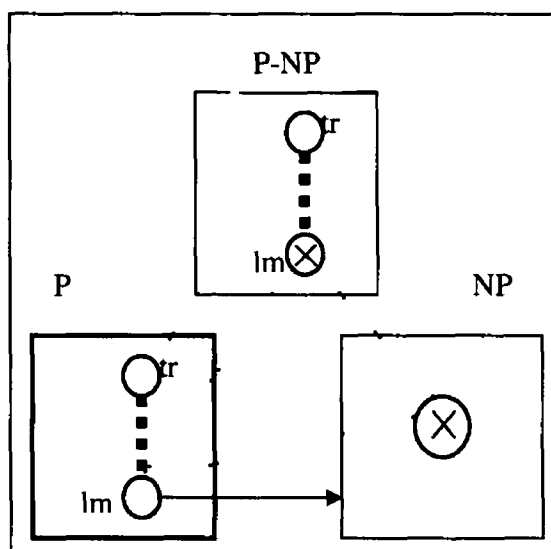
Figure 3-5· **Graphical representation of a possible construction schema for English prepositional phrases. Based on Figure 12 (Langacker 1991b pg. 25).**

The composite structure's phonological pole organises the linear sequence of the phonological structures of its components, here, a preposition and a noun phrase Figure 3-5 depicts schematically the integration of the component structures resulting in the composite semantic pole Here, a horizontal correspondence (represented by dotted lines) between the preposition's (schematised as a stative relation) landmark (designated by lm), and the noun phrase (schematised as a thing) is established Furthermore, there is an asymmetry in the degree of specificity with which the prepositional phrase and the noun phrase characterise the component structures The landmark in the preposition's predication is schematic relative to the thing profiled by a noun phrase The arrow between the preposition landmark and the whole noun phrase predication represents a relationship of schematicity The integration process, based on the horizontal correspondence, results in a vertical correspondence between the elements in the composite structure and the component structure The composite structure profiles the relational predication representing the preposition Langacker uses the term profile determinant to label a component structure that lends its profile to its composite In the above example the preposition's predication is the profile determinant for the composite prepositional phrase This is indicated in the diagram by outlining the box enclosing this

93

predication in heavy lines Within the model, these schemas are used for encoding and decoding novel expressions and as structural descriptions toward categorising processes A substructure that elaborates a component of another substructure within the same composite structure is described as being conceptually autonomous, while the elaborated component is described as being conceptually dependent In an instantiated case of the above schema, the noun phrase is conceptually autonomous as it elaborates an element of the prepositional schema A composite structure resulting from a constructional schema at one level of organisation can be used as a component at a higher level It may or may not be the profile determinant of the higher level composite structure

Using Langacker's approach, a constructional schema for locative expressions, with the structure NP-P-NP, is developed here This involves integrating the composite structure resulting from the preposition phrase construction schema described above in Figure 3-5 with a noun phrase A diagram of the overall schema is shown in Figure 3-6 This diagram follows the same conventions as Figure 3-5 The correspondences between entities are represented by dotted lines, the arrows represent relationships of schematicity, and the profile determinants at each layer of integration are explicitly indicated by outlining the box enclosing the component predication in heavy lines

As in the above example, the composite structure phonological pole organises the linear sequence of the components Another parallel between this schema and the earlier example is that the resulting composite schema describes a stative relation However, in contrast with the prepositional phrase constructional schema, where there is only one layer of integration, the schema in Figure 3-6 has two layers of integration forming a hierarchy This extra layer of integration allows the resulting expressions to fully specify the stative relations they describe The lowest layer in this hierarchy describes the integration of a preposition and an object noun phrase, this process is identical to Figure 3-5 The next layer in the hierarchy integrates a subject[21] noun phrase with the prepositional phrase Here, a horizontal correspondence between the prepositional phrase's trajector (designated by tr) and the noun phrase (schematised as a thing) is

---

[21] The terms subject noun phrase and object noun phrase follows Herskovits' (1986) terminology, see Section 2 3 1 for a definition of their respective syntactic positions in a locative expression

94

established There is a relationship of specificity between these two things which is illustrated by the arrow The integration process, based on this horizontal correspondence, results in a vertical correspondence between the elements in the composite structure and the component structures The profile of the resulting composite structure is identical to the profile of the object noun phrase, consequently, the object noun phrase is designated as the profile determinant for the composite structure This is indicated by the heavy lines outlining its predication
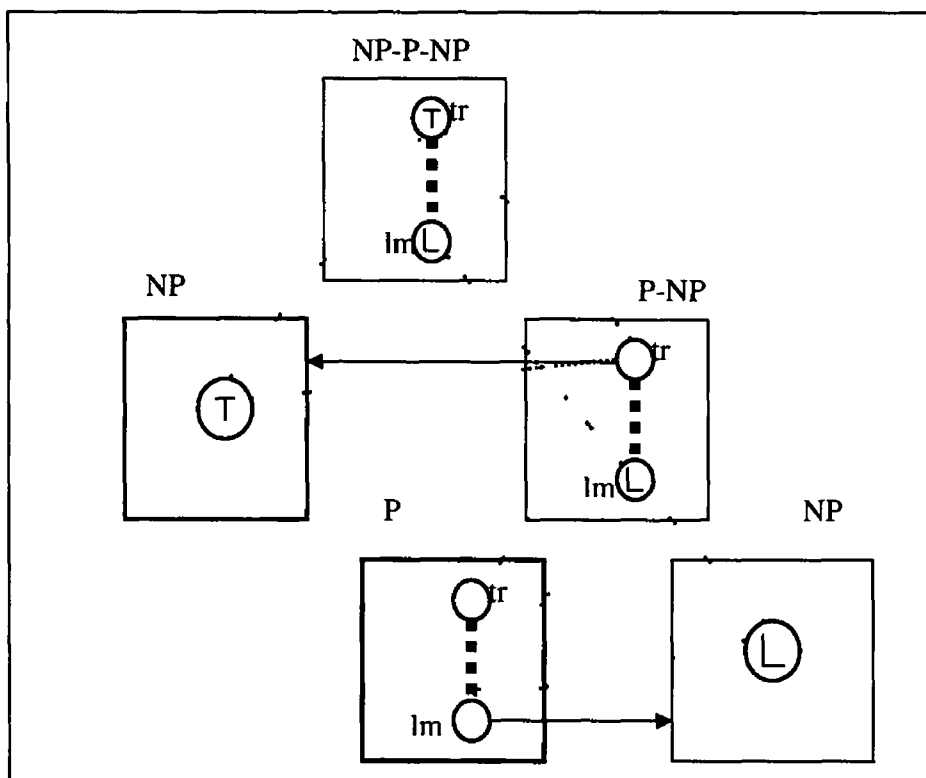
Figure 3-6 Representation of a possible construction schema for English locative expressions with an NP-P-NP structure[22]

---

[22] The use of the label P-NP to describe a prepositional phrase in Figure 3-6 follows Langacker's terminology In structuralist linguistics this structure would be analysed as NP => NP + PP, PP => P + NP

Figure 3-7 illustrates this analysis when applied to the phrase *the man to the left of the car* For diagrammatic convenience, the objects symbolised by the noun phrases *the man* and *the car* are represented by an idealised shape specification The structure and analysis of this diagram is identical to that of Figure 3-6
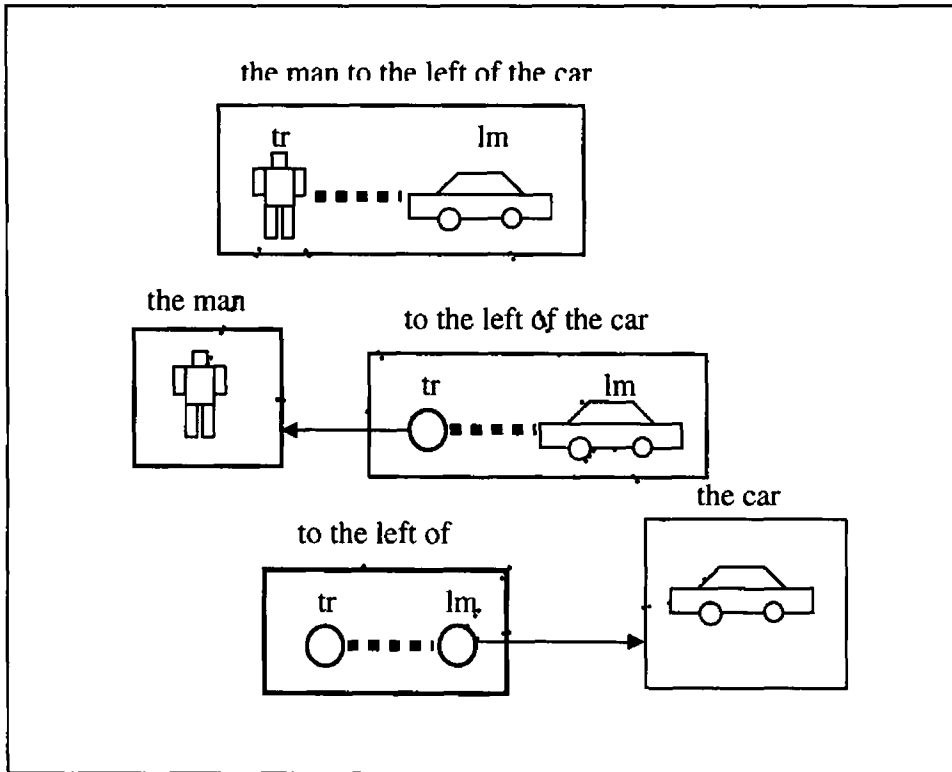
Figure 3-7. Graphical representation of a construction schema for *the man to the left of the car*.

Using the notion of conceptually autonomous and conceptually dependent constituents, Langacker gives definitions for the problematic grammatical notions of modifier and complement, "a 'modifier' is a conceptually dependent predication that combines with a head, whereas a 'complement' is a conceptually autonomous predication that combines with a head a construction's head can be identified with its profile determinant" (Langacker 1991b pg 29)

3.3 Cognitive Grammar Summary

By grounding meaning in conceptualisations, cognitive grammar admits perceptual factors to impact on the context of the discourse

"Meaning is equated with conceptualisation Linguistic semantics must therefore attempt the structural analysis and explicit description of abstract entities like thoughts and concepts The term conceptualisation is interpreted quite broadly it encompasses novel conceptions as well as fixed concepts, sensory, kinaesthetic, and emotive experience, recognition of the immediate context (social, physical, and linguistic), and so on " (Langacker 1991b pg 2)

This approach is convergent with the basic premise of this thesis, that to interpret spatial language, computational systems must model the perceptual context of the discourse However, the lack of formalisation in Langacker's theory is a major drawback for computational systems adopting this framework To be more specific, while the meaning of an expression is based on the relationship between a profile and a domain, both of which are mental conceptualisations, no formal characterisation of these conceptualisations is given In Section 4 4, the work of Salmon-Alt and Romary (2001), who have developed a context model which attempts to refine Langacker's theories by formally defining the constituency of referential domains, will be reviewed The SLI discourse model developed in Chapter 9 adapts and extends Salmon-Alt and Romary's (2001) work by integrating it with a model of visual perception and a perceptually based computational framework for interpreting projective locative expressions

4 Linguistically Inspired Models of Context

4.1 Introduction

In Section 2 4, the need for a discourse model as a component of any NL interface that aims to usefully interpret a user's commands was described The basic issues such a model must address were also delineated

- What information should the model carry forward that may be useful in resolving future references?
- How does the model attribute a referent to a given referring expression?

Some models try to address these issues through the use of rules based on discourse organisation Usually, this entails a procedure for linking pronominal reference to prior entities that have been explicitly mentioned in the preceding discourse This can be attributed to the linguistic focus of the models that propose it While this approach is suitable for many linguistic applications, the weakness of this approach is that often the information required to compute a unique referent for an expression is not always available from the linguistic record at the time the expression occurs in the discourse Examples are instances of one-anaphor (e g , *the other one, the blue one*) or locative expressions (e g , *the tree in front of it*) in situations where the referent is provided by the visual context but not represented in the linguistic context

Chapter 3 reviewed Langacker's (1987, 1991b, 1994) cognitive grammar The linguistic approach adopted by this thesis is framed within this paradigm The motivation for adopting this linguistic approach is the emphasis that cognitive grammar places on situating language within wider general cognitive faculties This emphasis makes the linguistic framework amenable to integration with models of other cognitive processes such as visual perception This chapter continues the theme of linguistic models by reviewing linguistically inspired models of context

98

While there have been many models of discourse, reviewing all of these is beyond the scope of this thesis Here, two of the best known discourse models are described – in Section 4 2, Discourse Representation Theory (DRT) (Kamp and Reyle 1993) and in Section 4 3 Centering Theory (Grosz *et al* 1995) In Section 4 4, a more recent cognitively based (as distinct from purely linguistic) model (Salmon-Alt and Romary 2001) is described Section 4 5, contains a critical review of the models described and an explanation of how the (Salmon-Alt and Romary 2001) model will be extended for this thesis

4.2 Discourse Representation Theory (DRT)

The term Discourse Representation Theory (DRT) originated in the area of dynamic interpretation of natural language Initially, it referred specifically to the work of Hans Kamp, however, it has gradually become a generic term encompassing several frameworks of dynamic natural language interpretation The core features of this approach are (1) the construction of a global context comprising all potential referents introduced into the discourse and (2) that each new sentence in the discourse is interpreted against and in terms of the contribution it makes to the context In DRT, the global context model is represented as a discourse representation structure (DRS) Importantly, the result of processing a discourse entity in the context of a representation structure R is a new representation structure R' which may be viewed as an updated version of R (van Eijck 1994) The processing of a discourse is incremental, with the context for each new sentence being taken as the representation structure that resulted from processing the previous sentences

A discourse representation structure consists of two parts a finite list of reference markers and a finite list of conditions on reference markers (van Eijck 1994, Salmon-Alt and Romary 2001) Reference markers are similar to variables In the simplest case, these are introduced into the context by the processing of nominal phrases Each discourse marker has a scope which (a) depends on the form of noun phrase that introduced it and (b) limits the accessibility of later discourse entities to take it as an antecedent Indefinite

noun phrases introduce new reference markers Definite noun phrases and anaphoric pronouns also introduce new reference markers, however, these markers are always linked to appropriate antecedent discourse markers Any reference marker accessible to a pronoun may serve as its antecedent The set of reference makers which may act as the antecedent for an element in the current discourse structure may be roughly approximated to the markers in the current structure, plus the markers in any encompassing structure (van Eijck 1994 pg 976)

4.3   Centering Theory

Centering Theory (Grosz *et al* 1995) is based on the observation that task-orientated dialogs have an inherent structure based on the task, each subtask having in effect its own sub-dialog Moreover, the uses of anaphora within these dialogs reflect this structure, "anaphors often found their antecedents in much earlier dialogs on the same subtask, even if another subtask intervened" (Hirst 1994 pg 3489)

In Centering Theory, the overall dialog with its principle goal is decomposed into a hierarchy of subdialogs or discourse segments, each with an associated subgoal The segmentation of the dialog is based on linguistic cues such as keywords (e g , *OK*, *So*, *Anyway*) or the location of the anaphoric antecedents within the focus area encompassing the current segment Each subdialog is viewed as a focus space making available possible antecedents for anaphora within the dialog The completion of a subdialog enables "anaphoric reference to entities on 'the next level up'" (Hirst 1994 pg 3489)

The theory attempts to relate three components of discourse linguistic structure, intentional structure, and attentional structure The linguistic structure is based on the segmentation of the discourse into constituent discourse segments An embedded relationship may hold between two of these segments The intentional structure is based on the relationship between the overall goal of the dialog and the subgoals of the sub dialogs The attentional state represents the discourse participants' "focus of attention at any given point in the discourse" (Grosz *et al* 1995 pg 4)

The notion of discourse coherence is at the core of Centering Theory Indeed, Grosz *et al* view coherence as the defining characteristic of discourse "for a sequence of utterances to be a discourse, it must exhibit coherence" (1995 pg 5) Coherence may approximately be defined as the amount of inference required of a hearer or reader to successfully interpret the discourse There are two factors that affect it (1) changing of 'aboutness' make discourse less coherent, (2) different types of referring expressions and different syntactic forms make different inference demands on a hearer or reader (Grosz *et al* 1995)

The model distinguishes two forms of coherence local and global Local coherence pertains to coherence among the utterances in that segment Global coherence describes the coherence between segments of the discourse The attentional state models these forms of coherence by separate components Global coherence is modelled by a stack, "pushes and pops of focus spaces onto the stack depend on intentional relationships" (Grosz *et al* 1995 pg 4) Each element in the stack includes the salient objects in the discourse segment New items are pushed on the stack whenever a discourse segment is begun, items are popped off the stack when a segment's goal is completed Local coherence is modelled by the partial ordering of entities with a particular focus space and the links between them The ranking of entities reflects their relative prominence and is primarily based on their grammatical role

The linking between utterances is based on the concept of an utterance's center Centers are semantic objects that are constructed through discourse

"Each utterance U in a discourse segment (DS) is assigned a set of *forward-looking centers*, $C_f(U, DS)$, each utterance other than the segment initial utterance is assigned a single *backward-looking center*, $C_b(U, DS)$ The backward-looking center of utterance U_{n+1} connects with one of the forward-looking centers of utterance U_n " (Grosz *et al* 1995 pp 8-9)

Following on from this linking mechanism, a set of centering constraints is proposed These constraints govern aspects such as the form of referring expressions used in an utterance and the type of preferred transitions between utterances It is claimed "that

101

to the extent a discourse adheres to centering constraints, its coherence will increase and the inference load placed upon the hearer will decrease" (Grosz *et al.* 1995 pg. 11).

4.4   Salmon-Alt and Romary

Salmon-Alt and Romary (2001) propose a cognitive approach (as opposed to purely linguistic models) to reference resolution. The cornerstone of this framework is the "assumption that all reference is accomplished via access to domains of reference – restricted sets of contextually available entities, structured by contrasts – rather than by direct linkage to the entities themselves" (Salmon-Alt 2001 pg. 1). Adopting such an approach transforms the scope of a referential expression from the list of previous discourse referents to the set of elements within a local contextual domain. In parallel with this transformation in form, the purpose of the context model changes from maintaining the list of previous discourse entities to creating these contextual domains and furnishing them to the interpretation module. The advantage of such an approach is the ability of the model to handle different types of referring expressions with a single access mechanism (Salmon-Alt 2001).

4.4.1   Context Model

The basic unit of the context model is the reference domains. Salmon-Alt and Romary describe these as "mental representations for entities to which it is possible to refer" (2001 pg 289). Discourse or perceptual information may trigger a creation of a reference domain. A reference domain minimally contains type information. This information is derived from generic domains containing knowledge that is assumed to exist prior to discourse (Salmon-Alt and Romary 2001). This minimal configuration may be extended by one or more partitions.

The partitions within a domain represent possible decompositions of the domain. Each partition is unique within a domain with respect to the perspective of the domain the partition models. The term differentiation criterion is used to describe the particular

102

perspective represented by a partition These criteria may be based on previous discourse information, perceptual information, or conceptual knowledge Salmon-Alt and Romary claim that, as each partition models a particular perspective on the elements within a reference domain, each partition "predicts a particular referential access to its elements" (Salmon-Alt and Romary 2001 pg 289)

One element within each partition may be profiled This profiling of a reference domain's element is inspired by Langacker's Cognitive Grammar (Section 3 2) Profiling an element denotes it as the primary element in the partition with respect to perceptual or discursive salience There are two operations that result in an element being profiled grouping and extraction The first of these, grouping, results in the creation of a new reference domain within the context model The grouping operation combines two or more domains into a more complex domain with a partition for the grouped elements The goal of this operation is to create new domains and make these available to the interpretation process The grouping operation may or may not result in profiling an element The other operation, extraction, is best explicated within the description of the interpretation process in Section 4 4 2 below

The above context model, thus, supplies the interpretation process with one of three fundamental structures (a) domain without any partition, (b) domain with a partition but without a profiled element, (c) domain with a partition containing a profiled element

4.4 2   Interpretation Process

Salmon-Alt and Romary's (2001) reference resolution framework defines an interpretation process for indefinite, definite, pronominal, and demonstrative referring expressions There are three stages to the interpretation process The first stage is the calculation of an underspecified domain that represents the referring expression that is to be interpreted There are two criteria which impact on the construction of the underspecified domain for a given noun phrase (a) an elaboration of an abstract semantic schema for determiners by the semantics of the current determiner and (b) an elaboration

of the abstract semantic schema for nouns, by the semantics of the current expression (Salmon-Alt and Romary 2001)

The second stage of interpreting an expression is a selection process that extracts one or more domains according to their activation level from the context model based on their compatibility with the current expression's underspecified domain This compatibility depends on criteria such as type, cardinality, differentiation criteria, etc (Salmon-Alt and Romary 2001)

Once a suitable domain has been identified within the context model, the referent of the expression must be extracted and profiled (Salmon-Alt and Romary 2001) This is the final stage of interpretation and is achieved through a restructuring of the domain The determiner of the expression being evaluated guides this operation For example, in the case of a "definite expression *the N*, the item of the suitable type N is identified and profiled within the existing partition of the domain of reference" (Salmon-Alt and Romary 2001 pg 295)

In summary, the interpretation process consists of a unification of the underspecified domain, representing the expression, with a compatible context dependent domain of reference and a profiling of an element within that domain (Salmon-Alt and Romary 2001) The profiling of an element denotes it as the referent of the expression

4.5 Linguistically Inspired Discourse Models Summary

Both DRT and Centering Theory can be described as predominantly linguistic approaches to modelling discourse This is evident in the omission of explicit references to extralinguistic contextual sources of information in the models Salmon-Alt and Romary (2001) describe several consequences of this assumption

1 These models have an intrinsic preference for locating the referent of an expression in the previous discourse entities, such an approach has difficulty handling references which draw on perceptual information (e g , demonstratives, one-anaphora, other-expressions, etc)

2 Their context models are primarily the set of previous referents within the discourse, consequently, they omit referents that are visually or conceptually accessible to the interlocutors

3 Their context models are only updated for each utterance in the discourse, thus, they neglect changes in the perceptual context of the discourse which may introduce new referents into the purview of the interlocutors

Here, in this thesis, models such as DRT or Centering Theory, which neglect visual perceptual information, are considered sub-optimal as discourse models for NLVR systems Although the conceptual domain of a speaker subsumes such a complex and vast network of knowledge that is impossible to model computationally in its entirety, it is conjectured that computational systems, which situate a user's linguistic interaction with the system in a visual 3-D environment, ground the domain of this interaction within the visual model Given this, it is possible to model the perceptual domain to a certain granularity and abstraction Moreover, to neglect to do so omits an important knowledge source from an interpretive language system

The cognitively based approach of Salmon-Alt and Romary (2001) admits visual perceptual information into their context model by allowing it to trigger certain events (i e , domain creation, the grouping operation, and partitioning) As a result, this model is compatible with this thesis However, the model gives no description of how to computationally gather visual perceptual information or how this visual perceptual information is to be combined with the linguistic information when resolving references In Chapter 9, a discourse model is developed which adapts and extends (Salmon-Alt and Romary 2001) in order to overcome these shortcomings

5   Previous Work

5.1   Introduction

This chapter provides a critique of previous computational work relevant to this thesis Section 5 2 critically reviews previous models of visual attention, beginning with an examination of approaches to vision developed for robots and continuing with models of vision based on graphics techniques, including ray casting models and false colouring models

Section 5 3 focuses on previous work related to resolving locative expressions There are two parts to this section Section 5 3 1 introduces previous work on frames of reference, Section 5 3 2 examines prior approaches to modelling the spatial template of a preposition Section 5 3 1 begins by reviewing the literature on frames of reference, including a selection of linguistic and psycholinguistic work (Carlson-Radvansky and Irwin 1993, Carlson-Radvansky and Irwin 1994, Carlson-Radvansky 1996, Levelt 1996, Taylor *et al* 2000) In this review of the frame of reference literature, a novel analysis of the findings in (Carlson-Radvansky and Irwin 1993) is proposed This analysis allows the definition of a threshold for resolving the competition between reference frames along the vertical axis Section 5 3 2 begins with a critique of neat models and the issues affecting them and concludes with a critical review of Herskovits' (1986) multiple relational model Subsequent to this, the scruffy or continuum models are presented, including models proposed by (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994, Fuhr *et al* 1998, Mukerjee *et al* 2000)

Section 5 4 critically reviews previous computational systems that have integrated language and vision Within this critique, a detailed description and review of the scruffy models introduced in Section 5 3 2 2 is given The systems reviewed in Section 5 4 include SHRDLU (Winograd 1973), CITYTOUR (Andre *et al* 1986, Andre *et al* 1987), CSR-3-D (Gapp 1994a), SPRINT (Yamada 1993), WIP (Olivier *et al* 1994, Olivier and Tsuji 1994), the Situated Artificial Communicator (Socher and Naeve 1996, Socher *et al*

1996, Vorwerg *et al* 1997, Fuhr *et al* 1998), Virtual Director (Mukerjee *et al* 2000), and CommandTalk (Dowding *et al* 1999, Stent *et al* 1999, Goldwater *et al* 2000) None of these systems provide an adequate semantic model for locative expressions since they lack an algorithm for selecting a frame of reference which is congruent with the psycholinguistic evidence and a semantic model for prepositions which address the issues highlighted in Chapter 2, Sections 2 3 4 and 2 3 5 Furthermore, none of these systems propose a suitable model of user visual perception and consequently do not provide a discourse model that integrates visual perception with linguistic knowledge

5.2 Spatial Attention and Models of Visual Perception

Section 2 2 2 examined some of the aspects of perception that pertain to modelling vision, in particular how attention affects the human awareness of what people perceive It was noted that spatial attention is the most commonly used visual filtering mechanism There are many computational models of vision that use this abstraction, most have been developed for robot navigation These systems proliferated because, until recently, computers were not able to process an entire camera frame within the relevant time constraints and the best optimisation was to select only a portion of the input to process (Hewett 2001) More importantly, however, the proliferation of developed systems based on the abstraction of spatial attention attests to the viability of using this generalisation of perception

The following sections will focus on some of the previous models of vision and visual attention As noted above, one field of research where several models of vision have been developed is robotics However, although it might appear that the models of vision developed for intelligent robots should be similar to those for avatars in 3-D graphics systems, they are quite different This is because in a rendered environment the visual scene is already modelled A consequence of this is that, when developing a computational model of vision that is to be used in a rendered environment, the issues of pattern recognition, distance detection, and the binding problem (which are some of the most difficult issues facing robotic vision (Renault *et al* 1990)) may be ignored

Nonetheless, this review of vision models will begin with a general overview of these robotic systems, the purpose of which is to highlight the differences between the approach to vision adopted by robotic systems and the type of architecture required by 3-D simulated graphic environments Following this, the review will focus on models of visual perception that use 3-D graphics techniques These models will be classified based on the graphics techniques they use ray casting and false colouring

Finally, most of the graphics based models of vision were developed as part of behavioural animation frameworks, the idea being that if virtual characters are enabled to react to their environment with intelligent appearing behaviours, the input required by the animator in creating the simulation would be reduced The general approach adopted by these frameworks is to use the output of their visual model to create a visual memory or cognitive map of the environments Because the SLI framework proposed in this thesis uses the results of its visual model to create a visual memory of the environment that functions as part of its context model, the approaches to visual memory developed by these animation systems is of some relevance Consequently, where a system has a visual memory as part of its design, the review is extended to include this component

5 2 1   Robotic Vision

Hewitt (2001) reviews several of the robotic attention systems Nearly all of these systems have a connectionist or neural net architecture This form of system requires training As a result, these models are restricted to the domains described by the training set given to the system For example, connectionist navigational systems trained with images from the inside of a factory would need to be retrained to handle a forest environment A system that requires retraining when shifting from one visual domain to another is not suitable as a model of rendered environments which may change drastically from program to program or even within the one application

5 2 2   Ray Casting Models

Tu and Tersopoulos (1994a, 1994b) implemented a realistic virtual marine world inhabited by autonomous artificial fish These fish exploited a rudimentary perceptual system which was part of an overall animation framework that enabled them to autonomously generate realistic patterns of behaviour While Tu and Tersopoulos's (1994a, 1994b) animation framework dealt with physics, locomotion, behaviour, and perception, it is their model of visual perception is most relevant for this thesis

Each fish in the simulation was equipped with a cyclopean vision sensor with a visual area extending to a 300° spherical angle The depth of the visual field was dependent on the visibility of the water in the simulation An object was deemed to be visible to a fish if any part of it entered the view volume of the fish's visual sensor and was not fully occluded by another object The model used a graphics technique called ray casting to determine if an object met the visibility conditions Ray casting is widely used in offline rendering of graphics However, it is not used in real-time rendering due to its computational cost It allows a programmer to calculate the relationship between a location on a 2-D scene and the contents of the 3-D world as viewed from a particular camera location While the maths behind **ray casting** is complicated, its operation can be described as casting a ray (i e , drawing an invisible line) from one point in the 3-D world in a certain direction and then reporting back all the models that the line intersected Figure 5-1 illustrates ray casting, in this figure, one of the rays cast intersects with the house while the other ray intersects with the tree
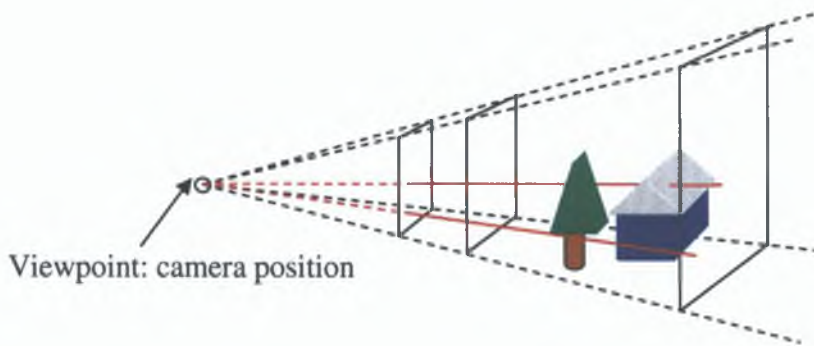
Figure 5-1: Two rays are cast from the viewpoint and intersect with objects in the view volume.

5.2.3 False Colouring

Another graphics based approach to modelling vision was proposed by Renault *et al.* (1990). As with Tu and Tersopoulos's (1994a; 1994b) model, Renault *et al.*'s approach was designed as an aid to behavioural animation of synthetic actors. Their visual model was dependent on the functionality offered by the IRIS 4D Workstation architecture and graphics engine, in particular the z-buffer[23] and double frame buffering[24]. Their idea was to use a perspective projection[25] of the actor's view volume to create a 2-D array of pixels representing the actor's view of the world. Each pixel contains the distance from the actor's eye to the point drawn at that pixel and an identity number for the object drawn at that pixel. The distance between the eye and the point of the object that the pixel represents is extracted from the z-buffer. The front buffer of the

---

[23] The z-buffer or depth buffer is an area in graphics memory reserved for storing the depth value of each pixel. A pixel's depth value is the depth that the object being drawn at that pixel is from the viewpoint.

[24] Double frame buffering requires two drawing frames to be available to the graphics engine, conventionally called the front and back buffers. The front buffer is the one that is displayed while the back buffer is where the images to be displayed are updated. At each frame the content of the buffers are swapped. This architecture speeds up the rendering by allowing a system to refresh the displayed image and render the next image at the same time.

[25] In a perspective projection distant objects appear smaller than nearby objects on the projection plane.

double frame "is used to display the projection objects, which allows the animator to know what the synthetic actor sees In the back buffer the object identifier for each pixel is stored" (Renault *et al* 1990 pg 6) At the end of the process, this array contains the identity numbers for all the objects currently visible to the actor and the distance between the actor's eye and the different points of the object that are currently visible The array was then passed to a Displacement Local Automata (DLA) which used this information to navigate the synthetic actor DLA are similar to scripts in Schank's (1973) framework The purpose of a script is to contain specific knowledge corresponding to frequent situations For example, the restaurant script which describes the actions people perform when they go to a restaurant Examples of the type of DLAs developed by Renault *et al* (1990) were *follow-the-corridor* and *avoid-the-obstacle* The DLAs were strictly algorithmic, they corresponded to reflexes and their information source was restricted to the visual information available in the current view While the current view could be described as a short term visual memory, there was no data structure in the system that functioned as long term visual memory

Noser *et al* (1995) extended this model by using **false colouring** to represent the identity numbers of the objects in the world and an octree[26] data structure as a long term visual memory for their characters As with Renault's earlier work (Renault *et al* 1990), this model used graphics hardware of the Silicon Graphics IRIS architecture to implement their synthetic vision Their goal was to allow actors to explore an unknown environment and to build a cognitive map from this exploration based on their visual experiences "While or after the maps are built, the actor can do path-planning, navigation, and place-finding" (Noser *et al* 1995 pg 143) The vision module is comprised of a modified version of the world being fed into the system's graphics engine and scanning the resulting image In brief, each object in the world is assigned a unique colour or "vision-id" (Noser *et al* 1995 pg 149) This colour differs from the normal colours used to render the object in the world, hence the term false colouring An object's false colour is only used when rendering the object in the visibility image off-screen, and

---

[26] An octree is a data structure that divides a 3-D space into different regions The partitioning of the space is done by planes parallel to the coordinate axes, and each step of the partitioning subdivides space into 8 octants

does not affect the renderings of the object seen by the user, which may be multi-coloured and fully textured Then, at a specified time interval, an unlit model – **flat shading**[27] – of the character's view of the world, using the false colours, is rendered The motivation for flat shading the false colour image is to avoid shadows changing the vision-id colour Once the drawing is finished, the viewport[28] is copied into a 2-D array along with the z-buffer values By scanning the array and extracting the pixel colour information, a list of the objects currently visible to the actor can be obtained The z-buffer values allow the system to extract the distance to each of the visible objects Noser *et al* (1995) used an octree data structure or occupancy grid as a visual memory store for their synthetic actors Each node in an octree data structure contains a list of eight pointers to child nodes Each of these child nodes represents a subdivision of its parent node Noser *et al* (1995) used this octree structure to create a topological map of the environment based on the information supplied by the vision model To do this, they took each pixel in the 2-D array, calculated its 3-D position in the octree space and then inserted the object ID into the octree at that location While this form of data structure is extremely efficient for topologically based applications (such as path finding or place finding) an octree approach does not easily lend itself to the structuring of information in a non-topological manner A pertinent example is linguistic applications knowing the Cartesian location of an object does not help when resolving anaphoric references to the object

---

[27] The shading that is applied to a polygon defines what the surface of a polygon looks like when it is rendered It defines the colour or texture of the polygon and how the polygon's surface reflects light In flat shading only one vector is used to compute how the surface of the polygon reflects light This means that the colour of the surface of the polygon will be constant across the whole polygon, i e , there will be no shadows computed If the false colour rendering did not use flat shading the shadows computed by the rendering engine and applied to the surfaces of the model's polygons would distort the colour each model was drawn with This would interfere with the computation of the number of pixels the model covered as the colour of these pixels would vary across the model

[28] A viewport is the rectangular area of the display window It can be conceptualised as the window onto the 3-D simulation

Another navigation behavioural system that used false colouring synthetic vision was proposed by Kuffner and Latombe (1999) As with the previous systems, each object in the world is assigned a unique ID "To check which objects are visible to a particular character, the scene is rendered offline from the character's point of view using flat shading and the unique colour for each object as defined by the object ID" (Kuffner and Latombe 1999 pg 120) The resulting image is scanned and a list of currently visible objects is obtained by analysing the picture colour information Where Kuffner and Latombe's (1999) work differs from Noser *et al* 's (1995) is in their model of visual memory While Noser *et al* (1995) used an octree to store the location of perceived objects, Kuffner and Latombe (1999) used an array indexed by the objects unique ID for faster performance Each character maintains an array containing a set of observations built incrementally from the output of the vision module Each element in this array describes one object and consists of a **tuple**[29] with components for the object ID, properties of the object, 3-D transformation of the object, velocities of object, and observation time This array maintains a list of the last observed state of objects which have been perceived by the character, representing the characters' visual memory of the environment After each update, the character invokes a navigation path-planning module that uses the array as its only information source for obstacles Kuffner and Latombe (1999) describe several update rules that allow the system to recognise when an object has been moved or removed and updates the memory array accordingly This system allows the characters to remember all the perceived objects until observational data indicates that the object has been removed from the world While this memory model is efficient in terms of access speed for individual objects or for the location of all the objects in the world, its flat organisation of information in an array is not suited to linguistic applications where objects may be identified by type, physical attribute, or linguistic context, e g , *the house*, *the tall red house*, or *the other one*

The final model of vision reviewed was proposed by Peters and O'Sullivan (2002) Peters and O'Sullivan integrated their vision model as part of a goal driven memory and attention model which directed the gaze of autonomous virtual humans Its primary

---

[29] An entity or set with a given number of elements

113

function is to locate particular objects in the scene and to direct the gaze of the virtual viewer at the located object The vision model uses false colouring, however, it extends previous false colouring vision models by providing multiple vision modes Each object is assigned a unique ID, however, a different palette is used to map the ID to colour for each vision mode The different vision modes are useful for capturing different types of information about the environment The two main vision modes are distinct mode and grouped mode Distinct mode is identical to the approaches described above Each object is rendered in a unique colour, and the image is then scanned and the pixel colour information used to look-up the object's globally unique identifier in the world model This mode allows the system to update the information about individual objects in the world The list of identified object IDs is passed to the memory model In grouping mode, objects are false coloured with group colours rather than individually Objects may be grouped according to a number of different criteria shape, proximity, or type

The system's memory model consists of three different modules short-term sensory storage (STSS), short-term memory (STM), and long term memory (LTM) Each module consists of a list of memory entries These consist of an observation and how many times the memory has been rehearsed An observation is one output from the vision module It may describe a single object or a group of objects and is represented by a tuple that is composed of an object ID, the azimuth of the object, the elevation of the object, the distance to the object, and a time stamp A rehearsal occurs if an observation of an object is in STM and the object is observed again An agent maintains at most a single observation per object This observation corresponds to the last perceived state of the object The STSS is updated at each refresh of the viewpoint rendering It comprises the observations extracted from the synthetic vision module The system permits a large number of items to be stored in the STSS It should be noted, however, that these items may represent groups of objects in the world rather than individual items This chunking of information in the STSS reduces the throughput of objects into the STM The STM is limited to storing eight observations "Memory entries are removed from the STM under two conditions they are displaced by newer memories when the STM is full, and they also decay over time (forgetting)" (Peters and O'Sullivan 2002 pg 24) A default time of 20 seconds is allotted to each observation in STM, however, in instances where the

114

memory entry is rehearsed, the allotted time may be extended to 20 minutes Memory entries that are in the LTM module do not expire Memories are transferred from the STM module to the LTM module base on repeated exposure "The LTM contains encode (add memory), decode (retrieve memory), and recall (query memory) functions When an item is retrieved from LTM it is moved into the STM, overwriting anything currently in the STM" (Peters and O'Sullivan 2002 pg 25)

Peters and O'Sullivan's memory and attention process is initiated by giving a location task to the virtual human This task consists of a command that contains an object ID to the virtual human If an observation of the object is already memorised in either the STM or the LTM, the observation information is extracted and the virtual human looks at the object and updates its perception of the object using distinct mode rendering If the object is not in the STM or the LTM, then the virtual human's perception of the environment is rendered using grouping mode based on proximity The virtual human then renders each of the objects in the resulting observations in group mode based on type If an object of the same type as the goal object is found, the virtual human checks to see if it is the goal object If it is the goal object, the perceived state of the object is loaded into the STM, if not, the search continues through other objects of similar type in the group and, in the case where there are no more, the search proceeds to other groups

While this framework allows the system to identify objects in the scene, it is restricted to locating objects which are identified by their unique ID There is no description of object attributes such as colour or relative height in the memory model Moreover, Peters and O'Sullivan give no mechanism for integrating a linguistic context model into the system

5 2.4 Spatial Attention and Models of Visual Perception Summary

The review of work related to computational models of vision began by noting that connectionist architectures developed as vision modules for robots are less suitable for avatars in rendered 3-D environments There are two reasons for this firstly, the differences between the issues facing robotic vision and those pertaining to synthetic vision in simulated environments, secondly, the training required by connectionist architectures makes them impractical for applications which have a broad range of inputs After reviewing the connectionist models of vision, the focus of the review of visual models shifts to those models that are based on graphics techniques First Tu and Tersopoulos' (1994a, 1994b) ray tracing model of vision for synthetic fish was reviewed In the description of this model, it was noted that ray tracing is a computationally expensive function and is not used in real-time rendering Next, models of vision that use a false colour approach were examined The kernel of this technique was developed by Renault *et al* (1990) This initial framework was further developed by Noser *et al* (1995) by adding the false colouring rendering and an octree data structure that functioned as a visual memory for avatars More recently, this model of synthetic vision was adopted by Kuffner and Latombe (1999) and Peters and O'Sullivan (2002) It is important to note that although several of these models (Noser *et al* 1995, Kuffner and Latombe 1999, Peters and O'Sullivan 2002) use the output of the visual module to create a visual memory or cognitive map of the simulated environment based on what the avatar has observed, none of these systems use this information as an aid to interpreting language Indeed, the data structures used in these systems are not suitable as inputs to a linguistic context model Furthermore, although Peters and O'Sullivan's (2002) approach has a limited model of attention based on the number of times an object has been observed, the majority of the false colouring models make no attempt to rate the saliency of the observed objects Renault *et al* (1990) simply notes what objects have been observed and the distance to each observed object from the viewpoint, Noser *et al* (1995) stores a list of the observed objects and their Cartesian location in the world, Kuffner and Latombe (1999) note the observed objects and some of their properties location, velocity,

116

observation time, type, etc Moreover, the model of attention proposed by Peters and O'Sullivan's (2002) is only updated when the avatar attends directly to a goal object that it has searched for and only updates the observations pertaining to that goal object This approach to attention is not ideally suited as a method for creating a visual memory of an environment which may be used as part of a linguistic context model because it requires the avatar to search the environment for objects which it may have already perceived but not attended to directly and thus not noted

In Chapter 7, a model of vision is proposed that uses the false colouring approach described above The model is designed for use as an interface between rendered environments and a linguistic interpretive module In order to adapt the false colouring model of synthetic vision to this novel application, it was necessary to extend the model to rate the observed objects based on their saliency within the viewed scene Furthermore, the output of the proposed visual module is organised in a manner that is optimal for integration with the SLI discourse model developed in Chapter 9 Finally, it should be noted that the separation of the perceptual mechanism from the linguistic interpretive module admits the possibility of replacing the perceptual module with a more refined version at some later date

5.3 Locative Expressions

Section 2 3 introduced the concept of a locative expression After highlighting the importance of this linguistic construct within spatial language a general algorithm for interpreting this form of expression was defined The first stage of this algorithm consisted of selecting the object from the environment that the speaker intended as the landmark Section 2 3 2 reviewed the key issues at this stage of the interpretive process To summarise briefly, the process of landmark selection can be described as extracting the most salient object matching the noun phrase in the object position of the expression However, this is exactly what a general model of reference should achieve A consequence of this position is that the review of previous approaches to reference

117

resolution covers the most relevant work pertaining to landmark selection This section reviews previous work related to the second and third stages of the interpretive algorithm

- the selection and imposition of a frame of reference on the landmark
- the modelling of the spatial template of a preposition

5 3 1 Frame of Reference

If a locative expression contains a projective preposition, the second stage of the interpretive process is the selection of a frame of reference (see Section 2 3 3) In some instances, the frame of reference is made explicit in the linguistic input More often, however, the intended frame of reference is implicit in the statement In these situations, a process for selecting a frame of reference is required In general, previous linguistic interpretive systems adopted one of four approaches to handling issue of frames of reference

- Situate the discourse in domains where only simple objects with no intrinsic reference frame associated with them are modelled, e g , the SHRDLU system (Winograd 1973)
- Assume a default frame of reference and force the user to adopt this for all input, e g , the Virtual Director system (Mukerjee *et al* 2000) defaults to the intrinsic frame of reference if the landmark has one associated with it
- Allow the user to switch between frames of reference if they use an explicit marker in the input, e g , the CITYTOUR system (Andre *et al* 1986, Andre *et al* 1987)
- Assume that the frame of reference is explicitly supplied, e g , the Situated Artificial Communicator system (Fuhr *et al* 1998)

118

These approaches either restrict the domain of the discourse or impose restrictions on the user In Chapter 8 an algorithm is developed based on linguistic and psycholinguistic work that attempts to select the user's intended frame of reference In preparation for this, the linguistic and psycholinguistic literature this approach draws upon is reviewed

5 3 1 1 Frame of Reference Activation During Spatial Term Assignment

Several studies in psycholinguistics have examined the biases of English speakers during reference frame selection Carlson-Radvansky and Irwin (1994), examined how reference frames, or mental schemata of space, guide the process of spatial term assignment

> This process "takes place within a mental representation of space in which a
> reference frame is imposed on the representation of the perceptual event, the
> reference frame's axes are oriented, and the spatial term is assigned to the
> direction indicated by the relevant axis" (Carlson-Radvansky and Irwin 1994 pg
> 647)

The goal of their experiments was to examine how the orientation of the vertical axis was selected when different sources of perceptual information dictated competing orientations, more specifically, the research analysed the online activation of conceptual reference frames during spatial term assignment The work contrasts two possibilities for how this process proceeds

119

(1) Single Frame Activation Hypothesis

One reference frame is selected and the orientation of its axes serves as the basis for interpreting a spatial term presented in a sentence and verifying a relation between two objects in a picture "Note that this view asserts that only a single frame will be active at any one time, but it makes no claims about which frame it will be" (Carlson-Radvansky and Irwin 1994 pg 651)

(2) Multiple Frame Activation Hypothesis

According to this hypothesis, when the perceptual cues indicate different orientations, multiple reference frames are initially activated with multiple orientations available for axis alignment, one orientation is subsequently selected for assigning a direction to the spatial term and is correspondingly used as a basis for a response (Carlson-Radvansky and Irwin 1994)

The results of the research were consistent with the multiple frame activation hypotheses, indicating that more than one reference frame is initially activated These active frames compete when their axes are dissociated and assign different directions to the same spatial term However, later work by Carlson-Radvansky and Irwin (cited in Carlson-Radvansky 1996) showed that when the reference frames are aligned and assign the same direction to a spatial relation, there is no competition and multiple reference frames do not seem to be active While these experiments indicate that there is competition between dissociated reference frames, the work did not specify how this competition is resolved However, in concluding, the authors noted that subjects showed a "strong preference to use extrinsic[30] above, indicating that the environment-centred frame was used for the vertical axis alignment most often" (Carlson-Radvansky and Irwin 1994 pg 669) This suggests a bias in the competition between reference frames This phenomenon is examined in detail in the next section

---

[30] Carlson-Radvansky uses the term extrinsic to describe what this thesis calls the absolute frame of reference

5.3.1.2 Biases in Frame of Reference Competition

The experiments reported above show that when the perceptual cues indicate different orientations, multiple reference frames are activated and compete for selection as the frame of reference used for assigning direction to the spatial term. Moreover, there appeared to be a bias towards the absolute frame of reference along the vertical axis. Another set of experiments by Carlson-Radvansky and Irwin (1993) illustrated that the vertical axis is dominated by the absolute frame of reference. Such a finding clearly indicates that the competition between reference frames is not equal.

Figure 2-6 in Section 2.3.3.1 (repeated here as Figure 5-2 for convenience) is based on a formal representation by Levelt (1996) of the scenes used by Carlson-Radvansky and Irwin in these experiments.

	(a)		(b)		(c)		(d)	
Absolute	+		+		+		+	
Viewer-centred	+	(.95)	+	(.63)	-	(.93)	-	(.76)
Intrinsic	+		-		+		-	

	(e)		(f)		(g)		(h)	
Absolute	-		-		-		-	
Viewer-centred	+	(.28)	+	(.01)	-	(.30)	-	(.00)
Intrinsic	+		-		+		-	
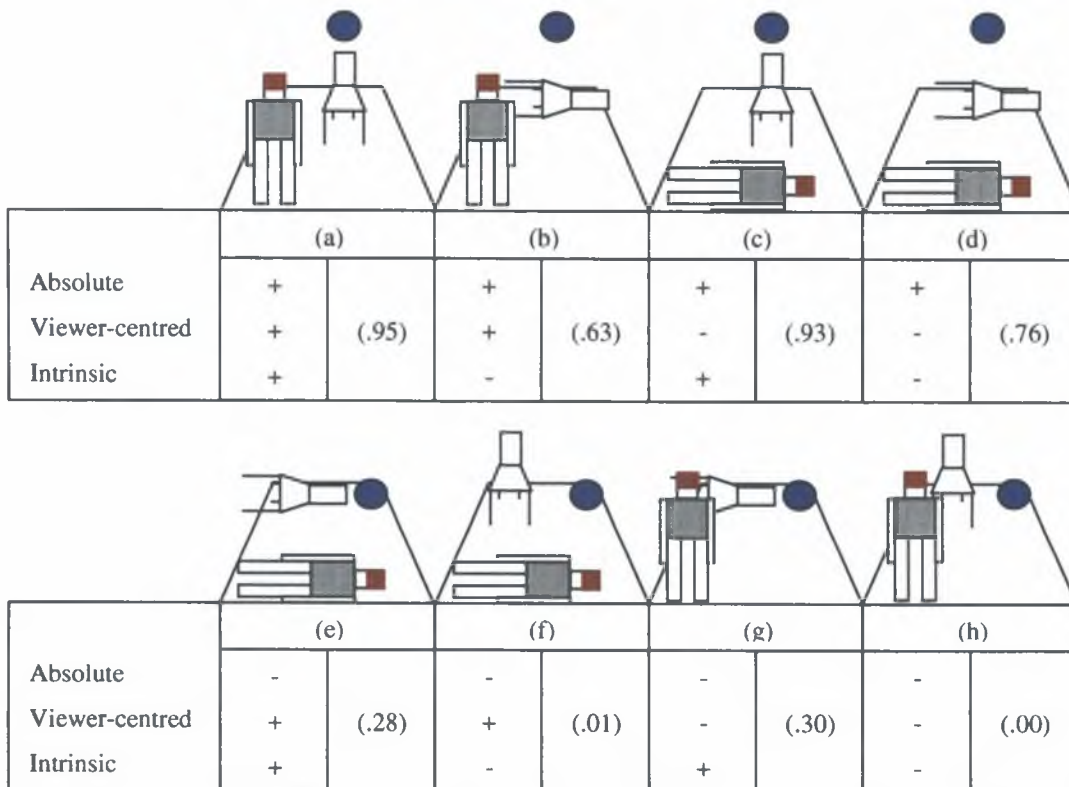
Figure 5-2: Formal representations based on an image from (Levelt 1996) of scenes used by (Carlson-Radvansky and Irwin 1993) to analyse *"the ball is above the*

chair" The + and – signs indicate for each scene which perspective this description is appropriate for

The experiments examined how subjects use the spatial term *above* when the viewer-centred, intrinsic, and absolute reference frames agree or conflict with respect to the vertical The goal of the work was to discover if there was a dominant reference frame in the spatial term assignment for *above* The numbers below each scene show the percentage of *above* responses for each configuration The absolute perspective system is quite dominant here, scenes (a) to (d) are *above* cases in absolute perspective In the absence of an absolute perspective, intrinsic *above*, scenes (e) and (g), was sufficient to elicit *above* responses approximately 30% of the time, even when it was not aligned with the viewer-centred frame of reference Scene (f) isolated the viewer-centred reference frame By itself, this is not sufficient to elicit *above* responses In summary, "an environment-centred[31] frame of reference appears to dominate vertical alignment and assignment of the spatial term above, but the object-centred[32] frame of reference also contributed to the process" (Carlson-Radvansky and Irwin 1993 pg 241) These experiments also revealed several factors that influence subjects adopting an intrinsic reference frame

- Decreasing the distance between the trajector and the landmark increased the applicability of intrinsic descriptions
- The applicability of the intrinsic frame of reference increased if more than one object shared the orientation of the landmark
- The authors also suspected that if the two objects were functionally related, the adoption of an intrinsic reference frame might increase

---

[31] Carlson-Radvansky and Irwin use the term environment-centred to describe the orientation denoted in this work by the term the absolute frame of reference

[32] The term object-centred frame of reference describes the orientation referred to in this thesis by the term the intrinsic frame of reference

122

Nonetheless, even in situations where the factors favouring the intrinsic frame were heightened, the absolute frame of reference still dominated spatial term assignment along the vertical axis

Carlson-Radvansky and Irwin's (1993) analysis of their experimental results was macroscopic in so far as it accounted for all possible spatial configurations Their analysis resulted in a useful but general guide to reference frame selection along the vertical in general, the absolute frame of reference is dominant However, by disregarding the spatial configurations that do not occur in computer use (user lying down, scenes (c), (d), (e), (f)) or situations that can be dealt with using a canonical encounter strategy (user and landmark in canonical position, scenes (a), (h)), the situations that are problematic for this work are highlighted and a more specific guide can be constructed In scenes (b) and (g), the viewer is upright but the landmark's canonical orientation has been toppled It is evident that, in these situations, the combined absolute and viewer-centred reference frame is twice as dominant as the intrinsic frame for reference (absolute b = 0 63, intrinsic g = 0 3) Based on this comparison, a threshold for resolving the competition between reference frames along the vertical axis can be set in this thesis, if the rating of candidate trajector within the intrinsic meaning of the spatial term's spatial template is greater than twice that of the viewer-centred candidate object rating, then the intrinsic object is selected, otherwise the absolute and viewer-centred frames are dominant

In the closing section of the paper, Carlson-Radvansky and Irwin discussed the alignment of the horizontal axes during spatial alignment In the light of later research their comments on this topic are revealing they allowed the possibility "that the orientation of the two horizontal axes (front-back and left-right) may be set with respect to different reference frames" (1993 pg 243)

Research carried out by Taylor et al (2000) used behavioural tests and an electrophysiological measure, event-related potentials (ERPs), to analyse spatial description processing ERPs allow researchers to examine the time course of cognitive processing by measuring electrophysiological changes in the brain in response to specific stimuli The work had two aims

123

(1) ascertain whether processing the spatial location of an object in a scene differed from processing the object's attribute information

(2) establish whether independently available reference frames are processed differently

The authors asserted that their use of ERPs to examine spatial processing was novel Indeed, the motivation behind their analysis of the differences in locational and attribute processing was the need to determine which ERP components were relevant to spatial description processing

ERP components are named with respect to their polarity and latency Components that occur 80 milliseconds or sooner after the stimulus onset are in reaction to the stimulus's physical properties, while later components are associated with cognitive processing After reviewing the literature on neurocognition, the authors concluded that the N400 and P3 components were the most relevant to the spatial processing Previous work in the field has suggested that the N400 reflects integration of new information into a developing mental representation, the more difficult the information is to integrate, the larger the N400 The P3, P300, or Late Positive Component (LPC) amplitude is known to be sensitive to stimulus probability and task relevance (low probability, task relevant stimuli produce the largest P3s) This component has been associated with memory updating with larger amplitudes indicating increased updating requirements The latency of this component is correlated with decision processes

Understanding spatial language requires, "semantic integration, working memory updating and decision making" (Taylor *et al* 2000 pg 4) Using this basis the authors posited that the N400 and P3 would prove to be components of interest The experiments came in two forms, each designed to examine one of the two goals and each analysed using two different approaches measuring ERPs and behavioural studies Both forms of experiments involved showing subjects stimulus pictures containing two objects, one intrinsically orientated, one not Following the picture, a three-word descriptor of the displayed scene was presented The participants then answered whether this descriptor was true or false

124

The descriptors had a constant format, although their content differed across the experiment types In both forms of the experiment, the first word of the descriptor named the non-intrinsically oriented object The second word was a relative descriptor In the experiments designed to study the differences in attribute versus location processing, this descriptor presented either a colour comparison (*redder* etc) or size comparison (*bigger* etc) The other form of experiments examined whether different frames of reference were processed differently The descriptor here was a location term (*front, right*, or *left*) The fact that the location descriptors used in these trials were all aligned with the horizontal axes magnifies the importance of these experiments to the work presented in this thesis Carlson-Radvansky and Irwin (1993) had allowed for the possibility that the dominance of the absolute reference system might only extend along the vertical, Taylor, and Naylor *et al* 's (2000) works, although not specifically designed to, examined this hypothesis The final word in both types of experiment was either the name of the intrinsic object or the word *you* For locational trials, the final word defined the reference frame of use, the object's intrinsic frame if the last word named it, or a viewer-centred reference frame if the last word was *you* For attribute trials, the final word indicated the comparison type

Both the ERP and the behavioural data indicated that attribute processing required more integration However, it was the study's other goal that was of primary importance to this thesis The ERP results from the spatial frame processing experiments found some evidence of multiple reference frame activation These results were convergent with (Carlson-Radvansky and Irwin 1994) Furthermore, the results indicated an easier decision-making process for correct intrinsic configurations than for correct viewer-centred or neither frame correct trials The behavioural results, however, were not completely coincident with the ERP data although the data from the behavioural trials most closely matching the ERP did show a strong convergence, with the response times for trials requiring intrinsic frame processing the fastest when the spatial term predicted the intrinsic frame The results indicated "strategic processing, with priority selection of the intrinsic frame" (Taylor *et al* 2000 pg 11) The authors qualify their analysis with the caveat that their findings may indicate a task requirement influence on spatial frame processing, rather than a plane-based bias The structure of the experiments required the

125

participants to determine the accuracy of the descriptor after the picture was removed This sequence necessitated a heavy memory load that may have caused the participants to focus on the reference frame requiring most work when the picture was perceptually available and to rely on memory when processing the easier frame of reference The advantage of this strategy is that if the descriptor does not specify an intrinsic frame of reference, the participant need only process the easier viewer-centred frame from memory "In other words, they are solving the more difficult problem perceptually and relying on memory, if necessary, to solve the easier problem" (Taylor *et al* 2000 pg 11)

So far, the discussion has centred on the question of how to select a reference frame This is the crucial stage in decoding a locative because it is fundamental to the alignment of spatial templates, in effect, selecting a reference frame locates the search area for the trajector Alignment, however, is not the only impact that reference frames have on spatial template construction Psycholinguistic experiments have shown that the process of reference frame selection affects the shape of the spatial template defining the region described by the preposition (see Section 2 3 4)

5 3 1 3   The Effect of Frame of Reference Selection on Spatial Templates

Carlson-Radvansky (1996) examined two facets of the relationship between reference frame selection and spatial template construction The first question addressed was whether the spatial template associated with a preposition was independent of the type of reference frame used to align it The results indicate that this was indeed the case, the spatial template for a preposition describes a similar shaped area regardless of whether it is constructed in an absolute, viewer-centred, or intrinsic frame of reference These results are examined in Section 8 4, including their validity, because the experimental procedure used meant that visual cues such as object occlusion did not occur in the test Moreover, these experiments focused on the spatial template for the preposition *above* which is canonically aligned with the vertical axis Object occlusion would have little impact on this preposition and as a result may not have been a major consideration in the design of the experiments Given this, it is not surprising that the

126

experimental results indicated the preposition's spatial template was consistent across the frames of reference

However, for the current discussion the second issue examined by Carlson-Radvansky is most relevant That is, whether the competition between frames of reference during the selection process affected the shape of the spatial template The basis for this proposition was the possibility that when multiple reference frames are active, multiple spatial templates are constructed "This would mean that the parsing of space into good, acceptable and bad regions will necessarily reflect some mixture of the two spatial templates, with the ratings predicted by some combination of the corresponding cells in the constructed template" (Carlson-Radvansky 1996 pg 3)

Figure 5-3, Figure 5-4, and Figure 5-5 are representations of spatial templates based on images given in (Carlson-Radvansky 1996) The house in the middle of the figures represents a landmark that has been toppled out of its canonical position Figure 5-3 depicts a schematic representation of the spatial template for *above* constructed using a viewer-centred reference frame The letters G, A, and B label the applicability of points within the spatial template as good, acceptable, and bad, respectively Figure 5-4 depicts a schematic representation of the spatial template *above* constructed using the landmark's intrinsic frame of reference using a similar labelling scheme Figure 5-5 shows a mixed spatial template Here, V prefixes the viewer-centred rating and I the intrinsic rating of the point Carlson-Radvansky conjectured that the area acceptable in both frames of reference (labelled by VA + IA) "should be somewhat privileged (in terms of higher acceptability ratings or easier and more accurate access) relative to the acceptable regions defined by a single template" (Carlson-Radvansky 1996 pg 4)

A	A	A	G	A	A	A
A	A	A	G	A	A	A
A	A	A	G	A	A	A
B	B	B		B	B	B
B	B	B	B	B	B	B
B	B	B	B	B	B	B
B	B	B	B	B	B	B

Figure 5-3 Schematic Template based on viewer/absolute reference frame

A	A	A	B	B	B	B
A	A	A	B	B	B	B
A	A	A	B	B	B	B
G	G	G		B	B	B
A	A	A	B	B	B	B
A	A	A	B	B	B	B
A	A	A	B	B	B	B

Figure 5-4 Schematic Template based on intrinsic reference frame

VA + IA	VA + IA	VA + IA	VG + IB	VA + IB	VA + IB	VA + IB
VA + IA	VA + IA	VA + IA	VG + IB	VA + IB	VA + IB	VA + IB
VA + IA	VA + IA	VA + IA	VG + IB	VA + IB	VA + IB	VA + IB
VB + IG	VB + IG	VB + IG	◁▤	VB + IB	VB + IB	VB + IB
VB + IA	VB + IA	VB + IA	VB + IB	VB + IB	VB + IB	VB + IB
VB + IA	VB + IA	VB + IA	VB + IB	VB + IB	VB + IB	VB + IB
VB + IA	VB + IA	VB + IA	VB + IB	VB + IB	VB + IB	VB + IB

Figure 5-5: Mixture of spatial templates from the different frames of reference.

Carlson-Radvansky postulated that the conjecture of multiple reference frame activation influencing spatial template construction could be validated by comparing the spatial template of a preposition when applied to an object in its canonical and non-canonical position. "If context did not matter, then the non-canonical spatial template should look similar to the canonical template" (Carlson-Radvansky 1996 pg. 5). Experiments were run to test this theory. The procedure for these experiments was to present a sentence of the form *the box is above the tree* to a participant followed by a picture containing a tree and a box. The tree was either in its canonical upright position (canonical trial) or it was rotated onto its side (non-canonical trials).

A plot of the results for the canonical trials revealed a spatial template for above which was convergent with the pattern found by Logan and Sadler (1996) (see Section 2.3.4). There was "a good region along the vertical axes of the reference frames, two acceptable regions sloping downwards and symmetrical about the good region, and bad regions corresponding to non-acceptable uses of 'above'" (Carlson-Radvansky 1996 pg. 5). However, plotting the results of the non-canonical trials revealed the predicted amalgamated spatial templates. There was no good region in the non-canonical template, the acceptable regions were bigger and asymmetrical, while the bad region was smaller. Furthermore, the points located at the intersection of competing frames of reference had a higher applicability relative to the points applicable to a single template. Based on these

results Carlson-Radvansky argued that "clearly the context in which the spatial template is constructed greatly influences its shape when reference frames were dissociated, a very different spatial template emerged than when reference frames were aligned" (1996 pg 5)

A refined analysis of the work in (Carlson-Radvansky 1996) is given in (Carlson-Radvansky and Logan 1997) The substance of this later paper is congruent with (Carlson-Radvansky 1996), however, there are two points of note in this later paper One is that in these experiments the intrinsic frame of reference was relatively more dominant that the viewer-centred frame of reference along the vertical axis This is at odds with earlier work that examined frame of reference bias (see (Carlson-Radvansky and Irwin 1993, 1994) Section 5 3 1 2) Carlson-Radvansky and Logan attributed this anomaly to the difference in the visual stimuli used in the experiments

"One difference was that the current experiments used displays containing only the reference and located objects, whereas the Carlson-Radvansky and Irwin studies used displays containing whole scenes with multiple objects and typically a horizon line, thus emphasizing the environment Such display characteristics could influence the preferences for using different reference frames " (Carlson-Radvansky and Logan 1997 pg 435)

The second point of note is the conclusions that Carlson-Radvansky and Logan drew on the impact of their findings on the process of reference frame selection

"When multiple spatial templates are constructed, the parsing of space around the reference object[33] is best represented as a composite template that is a simple weighted sum of all the existing template for a given spatial relation [ ] Specifically, we believe that preferences for using a particular reference frame are exhibited through the weights assigned to the spatial templates, such that when they are combined a composite map of space surrounding the reference object reflects such biases " (Carlson-Radvansky and Logan 1997 pg 435)

5 3 1 4   Frame of Reference Summary

This review of research relating to the issue of frames of reference began with a general description of the approaches adopted by previous systems that interpreted language within the context of a rendered environment In this thesis these approaches are rejected because of the restrictions they place on the domains the systems can model and the input from the user In Chapter 8, an algorithm for selecting a frame of reference is developed To lay the foundation for this algorithm, the psycholinguistic work on which this approach is based was reviewed In Section 5 3 1 1, (Carlson-Radvansky and Irwin 1994) were reviewed, their results indicated that when frames of reference are dissociated more than one reference frame is initially activated and these active frames compete The work of Carlson-Radvansky and Irwin (1993) and Taylor *et al* (2000) , illustrated the biases present in competition between reference frames (Section 5 3 1 2) These biases are dependent on the orientation of the plane that a given spatial term is canonically aligned with Briefly, the absolute frame of reference dominates the alignment of spatial terms that are usually aligned with the vertical axis, while the intrinsic frame of reference is prioritised with respect to aligned spatial terms associated with the horizontal axis Carlson-Radvansky (1996) and Carlson-Radvansky and Logan (1997) investigated the influence of frame of reference selection on the construction of a preposition's spatial template The findings of this research impact on the modelling of spatial templates

---

[33] Carlson-Radvansky and Logan use the term reference object to describe the object called the landmark in the terminology used in this thesis

131

because they indicate that, if there is a competition between reference frames, the spatial templates constructed for each of the competing reference frames should be amalgamated using a weighting that reflects the bias towards a particular reference frame for a given preposition Keeping with the theme of spatial templates, previous computational models of prepositions are reviewed

5.3 2 Computationally Modelling Prepositions

Mukerjee (1998) surveys computational models of spatial expressions based on their discretisation of space He classifies previous work into two categories which he labels neat and scruffy.

5 3 2 1 Neat Models

The **neat** paradigm was the approach adopted by early research attempting to represent prepositions The major frameworks proposed using this approach include the works of Cooper (1968), Leech (1969), Bennett (1975), and Miller and Johnson-Laird (1976) The defining characteristic of a neat framework is the proposal of definitions for spatial prepositions that may be expressed in first-order logic They attempt to locate entities in a geometric space that has been separated into discrete cells with any set of entity coordinate values resulting in a unique location for the entity within a cell Each distinct meaning of a preposition is semantically defined as intending on a cell, the object intended on by a preposition is the object located in the region associated with the selected meaning of the used preposition For example, the meaning of *in* is defined by each of these models as

132

1   "x in y x is located internal to y with the constraint that x is smaller than y" (Cooper 1968) cited by (Miller and Johnson-Laird 1976 pg 384) and by (Vandeloise 1991 pg 9)

2   "x in y x is enclosed or contained either in a two-dimensional or in a three-dimensional place y" (Leech 1969) cited by (Miller and Johnson-Laird 1976 pg 384) and by (Vandeloise 1991 pg 9)

3   "in locative interior(y)" (Bennett 1975 pg 67)

4   "in (x, y), referent x is in a relatum[34] y if (i) [PART (x, z) & INCL (z, y)]" (Miller and Johnson-Laird 1976 pg 385)

Part of the problem with such an approach to semantics is that "no spatial word lends itself easily to such a strict definition, and counterexamples may be found for every proposed definition" (Vandeloise 1991 pg 9) Figure 5-6 illustrates some of the trajector-landmark relationships that the preposition *in* may be used to describe Diagram (a) illustrates the enclosure of an object by a country Although the container is delimited by a physical boundary in this example, this is not invariant For example, mathematical sets are abstract entities with no obvious physical boundaries However, an element of a set may be described as being in the set This illustrates that language also defines enclosure for strictly conceptual entities Contrasting with this, the landmarks in (b) and (c) are both canonical physical containers, nevertheless, defining the semantics for *in* for these situations is complicated by the linguistic tolerance manifest in (b) that allows an object protruding from a container to be described as in it Furthermore, this tolerance admits situations such as (d) where "a container need not always be larger than the object it has 'in' it, as in 'the club in the hand'" (Miller and Johnson-Laird 1976 pg 385) Example (e) illustrates a part-whole or meronymic relationship which contrasts with example (f) where the bird is not in the interior or part of the landmark, as for instance *the wood in*

---

[34] Miller and Johnson-Laird use the term relatum to describe the landmark

the tree is, but in the fuzzy volume outlined by the tree branches It should be noted that example (d) refutes Cooper's definition (1 above) which requires that the trajector be smaller than the landmark Moreover, the tolerance illustrated in diagram (b) which allows an object that protrudes from a landmark to be described as in the landmark contravenes Leech's definition (2 above) which requires that the trajector should be enclosed or contained by the landmark and Bennett's definition (3 above) which constrains the trajector to the interior of the landmark

(a) The city in Ireland.

(b) The flower in the vase.

(c) The coffee in the cup.

(d) The club in the hand.

(e) The bone in the leg.

(f) The bird in the tree.

Figure 5-6: Diagrams illustrating the range of meanings that may be adopted by the preposition *in*.

135

Another issue with these neat definitions is their use of relations such as *enclosure* that "presume geometric invariants which prove difficult to define in standard point and line geometries" (Garrod *et al* 1999 pg 170) Consequently, it is it difficult to say precisely what is meant by enclosure in these definitions For example, definition 4 above proposed by Miller and Johnson-Laird is compatible with all the examples in Figure 5-6 However, this flexibility is only achieved by incorporating uncertainty into the definition Miller and Johnson-Laird provide no formal definition for the functions PART(x, z) and INCL(z, y) Rather, they argue that "the question here is whether part (some or all) of the referent is included in the relatum[35] The schema leaves uncertain how much of the referent must be inside the relatum before one is willing to say it is 'in' it" (Miller and Johnson-Laird 1976 pg 385) Moreover, even with this uncertainty, these neat approaches are not immune to counter examples Figure 5-7 illustrates a situation where an object X, in this example a triangle, may be described as being in object Y, here a box, even though none of X is located within the interior of Y
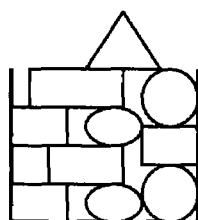
Figure 5-7 The triangle is in the box.

Finally, simple logical definitions (such as those given on page 133) cannot account for the specificity or vagueness that may occur within a particular use of a preposition Figure 5-8, drawn from an image in (Mukerjee 1998), depicts a bird's eye view of a desk with the boxes A, B, C, D, and E representing possible locations of *the chair in front of the desk* The chair at location A is more *in front of the desk* than the chair at location E

---

[35] As noted above (see Footnote 34on page 133), Miller and Johnson-Laird use the term relatum to describe the landmark

This example illustrates that distance from the desk and the angular deviation of the chair from the canonical direction of *in front of* both factors in the *in-front-ness* of the chair. Neat models cannot distinguish between these shades of meaning.
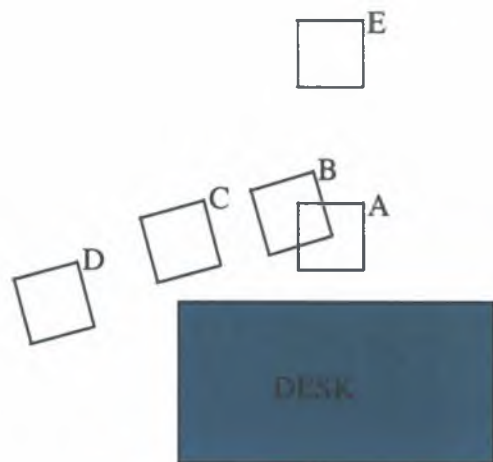
Figure 5-8: Figure illustrating the gradation of in-front-ness for positions A through D. The chair at position A is more *in front of* the desk than the chair at position E. Figure based on an illustration in (Mukerjee 1998).

Clearly, there are many issues with logical definitions of prepositions. One approach to handling these problems is to treat prepositional meanings as natural categories. This allows us to treat certain usages of prepositions as more prototypical or representative. The advantage of this is that it mollifies the impact of any particular counter examples to the proposed definition by allowing it to be categorised as a deviation from the prototype the logical definition defines. Herskovits's (1986) adopts this approach, proposing a multiple relational model that takes logical definitions similar to those given above as prototypical of a preposition's meaning and then defines functions that attempt to explain how deviations from these ideals occur. As Herskovits's (1986) framework can be seen as a culmination of the theoretical analysis of prepositions based on a logical semantics, it is reviewed in detail below.

Herskovits' semantic approach is based on the notions of ideal and deviation from the ideal   Word meanings have an ideal form from which use types can be obtained via transformations

"I suggest two levels of abstraction   ideal meaning and use type   The ideal meaning abstraction is not sufficient to build truth conditions, but it is a necessary anchor that organises the overall set of uses of the preposition   The use type abstraction, with several types derived from the same ideal meaning, is much richer and provides material that brings us much closer to a definition of truth conditions " (Herskovits 1986 pg 18)

The ideal meaning of a preposition describes a salient relation (e g , parallelism of lines, enclosure, contiguity) between two or three ideal salient geometric objects (e g , points, planes, and the vertical direction)   Herskovits uses as her ideal meanings the type of logical relations that neat or simple relational models, critiqued in Section 5 3 2 1 above, propose as the meaning of a preposition   Herskovits does not give a precise procedure for defining these ideal meanings, instead this characterisation of a preposition emerges through a fitting process between the regularities that are revealed by a careful examination of the uses of the preposition (Herskovits 1986)   The following are the ideal meanings adopted by Herskovits for the three basic topological prepositions

- *"at* for a point to coincide with another" (Herskovits 1986 pg 128)
- *"on* for a geometrical construct X to be contiguous with a line or surface Y, if Y is the surface of an object O_y, and X is the space occupied by another object O_x, for O_y to support O_x" (Herskovits 1986 pg 140)
- *"in* inclusion of a geometric construct in a one-, two- or three- dimensional geometric construct" (Herskovits 1986 pg 149)

138

In order to communicate about the imperfect world they experience, people "bend and stretch these ideal concepts" (Herskovits 1986 pg 3) These deviations are constrained by the need to maintain understanding between the speaker and the receiver There are two forms of transformation that can be applied to the ideal meaning, called sense shifts and tolerance

Sense shifts are based on convention and give rise to polysemy A sense shift is a discontinuous shift from one relation to another conceptually close relation Herskovits does not give a general principle that regulates sense shifts Instead, she provides a survey of some of the possible processes by which an ideal meaning may be sense shifted

1 An ideal meaning which is the conjunction of two conditions may be sense shifted by dropping one of the defining conditions (Herskovits 1986 pg 94)

2 An ideal meaning may be sense shifted by the addition of a condition (Herskovits 1986 pg 94)

3 A sense shifted ideal meaning may be related to the ideal meaning by a process involving resemblance (Herskovits 1986 pg 94)

4 The ideal meaning and the transformed ideal meaning may generally co-occur in the everyday world "Thus *on* is used to mean attachment, but attachment most often co-occurs with contiguity and support" (Herskovits 1986 pg 94)

5 An ideal meaning may be sense shifted by uses which are generalisations of the ideal meaning to higher dimensions (Herskovits 1986 pg 94) An example of this form of sense shifting is the statement *"The temperature is highest at the equator"* (Herskovits 1986 pg 51) While this statement is a valid use of the preposition *at*, it violates the dimensional restrictions on the landmark defined in Herskovits' ideal meaning for the prepositions (recall *"at* for a point to coincide with another" (Herskovits 1986 pg 128)) The landmark in this example, *the equator*, cannot be schematised as a one-dimensional point Herskovits' framework accommodates uses of prepositions where the schematisation of the landmark and/or trajector objects do not fulfil the dimensional restrictions of the ideal meaning by

139

allowing generalisations of the ideal meaning to be derived through sense shifts.

6. "The ideal meaning may be, so to speak 'embedded' in the relation of the use type" (Herskovits 1986 pg. 94). Herskovits gives *"The target is at 10 feet"* (1986 pg. 94) as an example of this type of sense shift. In this example, the use type relation of *at* describes a spatial configuration between the target and an implied observer. However, the ideal meaning of the preposition *at* describes the relationship between the target and a point 10 feet from the observer.

A useful example of a sense shift can be demonstrated using the ideal meaning for *on*. This relation is an example ideal meaning that is a conjunction of two conditions; in this instance contiguity and support. Using this meaning for *on*, it is incorrect to describe the ellipse in Figure 5-9 as being *on the table*.
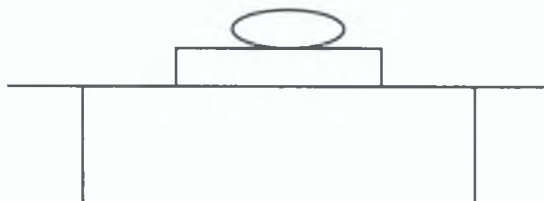
Figure 5-9: The ellipse is on the table.

However, the ellipse-table relationship does resemble one of support. This resemblance allows the use of a sense shifted ideal meaning for *on* which drops the contiguity condition to describe this configuration. Therefore, it is quite acceptable to say *the ellipse is on the table*.

Tolerance is a pragmatic process allowing an ideal meaning or a sense shifted ideal meaning to be approximately true. Although Herskovits describes these tolerance shifts as "gradual deviations measurable as an angle or distance" (1986 pg. 41), she does not give a general predictive principle to characterise the permitted deviation. Instead, she argues that the permitted deviation is dependent on the nature of the objects, on

perception and on contextual relevance (1986 pg 81) The tolerance process can be illustrated using the above ideal meaning for *in* Clearly, this is not true of the position of the triangle in Figure 5-10[36] However, it is quite natural to describe the location of the triangle as *the triangle is in the box*
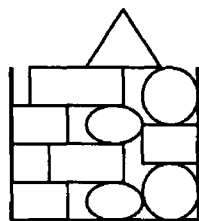
Figure 5-10. The triangle is in the box.

Table 3 lists some examples given by Herskovits of use types for *at, on,* and *in* Figure 5-11 depicts the relationship between a preposition and its set of use types within Herskovits's framework

---

[36] Figure 5-10 is identical to Figure 5-7, reproduced here for convenience

Table 3 Examples of the use types of at, on, and in (Herskovits 1986 pg. 107)

In	
N(spatial entity) in N(Container)	Spatial entity in container
N(gap/object A) in N(object B)	Gap/object (A) "embedded" in physical object (B)
N(person) in N(clothes)	Person in clothing
On	
N(spatial entity A) on N(object B)	Spatial entity (A) supported by physical object (B)
N(physical object) on N(area)	Physical object on edge of geographical area
N(physical object) on N(vehicle)	Physical object transported by large vehicle
At	
N(spatial entity) at N(place)	Spatial entity at location
N(person) at N(artefact)	Person using artefact[37]
N(physical object) at N(path)	Physical object on a line and indexically defined crosspath[38]

---

[37] An example of a use type for the preposition *at* that implicitly defines a person using the artefact is "*Maggie is at her desk*" (Herskovits 1986 pg 135) Herskovits' argues that although one could claim that *Maggie is at her desk* is not false when she is cleaning the floor, it is uncooperative "the speaker should know that the addressee will infer that Maggie is using the desk" (Herskovits 1986 pg 136)

[38] An example of this use type for *at* is *The gas station is at the freeway* "The gas station is at the intersection of the freeway with some indexically defined crosspath" (Herskovits 1986 pg 138) An example context where this use type is applicable is when a speaker and addressee are on a path that intersects with a linear (or schematised as linear) landmark and are some distance from the landmark
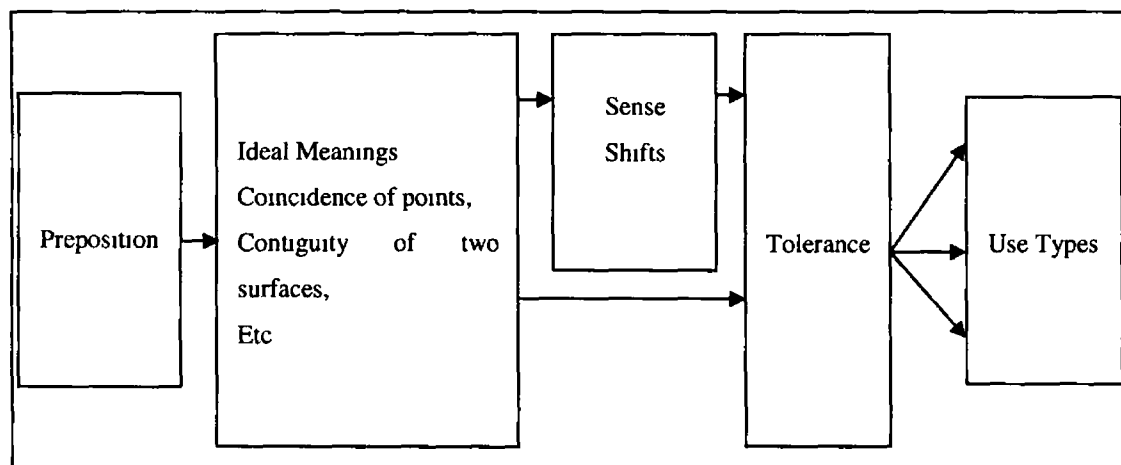
Figure 5-11. The relationship between a preposition and its set of associated use types within Herskovits's framework.

As previously mentioned, the ideal meaning of a preposition describes a relation between pairs or triplets of geometric elements This is also true of the use types or transformed ideal meanings, consequently, for "a particular expression used in a particular context, the arguments of the transformed ideal meaning will not be the objects referred to, but the geometric descriptions of these objects" (Herskovits 1986 pg 17) The geometric descriptions of these objects must match the geometric categories specified in the ideal meaning The matching of geometric descriptions onto the real world objects is achieved through a process called schematisation This concept was introduced in Section 2 3 4 1 2 as a proposed solution to the issue of differentiating between meanings of topological prepositions Herskovits adopts Talmy's definition of the process as "the systematic selection of certain aspects of a referent scene to represent the whole, disregarding the remaining aspects" (Talmy 1983 pg 225) Schematisation underpins Herskovits's framework, as such, it is appropriate to review it in detail However, it is important to remember that within the semantics of spatial language the importance ascribed to schematisation by Herskovits and other researchers is not universally accepted (for Vandeloise's functional approach to this issue see Section 2 3 4 1 3)

Schematisation allows us to reduce a rich physical scene to a very sparse sketchy semantic content There are three distinct processes within the ambit of schematisation

143

abstraction, idealisation, and selection The first of these, abstraction, is prevalent across all linguistic categories These categories abstract from details of their individual elements For example, when someone says

(15) *The car hit the wall*

they abstract away from the angle of impact between the car and the wall, the precise speed the car was travelling at, the size of the wall, etc The idealisation process when applied to the spatial domain takes a geometric form "Spatial expressions conjure up points, lines, ribbons, and so forth, but the scene described does not usually include them, we 'idealize' features of the real scene so they match these simple geometric objects" (Herskovits 1998 pg 150) Idealisation extends abstraction There is an implicit mismatch between the complex geometry of objects in the detailed physical scene and the simplified geometric categories used to characterise them The third process, selection, "involves using a part or aspect of an object to represent the whole object" (Herskovits 1998 pg 150) For example

(16) *The cat under the table* (Herskovits 1998 pg 150)

Here, the top of the table is selected to represent the whole table In Herskovits' approach, schematisation is the result of applying geometric description functions These functions are context dependent, they are applied to an object as it is located in space at the time of the utterance There are two types of functions Those that model (a) people's geometric conceptualisation of physical objects and (b) those that model how people map regions of space onto regions of space

There is only one function of type (a) Herskovits calls it "place" (1986 pg 33) This is the simplest form of function and applies when the object is perceived "as it is in the fundamental description of the world" (Herskovits 1986 pg 63) Here, the applicable

144

geometric description is the region of space occupied by the object[39] at the time Place "has for its domain the product of the set of all spatial entities by the set of time instants, and for its range the set of regions of space" (Herskovits 1986 pg 64) It is applied to the landmark, the trajector, and any other object referred to in the locative expression

Functions of type (b) take the resulting geometric descriptions of place (i e , the regions of space occupied at a particular time by the spatial entities) as arguments and map them onto other regions of space Consequently, these functions map regions of space onto regions of space "Defining elementary geometric description functions in this way avoids type inconsistencies, every elementary function always gets the right type of argument, and the global geometric description function maps a spatial entity taken at time t onto a part of space" (Herskovits 1986 pg 64) Herskovits divides the geometric functions other than place into six categories, Table 4 lists these categories and the main functions of each type

---

[39] In this instance Herskovits uses the term *object* to include anything that might be referred to in a locative expression, e g , geometric object, object parts, parts of space, environments, etc It is interchangeable with the more general term *spatial entity* See (Herskovits 1986 pg 63)

145

Table 4 Elementary geometric description functions (Herskovits 1986 pg. 64)

(1) **parts**	(3) **good forms**
- three-dimensional part	- outline
- edge	- completed enclosure
- base	- normalised region
- oriented total outer surface	(4) **adjacent volumes**
- oriented free top surface	- interior
- underside	- volume/area associated with vertex
- overside	- lamina associated with surface
(2) **idealisations**	(5) **axes**
- approximations to a point	- main axis
- approximations to a line	- associated point of observation
- approximations to a surface	(6) **projections**
- approximations to a horizontal plane	- projection on plane at infinity
- approximations to a strip	- projection on ground

Herskovits eschews defining a general algorithm for the selection and application of these functions to the spatial entities in a given predicate, instead she analyses a range of sentences that demonstrates how these function may be applied in particular instances To illustrate the use of these functions a few examples are cited

The function orientated free top surface is an element of the parts category in Herskovits's ontology of geometric description functions It returns "the surface composed of the set of points of a three-dimensional region which are in the highest horizontal plane Its orientation is defined so it faces the outside of the region" (Herskovits 1986 pg 66) Herskovits posits that the locative expression *the chopstick on the bowl* is an example of a phrase whose semantic interpretation involves this function

146

"The chopstick on the bowl

> *Contiguous(Place(Chopstick), OrientedFreeTopSurface(Place(Bowl)))*
>
> *And*
>
> *Support(Bowl, Chopstick)"* (Herskovits 1986 pg 66)

In this analysis, the terms *Contiguous()* and *Support()* represent the conditions that define the relationship between two objects described by the preposition *on* (see the ideal meaning for *on* above) The terms *Place(Chopstick)* and *Place(Bowl)* represent the regions of space returned by the place function when applied to the chopstick and bowl objects Herskovits analysis can be rewritten as the locative expression *the chopstick on the bowl* intends on the chopstick that occupies a region of space at the time of the utterance which is contiguous with the oriented top surface of the region of space that is occupied by the bowl at the time of the utterance, and is supported by the bowl

The city on the road to London is another example used by Herskovits to demonstrate these elementary geometric functions

"The city on the road to London

> *Contiguous(PtApprox(Place(City)),*
>
> *LineApprox(Place(Road)))"* (Herskovits 1986 pg 67)

In this example, the term *Contiguous()* again refers to a condition defining the relation described by the ideal meaning of the preposition in the phrase *on* The terms *Place(City)* and *Place(Road)* refer to the regions of space returned by the place function when it is applied to city and road objects *PtApprox()* refers to the geometric function that approximates a region to a point *LineApprox()* refers to the geometric function that approximates a region to a line Both the *PtApprox()* and *LineApprox()* functions are categorised by Herskovits as idealisation functions Rewriting Herskovits's formal specification of her interpretation results in the following analysis *the city on the road to London* intends on the city that occupies a region of space at the time of the utterance which can be idealised as a point that is contiguous with a linear idealisation of the region of space occupied by the road to London at the time of the utterance

147

Figure 5-12 depicts a diagrammatic summary of the schematisation in Herskovits' framework This process results in a geometric representation of real world objects

Spatial entities	Schematisation of spatial entities		Geometric
Landmark	Place	Elementary geometric functions Parts Idealisations Good forms Adjacent volumes Axes Projections	representation of spatial entities
Trajector			
Etc			
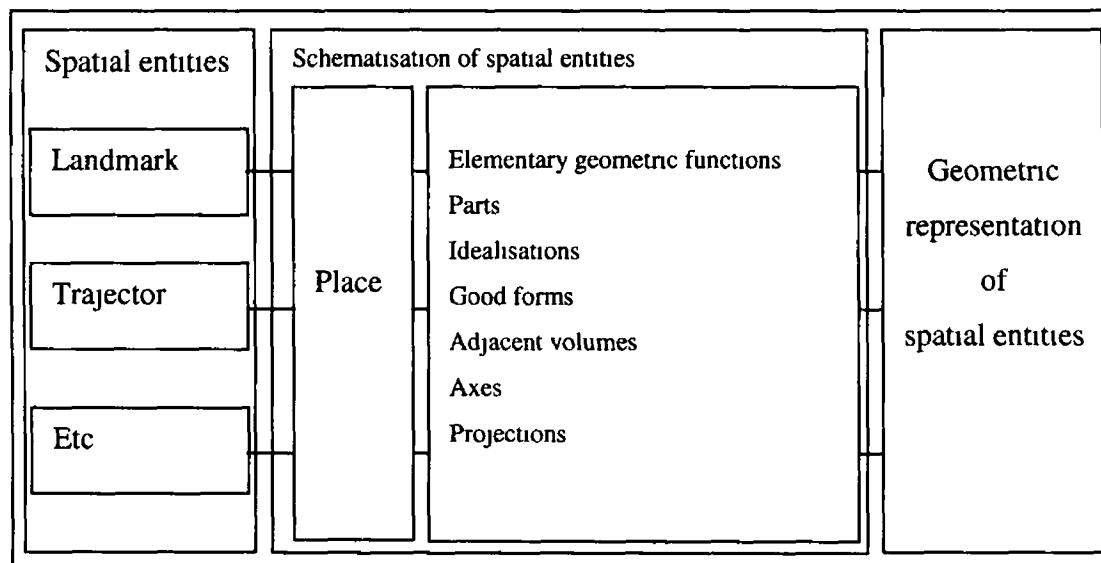
Figure 5-12 Schematic representation of the schematisation process proposed by Herskovits

In this review of Herskovits (1986) two processes have been described (a) the process which transforms ideal meanings into use types and (b) the process of schematisation that results in a geometric representation of real world objects In order to interpret a locative expression, these two processes must be combined and extended

At the core of Herskovits' proposed solution to the problem of interpreting a locative expression is the concept of a normal situation type A normal situation type is "a set of conditions for the true and appropriate use of an expression under normal conditions (there will be several such sets if the expression is ambiguous)" (Herskovits 1986 pp 97-98) There are two stages in this process

1   Generate "the normal situation type(s) associated with the expression"
     (Herskovits 1986 pg 98)

2   Use "the normal situation type(s) and the particular context of the utterance
     to specify the particular interpretation suggested by the context" (Herskovits
     1986 pg 98)

The set of use types associated with a given preposition is the domain of the
process for generating the normal situation type(s) of the preposition This process
involves finding a match between a use type for the preposition in the given locative
expression, the geometric descriptions resulting from schematising the objects talked
about, and what is known about the normal behaviour of these objects If a match can be
found between these three inputs, a normal situation type can be generated, if not, the use
type is rejected Herskovits claims that "all the constraints of a matching use type can be
assumed true of the objects referred to in the expression, given appropriate geometric
descriptions and tolerance" (1986 pg 98) Figure 5-13 shows the process for generating
the normal situation types for a given spatial locative

Spatial entities

Landmark

Trajector

Etc

Schematisation of spatial entities

Place

Elementary geometric functions
Parts
Idealisations
Good forms
Adjacent volumes
Axes
Projections

Geometric representations of objects

Spatial locative expression

Matchings between the geometric forms of the schematised spatial entities the arguments of the transformed ideal meaning and world knowledge of the normal behaviour of the objects talked about results in the creation of the set of Normal Situation Types for the locative expression

Normal situation type(s)

Preposition

Ideal meanings

Sense shifts

Tolerance

Use types

Figure 5-13· Schematic representation of the steps involved in generating the set of normal situation types for a given locative expression.

Using the normal situation type(s), resulting from the first stage of the interpretation process, the second stage of the interpretation process exploits contextual information to accomplish three tasks

1 If the initial step of decoding generated more than one normal situation type, select one from among the candidates

2 Instantiate the normal situation type selected, assigning values to its variables as suggested by the current context

3 Draw inferences allowed by the current context in conjunction with the instantiated normal situation type

150

The final step in the interpretation of an expression is the integration of any proposition derived by the decoding process and relating to the spatial arrangement of objects into the spatial representation scheme, that is, "the end results of the comprehension should be to create the representation of an imaginary scene, or modify that of a remembered scene" (Herskovits 1986 pg 99) Herskovits notes that this final step is dependent on one's assumptions about the form of such spatial representations

In summary, the essence of Herskovits's (1986) book is a descriptive framework of a model of comprehension and production of locative expressions From a traditional linguistic perspective, the work motivates many of the issues underpinning the use of spatial language However, for research attempting to create a man-machine dialog system based on linguistic and perceptual theory there are several drawbacks to using this work as a foundation

Firstly, as has been noted the validity of Herskovits's reliance on a geometric explanation of spatial language use is less than secure, Claude Vandeloise (1991) argues for the rejection of the primary role of schematisation (see Section 2 3 4 1 3) Moreover, even if Herskovits's approach were valid, implementing it requires a system to list for each model it contained the set of possible conceptual schematic forms the model can assume and a mechanism to select which of these forms to take for a given input Although this may be possible, it would be extremely difficult and cumbersome

Secondly, the relational-based ideal meanings do not admit gradation in the semantics of a preposition While Herskovits's attempts to allow for this phenomenon through the processes of tolerance and sense shifts from a computational standpoint, the description of these processes is vague No general regulatory principle for sense shifts is given Indeed, the sum description of this process is a survey of some of the possible examples by which an ideal meaning may be sense shifted Furthermore, no mechanism for selecting which sense shift should be applied to an ideal meaning in a given situation is given Herskovits describes these tolerance shifts as "gradual deviations measurable as an angle or distance" (1986 pg 41) However, she does not give a general predictive principle to characterise the permitted deviation Finally, although Herskovits's gives some heuristic rules (1986 pp 172-173) for selecting a frame of reference, she does not propose a general algorithm to resolve this issue

151

5.3.2.2 Scruffy Models

Scruffy models attempt to address the issue of gradation across a spatial template. Within this genre, spatial relations define fuzzy classes "over the quantitative space based on a *measure* defined on the continuum and not as a discrete set" (Mukerjee 1998 pg. 4). Figure 5-14 gives schematic representations of (a) a neat discretisation of space and (b) a scruffy or continuum parsing. This figure is based on an image in (Mukerjee 1998).
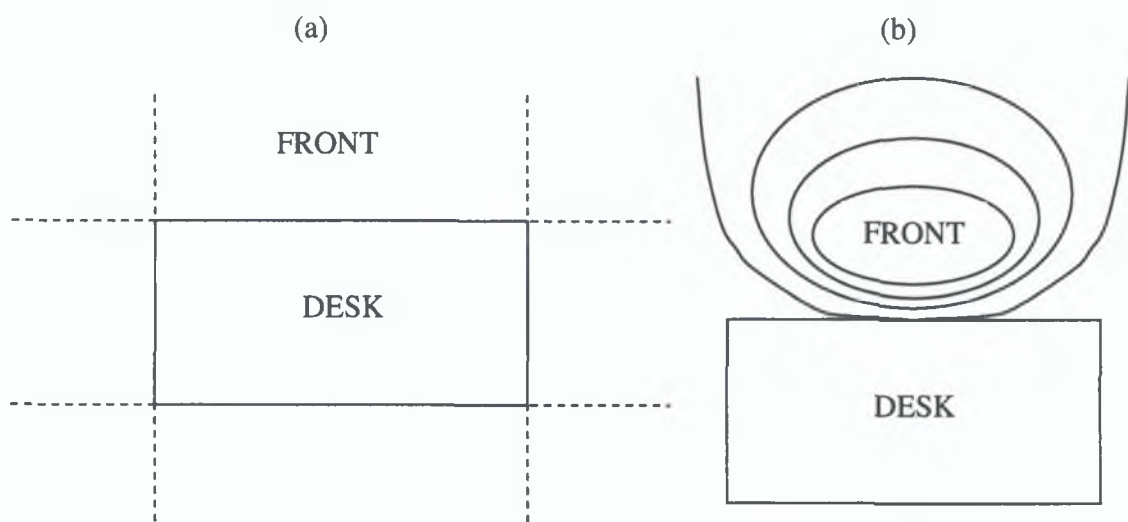
Figure 5-14: Diagram (a) is a schematic 2-D representation of a neat discretisation of space around a desk. Diagram (b) is a schematic 2-D representation of a scruffy discretisation of space around a desk.

A potential fields model is one form of continuum measure that is widely used (Yamada 1993; Gapp 1994a; Olivier and Tsuji 1994). In these models, a preposition's spatial template is constructed based on a potential energy function which returns a value for each location in the template indicating the cost of accepting that location as an interpretation of the preposition. The lower the value ascribed to a location, the higher its acceptability. By assigning a value to each point in a region, these functions allow scruffy semantic models to accommodate the gradation of acceptability across a spatial template within their framework. Another continuum model is proposed by (Mukerjee *et al.* 2000).

152

In this model, the continuum field is created by first defining the location of the field's global minimum. Following this, a set of concentric ellipses that use the **global minimum**[40] as a fixed focus are created by varying the eccentricity of the ellipse and the position of the second focus. These concentric ellipses define the different regions of applicability within the model. Fuhr *et al.* (1998) propose a hybrid approach which uses the degree of overlap of an object with discretised regions as its measure.

The advantage of these models is their ability to distinguish between different locations within a spatial template by assigning each point an applicability rating. This simplifies the trajector rating and selection process. However, some of these models only work in 2-D (Yamada 1993; Olivier *et al.* 1994; Mukerjee *et al.* 2000); one (Fuhr *et al.* 1998) has problems distinguishing between the position of trajectors that are fully enclosed within a region; and most use the centroid of the object's bounding box to represent the objects (Yamada 1993; Gapp 1994a; Olivier and Tsuji 1994; Fuhr *et al.* 1998) which is problematic (see Sections 2.3.4.2.4 and 2.3.5.1) and those that do not (Mukerjee *et al.* 2000) are dependent on locating the local minimum within the continuum field of a preposition which is problematic since the location of the local minimum varies from person to person (see Section 5.4.6). Furthermore, all of these models abstract to a purely topological analysis. This abstraction ignores perceptual information, in particular the issue of object occlusion. Moreover, none of these propose a cognitively plausible approach to the issue of reference frame selection. This by itself is a glaring gap within these models. However, a further consequence of this omission is that these models ignore the impact that the process of selecting a frame of reference can have on the construction of a spatial template (see Section 5.3.1.3). Finally, none of these models propose mechanisms for handling anaphoric references. Section 5.4 contains a survey of some of the previous systems that have integrated language and vision which provides a detailed review of all the potential field models mentioned in this section.

---

[40] The global minimum is the point in a continuum which has the lowest value in the field.

5.4   Language and Vision Systems

The review of semantic models of prepositions in Section 5 3 2 noted the deficiencies in logically based (or neat) approaches and the shift towards scruffy or continuum models The computational systems that implement these models are now reviewed The system developed in this dissertation, the SLI, is inspired by the systems reviewed below At the same time, however, the SLI system tries to address the shortcomings of previous systems identified in Section 5 3 2 2 and discussed in greater detail below The critique of these systems will examine four areas

1   Does the system ground the interpretive process in a model of user perception or is the interpretive module given complete access to all the objects in the simulated environment?

2   How does the system handle the general issue of reference resolution?

3   Do the systems allow the use of different frames of reference? If so, what mechanism do they use for reference frame selection?

4   How do the systems semantically model prepositions?

5 4 1   SHRDLU

SHRDLU (developed by Terry Winograd in 1971 at MIT) is one of the earliest and best-known systems in this genre The program carried on a dialog with a person concerning the activity of a simulated robot arm in a simple blocks world

The SHRDLU interpretive system was initially given a model of the current state of its environment This world model took the form of LISP relational tuples and contained information on object names, object attributes, inter-object relationships, and event causal relationships The elements of these tuples represented the conceptual categories available to the user The meaning of a category was based on its interconnections with all the other categories in the model Each user input was converted into a set of commands that modified the state of the world model

154

"We can think of any utterance as a program – one that indirectly causes a set of operations to be carried out within the hearer's cognitive system " (Winograd 1973 pg 170)

The instructions were then executed in order to model the user's meaning In effect, this approach attempts to satisfy the user's goal by dividing it into successive implicit sub-goals which must be accomplished in order to achieve the main goal

It is important to note that there was no model of human perception in the SHRDLU system The only inputs to SHRDLU's interpretive module were the initial world model and the user's linguistic inputs As a result, SHRDLU could not resolve pronominal references to objects which had not been explicitly referred to in the dialogue Moreover, the system could not handle other-anaphoric and one-anaphoric expressions where there was more than one object in the world that fulfilled the linguistic description of the referent

With respect to the issues of frames of reference use and selection, the user's fixed view of the world and the simplistic objects in the domain disguises the assumption of a viewer-centred reference frame Finally, while the literature currently available does not describe how SHRDLU modelled prepositions, the system's simple relational representation of meaning implies a simplistic neat modelling approach

5 4 2   Visual TRAnslator (VITRA)

The VITRA (Herzog and Wazinski 1994, Herzog 1995, Herzog 1997) project examined the relationship between natural language and vision Its main research topics included

- a referential semantics for spatial prepositions,
- representation and incremental recognition of motion events,
- incremental recognition and verbalization of plans, intentions, and plan interactions,

155

- listener modelling by means of anticipated imagination,

- simultaneous natural language descriptions of dynamic imagery,

- multimodal, incremental route descriptions

(Herzog 1997)

Several systems linking language and vision were developed during this project
The ones most relevant to this thesis were the CITYTOUR system (Andre *et al* 1986,
Andre *et al* 1987), the SOCCER system (Andre *et al* 1986, Andre *et al* 1988, Schirra
and Stopp 1993) and the CSR-3-D system(Gapp 1994a, Gapp 1996) These are reviewed
below

5 4 2 1 CITYTOUR

The CITYTOUR (Andre *et al* 1986, Andre *et al* 1987) project was a German
question-answer system whose domain of discourse was a simulated tour through a city
The system handled dynamic and static objects in 2-D It presented the user with a bird's
eye 2-D map view of the domain

The system represented static objects as centroids, closed polygons, prominent
fronts, and delineative rectangles Based on an image in (Andre *et al* 1987), Figure 5-15
illustrates these representational forms The perspective used in Figure 5-15 is the one
used by the CityTour system to represent the domain to the user, i e , a bird's eye view
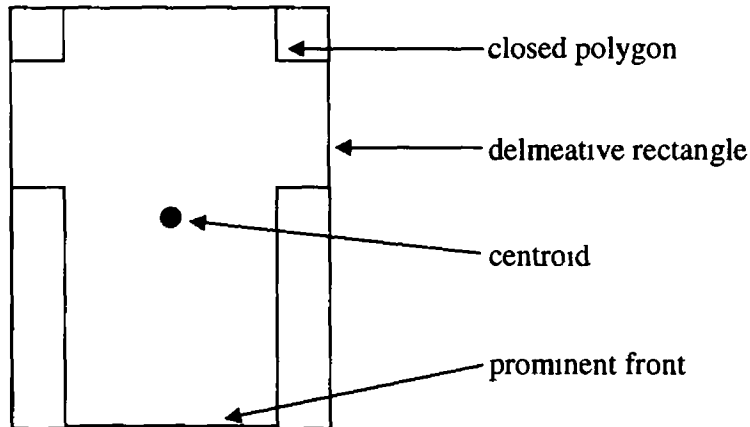
Figure 5-15: The representation of static objects in the CityTour system Based on an image in (Andre _et al._ 1987)

Dynamic objects are represented by a point whose path is a list of time-indexed positions A dynamic object called the sightseeing bus represented the user's location in the world This embodiment of the user allows the system to handle both viewer-centred and intrinsic reference frames In the CityTour system, the viewer-centred reference frame uses the bus's position as the origin and is generated by taking the line of sight from the bus to the landmark as its front axis For an intrinsic reference frame, the prominent front of the object orientates the reference frame with each edge of the delineative rectangle defining a half plane representing the region's front, left, right, and back Figure 5-16 illustrates the definition of the regions associated with each preposition using a delineative rectangle orientated using the prominent front of the object A bird's eye perspective is used in this figure
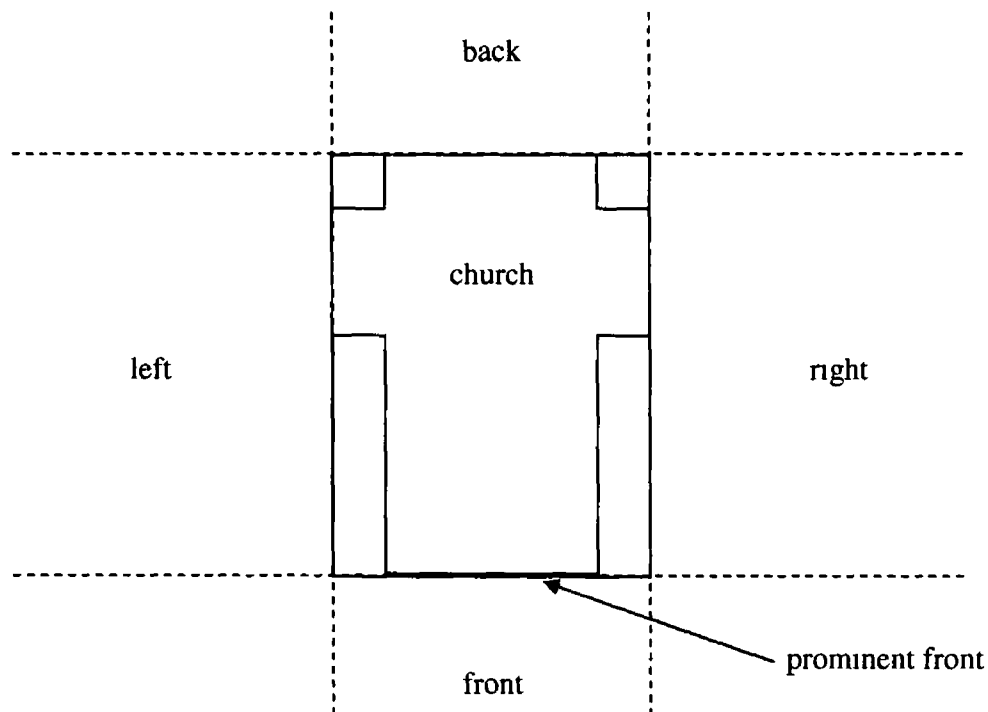
Figure 5-16 Definition of prepositional regions using the edges of a delineative box orientated on the object prominent front (Andre *et al.* 1987).

There is no reference to a model of visual perception in the CITYTOUR system An implication of this is that the interpretive module was given direct access to the model of the whole environment With respect to the reference resolution issue, there is no mention of a context model in the literature pertaining to the CITYTOUR project Moreover, in the example dialogs, all references to objects in the world are made using definite noun phrases

While the CITYTOUR system addresses the issue of frames of reference, the scope and solution implemented are very limited The 2-D map view of the domain excludes many situations where the frame of reference problem occurs in a 3-D environment For example, an object with an intrinsic reference frame may be encountered when it is toppled from its canonical position In such situations, the viewer-centred and intrinsic reference frames become disassociated Furthermore, the mechanism for frame of reference selection is simplistic The system assumes the intrinsic reference frame as a

158

default with the viewer-centred frame of reference treated as a marked case, the user must use an explicit linguistic cue such as *from here* to specify its use As was noted in Section 2 3 3 5, people rarely use such explicit markers to convey the intended frame of reference to a listener

Andre *et al* (1986) describe CITYTOUR's semantic characterisation of prepositions

The system partitions "the area around the reference object[41] into four half-planes With each half-plane, one of the relations *front, back, right* and *left* is associated If the reference object is represented by a polygon, our algorithm determines a delineative rectangle which serves as an extended origin [ ] a relation is applicable if the subject is within the corresponding half-plane [ ] different degrees of applicability can be determined by partitioning the half-planes into regions of the same degree of applicability" (Andre *et al* 1986 pg 11)

There are several problems with this approach Firstly, the use of a delineative rectangle to represent the landmark incurs the issues associated with a bounding box representation (see Section 2 3 4 2 4) a delineative rectangle in a 2-D environment is equivalent to bounding box in a 3-D environment Secondly, although the system computes a degree of applicability across the half-plane characterising a preposition, the only factor considered in this process is the distance of the trajector from the landmark Such a model ignores the angular deviation from a preposition's canonical direction Thirdly, CITYTOUR characterises the trajector by its centroid (Herzog 2001) The problems associated with such a representation were discussed in Section 2 3 5 1 Finally, the system's 2-D maplike representation of the domain ignores the issues of object occlusion and also fails to account for situations in 3-D where a trajector may be proximal to a landmark on the horizontal but distant on the vertical, e g , a plane flying over a building

---

[41]Andre Herzog et al use the term reference object to describe the landmark

The SOCCER system generates natural language descriptions of short soccer scenes in German "The listener is assumed not to be watching the scene, but to have prototypical knowledge about the static background" (Andre *et al* 1986 pg 2) Like its sister project CITYTOUR, the SOCCER system worked in a 2-D environment

As input, the system receives a geometric description of the scene, which was initially represented by an image sequence There are two parts to this input (1) a model of the static background consisting of the football pitch and its parts and (2) the mobile objects in the scene, perceived as points

The core of the system consists of three components (1) an event recognition component, (2) a selection component, and (3) a language generation component The event recognition "produces a set of propositions interpreting the given percepts as instances of spatial and spatio-temporal relations" (Schirra and Stopp 1993 pg 3) The selection component "selects relevant propositions, orders them, and passes them to the encoding component" (Andre *et al* 1988 pg 4) The generation component transforms the selected event propositions into German utterances which are coherent relative to the preceding descriptions (Schirra and Stopp 1993 pg 4)

The SOCCER system is a generative rather than interpretive language system As such it does not require a mechanism for resolving user references There is, however, a description of how the system generates references

"For referring to objects, their internal identifiers (e g *player#1*) are transformed into nominal phrases To this purpose, the system selects attributes enabling the listener to uniquely identify the intended referent whereby it must access the partner model and the text memory If an object cannot be characterised by attributes stored a priori in the partner model, it will be described by means of spatial relations, for example *der linke Elfmeterpunkt* (the left penalty spot), or by means of events already mentioned in which it was (is) involved, for example *der Spieler, der angegriffen wurde* (the player who was attacked) In order to increase

text coherency, anaphoric expressions are generated if the referent is in focus and confusion is excluded " (Andre *et al* 1988 pg 8)

A component of the SOCCER system called ANTLIMA – ANTicipation of the Listener's IMAgery – controlled the generation of noun phrases, anaphora, and ellipsis ANTLIMA was designed to "construct and maintain a model of the listener's knowledge of the events that have already been described" (Schirra and Stopp 1993 pg 4) The basic assumption underlying this system is that a listener understands an utterance by mentally representing the referents in it "Conceived as the mental representation of the understanding of the previous text, it [ANTLIMA] allows for explaining the success or failure of acts of reference an NP – be it anaphoric or elliptical - can only be used if it uniquely identifies its referent in the image" (Schirra and Stopp 1993 pg 4)

The philosophy behind the ANTLIMA module is congruent with the approach underpinning this thesis that computationally interpreting/generating language should be grounded in a model of the user's knowledge of the environment However, ANTLIMA's model of user knowledge is built solely on linguistic utterance Consequently, it does not tackle the issue inherent in modelling a visual domain for example, how to model visual attention Moreover, as ANTLIMA constructs the model of the listener's model based solely on the previous descriptions, the generation of anaphoric or elliptical referents are restricted to referents which have been previously introduced into the discourse Although this approach is suitable for the domain of the SOCCER system, the ANTLIMA listener model is not feasible as context model for systems where the referents of anaphoric or demonstrative references may be introduced solely by visual perception

The SOCCER system attempts to avoid generating output which requires a frame of reference interpretation Where it does, the system defaults to the intrinsic frame of reference

The computational semantics for prepositions implemented by the SOCCER system follows the algorithms proposed for its sister project CITYTOUR (Andre *et al* 1988) As such, the above critique of CITYTOUR's model applies to the SOCCER system

161

5 4 2 3   Computation of Spatial Relations in 3-D-Space (CSR-3-D)

In Gapp (1994a), the CSR-3-D system, a "computational model for the basic meanings of spatial relations which propositionally describes the relationship between geometrical objects in 2-D and 3-D space" (1994a pg 1), was presented The essential idea of the model is to measure the distance and the angle of deviation from the trajector's centre of gravity to the landmark, in a local coordinate system which is scaled by the extension of the landmark, and to account for the vagueness of spatial relations by mapping these values to specific evaluation functions (Gapp 1996) Three classes of spatial relations were considered topological, projective, and the relation *between* For topological relations (*at, near*, etc), a measure based on distance between trajector and landmark was developed The procedure for modelling projective relations extended the topological algorithm to account for the angular deviation between the trajector's position and the preposition's canonical direction

The first step in the algorithm is the schematisation of the objects In a 3-D domain, the landmark is approximated by its **bounding right parallelepiped** (BRP) and the trajector by its centre of gravity "The bounding right parallelepiped of an object with respect to a direction vector v→ is the minimal right parallelepiped which is aligned to v→ and contains the 3-D representation of the object" (Gapp 1994b pg 7) The problem associated with approximating the trajector by it centroid was discussed in Section 2 3 5 1

The landmark's BRP is aligned using the direction of gravity and the object's front which is dependent on the frame of reference The model does not include a mechanism for the selection of the appropriate reference frame This decision is left to an unspecified conceptual system that "must determine which perspective to select in a specific case" (Gapp 1994a pg 4)

Once a frame of reference has been imposed on the landmark's BRP, a local coordinate system is defined This is achieved by aligning the landmark's BRP to the positive y-axis of the world coordinate system with the help of a rotation around the z-axis and then translating the landmark's BRP to the origin Following this, the world coordinate system axes are scaled relative to the extension of the landmark's BRP along

the prevailing dimension In order to compute the trajector's coordinates in the landmark's local coordinate systems, the same rotation and transformation are applied to the trajector's centre of gravity as were applied to the landmark's BRP The trajector's transformed coordinates are its position in the local coordinate system These coordinates are also the distance between the landmark and the trajector in the local coordinate system The local coordinate system ensures a scaling of the distance between the trajector and the landmark depending on the extension of the landmark's BRP (Gapp 1994a)

In evaluating a topological preposition, the local distance between the landmark and trajector is used as the parameter to a cubic spline function that maps the distance to the interval [0 1] The applicability of the spline result depends on the definition associated with the prepositions This allows a continuous gradation of the prepositions applicability rating based on distance the greater the local distance the lower the applicability

For projective prepositions, the topological algorithm is extended to allow for the canonical direction constraint implied by the preposition The angular deviation of the vector between the trajector location in the landmark's local coordinate system's centre of gravity and the direction vector implied by the preposition is computed Similar to the local distance, this angular deviation is mapped to a spline function which results in a value between 0 and 1 The overall degree of applicability of a projective relation is simply the product of the results of the distance and angle spline functions

The CSR-3-D system (Gapp 1994a) is the only system, prior to the SLI system developed in this thesis, that defines a 3-D spatial template that accommodates a measure of the trajector's angular deviation from the canonical direction of the projective preposition's search axis and a measure of the distance of the trajector from the spatial template's origin There are, however, several weaknesses in this system Firstly, there is no reference to a model of user perception in this system Secondly, there is no description discourse framework Consequently, the CSR-3-D system has no mechanism for anaphoric or demonstrative reference resolution Thirdly, there is no algorithm given for selecting a frame of reference Fourthly, although the use of a local coordinate system which is scaled relative to the extension of the landmark's BRP ensures a scaling of the

163

trajector's angular deviation and distance within the preposition's spatial template, it also forces the CSR-3-D system to use the landmark's BRP centroid to represent the landmark The problems with representing the landmark by its centroid were described in Section 2 3 4 2 4 Finally, the CSR-3-D system schematises the trajector by its centroid, the problems associated with this representation of the trajector were described in Section 2 3 5 1

5 4 2 4 VITRA Summary

Of the three systems in the VITRA project, reviewed for this thesis, only the SOCCER system proposed a model of user perceptual knowledge This model, however, is solely based on previous discourse Consequently, it does not address the issue attending to modelling visual perception how to model visual attention The SOCCER system was also the only project in the VITRA project that constructed a model of discourse, thus allowing the use of anaphoric references This model, however, is based solely on previous discourse and consequently is prone to the same limitations as DRT and Centering Theory (see Section 4) Furthermore, none of these systems proposed a suitable mechanism to select a frame of reference

Focusing on the proposed semantic models for prepositions, it is evident that none are satisfactory While the CSR-3-D model works in 3-D, the CITYTOUR and SOCCER systems are restricted to use in 2-D environments Moreover, in CITYTOUR and SOCCER the process of computing the applicability ratings across the area associated with a given projective prepositions only accommodates the distance between the trajector and the landmark, thus ignoring the angular deviation of the trajector from the preposition's canonical direction Furthermore, in both the CITYTOUR and SOCCER systems the applicability ratings associated with a preposition are fixed, i e , they cannot be adjusted to accommodate the variation of the preposition's spatial template when it is applied to different-sized landmarks Also, all of these systems, including the CSR-3-D system, schematise the landmark by its centroids Indeed, the use of a local coordinate system which is centred on and scaled relative to the landmark's BRP forces the CSR-3-

164

D system to schematise the landmark by its BRP centroid. The problems with representing of the landmark by its centroid were described in Section 2.3.4.2.4. Furthermore, all of the VITRA systems (CITYTOUR, SOCCER, and CSR-3-D) schematise the trajector by its centroid. The problems associated with this representation of the trajector were described in Section 2.3.5.1. Finally, all these models are based on a purely topological approach with no attention paid to the issue of object occlusion (as in Section 2.3.5.2).

5.4.3 SPRINT

The SPRINT (SPatial Representation INTerpreter) system is described in (Yamada 1993). The SPRINT system constructs a 3-D model from a description of a world in Japanese. The system can be divided into two parts: part (a) consists of the component which extracts qualitative geometric constraints among the spatial attributes of the entities in the input and part (b) interprets each of these constraints into "potential model-based representations which are the numerical constraints on the entity parameters" (Yamada 1993 pg. 111). Following this, the system uses a gradual approximation technique based on a gradient descendent method to compute the solution with minimum energy. Once this has been calculated, SPRINT draws an image of the described scene.

Because the user supplies a description of the world to the SPRINT system, it can be assumed that they have complete knowledge a priori thus obviating the need to model the user's knowledge of the world. However, there is no mention of a model of visual salience or discourse in the SPRINT system. As a result, the system cannot resolve anaphoric references or references which are not completely disambiguated by the linguistic content. This is evident in the simple mechanism the SPRINT system uses to resolving reference; it assumes each entity is described by a noun which is associated with a predefined graphic object. A similarly simplistic approach is taken with the issue of frame of reference selection. The SPRINT system always aligns the canonical direction of a projective preposition relative to the viewer-centred frame of reference of a viewpoint embodied in the world. This is evident in Yamada's analysis of how the

systems analyses the input *"The marine tower stands to the right of Hikawa-maru* [a ship]" (Yamada 1993 pg 117)

"if you do not know which direction the observer sees, you would not be able to figure out the direction 'to the right' and could not imagine where the tower was (Another way to figure out the direction 'to the right' is to calculate it only from the orientation of the ship, but we do not think it is usual) " (Yamada 1993 pg 119)

As with the previous systems in this review, the SPRINT system schematises the landmark and trajector by their centroids Furthermore, the potential field models used in SPRINT to characterise the prepositions only work in 2-D Moreover, these models are dependent on the definition of a constant which fixes the scale of the potential field associated with a given preposition, thus, it cannot be adjusted relative to the dimensions of the landmark

Although there is some discussion relating to the issue of object occlusion in (Yamada 1993), this is restricted to computing the location of objects which are described in the input as occluding other objects in the scene For example, in the context of analysing the input *"As you walk along the street, there is a tree hidden behind a building"* (Yamada 1993 pg 128), Yamada reviews different strategies for locating the occluding building relative to the embodied viewpoint He does not discuss how object occlusion may impact on the interpretation of a locative

5.4.4 Words In Pictures (WIP)

The WIP (Olivier *et al* 1994, Olivier and Tsuji 1994)[42] system utilised an object schemata to represent the dimensional and perceptual properties of the objects The perceptual attributes of an object includes its intrinsic and viewer-centred frame
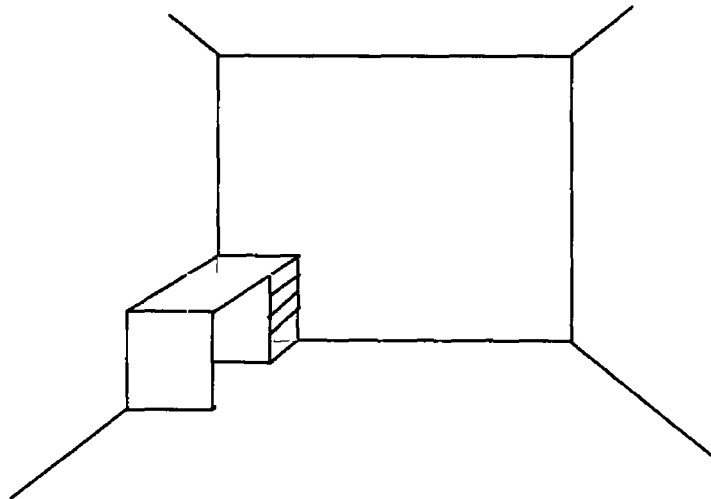
Figure 5-17 A desk in a room, based on Figure 1 in (Olivier and Tsuji 1994)

An example schema for a desk similar to the one in Figure 5-17 is given in (Olivier and Tsuji 1994) and is reproduced in Table 5 Figure 5-18 illustrates how this example object schema applies to the desk in Figure 5-17

---

[42] Olivier and Tsuji (1994) use the term deictic frame of reference to describe the viewer-centred frame of reference and the prefix d- in their tables and figures to denote deictic elements For the sake of consistency in terminology across the dissertation in the tables and figures presented in this section the prefix vc- is used in place of the prefix d- in the original

Table 5. An example object schema for a desk (Olivier and Tsuji 1994). A, B, and C label three orthogonal axes centred at the object. A1/A2, B1/B2, and C1/C2 are corresponding half axes Intrinsic axes are prefixed by i- and viewer-centred axes by vc-.

A	Maximum	B	Vertical	C	Across
A1	i-left	B1	i-bottom	C1	i-front
	vc-front		vc-bottom		vc-right
A2	i-right	B2	i-top	C2	i-back
	vc-back		vc-top		vc-left
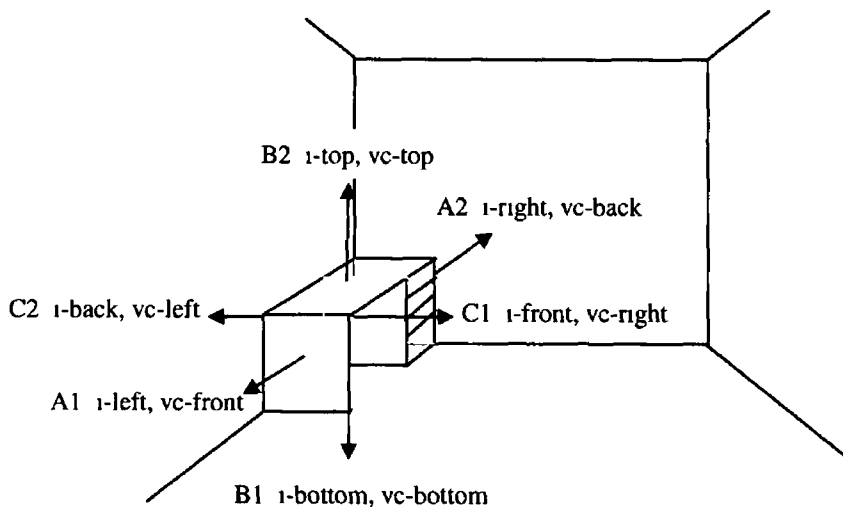
Figure 5-18. A diagram of a desk in a room with the example object schema for a desk given in (Olivier and Tsuji 1994) overlayed

"The preposition 'behind' references the region of space projecting out from the object in direction C2 under the intrinsic interpretation and A2 under the deictic[43] interpretation" (Olivier and Tsuji 1994 pg 152) By explicitly representing the viewer-

---

[43] Olivier and Tsuji (1994) use the tem deictic frame of reference to describe the orientation referred to in this thesis as the viewer-centred frame of reference

168

centred and intrinsic directions associated with a given preposition relative to each object in the system's visual domain, the WIP system was able to identify "the actual face of an object relative to which a spatial constraint is constructed" (Olivier *et al* 1994 pg 1407) in a given frame of reference

The system used a potential field model to represent proximal and directional constraints associated with a preposition The proximal potential function used was

$$P_{prox} = (K_{prox}/2) \left(((x\text{-}x_0)^2 + (y\text{-}y_0)^2)^{1/2} - L_{prox} \right)^2$$

Equation 1. The Proximal Potential Function used in the WIP system (Olivier and Tsuji 1994) This function is a simple elastic function The greater the distance between the landmark's position (x_0, y_0) and the position of the object being located in the field, (x, y), the higher the potential value returned by the function P_{prox} K_{prox} is a constant defining the elasticity of the function L_{prox} is the original length of the function.

The potential function used to represent directionality was

$$P_{dir} = (K_{dir}/2) (x\text{-}x_0)^2$$

Equation 2 The potential function used to represent directionality in the WIP system (Olivier and Tsuji 1994) Here, K_{dir} is a constant defining the elasticity of the directional constraint; x defines the position of the object being located in the field on the x-plane, x_0 defines the position of the landmark on the x-plane, and P_{dir} is the potential directional score ascribed to the object being located in the potential field

The overall value assigned to a point in this potential field is then

$$P = P_{prox} + P_{dir}$$

Equation 3: The equation defining the overall value assigned to a point in the potential field created by the WIP system (Olivier and Tsuji 1994). P_{prox} is computed using Equation 1 above and P_{dir} is computed using Equation 2 above

169

The values for K_{prox}, L_{prox}, and K_{dir} were linearly dependent on the dimension of the landmark and the trajector

While the WIP system encoded information for intrinsic and viewer-centred frames, there is no process given for the selection of a reference frame The equations for the potential field are 2-D in nature This is problematic when translated into a 3-D environment Dropping the z dimension from the coordinate system causes all objects to be translated onto the xy-plane This can have unwanted effects For example, if a bird is flying over a house, a system that only uses the xy coordinates of the bird may locate the bird in the house or in front of the house, etc The system's reference resolution capabilities were restricted to definite descriptions (Olivier 2001) Furthermore, although the system used the dimensions of the landmark to parameterise the potential field, the landmark and trajectors were schematised by the centroids (Olivier 2001) The problems with such an approach were discussed in Section 2 3 4 2 4 and Section 2 3 5 1 Finally, although explicit references to the issue of occluded trajectors are made in the literature describing this system "(e g interpretation that leads to the located object being hidden from view should in general be disallowed)" (Olivier and Tsuji 1994 pg 157), no solution to this issue (such as a model of user visual perception) is proposed

5.4 5 Situated Artificial Communicator

Fuhr *et al* (1998) describe a system that "observes a scene with a stereo camera and communicates with a human partner via speech in order to solve a construction task" (1998 pg 1) The paper presents a model for interpreting six projective prepositions *left, right, in-front, behind, above,* and *below* The objects in the domain of the system were from a children's toolbox

The construction of the world model is described in (Socher *et al* 1996) While the details of this process are not relevant to this work, it may be briefly described as iteratively fitting object hypotheses to CAD like object models The object hypotheses are generated by a hybrid module, combining neural and semantic networks, that analyses the video images The CAD-like object models are supplied to the system a priori

170

The system's reference resolution capability allows a user to intend on an object by specifying its type, colour, size, and shape and spatial relation relative to other objects In contrast with the other systems reviewed so far, here reference resolution is based purely on the visual domain Interestingly however, this approach ignores situations where a discourse history is required Furthermore, the language interpretation process assumes that the reconstruction of the 3-D world model is complete From the perspective of this thesis there are two drawbacks with this approach Firstly, in 3-D rendered environments it is normal for the user viewpoint and world objects to move around the environment In these scenarios, the Situated Artificial Communicator would need to reconstruct a 3-D model of the world after each movement Secondly, as with previous systems, the language interpretation process is given direct access to a complete model of the world Moreover, it should be noted that world model construction process only attempts to recognise the locations and types of objects in the video image It does not attempt to rate the saliency of the objects in the image Consequently, if there is more than one object in the scene which matches the description given by the user, the system is forced to ask the user for clarification (Socher and Naeve 1996)

In Socher and Naeve (1996), an end-to-end evaluation of the system is described The system's object identification results were obtained by running the system on 270 spontaneous utterances to objects in eleven different scenes The results were grouped into four categories correct, unique, false, or nothing The correct category contained all utterances where the intended object was among the ones identified by the system The unique category contained all utterances where the system correctly and uniquely identified the intended object The false category contained utterances where the wrong object was identified and the nothing category contained all the utterances for which the system made no selection The results from the text input showed that 72% of the utterances were classified as correct, 27% as unique, 6% as false, and 22% as nothing While the amount of correct identifications is impressive, the relatively low number of unique identifications (27%) highlights the lack of a mechanism for resolving linguistically ambiguous references

The Situated Artificial Communicator used a hybrid approach to modelling prepositions The system uses the degree of overlap of an object with discretised regions

171

as a measure of how well the object fulfils a preposition's meaning The algorithm for interpreting prepositions contained two steps

1 An object-centred relational representation is computed by partitions of 3-D space relative to the landmark These relations are independent of any frame of reference

2 Reference frames are superimposed to derive semantic definitions for prepositions relative to the landmark

The generation of the set of object-centred relations was "based on acceptance relations that are induced by acceptance volumes partitioning the 3-D space in an object-specific way" (Fuhr et al 1998 pg 3) It should be noted that the geometric object models created by this world reconstruction process were not used in the language interpretation module Instead, objects were approximated by their bounding boxes There were 79 acceptance volumes for each object the object's bounding box, one for each side of the box, two bound to each edge, and six bound to each vertex (see Figure 5-19) The motivation for defining 79 acceptance volumes is not given in (Fuhr et al 1998)

(a)   (b)   (c)
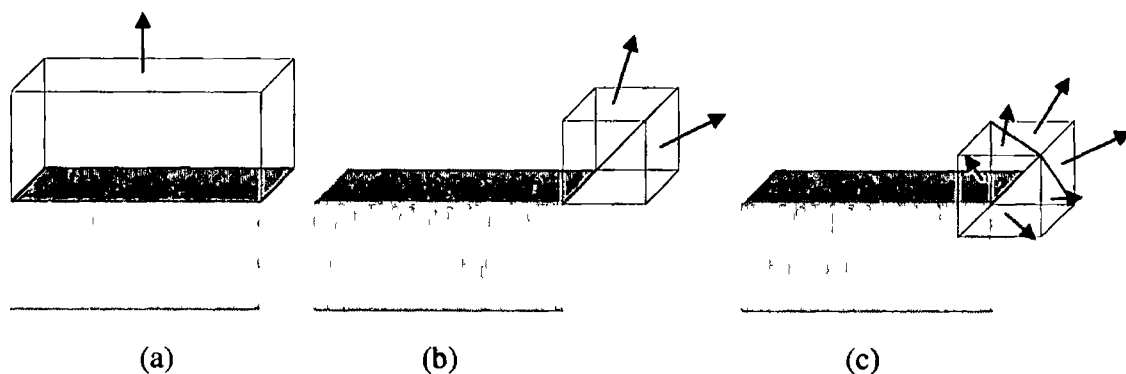
Figure 5-19 3-D acceptance volumes attached to an object's bounding box: (a) the acceptance volume defined by the top side of a bounding box, (b) the two acceptance volumes at an edge, (c) the six acceptance volumes bound to a vertex This illustration is based on an image in (Fuhr et al. 1998).

172

Let[44] *OBJECTS* denote the set of objects in the scene and O an object in the scene Based on each object's bounding box B_O a set of acceptance volume AV_t^O is defined for each object Each acceptance volume AV_t^O has a direction vector $d(AV_t^O)$ associated with it Each acceptance volume's direction vector approximates the direction that the edge, vertex, or face of the object's bounding box that the acceptance volume is associated with faces in space Next, for each object O, a set of acceptance relations r_t^O is defined Each element in the set of object O's acceptance relations r_t^O describes an intersection between the bounding box of one of the other objects in the scene, P, and one of object O's acceptance volumes, AV_t^O Formally, this is defined as $r_t^O \subseteq OBJECTS \times \{O\}$ with $(P,$ $O) \in r_t^O \Leftrightarrow B_p \cap AV_t^O \neq \emptyset$ A degree of containment for each object P - acceptance volume AV_t^O pair is defined as $\gamma(P, r_t^O) = vol(B_p \cap AV_t^O) / vol(B_p)$ with $\gamma(P, r_t^O) \in$ $[0,1]$

This containment measure represents the relative part of the object P in acceptance volume AV_t^O This representation of the scene is called the reference-independent representation and is stored as a relational network

Once the reference-independent representation has been computed, reference frames are superimposed to derive meaning definitions for prepositions relative to the landmark

A pair of inverse vectors represents each axis in a reference frame, for example, the front-back axis is given by the vectors *fb* and *bf* pointing from front to back and back to front, respectively Each of these vectors is associated with a particular preposition

For a given landmark[45] L, preposition *prep* and reference frame *ref*, the derivation of the semantic definitions of the prepositions is computed through a labelling procedure

---

[44] All the equations and numerical examples in this review of the Situated Artificial Communicator are quoted from (Fuhr *et al* 1998)

[45] Fuhr, Socher et al (1998) use the term reference object to describe the landmark and local object to describe the trajector In this section for the sake of consistency in terminology across the dissertation the terms landmark and trajector are used A consequence of this is that the equations in this section use different terminology to those given in (Fuhr *et al* 1998) the terms *RO* and *LO*, used in the original to denote the reference object and the local object respectively, have been replaced with *LM* and *TR*, which symbolise the landmark and trajector, respectively

173

that determines the set of acceptance relations that captures the meaning of the preposition This is called the definition set of the preposition and is denoted by

$$def(ref, prep, LM)$$

The condition for an acceptance relation's inclusion in this set is that the inner product of its directional vector and the vector associated with the preposition in the assumed reference frame is greater than 0 This is given by the equation

$$\{d(AV_i^{LM})| \, prepvector\} > 0$$

Equation 4. The equation defining the condition for an acceptance relation's inclusion in the definition set of a preposition in the Situated Artificial Communicator (Fuhr *et al* 1998). In this equation $d(AV_i^{LM})$ represents the direction vector associated with the landmarks acceptance relation i, and *prepvector* represents the vector associated with the preposition in the assumed frame of reference

Each acceptance relation r_i^O in the set *def(ref, prep, LM)* is associated with an applicability degree $\alpha(\text{ref}, \text{prep}, r_i^O)$ representing the strength of the compatibility of the acceptance relation with the meaning of *prep*

$$\alpha(ref, prep, r_i^O) = 1 - 2 \bullet arcos(d(AV_i^{LM})| \, prepvector\}) / \pi$$

Equation 5: The applicability degree of an acceptance relation in the Situated Artificial Communicator (Fuhr *et al* 1998)

An acceptance relation's degree of applicability has a linearly inverse relationship with the inner angle between the acceptance volume's direction vector and the vector associated with the preposition in a particular frame of reference The applicability of a relation approaches 0 as the inner angle approaches $90°$ Figure 5-20 shows the

calculation of the definition set for the preposition *behind* (a) and *left* (b) in a 2-D scene
(Fuhr *et al* 1998)

$$\Upsilon(TR, r_8^{LM}) = 0\ 37,\ \Upsilon(TR, r_9^{LM}) = 0\ 63$$

(a) def(ref, behind, LM) = {$(r_8^{LM}, 0\ 25), (r_9^{LM}, 0\ 75), (r_{10}^{LM}, 1\ 0), (r_{11}^{LM}, 0\ 75), (r_{12}^{LM}, 0\ 25)$}

(b) def(ref, left, LM) = {$(r_5^{LM}, 0\ 25), (r_6^{LM}, 0\ 75), (r_7^{LM}, 1\ 0), (r_8^{LM}, 0\ 75), (r_9^{LM}, 0\ 25)$}
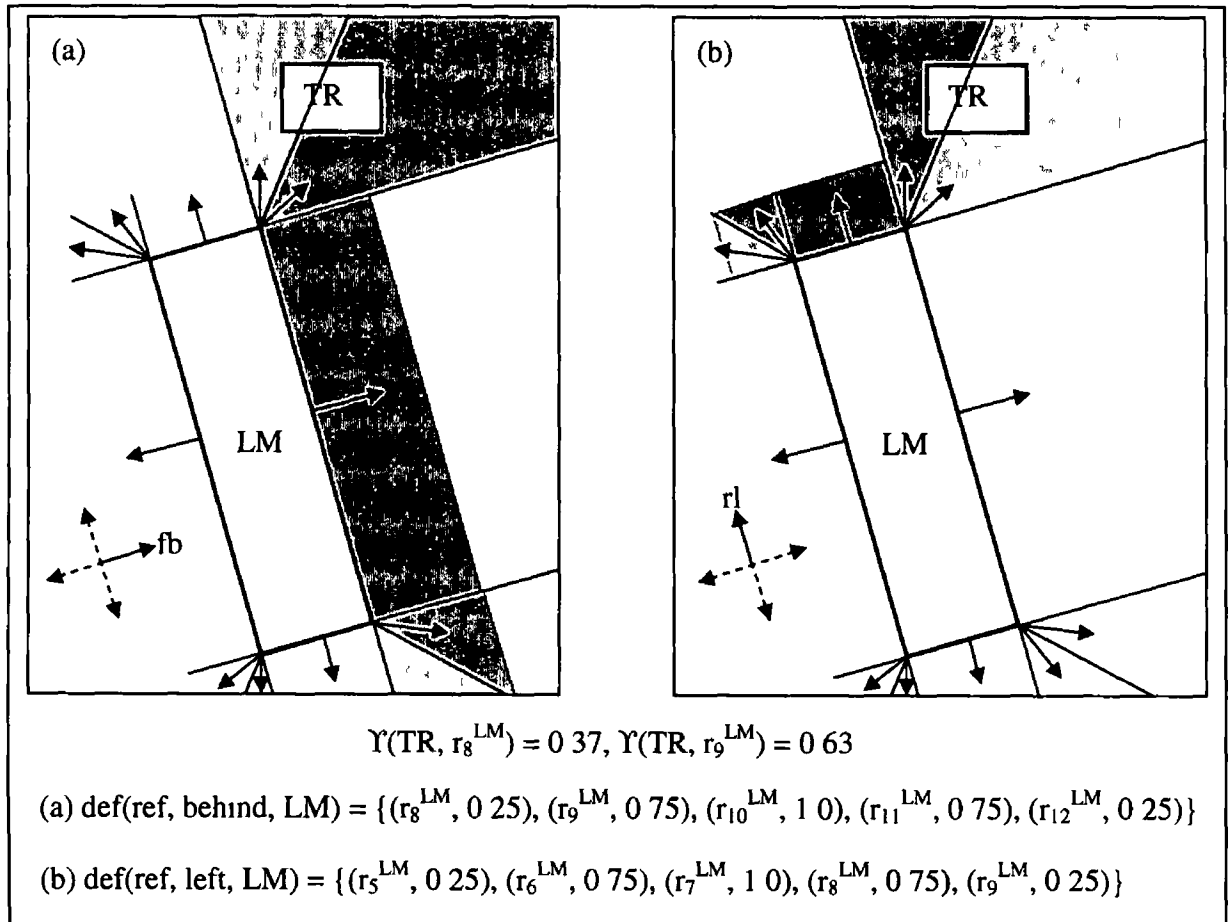
**Figure 5-20: Meaning definitions and trajector TR degree of containment for *behind*
and *left* in a give frame of reference. This drawing is based on an image in (Fuhr *et
al* 1998).**

A trajector fulfils a preposition with respect to a landmark in a given reference
frame if *def(ref, prep, LM)* is contained in the relational network for the reference
independent spatial representation

The measure of fulfilment of a trajector position with respect to a preposition
applied to a landmark in a given reference frame is calculated by

$$\delta(ref,prep, TR, LM) = \sum_{r_i^M \in def(ref, prep, LM)} \alpha(ref, prep, r_i^{LM}) \bullet \eta(TR, r_i^{LM})$$

Equation 6· The measure of fulfilment of a trajector's position with respect to a preposition applied to a landmark in a given reference frame in the Situated Artificial Communicator (Fuhr *et al.* 1998).

Using the above equation, the fulfilment of the trajector's position in Figure 5-20 (a) and (b) are

a $\delta(ref, prep, TR, LM) = (0\,25 \bullet 0\,37) + (0\,75 \bullet 0\,63) = 0\,57$

b $\delta(ref, prep, TR, LM) = (0\,75 \bullet 0\,37) + (0\,55 \bullet 0\,63) = 0\,43$

The computed results seem intuitively correct However, the system definition of a spatial preposition is based on the disjunction of acceptance relations Indeed, an object's fulfilment of a preposition is based on its overlap with a neatly discretised space The coarseness of the discretisation impairs the system's ability to grade object positions Consequently, in situations where candidate trajectors are contained in the same acceptance relation, no gradation in the fulfilment of the preposition is possible In Figure 5-21, TR1 and TR2 would be judged equal in fulfilling *to the right of LO* using the reader's viewer-centred perspective Clearly, this equality is not the case
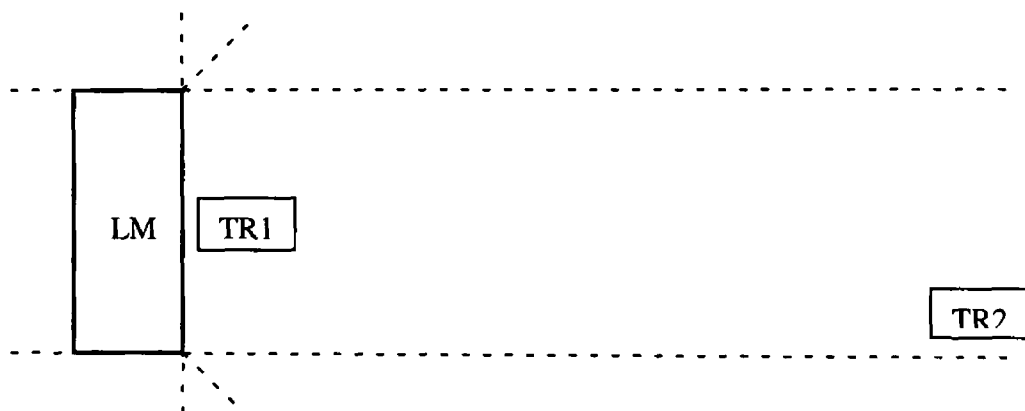
Figure 5-21 **Illustration of a scene where two trajectors TR1 and TR2 are fully contained within a single acceptability region. In this situation, both trajectors would be judged to be equal in fulfilling to the right of LM, using the reader's viewer-centred perspective**

Scenes with relatively large objects also cause problems for the system Fuhr *et al* give an example of such a situation Figure 5-22 illustrates the relative position of two elements in the test data used in (Fuhr *et al* 1998)
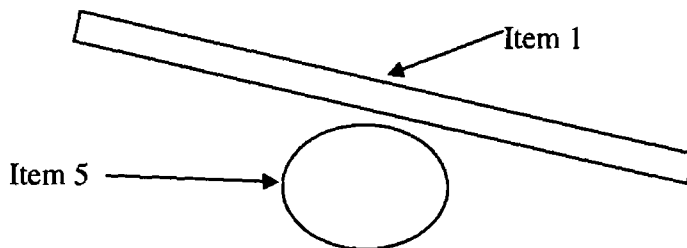
Figure 5-22 **An illustration of the relative position of two objects in a test scene observed by the Situated Artificial Communicator The relative position, shape, and labelling of the two elements in this drawing are based on Figure 5 in (Fuhr *et al* 1998)**

The Situated Artificial Communicator judged item number 5 in the scene as being *in front of* and *left of* item 1 "This results from the fact that the acceptance region

associated to the long side of the bar is contained in the definitions for the prepositions *front* and *left"* (Fuhr *et al* 1998 pg 7) Moreover, the system schematises objects by their bound box This introduces problems when dealing with convex or concave shaped objects (see Section 2 3 4 2 4)

Although the system claims the ability to handle different frames of reference, it is far from clear how this is achieved The process for defining the viewer-centred reference frame (which (Fuhr *et al* 1998) call deictic) is described

"In the case of deictic reference – and this is what we actually apply because our objects have no intrinsic orientation – the front-back axis is calculated dynamically as the line connecting the speaker's vantage point to the reference object's[46] centre of mass The bottom-top axis coincides with the speaker's vertical axis, and the left right axis is defined to be orthogonal to both other axes " (Fuhr *et al* 1998 pg 4)

In order to handle an intrinsic frame of reference, Fuhr *et al* grant that the vectors associated with the prepositions "should be known a priori and given with the object model" (1998 pg 4) However, there is no description of how this is done, moreover, the problematic issue of selecting a reference frame is not addressed The worked example assumes a viewer-centred frame of reference on the basis that the objects in the scene have no intrinsic orientation, and any general discussion of the overall algorithm assumes the appropriate reference frame is in some way given to the system

---

[46] As noted above (see Footnote 45 on page 173) Fuhr Socher et al (1998) use the term reference object to describe the landmark

5.4 6 Virtual Director

The Virtual Director system is described in (Mukerjee *et al* 2000) The Virtual Director system reconstructs a scene based on natural language input The linguistic input is restricted to a set of linguistic descriptions related to a limited domain an urban park Similar to the SPRINT system (in Section 5 4 3 above), the fact that the user supplies the Virtual Director system with a description of the world obviates the need to model the user's knowledge of the world since, a priori, they have complete knowledge of the world Nonetheless, however, there is no mention of a model of visual salience or discourse in the Virtual Director As a result, the system cannot resolve anaphoric references or references which are not completely disambiguated by the linguistic content

There are two principle components in the system a large database of objects and actions and a set of constraints corresponding to default dependencies in the domain The system uses a continuum approach to model the spatial constraints specified by prepositions in the input An object whose location is described by a prepositional phrase is instantiated at the global minimum of the continuum Where multiple constraints involving the same trajector are given, the continuum fields are combined to provide a resultant field The object is placed at the minima of the combined continuum The continuum field is created by first defining the location of the field's global minimum Next, a set of concentric ellipses that use the global minimum as a fixed focus are created by varying the eccentricity of the ellipse and the position of the second focus "The eccentricity of the ellipse and the location of the second focus follow a constraint such that the field does not intersect the object[47]" (Mukerjee *et al* 2000 pg 10) There are several weaknesses in this model One problem with this approach is that the continuum field is only 2-D Moreover, defining the global minimum for a preposition is problematic Although experiments carried out by Mukerjee *et al* (2000) identified a dependence between the distance of the global minimum from the boundary of the landmark and the space available to the preposition, there was a wide interpersonal

---

[47] The term object in this quotation refers to the landmark of the preposition phrase being modelled

variation in the results, i e , the location of a preposition's global minimum varies from person to person irrespective of the space available for its location Finally, the Virtual Director uses a very simple and cognitively implausible process for handling the issue of frames of reference the system defaults to the intrinsic frame of reference If the landmark does not have an intrinsic frame of reference associated with it, the system then uses the viewer-centred frame of reference

5 4 7   CommandTalk

The CommandTalk system (Dowding et al 1999, Stent et al 1999, Goldwater et al 2000) is an NL interface to the ModSAF (Modular, Semi-Automated Forces) battlefield simulator "The goal of the system is to allow military commanders to interact with simulated forces in a manner as similar as possible to the way they would command actual forces" (Goldwater et al 2000pg 1) The system allows the user to use NL commands and mouse gestures to

- Create forces and control measure (points and lines)
- Assign missions to forces
- Modify missions during execution
- Control ModSAF system functions, such as map display
- Get information about the state of the simulation

(Goldwater et al 2000)

The CommandTalk dialogue component adopts Centering Theory (see Section 4 3) as its theoretical basis (Stent et al 1999) "The system supports natural, structured mixed initiative dialogue and multimodal interactions" (Stent et al 1999 pg 186) Similar to Centering Theory, CommandTalk uses a dialogue stack to keep track of the current discourse context A stack push operation corresponds to the onset of a discourse segment, while a stack pop operation corresponds to the conclusion of a discourse segment Each stack frame corresponds to a sub-dialog in the discourse In (Stent et al.

1999), these stack frames are described as finite state machines Indeed, the dialogue manager is compared to recursive transition network

"The dialogue stack is reminiscent of a recursive transition network, in that the stack records the system's progress through a series of FSMs (Finite State Machines) in parallel " (Stent *et al* 1999 pg 187)

It should be noted that while such an approach is suitable for structured discourses, the use of finite state machines to model a dialogue constrains the permissible input at any given time in the dialogue Consequently, systems using this type of technology will find it difficult to handle unpredictable input or to correlate the dialogue model's expectations with a user's agenda after an unpredicted input

Most of the stack frames in CommandTalk have very simple structures CommandTalk implements 22 different simple sub-dialogs and three more complex dialogs (Stent *et al* 1999) Examples of the simple subdialogs include stack frames representing discourse segments for clarification questions, references failures, corrections, etc The three more complex stack frames represent dialogs for the embark/debark command, the infantry attack command, and a form filling dialog

CommandTalk supports both singular (such as proper names, definite descriptions, and pronouns *it*), and plural (such as plural descriptions, quantified descriptions, conjunctions, and pronouns *them*) references (Dowding *et al* 1999) CommandTalk uses two mechanisms for maintaining a local context in which to resolve references Firstly, the ModSAF system provides the CommandTalk interpretive module with a representation of events in the simulated world Secondly, CommandTalk uses focus spaces to model entities realised in linguistic utterances, including objects not directly represented in the simulation (Stent *et al* 1999) There is one focus space associated with each utterance Each focus space contains a reference to all the items referred to in its associated reference A focus space represents what was known at the time its associated utterance was input to the system Focus spaces are used during the interpretation of user responses to system questions and when a user corrects a previous input

There are several limitation of CommandTalk for the research pursued in this thesis Firstly, as was noted above, CommandTalk's dialogue management component uses data structures that are similar to finite state machines to model sub-dialogs Such an approach presupposes a structured dialogue where a designer can predict the set of possible branches within a dialog This is not surprising when one considers that this presupposition is a basic assumption of the theoretical dialogue model that was adopted by the designers of CommandTalk Centering Theory As a result, CommandTalk will have difficulties in handling unpredictable input Moreover, porting the interface to other domains is non-trivial Secondly, the interpretive process is grounded in a world model Consequently, the system has no mechanism for adjudicating between two or more candidate referents As a result, when there is more than one world object that fulfils the description of an expression's referent the system is forced to ask the user for clarification This is illustrated by one of the sample system dialogues

U Create a CEV at 72 69

S ☺

U CEV, conduct a crater breach facing south

S ☒ There are two CEVs Do you mean 100A11 or 100A12?

(Goldwater *et al* 2000 pg 3)

In this dialog, U labels user inputs and S CommandTalk's responses The ☺ symbol represents the system outputting a rising tone, which indicates the successful interpretation and execution of the user's command The symbol ☒ represents the system outputting a falling tone which indicates that the system was unsuccessful in interpreting the user's input Finally, the CommandTalk system does not handle locative expressions

182

5 4.8 VIENA. Virtual Environment and Agents

The VIENA project (Cao *et al* 1995, Wachsmuth and Cao 1995, Jordmg and Wachsmuth 2002) was developed at the University of Bielefeld The aim of the project was "to provide a way of intelligent communication with a technical system for designing and generating 3-D computer graphics" (Wachsmuth and Cao 1995 pg 1) The system allowed the user to use natural language commands to interact with an interior design application

As interesting facet of the VIENA system was the development of an agent that represented the system in the visual domain This agent was called Hamilton and had a humanoid appearance The VIENA project used the Hamilton agent to allow the user to change perspective The user could switch their view from an external view where Hamilton is visible in the scene to an internal view or involved view where the user viewpoint is located in Hamilton's forehead In either perspective, the user can direct the agent to move in the scene (Jordmg and Wachsmuth 2002)

The VIENA project used a multi-agent approach – where an agent is defined as "an entity consisting of a structural definition, a set of functional units that defines its behaviour repertoire, and some means of selecting and sequencing (possibly concurrent) behaviours" (Wachsmuth and Cao 1995 pg 6) There were four core agents in the system involved in the interpretation process

- Augmented database / bookkeeper agent The rendering system maintained a graphics data base which held the information necessary for drawing the scene The graphics database was mirrored in the agent environment, as an augmented graphics database Besides the current scene description, the augmented graphics database also held information about previous scenes This scene history was time stamped and was used to evaluate elliptic discourse (e g , *a little more)* The bookkeeping agent controlled access to and modifications of the augmented database Each time a new scene was drawn, the augmented database was updated accordingly

183

- Parser agent The parser agent translated user instructions into a structured representation that the interpreting agents could use If an instruction could not be resolved by the agents or could not be parsed, the parser agent would request clarification from the user

- Interpret agent The interpret agent functioned as a router It took the structured representation of the user input, created by the parser agent, and decided to which agent the instruction should be sent to

- Space agent The space agent had two responsibilities "(1) to identify mentioned objects and (2) determine where and how an object will be moved – in relation to other objects and avoiding collisions – in order to satisfy a user input" (Wachsmuth and Cao 1995 pg 11)

It is important to note that the VIENA project's approach to interpreting language in a simulated visual environment is similar to that used in this thesis That is, the interpretation of an utterance should be grounded in information available in the visual context of the utterance This approach is based on the assumption that "as the user gets immersed in the visual scene, verbal statements likely make reference to what can be seen in the current situation" (Cao *et al* 1995 pg 1)

Nonetheless, the situational information used in the VIENA system is restricted relative to the framework proposed in this thesis In particular, the VIENA system had no model of visual salience Rather, the VIENA system relied on a temporally based rating system to resolve linguistically ambiguous references This temporal salience was based on a time stamping mechanism, ascribing to each object the time of the last rendered scene the object was in Wachsmuth and Cao (1995) used the example input *move the table right* to describe how these time stamps were used to resolve references

"the Bookkeeping agent determines first which object named 'table' is addressed by the instruction, and then reads the geometry data of this object to the Space agent If more than one object is named 'table', the 'table' object with the most current time stamp is selected, or further input is requested" (Wachsmuth and Cao 1995 pg 11)

184

The problem with this approach is that the VIENA system could not resolve a referring expression if there was more than one object in the scene that matched the linguistic description of the referent without seeking clarification from the user, irrespective of the relative visual salience of the candidates.

To illustrate this shortcoming, a scene taken from the SLI system is shown in Figure 5-23 below. Given this visual context, if a user input the command *make the red house taller*, it is evident that they would be referring to the red house in the foreground and not the red house on the periphery of the scene. The SLI system is able to resolve this reference using the proposed model of visual salience without requesting a clarification from the user. In contrast, the VIENA project would treat the red houses in this scene as equally likely candidates as they would have the same time stamp associated with them.

Figure 5-23: A scene taken from the SLI system that illustrates the importance of visual salience in the resolution of linguistically ambiguous references.

The VIENA system implemented a very simple model of discourse

"In our setting we think it adequate that the user input is kept to the minimum significant information Thus we are not trying to process very complex sentential structures " (Wachsmuth and Cao 1995 pg 9)

Indeed, the dialog model was restricted to handling consecutive inputs Furthermore, the range of linguistic constructions that this dialog model could link to the previous input was restricted to modifying expressions and user responses to system requests for clarification In the context of the VIENA system, a modifying expression is an expression that parameterised the semantics of the adverb in the previous input For example, if the user inputs the instruction *move the chair right*, in the following input they could use a modifying expression such as *a little more* In this thesis, it is important to note that the only type of referring expression used in the example dialogs in the VIENA system is non-anaphoric definite descriptions Moreover, there is no description of how the system would handle anaphoric definite descriptions, indefinite descriptions, pronouns, conjunctions, or locative expressions

Although the VIENA system did allow the user to use prepositions, this was restricted to adverbial prepositions Consequently, the project did not propose a spatial template model nor could it handle locative expressions However, the VIENA system did implement an approach to resolving frames of references, which stems from the need to resolve in which direction an object should be translated in response to a user instruction that used a projective preposition as an adverb This issue is further exacerbated in the VIENA domain in a situation where the user is using an external perspective (e g , the Hamilton agent is visible in the scene) since this adds a third frame of reference Hamilton's intrinsic frame of reference For example, when interpreting the instruction *move the chair left*, the system needs to decide whether to use the chair's intrinsic frame of reference, Hamilton's intrinsic frame of reference, or a viewer-centred frame of reference when computing the transformation applied

The VIENA system adopted a simple approach to this issue The system assumes the viewer-centred frame of reference as a default and computes the translation using this frame of reference first "If this realisation does not match the expectation of the user, s/he can correct the system by stating 'wrong'" (Jording and Wachsmuth 2002 pg 11) The system then computes the transformation using either Hamilton's or the object's intrinsic frame of reference For this thesis, there are several problems with this approach Firstly, it ignores the psycholinguistic evidence which indicates that when the frames of reference are dissociated multiple frames of reference are activated, and this multiple activation interferes with spatial template construction (Sections 5 3 1 1 and 5 3 1 3) Secondly, when processing a locative expression, there may be more than one candidate object within the spatial template of a preposition in a given frame of reference Consequently, if the user corrects the system's selection of the referent this does not necessarily indicate that the wrong frame of reference was selected Rather, the user could be referring to another object in the same reference frame's spatial template, but with a lower spatial template rating In order to handle this possibility, a system that processes locative expressions must be able to adjudicate between each of the candidate referents based on an individual overall rating based on all the relevant frames of reference, rather than on a rating based on the assumption of a default frame of reference

5.5   Chapter Summary

The purpose of this chapter was to critically review work related to this thesis and to highlight their strengths and weaknesses

In Section 5 2, previous models of visual attention were reviewed It was noted that connectionist architectures developed as vision modules for robots are not suitable for avatars in rendered 3-D environments for two reasons Firstly, the major difficulties facing robotic vision (pattern recognition, distance detection, and the binding problem) do not affect models vision for simulated environments Secondly, the training required by connectionist architectures makes them impractical for applications that have a broad range of inputs Next, the models of vision that utilise graphics techniques were

examined First, models of vision using the ray tracing technique were examined It was noted that ray tracing is a computationally expensive function and not used in real-time rendering Following this, the models of vision that used a false colour approach were examined The false colouring approach avoids the computational expense associated with ray tracing Several of the systems that used this model used the output of the visual module to create a visual memory of the simulated environment based on what the avatar has observed However, it was found that none of these systems use this information as an aid to interpreting linguistic input Indeed, the data structures used in these systems are not suitable as inputs to a linguistic context model Furthermore, although Peters and O'Sullivan's (2002) approach has a limited model of attention based on the number of times an object has been observed, the majority of the false colouring models make no attempt to rate the saliency of the observed objects Moreover, Peters and O'Sullivan's (2002) model is only updated when the avatar attends directly to a goal object that it has searched for and only updates the observations pertaining to that goal object This approach to attention is not suitable as a method for creating a visual memory of an environment which may be used as part of a linguistic context model because it requires the avatar to search the environment for objects which may have already been seen but not attended to directly, and thus not noted

Section 5 3 examined work related to the issue of locative expressions, beginning with a review of the literature on frames of reference It was found that when frames of reference are dissociated more than one reference frame is initially activated and these active frames compete Moreover, there are biases present in this competition between reference frames These biases are dependent on the orientation of the plane that a given spatial term is canonically aligned with The process of selecting a frame of reference impacts on the construction of a preposition's spatial template if there is a competition between reference frames, the spatial templates constructed for each of the competing reference frames should be amalgamated Following this, previous work related to computational modelling the semantic models of prepositions was reviewed This review began by describing neat models and the issues affecting them Next, the scruffy or continuum models were reviewed The review of scruffy models was quite brief as they were reviewed in detail during the review of language and vision systems in Section 5 4

It was noted that some of these models only work in 2-D (Yamada 1993, Olivier *et al* 1994, Mukerjee *et al* 2000) One (Fuhr *et al* 1998) has problems distinguishing between the position of trajectors that are fully enclosed within a region Most use the centroid of the object's bounding box to represent the objects (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994, Fuhr *et al* 1998) which is problematic (see Sections 2 3 4 2 4 and 2 3 5 1) Those that do not (Mukerjee *et al* 2000) are dependent on locating the local minimum within the continuum field of a preposition This is problematic as the location of the local minimum varies from person to person (see Section 5 4 6) Furthermore, all these models ignore perceptual information For the purposes of this thesis, none of these models propose a cognitively plausible approach to the issue of reference frame selection Consequently, they ignore the impact that selecting a frame of reference can have on the construction of a spatial template (see Section 5 3 1 3)

Section 5 4 examined previous computational systems that integrated vision and language None of the systems propose a cognitively plausible approach to the issue of selecting a frame of reference Also, as noted in Section 5 3 2 2, none of these systems proposed a semantic model for prepositions that addresses the issues highlighted in Sections 2 3 4 and 2 3 5 In particular

1   How to model the affect of perceptually based cues, such as object occlusion, on a preposition's spatial template? (Section 2 3 4 2 3)

2   How to locate the origin of the spatial template? (Section 2 3 4 2 4)

3   How to model the trajector? (Section 2 3 5 1)

4   How to handle the issue of occluded trajectors? (Sections 2 3 5 2)

Furthermore, the SOCCER system is the only system that attempts to model the user's knowledge of the environment However, the model of user's knowledge built by the SOCCER system is based solely on linguistic utterances and consequently it does not tackle the problems inherent in modelling a visual domain, such as how to model visual attention and how to integrate this knowledge with linguistic knowledge Finally, the SOCCER system is also the only system that models previous discourse Again, however, the only input to this discourse model is the previous linguistic utterances and

189

consequently, it is not feasible as a context model for systems where the referents of anaphoric or demonstrative references may be introduced by visual perception

6 The Situated Language Interpreter

6.1 Introduction

This chapter introduces the SLI system, which implements the interpretive framework developed in this thesis In this framework the interpretive modules of an NLVR system models the linguistic context _and_ the user's perceptual context This approach differs from previous NL systems that attempted to interpret language grounded in a visual domain, (i e , SHRDLU (Winograd 1973), CITYTOUR (Andre _et al_ 1986, Andre _et al_ 1987), CSR-3-D (Gapp 1994a), SPRINT (Yamada 1993), WIP (Olivier _et al_ 1994, Olivier and Tsuji 1994), Situated Artificial Communicator (Socher and Naeve 1996, Socher _et al_ 1996, Vorwerg _et al_ 1997, Fuhr _et al_ 1998), Virtual Director (Mukerjee _et al_ 2000), and CommandTalk (Dowding _et al_ 1999, Stent _et al_ 1999, Goldwater _et al_ 2000)) as these systems – rather than modelling the user's knowledge of the simulation – gave their interpretive module complete access to all the objects in the simulated world This is phenomenologically unrealistic and impractical in large environments with many objects where such an approach can result in a multiplication of the possible referents

This chapter gives an overview of the SLI system's architecture and provides an example user-system interaction scenario that illustrates some of the system's functionality The scope is restricted to a general introduction and a high-level description of how the systems functions Chapters 7, 8 and 9, give a detailed description of the components of the SLI architecture and the interpretive framework underlying it

6.2   The SLI System Architecture

There are six components in the SLI system architecture the parser, the rendering engine, the world model, the visual saliency module, the context model, and the interpretive module

6.2 1   The SLI Parser

The SLI parser is simple and was developed by the author for the system It is written in Lingo, an object-oriented programming language similar to C++ or Java The only aspects of the SLI parser specific to this project are the categorising of an input into different types of commands and explicit checking for linguistic cues that impact on the frame of reference selection, in particular the use of a genitive nouns phrase to describe the landmark of a locative Essentially there are two types of commands that the system handles

> 1   Avatar Commands   user commands to their avatar, For example, *walk forward* or *turn right*
>
> 2   Artefact Commands   user commands which change the state of the world by altering attributes of any world artefact apart from the user's avatar For example, *make the house taller* or *move the tree to the right of the blue house forward*

Categorising the input is based on the verb used and the syntactic structure of the command For example, if the verb is either *run* or *walk*, it is categorised as an avatar command However, if the verb is *move* it could be either, e g , *move forward* (avatar command) or *move the house back* (artefact command) Here, *move + adverb* is categorised as an avatar command and *move + np +* is categorised as an artefact command Once the parser has processed the input string, categorised the type of

192

command, and checked for frame of reference cues, it passes its results on to the interpretive module, described later

6 2 2   The SLI World Model

The world model is a list of the objects to be drawn on the screen by the 3-D rendering engine A similar form of data structure is present in all 3-D applications Each element in this list describes the geometry, surface appearance, world position, and frames of reference associated with each world artefact The 3-D geometry of each object is represented by a mesh, the world position by a vector, and the frames of reference orientations by a set of unit vectors A mesh is an interconnected network of points that defines a 3-D model's surface The only part of this model specific to this thesis is the inclusion of the frame of reference information This, however, is not novel, since several other systems (e g , (Olivier and Tsuji 1994)) use similar approaches

6 2.3   The Rendering Engine

The rendering engine used by the SLI system is the Shockwave 3-D Viewer A **rendering** engine takes a geometric model of a scene and camera viewpoint and draws the scene on the screen There are several rendering engines available at present, all of which provide similar functionality Two of the best known are DirectX and OpenGL The motivation for using the Shockwave 3-D rendering engine in developing the SLI system was to deploy the system over the Internet

The novelty of the system resides in the visual saliency module, the context module, the interpretive module, and how these components interact, which are described in detail in Chapters 7, 8 and 9 A high-level overview of their role in the SLI framework is given here

The purpose of the SLI visual saliency module is to capture the flow of visual information from the simulated environment to the user This module extends the false

colouring technique described in Section 5.2.3. The output of this module feeds into the context module.

The role of the SLI context model is to provide a context for the interpretation of referring expressions, integrating and structuring the perceptual information captured by the SLI visual saliency algorithm and the linguistic discourse history. Because the SLI context module extends the scope of context beyond the linguistic discourse, the SLI system can resolve references to visually accessible objects which have not been previously mentioned in the discourse.

The SLI interpretive module takes user commands from the parser and uses information from the context model to interpret these commands. This process results in the updating of the world model and the context model. It is important to note that, unlike previous systems that have interpreted language which is grounded in a visual domain, SHRDLU (Winograd 1973), CITYTOUR (Andre *et al.* 1986; Andre *et al.* 1987), CSR-3-D (Gapp 1994a), SPRINT (Yamada 1993), WIP (Olivier *et al.* 1994; Olivier and Tsuji 1994), Situated Artificial Communicator (Socher and Naeve 1996; Socher *et al.* 1996; Vorwerg *et al.* 1997; Fuhr *et al.* 1998), Virtual Director (Mukerjee *et al.* 2000), and CommandTalk (Dowding *et al.* 1999; Stent *et al.* 1999; Goldwater *et al.* 2000), the SLI interpretation module does not directly access the world model. Rather, the interpretive module's knowledge of the world is mediated through the SLI context model and, consequently, is founded on the information captured by the SLI visual saliency module.

Figure 6-1 is a schematic overview of the system. Each of the boxes in this figure represents and names a component in the system. The arrows between boxes represent the flow of information between the components: arrow (1) represents the user inputting commands to the system; arrow (2) represents the parsed input being passed to the interpretive module; arrow (3) represents the flow of contextual information from the context model to the interpretive module (this information provides a context for the interpretation of user input received from the parser); arrow (4) represents the updating of the context model after the interpretation module has processed a user input; arrow (5) represents the updating of the world module after the interpretive module has processed a user input; arrow (6) represents the rendering engine's use of the world model during the rendering process, represented by arrow (7); arrow (8) represents the interrogation of the

194

rendering engine by the visual saliency algorithm after each frame has been rendered; arrow (9) represents the visual saliency algorithm's use of the false colouring information stored in the world model and its interrogation of the world model; and arrow (10) represents the updating of the context model by the visual saliency module after a frame has been rendered.
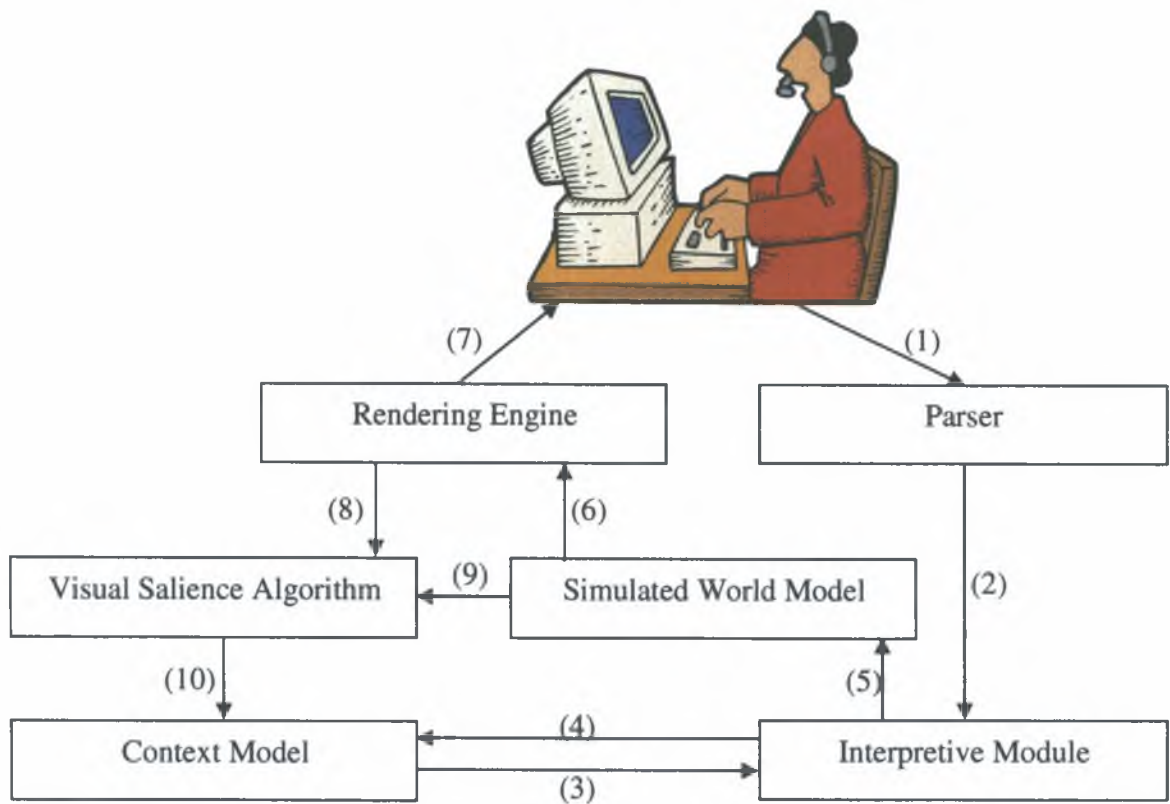
Figure 6-1: Schematic of the SLI system architecture and the data flow between the system components.

6.3 Example User-System Interaction Dialogue

This section contains a sample user-system interaction dialogue. The setting of the dialogue is a simple scene containing houses and trees. The numbered, italic and lower case text below was input by a user. The system's response to these inputs was displayed on the screen either as changes to the visual environment or text output in the SLI message window. All figures here are SLI screen shots which update the visual scene in response to an input. The dialogue was carried out in real-time.

Figure 6-2 shows the start-up screen. The major components of the SLI interface are named. Note the house and trees in the 3-D simulation window.
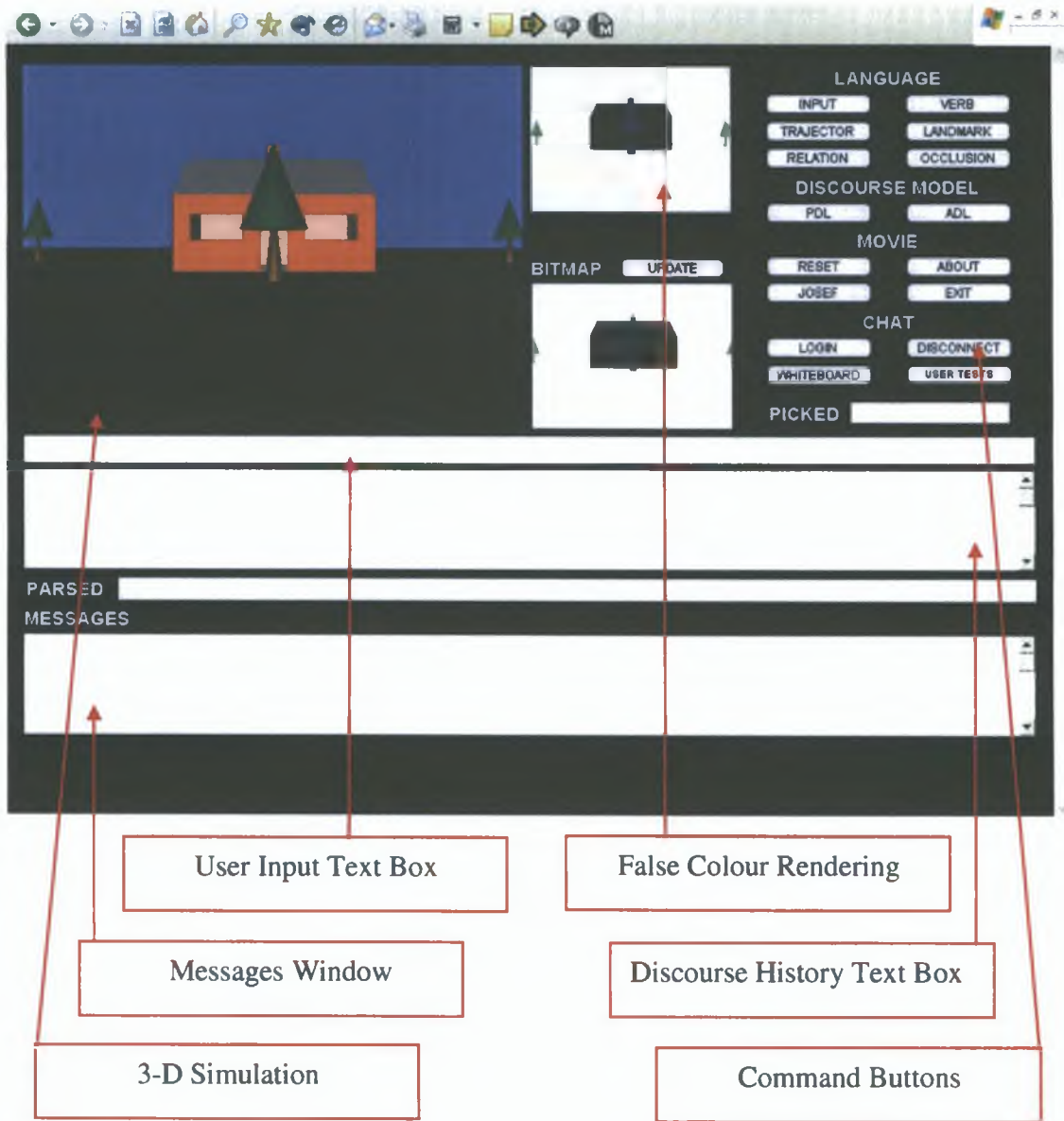
Figure 6-2: The SLI Interface.

The following is a sample input:

Input 6-1: *make the tree to the right of the house red*

In order to interpret Input 6-1, *make the tree to the right of the house red*, the system first resolves the landmark nominal *the house* It should be noted, that although there is more than one object in the world that fulfils the description *the house* (the complete 3-D simulation contains a number of houses currently not in the view, the SLI system is able to resolve this reference because its interpretive module does not directly access the world model Rather, its knowledge of the world, which is stored in the context model, is mediated by the visual saliency algorithm Consequently, despite the fact that there are multiple houses in the world, the system can use the knowledge that the user is currently only aware of one of these houses to resolve this definite description Importantly, systems that allow their interpretive module complete access to the world model would not be able to resolve a definite description if there was more than one object in the simulated world that fulfilled the description Furthermore, discourse models that are only updated in response to linguistic input would not have recorded the perceptual event of the user seeing the house Consequently, these discourse models would be of limited use in resolving this reference

Having resolved the landmark reference, the system constructs a model of the projective preposition's semantics using a potential field The system uses this potential field model to adjudicate between the elements of the set of candidate trajectors The set of candidate trajectors is the set of objects in the context model that fulfils the description of the subject noun phrase in the input Again, because the SLI's interpretive module does not have full knowledge of the world, the system can restrict the set of candidate objects to those that fulfil the linguistic description of the input and, crucially, that the user is currently aware of In this instance, there are three trees in the context model The system uses its model of the preposition's semantics to select between these candidates Once this selection is complete, the system updates the world model to reflect its interpretation of the user's input

Figure 6-3 illustrates the state of the simulated world after this input has been processed
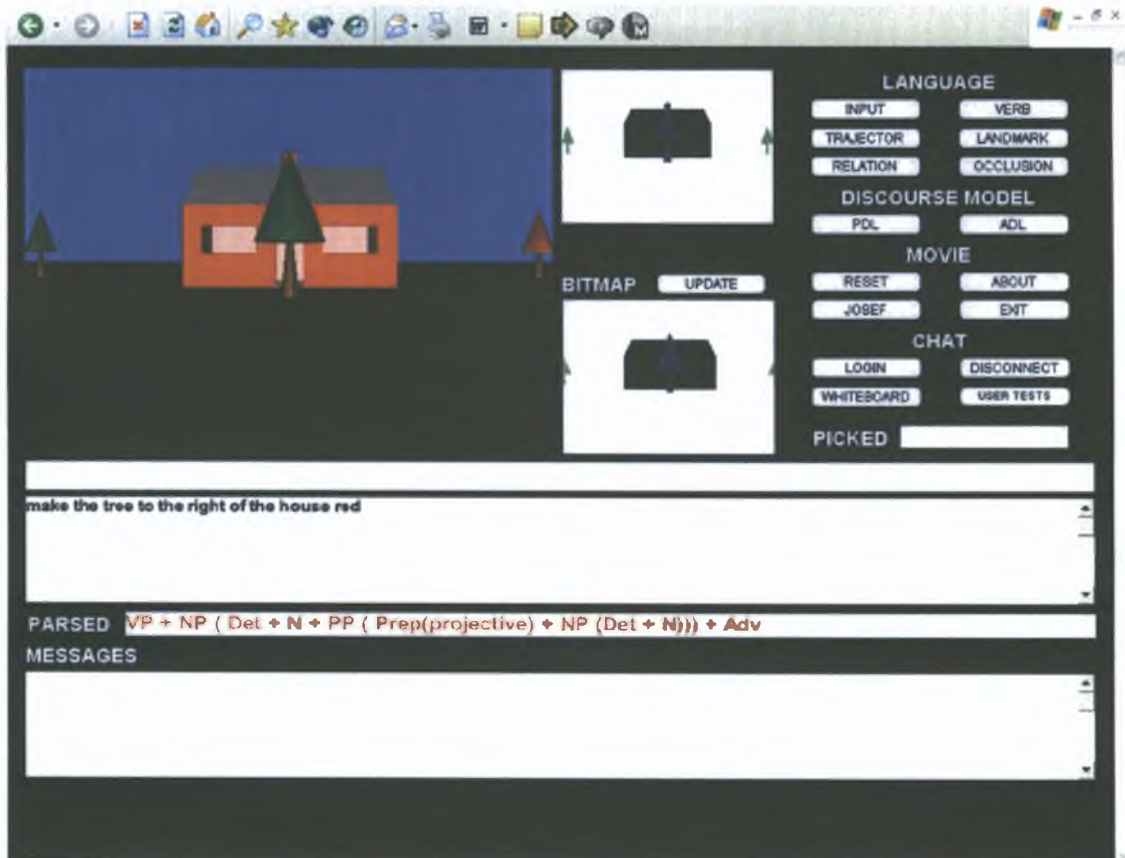
Figure 6-3: The state of the system after the input *make the tree to the right of the house red.*

Input 6-2: *make the blue house taller*

Although there are blue houses in the world, the user has not seen any of them during this interaction dialogue. As a result, there are no references to a blue house in the context model. Consequently, the system responds by outputting a message to the user in the message window:

No object of that type in memory!

Figure 6-4 illustrates the state of the system after this message has been output.

199
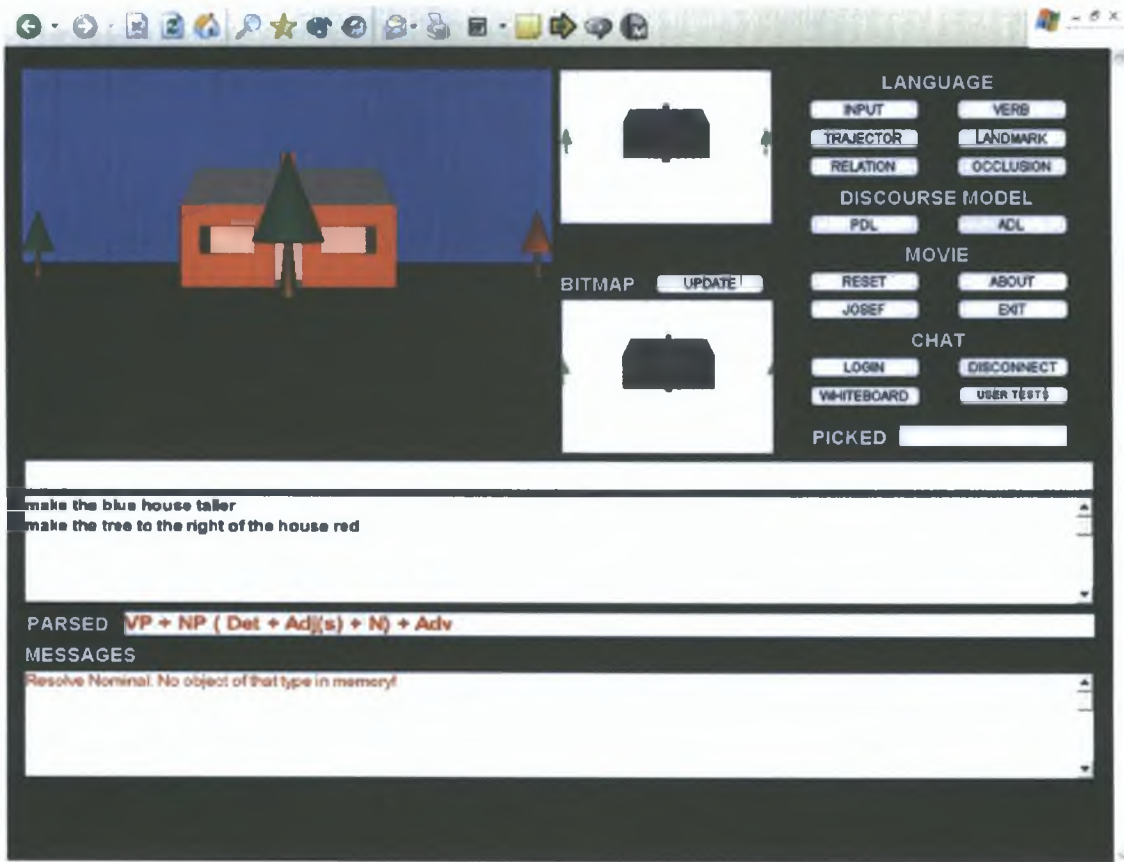
Figure 6-4: The SLI system after it has output a message to the user.

Input 6-3: *turn left*

Input 6-3 is an avatar command; i.e., it modifies the user's viewpoint. The system recognises these types of commands by examining the syntactic structure of the input (Verb + Adverb). Once the system determines that an input is an avatar command, it adjusts the user's viewpoint accordingly.

Figure 6-5 illustrates the user's view of the simulation after the input *turn left* has been processed.
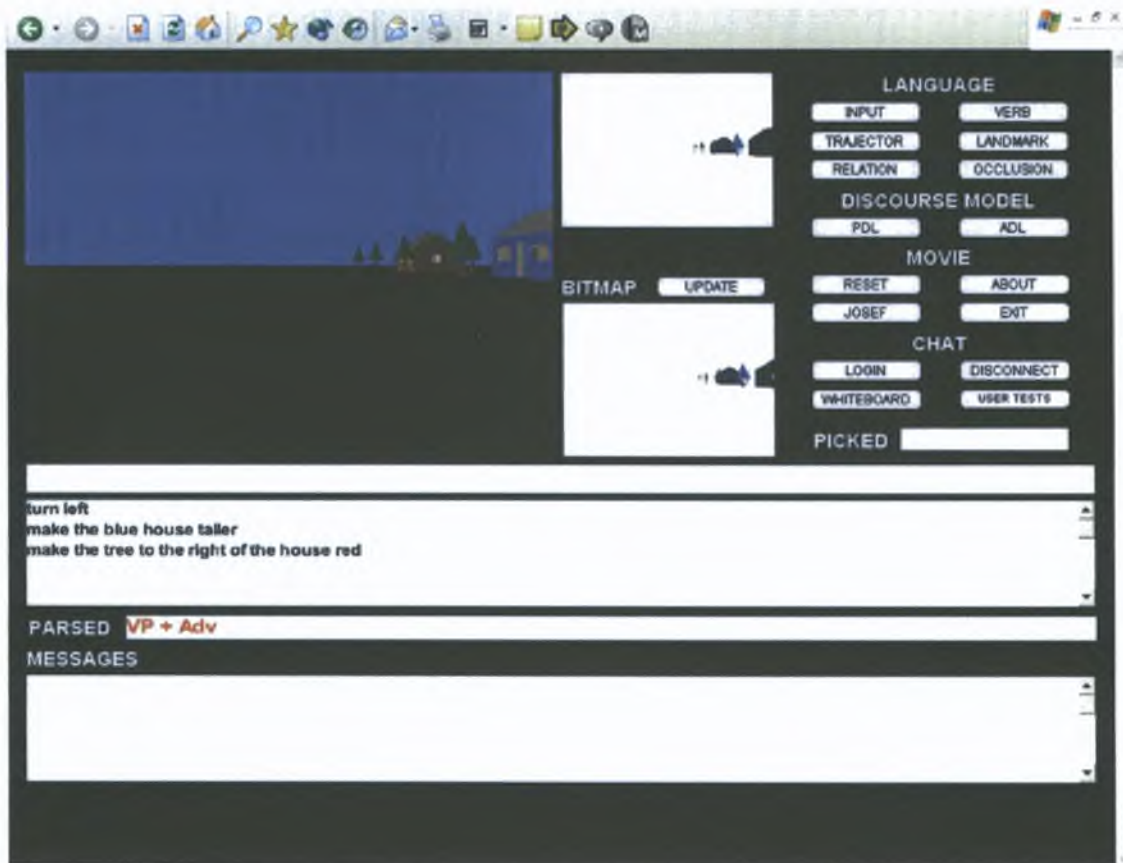
Figure 6-5: The user's view of the simulation after the input *turn left* **has been processed.**

Input 6-4: *make the blue house taller*

Input 6-4 is a repeat of Input 6-2. Earlier when this command was input, the system responded with: No object of that type in memory! However, now there is a blue house in the user's view volume. Recall that the system's context model is updated by the visual saliency module as well as the interpretive module (see Section 6.2), and as such, there is a blue house in the context model at this point. Therefore, the system can resolve the definite description *the blue house* in the input.

Figure 6-6 illustrates the visual context after this input has been processed. Comparing Figure 6-5 and Figure 6-6 highlights the change in the height of the blue building as a result of this input. More importantly, when this input was being processed there were three different houses in the context model: the brown and blue houses in the

201

current scene and the red house seen previously. However, the system was able to resolve the nominal *the blue house* by using the supplied adjectival description as a selectional restriction on the referent of the expression. The processing of this input also illustrates the system's ability to manipulate the properties of world objects in response to user input. The magnitude of the scaling of an object's dimensions is calculated by the system using simple heuristics. Note that changes to object properties in the main 3-D environment are mirrored in the false colouring rendering, evident when the false colour renderings in Figure 6-5 and Figure 6-6 are compared.
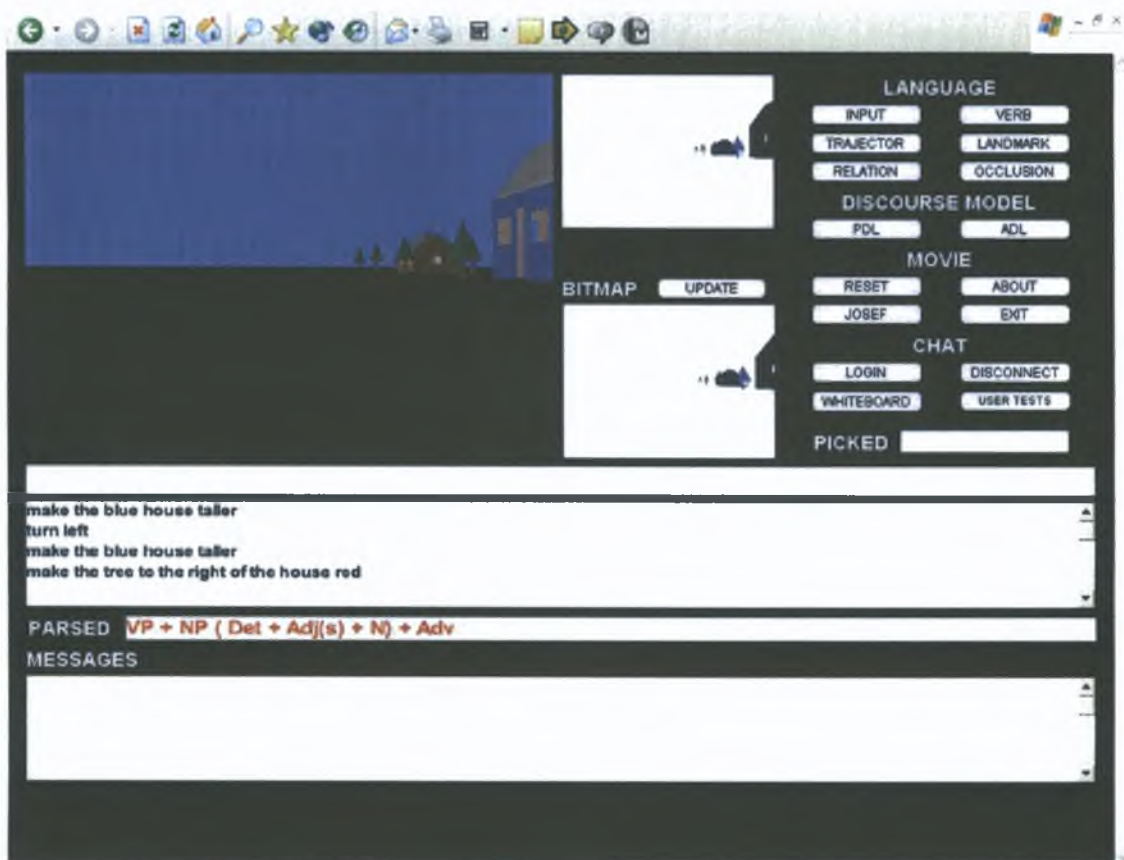
Figure 6-6: The SLI simulation after the input *make the blue house taller* has been processed.

Input 6-5· *make the other one yellow*

Interpreting Input 6-5 involves figuring out what it meant by *other* and *one* It is important to note that, although Input 6-5 uses a one-anaphoric expression, modified by *other* which is also anaphoric, the most likely referent of this expression (i e , the brown house that is currently in the view volume) has not been previously mentioned in discourse Again, however, because the SLI's discourse model is updated by the visual saliency algorithm, the object in question is in the context model

A detailed description of how the system interprets one-anaphora and other-anaphora will be given in Chapter 9 However, here, it will suffice to provide a brief overview The system interprets one-anaphora as intending on an object of the same type as the currently profiled object an object is profiled when it is selected as the referent for an expression The anaphoric modifier *other* is understood by the system to specify that the referent of this expression should not be currently profiled Following this, the SLI system interprets *the other one* to refer to the most salient object in the current context that is of the same type as the currently profiled object, but is not itself currently profiled

The currently profiled object is the blue house that was selected as the referent for the previous input, Input 6-4 There are two other (i e , not currently profiled) houses in the context model the brown house in the view volume and the red house the user saw earlier Neither of these are currently profiled, hence they both fulfil the linguistic descriptions of *the other one* as understood by the SLI system The system uses the current visual salience associated with each of these objects to select one as the referent In this example, the brown house is in the view volume while the red house is not Consequently, the brown house has a higher visual salience than the red house and is selected as the referent
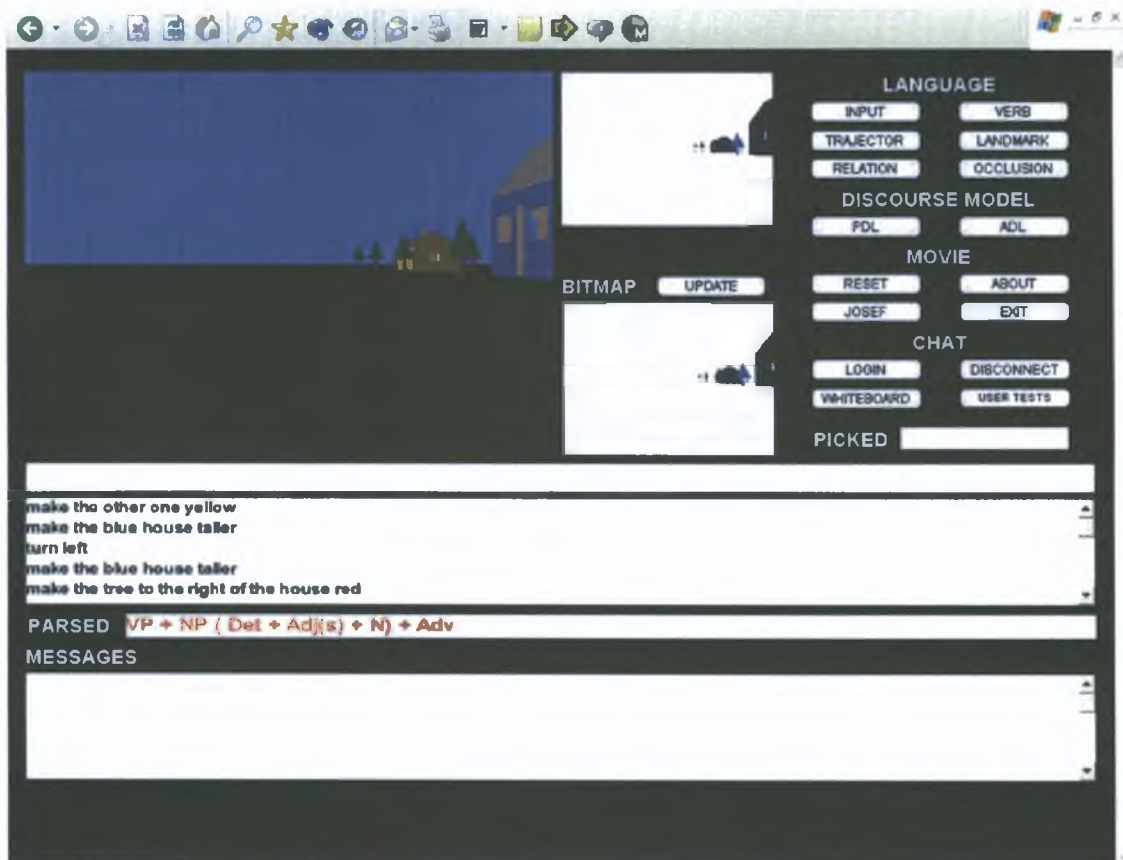
Figure 6-7: The state of the SLI simulation after processing the input: *make the other one yellow*.

Input 6-6: *look at the red house*

Although there is no red house in the user's current view, the system remembers that the user saw a red house earlier in the session. Crucially, this is recorded in the system's context model, even though no reference to a red house has been made in the discourse. This is because the context model is updated by the visual saliency algorithm after each frame has been rendered. As a result, the system can resolve the reference *the red house* to the red house previously seen by the user and rotates the user's avatar so that the user's view is centred on this house.

Figure 6-8 illustrates the user's view of the world after the input *look at the red house* has been processed.
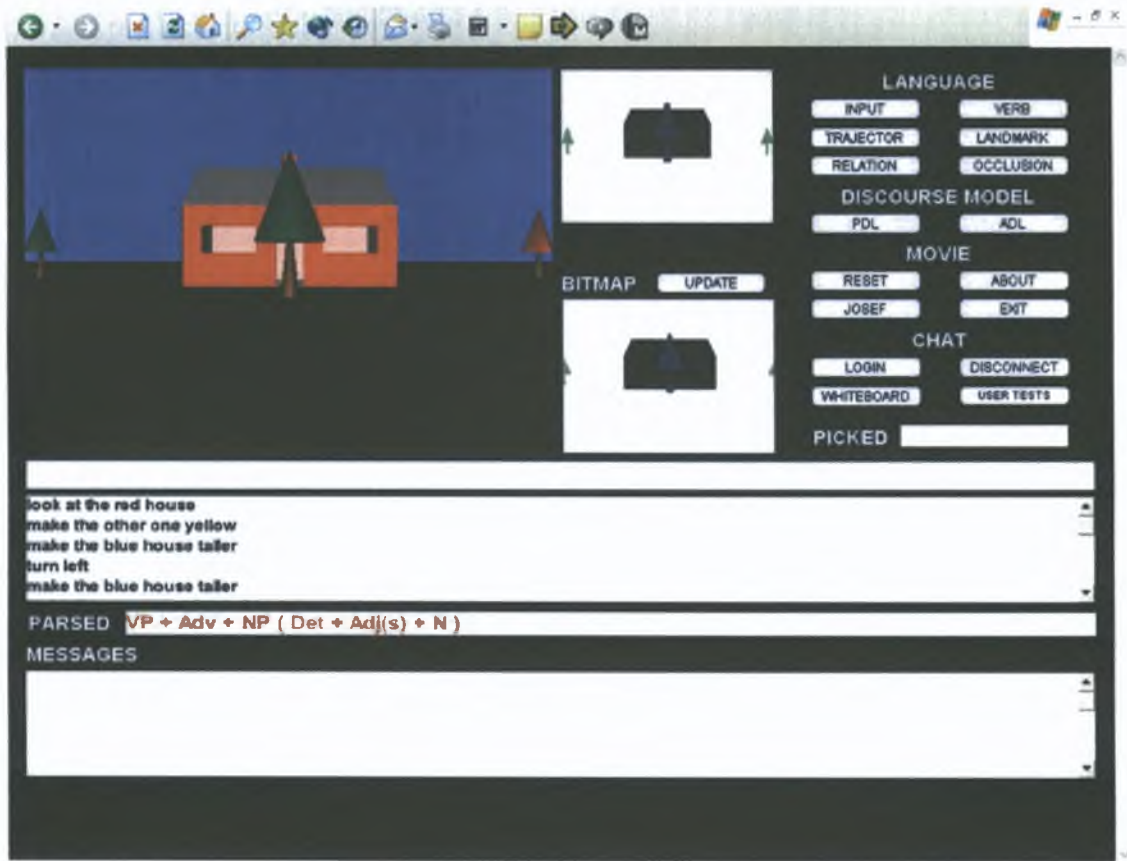
Figure 6-8: The user's view of the simulation after the input *look at the red house* has been interpreted.

Input 6-7: *make it taller*

This input uses the pronoun *it* to intend on its referent. In contrast to Input 6-5, whose referent had not previously been mentioned in the discourse, this input uses an anaphoric expression that links back to a previously mentioned object. The ability of the SLI system to resolve anaphoric expressions is based on the updating of the context model by the interpretive module after each user input has been processed.

Figure 6-9 shows the state of the world after the input *make it taller*, has been interpreted.
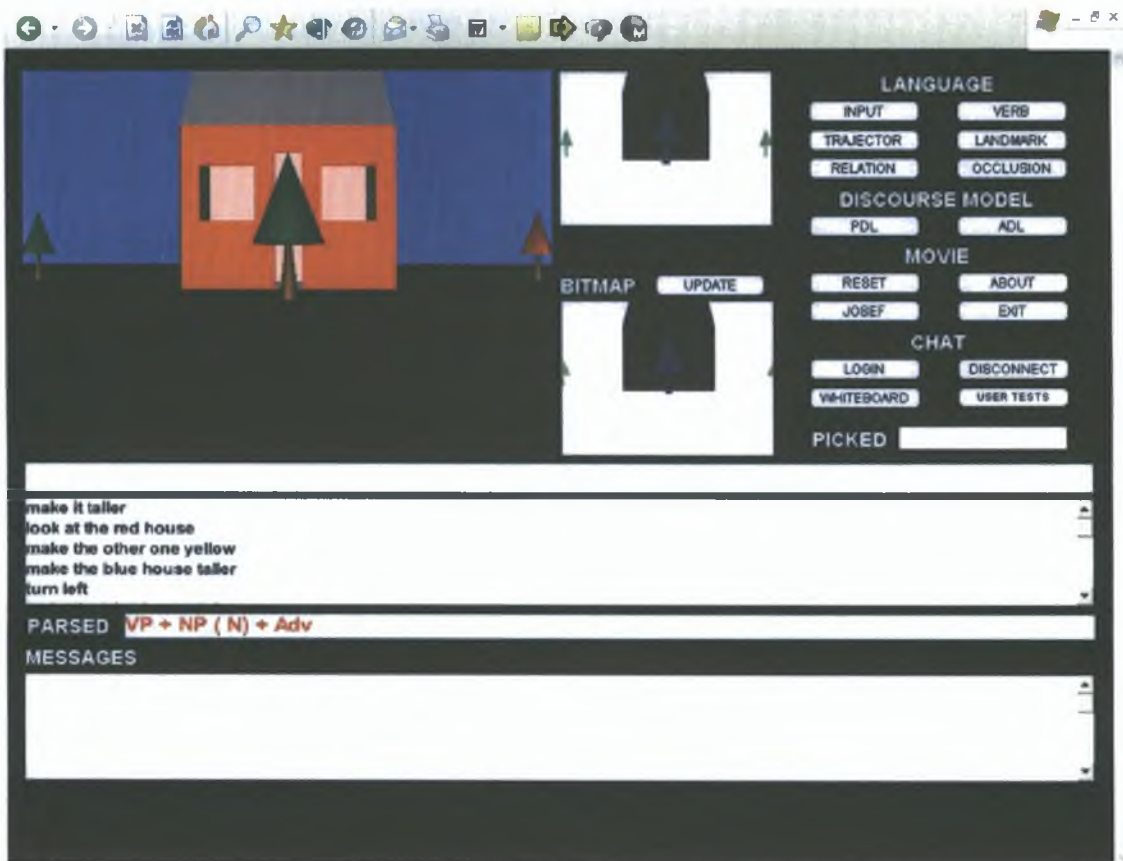
Figure 6-9: The SLI simulation after the input *make it taller* has been processed.

Input 6-8: *move backwards*

Input 6-9: *stop*

Input 6-8 and Input 6-9 are both avatar commands. These commands allow the user to navigate through the simulated world. Input 6-8 causes the user's viewpoint to move backwards from its current position. Input 6-9 stops the movement of the user's viewpoint.

Figure 6-10 illustrates the user's view of the simulation after these commands have been processed.
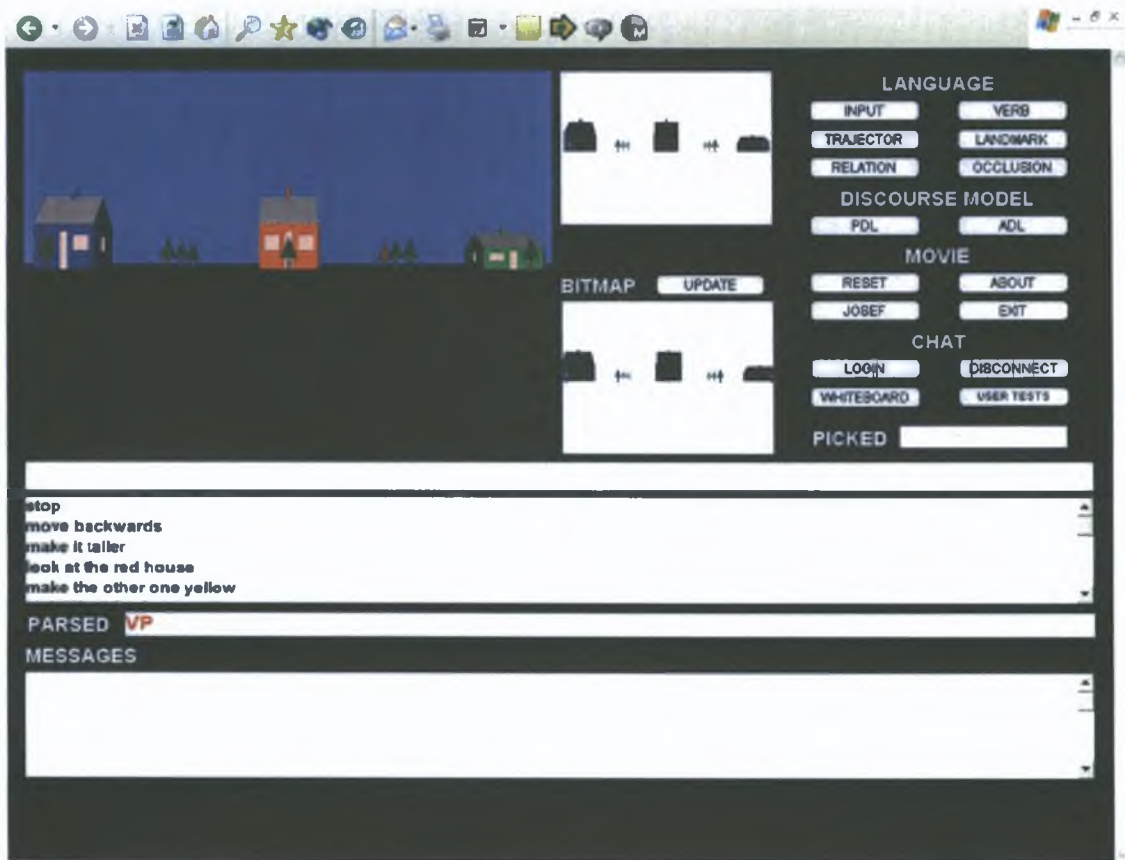
Figure 6-10: The user's view of the simulation after the avatar commands *move backwards* **and** *stop* **have been processed.**

It is worth noting that Figure 6-10 also illustrates that there are trees in the world which are more *to the right of the red house* than the tree which was selected as the referent to Input 6-1 *make the tree to the right of the house red*. However, at the time Input 6-1 was being processed, the user was not aware that these other trees existed. Because the SLI system used its model of the user's world knowledge (created by the visual saliency module and stored in the context model), rather than the world model itself, as the context for interpreting Input 6-1, it was able to exclude the trees that the user was not aware of from the interpretation process, even though they were more to the right of the house than the tree selected as the referent for Input 6-1.

Input 6-10. *make the red tree and the short house taller*

Input 6-10 illustrates the system's ability to handle coordinating expressions It also illustrates the system's ability to resolve definite description using both qualitative and quantitative object properties In resolving the expression *the red tree* the system uses a supplied qualitative adjectival description, *red*, as a restriction on an object being considered as a candidate referent In this instance, there is only one red tree in the context model and it is extracted as the referent for this expression In order to resolve *the short house*, the system must compare the dimensions of candidate referents to select the candidate that fulfils the supplied adjectival description to the greatest extent It first creates a list of all the objects in the current context that are off the right type, *house* At this point in the dialogue, there are four houses in the context model the three houses currently in the user's view (the red, blue, and the green houses) and the yellow house (currently not in view) to the left of the blue house that was selected as the referent to Input 6-5 Once the system has created the list of candidate referents, it then searches the list for the shortest object and selects this object as the referent for the expression In this instance, the green house and the yellow house have an equal score with respect to the shortness criteria In order to adjudicate between these two candidates, the system uses the visual saliency scores associated with each object Here, the green house is in the view volume and the yellow house is not, consequently, the green house has a higher visual saliency score and is selected as the referent for the nominal *the short house* Again, it should be noted that the referent of this expression has not previously been mentioned in the discourse

Figure 6-11 illustrates the state of the simulation after this coordinating expression has been resolved
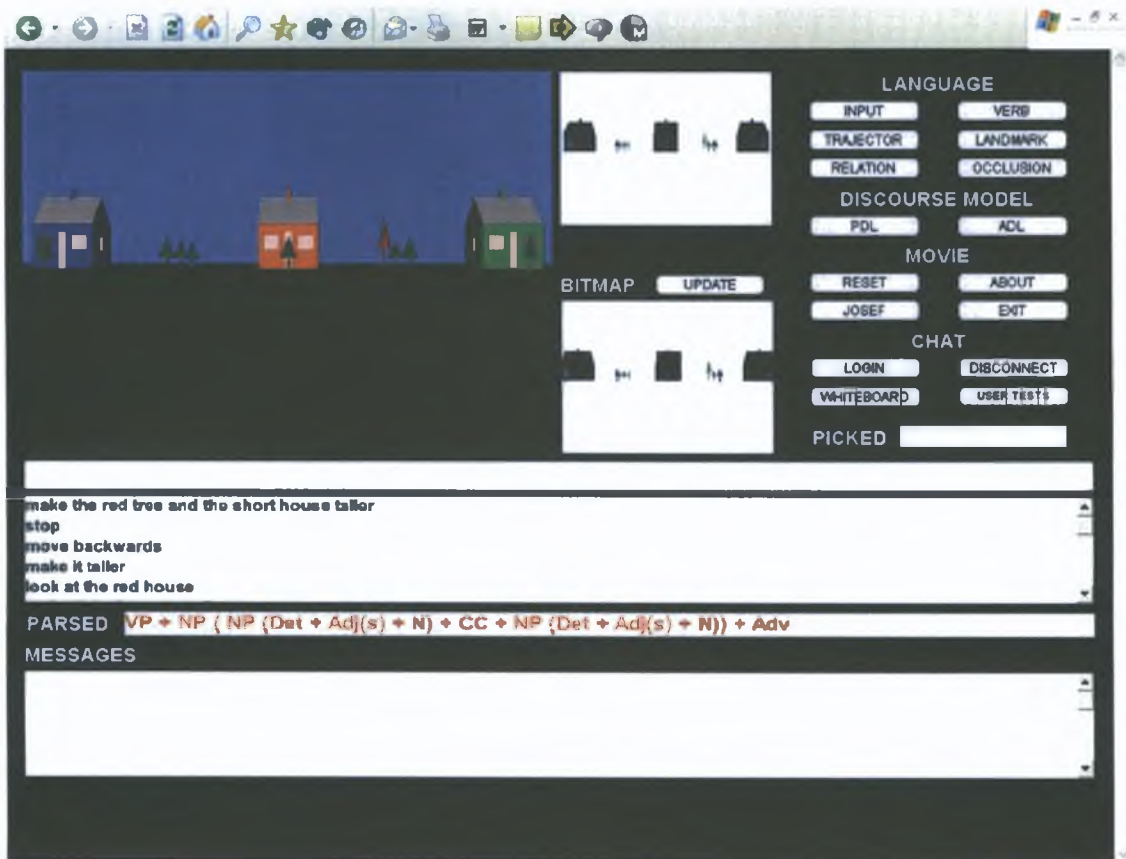
Figure 6-11: The state of the SLI simulation after the input *make the red tree and the short house taller* **has been processed.**

Input 6-11: *make a house yellow*

Input 6-11 illustrates the system's ability to resolve indefinite descriptions. In the context of the SLI system, an indefinite expression, *a N,* may be used to arbitrarily refer to one of the elements of type *N* in the spatio-temporal context, or to refer to the generic type *N* in commands that create new objects in the world (discussed in detail in Chapter 9). Here, it is sufficient to point out that the system decides which type of *N* an indefinite expression refers to based on syntactic cues. If the indefinite expression is followed by an adjective functioning as an adverb, as in Input 6-11, the system treats the input as referring to one of the elements of type *N* in the spatio-temporal context. Otherwise, the system assumes that the indefinite expression refers to the generic type *N*. Consequently, Input 6-11, is understood to refer to an object in the spatio-temporal context, and is

interpreted as the user wanting the system to arbitrarily select object of type *N* that is already in the simulation and manipulate it. Accordingly, the system randomly selects a referent from the set of objects in the view that match the type restriction and any supplied adjectival restrictions in the input. In this instance, the system selected the red house and changed its colour to yellow.

Figure 6-12 illustrates the system after this input has been processed.
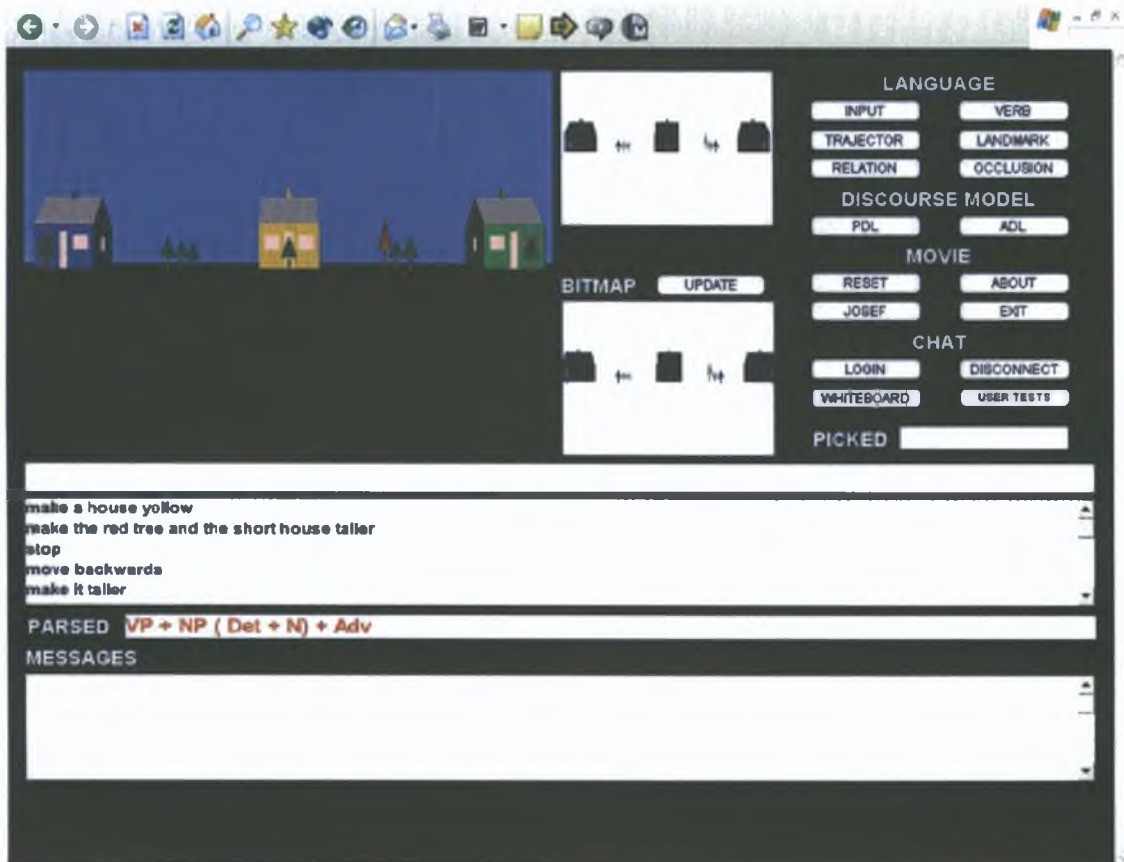
Figure 6-12: The state of the simulation after the input *make a house yellow* **has been processed.**

Input 6-12: *add a red house*

Input 6-12 illustrates how the system handles indefinites that refer to a generic object, rather than an arbitrarily selected object already in the context. It also illustrates the system's ability to extend the world by adding new objects. Note that it is possible to parameterise the creation of an object by using adjectives in the input. The system uses simple heuristic rules to calculate where the newly created object should be added.

Figure 6-13 illustrates the user's view of simulation after a new red house has been added to the world.
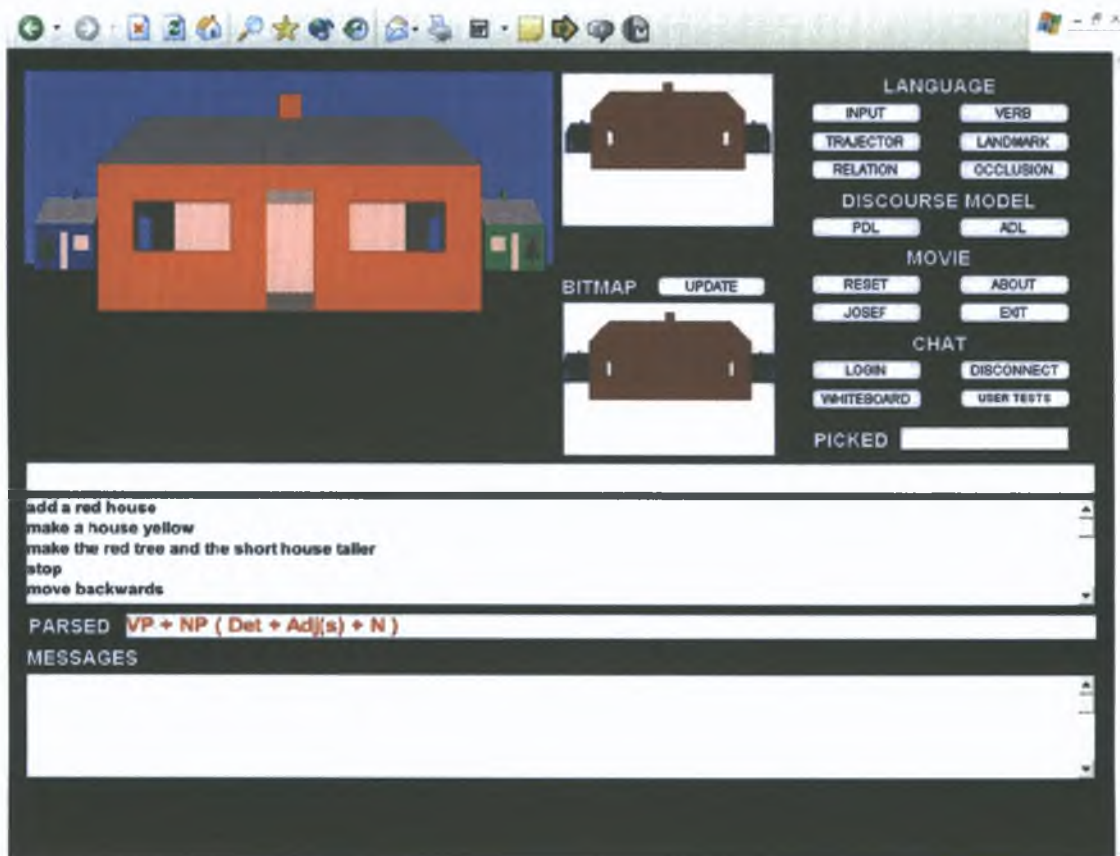
Figure 6-13: The state of the 3-D world after a new object has been added.

Input 6-13: *make this => brown*

The final input in this example dialogue, Input 6-13, illustrates how the SLI system resolves demonstratives that are accompanied by a pointing gesture. The arrow in the input (=>) symbolises the pointing gesture. In the SLI system, deictic gestures are simulated by mouse clicks on the intended object in the view volume.

Figure 6-14 illustrates the state of the SLI system after it has processed Input 6-13. The position of the user's mouse click is depicted in this image by the mouse cursor. Furthermore, the text box labelled PICKED lists the name of the object clicked on by the users.
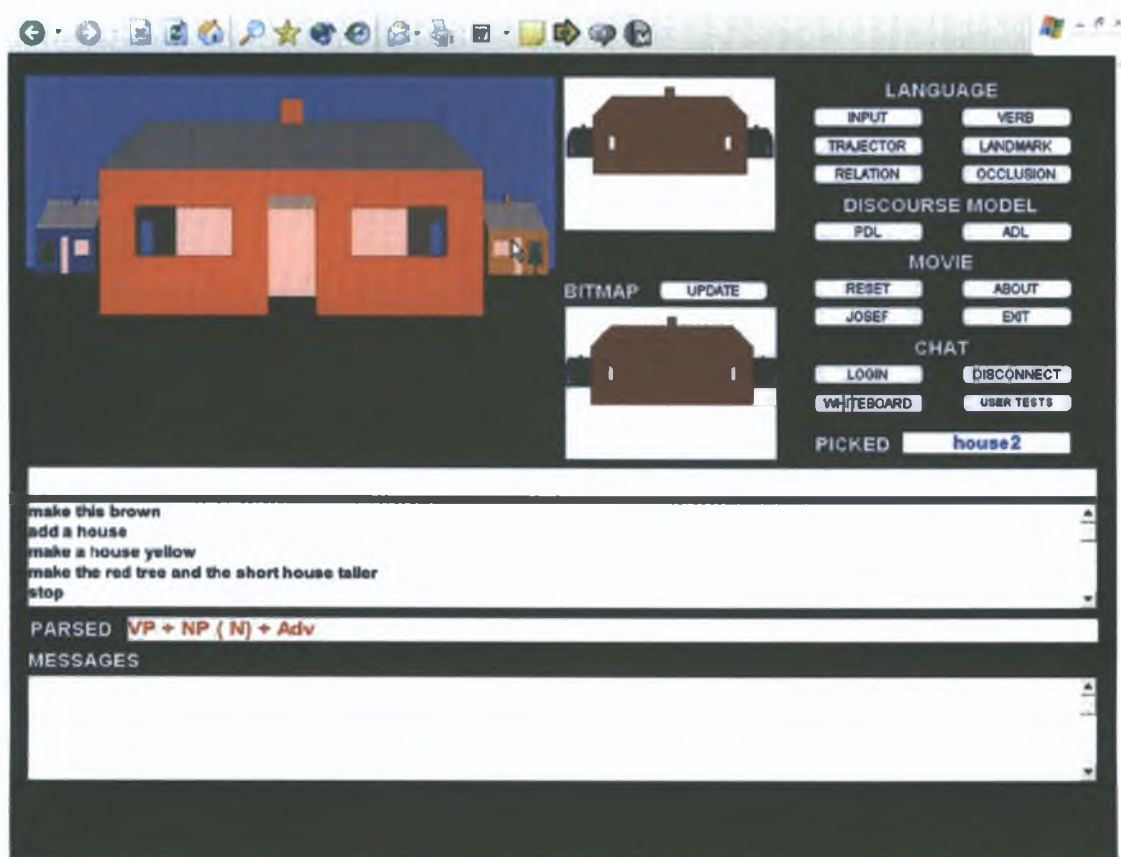
Figure 6-14: The state of the SLI system after the input *make this => brown* has been processed. The demonstrative *this* was accompanied by a pointing gesture (simulated by a mouse click). The position of the mouse is shown in the image. Also,

the text box in the SLI interface labelled PICKED lists the name of the world object that was clicked on.

6.4 Chapter Summary

The SLI system was implemented as a test bed for the framework developed in this thesis Illustrating the functionality of the system in this chapter before continuing with more details has hopefully provided the reader with an overview of the framework Moreover, describing the framework in the context of an implemented system also illustrates the tractability of the approach

7 Computing the Visual Context

7.1 Introduction

In Section 2 5 a brief description of the link between a computational model of visual salience and resolving a referring expression was given a computational model of visual salience permits a computational NL interface to a 3-D simulated environment to capture the perceptual events that cause the entry of an object into what the user considers as mutual knowledge In this chapter, a computational model of visual context that uses the false colouring technique introduced in Section 5 2 3 is developed This model of visual attention is novel because it extends the false colouring model of synthetic vision to rate the observed objects based on their saliency within the viewed scene Moreover, it is designed for use as an interface between a rendered environment and a linguistic interpretive module

The goal of this model is to allow the system to autonomously and incrementally develop a model of the user's knowledge of the environment based on what they have perceived This model of the user's knowledge may then be used as an input to the process of interpreting the user's NL inputs Because of time constraints in real-time systems, the visual saliency algorithm must be fast It must also be computationally efficient to keep from impacting adversely on the rendering of the environment Furthermore, it must be adaptable to different types of dynamic environments where objects can appear, move, disappear, or change their appearance

The function of this visual saliency model is to determine the set of objects currently visible and to rate them based on their saliency The inputs to the algorithm are the environment scene description along with a specification of the user's current view volume

Work related to this area was reviewed in Chapter 5, beginning with a general discussion about the approaches to vision developed for intelligent robots Recall that the majority of theses robot systems use connectionist architectures that require training This

training requirement makes these approaches unsuitable for graphics applications where the type of input may vary Furthermore models of vision based on graphics techniques were reviewed, including ray casting models and false colour models

In this chapter, some of the ray casting algorithms that were developed for earlier versions of the SLI system described in Chapter 6 are described with explanations of why they were rejected The false colouring model of visual salience is then described and how this model extends previous work in this area (Renault *et al* 1990, Noser *et al* 1995, Kuffner and Latombe 1999, Peters and O'Sullivan 2002) is explained

7.2   Ray Casting and Visual Salience Algorithms

The early models of vision implemented in the SLI system were similar to Tu and Tersopoulos's (1994a, 1994b) ray casting model However, (as was noted in Section 5 2 2), ray casting is a computationally expensive approach Furthermore, to get full coverage of an image, a ray must be cast from each pixel in the view Unfortunately, this process is to slow for a real-time system Following this, during the development of the SLI system other algorithms were developed that attempted to reduce the number of rays cast One algorithm divided the view into different regions and cast a ray from the centre of each region While this sampling of the scene increased the speed of the ray casting process, in order to fulfil the real-time constraint the scene had to be segmented in such a coarse manner that the accuracy of the algorithm was affected, i e , the system missed objects in the view volume Another algorithm only sent a ray at each pixel where a vertex of an object's mesh was drawn It was hoped that this would focus the rays in the areas of the scene where there were objects and allow the system to ignore areas of the scene that were empty However, there were several drawbacks to this approach Firstly, the rendering speed of the system varied, depending on the number of objects in the scene, an increase in the objects in the scene resulted in an increase in the number of rays cast which resulted in the system slowing down Secondly, an object that was close to the viewpoint could appear very large in the scene and yet have no vertices in the view In these instances, it was possible for the system to miss the object entirely

215

This last issue illustrates a problem that impacts on visual salience algorithms that use the size of an object in the view as an indicator of its salience and don't process all the pixels in the scene how does one judge how big an object appears to a viewer? While the absolute dimensions of an object can be obtained from its geometric model, its size within a particular scene depends on many other factors the observer's perspective of the object, the distance of the object from the viewpoint, the occlusion of parts of the object by other objects, or by the edges of the view volume, etc Satisfactorily weighting and combining these factors is difficult An easier solution would be to count the number of pixels in the scene that the object covers

Having tested the different ray casting approaches, it was concluded that at current processor speeds ray casting is still too computationally expensive for real-time interactive systems

7.3 A False Colouring Visual Salience Algorithm

The SLI's model of visual salience is based on the false colouring approach to synthetic vision described in (Noser *et al* 1995) and later adopted by (Kuffner and Latombe 1999, Peters and O'Sullivan 2002) Similar to these previous systems, each object is assigned a unique ID In the current implementation, the ID number given to an object is simply 1 + the number of elements in the world when the object is created A colour table is initialised to represent a one-to-one mapping between object IDs and colours Currently, this table contains 256 entries Although this restricts the number of objects that can be added to the world, this restriction is more a matter of convenience than necessity as the colour table can be extended without affecting the rest of the system Each frame is rendered twice firstly using the objects' normal colours and textures and normal shading This is the version that the user sees The second rendering is off-screen This rendering uses the unique false colours for each object and flat shading The size of the second rendering does not need to match the first Indeed, scaling the image down increases the speed of the algorithm as it reduces the number of pixels that are scanned In the SLI system the false colour rendering is 200 x 150 pixels, a size that yields

sufficient detail After each frame is rendered, a bitmap image of the false colour rendering is created The bitmap image is then scanned and the visual salience information extracted

To model the size and centrality of the objects in the scene, the SLI system assigns a weighting to each pixel using the distance from the centre of the scene as in Equation 7 In this equation, P^{48} equals the distance between the 2-D coordinates of the pixel being weighted and the 2-D coordinates of the centre of the image, and M^{49} equals the maximum distance between the 2-D coordinates of the centre of the image and the border of the image (in a rectangular or square image, M is equal to the distance between the 2-D coordinates of the centre of the image and the coordinates of one of the corners of the image)

$$Pixel\ Weighting = 1 - (\ P * (\ 1 / (M + 1)))$$

Equation 7. The equation defining the weighting assigned to each pixel in the bitmap created from the off-screen rendering of the false colour scene P is the distance between the 2-D coordinates of the pixel being weighted and the centre of the image. M is the maximum distance between the centre of the image and the border of the image.

This equation normalises the pixels between 0 and 1 The closer a pixel is to the centre of the image, the higher its weighting After weighting the pixels, the SLI system scans the image and, for each object in the scene, sums the weightings of all pixels that are coloured using that object's unique colour This algorithm ascribes larger objects a higher saliency than smaller objects since they cover more pixels and objects which are more central to the view will be rated higher than objects at the periphery of the scene as

---

[48] The distance between the 2-D coordinates of the centre of the image, and the coordinates of the pixel being weighted is computed using the geometric equation for the distance between two points distance = sqrt((P1 x – P2 x)^2 + (P1 y – P2 y)^2) Note sqrt = square root function

[49] M is computed using the same distance equation as P

217

the pixels they cover will have a higher weighting. This simple algorithm results in a list of the currently visible objects, each with an associated saliency rating.

It is important to note that the scanning process in the SLI visual salience algorithm differs from those in the previous false colour synthetic vision models (Renault *et al.* 1990; Noser *et al.* 1995; Kuffner and Latombe 1999; Peters and O'Sullivan 2002). It is this difference that allows the SLI system to rate the objects based on their visual salience. The SLI model of visual salience assumes that objects that are larger or more central within the user's view are more prominent than objects that are smaller or on the periphery of their view (see Section 2.2.2).

In this thesis, it is not claimed that this model accommodates all the perceptual factors that impact on visual salience. However, this algorithm does define a reasonable model of visual saliency that operates fast enough for real-time systems. The output of the saliency algorithm is stored in a data structure that forms part of the system's context model. Furthermore, this architecture gives the system a form of visual memory which it uses to restrict the number of possible referents for any given referring expression. Finally, it should be noted that the separation of the perceptual mechanism from the linguistic interpretive module admits the possibility of replacing the perceptual module with a more refined version at some later date.

As was noted in the introduction to this chapter, integrating a computational model of visual salience gives the SLI framework the ability to capture the perceptual events that cause the entry of an object into what the user considers as mutual knowledge. Integrating the information created by such a model into a model of discourse gives an NL system the ability to resolve deictic references as well as anaphoric references. In the sample user-system dialog, presented in Chapter 6, the system's ability to resolve deictic reference was illustrated on several occasions. In particular, resolving the nominal references in input (1) *make the tree to the right of the house red* and input (10) *make the red tree and the short house taller* involved extracting a referent from the context model that had not previously been mentioned in the discourse. However, a further advantage of this approach is that the visual salience scores associated with the objects in the context model allows the system to adjudicate between candidate referents when resolving some ambiguous references. Section 7.4 below describes how the output of this visual saliency

algorithm is used by the SLI system to resolve these references Some of the limitations of this approach are also noted

7.4 Using Visual Salience to Resolve Ambiguous References

Since Russell (1905), there has been a debate concerning the singularity constraint associated with definite descriptions[50] The singularity constraint is given the use of a definite description there should be one, and only one, candidate referent in the context of the utterance An ambiguous or undetermined reference is a reference that breaks the singularity constraint, i e , there is more than one candidate referent It has been shown, however, in psycholinguistic experiments that people can easily resolve ambiguous or underdetermined references (Duwe and Strohner 1997) "In order to identify the intended referent under these circumstances, subjects rely on perceptual salience as well as on pragmatic assumptions about the speaker's communicative goals" (Duwe and Strohner 1997 pg 6)

An advantage of using a visual saliency model as an input to an NLVR system's context model is that the visual salience scores associated with the objects in the context model allows the system, in some instances, to adjudicate between candidate referents when resolving underspecified or linguistically ambiguous references, as illustrated below Given Figure 7-1 as the visual context, the referring expression *the house* in *make the house red,* is an example of an ambiguous visible situation use of a definite description This is because there is more than one object in the context that fulfils the linguistic description of the expression's referent

---

[50] See Section 9 4 1 1 for a more detailed discussion on the uniqueness of a definite description's referent
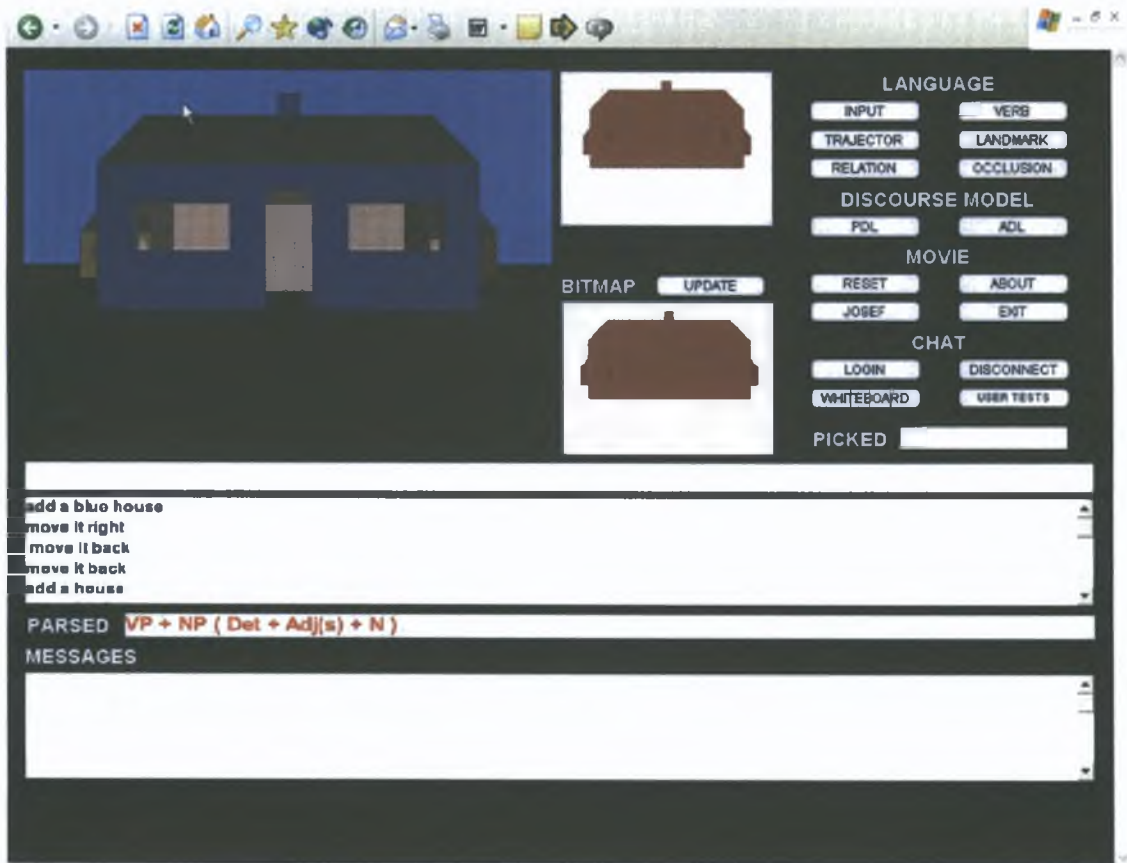
Figure 7-1: A scene containing three houses.

However, in this instance the SLI system can utilise the visual saliency scores associated with each of the candidates as a probability of the candidate being the referent for the expression. In this case, the SLI system ascribes the blue house in the foreground a normalised computed visual salience of 1.0000 and each of the houses in the background a normalised visual salience of 0.0117. Based on these visual saliency scores, the system decides that the user is referring to the blue house in the foreground and updates the simulation accordingly. Figure 7-2 illustrates the state of the system after this user input has been interpreted.
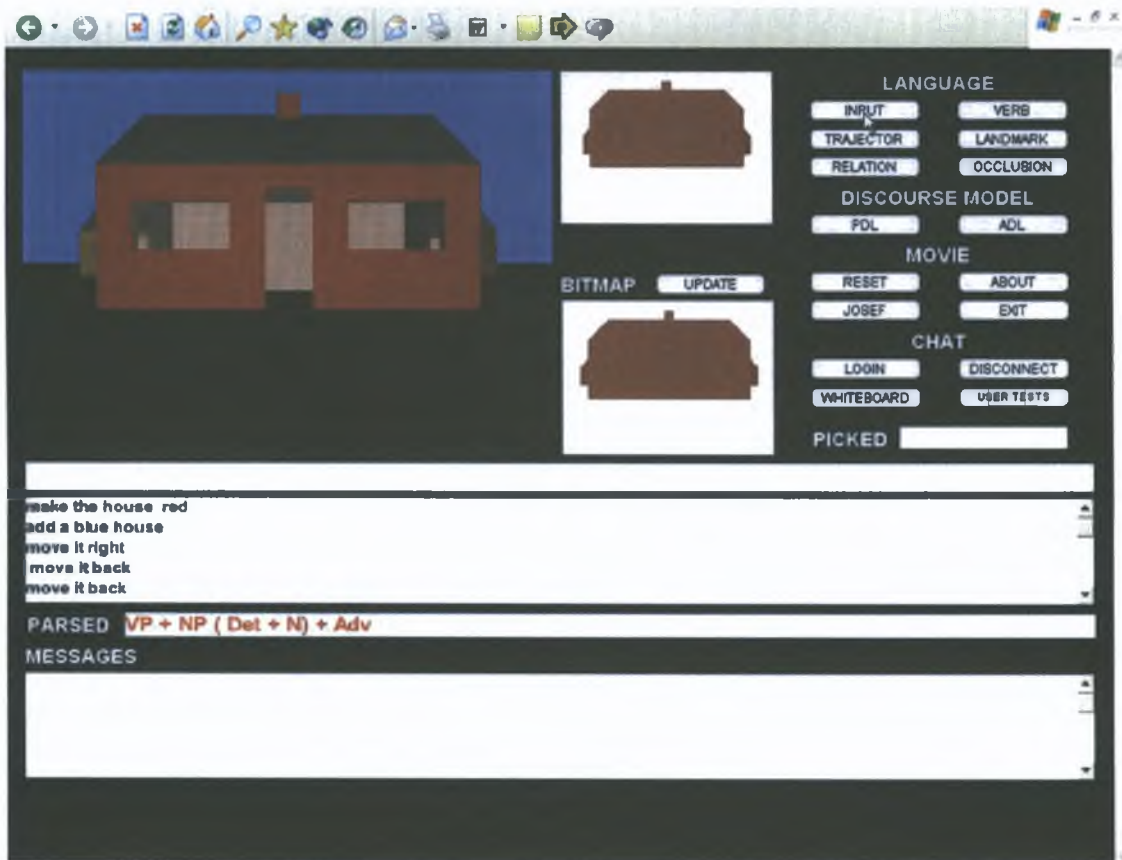
Figure 7-2: The state of the simulation after the SLI system has interpreted the underdetermined reference *the house* and processed the input *make the house red*.

Clearly, however, not all ambiguous references can be resolved based on visual salience. In some instances, the difference in the visual saliency scores associated with each of the candidate referents is not sufficient to allow the selection of a referent. Accordingly, as part of the interpretation process for resolving ambiguous references, the SLI system compares the saliency of the primary candidate referent and the other candidates. If the saliency difference does not exceed a predefined confidence interval, the system outputs a message to the user explaining that it is unable to resolve the reference. In SLI scenarios, it is found that when comparing normalised saliency scores, ranging from 0 to 1, a confidence interval of .4 works well. This of course can be adjusted to model a more or less stringent interpretation. Figure 7-3 illustrates a scene with two houses that have equal visual saliency scores. In this instance, both houses have a visual saliency rating of 1.0000.
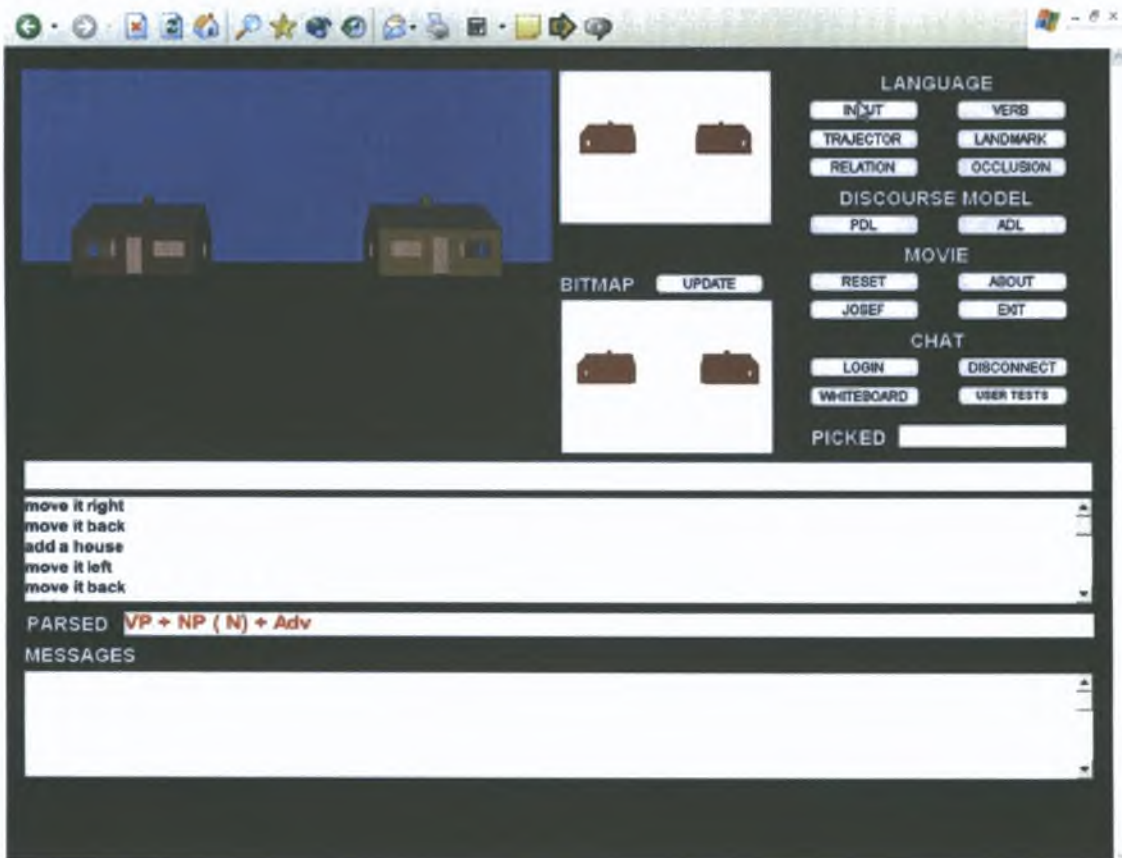
Figure 7-3: A scene with two houses that have equal visual saliency scores.

Taking Figure 7-3 as the visual context, if the user inputs an ambiguous referring expression, *make the house taller*, the system would be unable to resolve the reference. Figure 7-4 illustrates the state of the system after this command has been interpreted. Note that the visual scene has not changed and that the message text box contains a message to the user explaining why the system was unable to resolve the reference, as well as listing the candidate referents the system restricted its search to: *Required Saliency Interval Not Reached, Primary Candidates Saliency Confidence Insufficient, I think you mean: house 7 Normalised Salience = 1.0000 or house 6 Normalised Salience = 1.0000.*
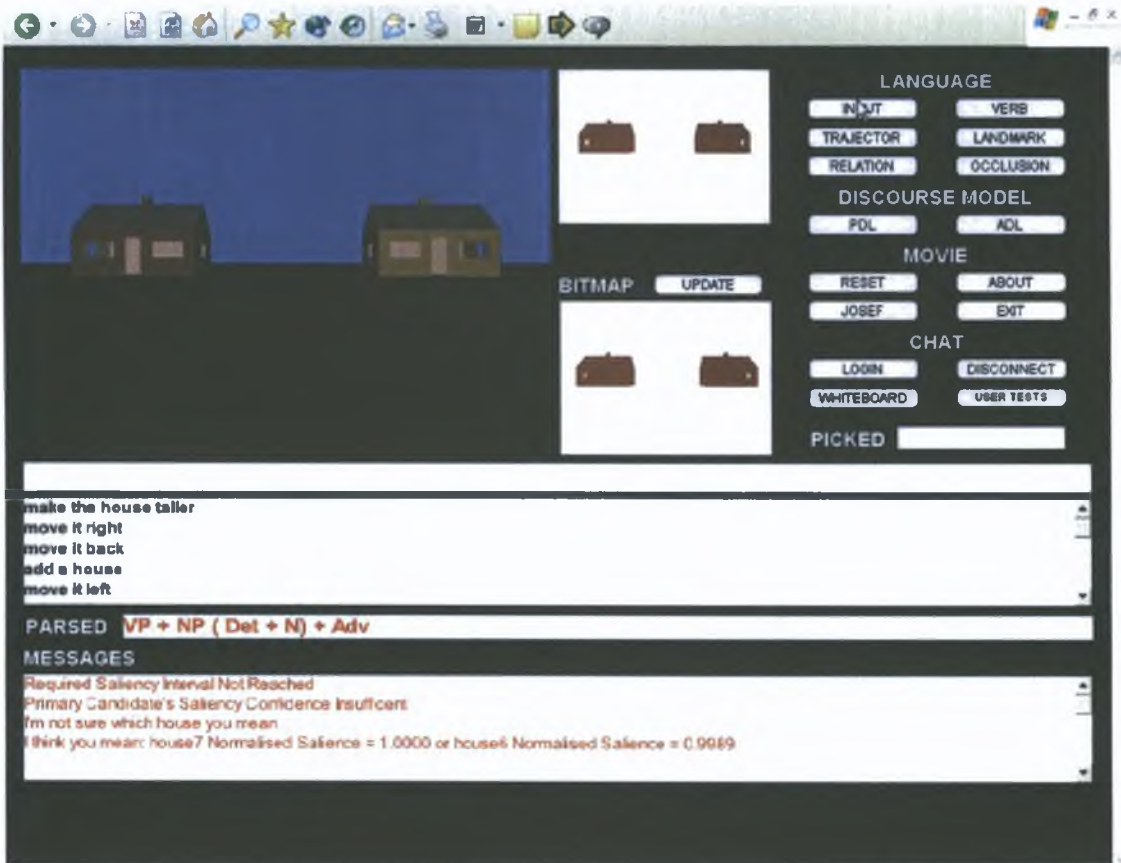
Figure 7-4: The state of the SLI system after the system has output a message to the user stating that the saliency differences between the candidate referents of an undetermined expression did not permit the system to resolve the reference.

7.5 Chapter Summary

In this chapter, a computational algorithm for modelling the visual salience of objects in the view volume was developed. This model of visual attention is a novel application and extension of a synthetic model of vision that uses a graphics technique called false colouring (Noser *et al.* 1995). The function of this visual attention model is to try to capture the perceptual information flowing from the environment to the user. The output of this module feeds into the SLI discourse model, which uses it to model the perceptual dialogue. How the saliency scores created by the algorithm can be used to

resolve undetermined references was illustrated In Chapter 8, the SLI algorithm for interpreting a complex form of referring expression – locative expressions – is developed

8 A Perceptually Based Computational Approach to Interpreting Projective Locative Expressions

8.1 Introduction

In Section 2 3 the concept of a locative expression was introduced and a general algorithm for interpreting locative expressions (Carlson-Radvansky 1996) described There are four stages to this algorithm

1 Identify the landmark

2 Select a frame of reference and superimpose it on the landmark

3 Define the area of search for the trajector as defined by the spatial template associated with the preposition

4 Identify the primary trajector within the search area

Following this, the major issues attending this process were introduced

1 Resolving the reference to the landmark (Section 2 3 2)

2 Computationally selecting a frame of reference (Section 2 3 3 6)

3 Locating the origin of the preposition's spatial template (Section2 3 4 2 4)

4 Computationally modelling the spatial template of a projective preposition Such a model should model the gradation of a preposition across its spatial template in a cognitively plausible manner (Section 2 3 4 2 1), be scalable to accommodate differently sized landmarks (Section 2 3 4 2 1), model both the angular deviation and distance of candidate trajectors from the landmark (Section 2 3 4 2 2), integrate perceptual cues into the spatial templates of certain prepositions (Section 2 3 4 2 3)

5 Modelling the location of a trajector (Section 2 3 5 1)

6 Handling the issue of occluded trajectors (Sections 2 3 5 2)

In this chapter, a solution to these issues is developed. In particular, in Section 8 3, an algorithm based on psycholinguistic work of (Carlson-Radvansky and Irwin 1993, Carlson-Radvansky and Irwin 1994, Carlson-Radvansky 1996, Taylor *et al* 2000) is developed, which attempts to predict a user's intended frame of reference. Following this, in Section 8 4, a novel semantic model of projective prepositions that defines prepositions in terms of topological and perceptual axioms is proposed.

The topological component of this model is defined in Section 8 4 1. There are two algorithms in this topological component: the first algorithm defines a process for dynamically locating the origin of a spatial template, the second algorithm defines a computational model of the semantics of projective prepositions which accommodates the topological considerations noted in Section 2 3 4 2 5. One of the most important aspects of the topological component of the SLI semantic model for projective prepositions is the dynamic location of the spatial template's origin in the viewer-centred frame of reference based on the user's location relative to the landmark. The algorithm for locating the spatial template's origin is developed in 8 4 1 1. This dynamic location of the spatial template's origin avoids many of the paradoxical definitions that occur with models that default to using the landmark's bounding box centre as the origin, (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994, Fuhr *et al* 1998) for a description of why the use of the landmark's bounding box centroid is problematic see Section 2 3 4 2 4. The second algorithm in the topological component of the SLI semantic model for projective prepositions is a computational model of the topological spatial template associated with projective prepositions. This algorithm is developed in Section 8 4 1 2. Although this model accommodates the topological considerations associated with projective prepositions, Section 8 4 1 2 concludes by highlighting some of the weaknesses of the topological model when it is used in the viewer-centred frame of reference.

In response to the weaknesses of the SLI topological model, in Section 8 4 2, a computational model of the semantics of projective prepositions that is based on perceptual axioms is developed. However, although this perceptually based approach can successfully handle the situations which the topological model defined in Section 8 4 1 found problematic, this perceptual model cannot handle some of the situations that the topological model could accommodate.

226

After noting these weaknesses in the perceptual model, it is posited that a model which integrates the topological model with the perceptual model provides a unified framework that can accommodate the issues affecting each of its components Following this, in Section 8 4 3, an algorithm for combining the topological and perceptual models if defined and an example illustrating how this integrated approach improves on previous models is given

In Section 8 4 4 a proposal to resolve the issues attending to the representation of the candidate trajectors within the preposition's spatial template and the selection of the referent from this set is described

Note, there is a large interdependence between the different solutions proposed in this chapter, e g , the approach taken to resolving the issue of reference frame selection impacts on the construction of a preposition's spatial template Furthermore, the selection of a referent from the set of candidate trajectors depends on the construction of the preposition's spatial template and the representation of the candidate trajectors in the spatial template This interdependence between the different stages in the process has complicated the presentation of a general solution to resolving a locative expression The approach taken here is to focus on each of the four main stages involved in the interpretation of a locative expression in sequence and develop a solution to each of the modules separately Based on this, in Section 8 6, an algorithm for resolving a locative expression that integrates these separate solutions is developed

8.2   Identifying the Landmark

Section 2 3 2 introduced the main issues at this stage in the interpretive process, and defined the process of identifying the landmark as extracting the most salient object from the visual context that matches the nominal expression in the object position of the expression However, this is exactly what a general model of reference should achieve (see Section 2 5) Consequently, identifying the landmark of a locative expression falls within the ambit of the SLI general model of reference resolution, which is developed in Chapter 9

227

8.3 Frames of Reference

If a locative expression contains a projective preposition, the second stage of the interpretive process is the selection of a frame of reference, (see Section 2 3 3) In some instances, the frame of reference is made explicit in the linguistic input In other instances, the landmark has no intrinsic frame of reference In such situations, the viewer-centred frame of reference is the only one applicable In other situations, the landmark's intrinsic frame of reference and viewer-centred frame of reference are aligned and there is no conflict However, if the landmark has an intrinsic frame of reference which is disassociated from the viewer-centred frame of reference and the intended frame of reference is left implicit in the utterance, a process for resolving the frame of reference is required

In Section 5 3 1, a brief overview of the approaches to the issue of frame of reference selection adopted by previous NLVR systems was given Furthermore, these approaches were criticised because of the restrictions they place on the domain of the discourse or on the user's interaction In this section, a procedure based on linguistic and psycholinguistic work that attempts to select the user's intended frame of reference is presented This algorithm does not claim to represent the cognitive processes used by a human in selecting a frame of reference It merely aims to model their general preferences

A possible approach to the frame of reference competition resolution issue is to combine the spatial template hypothesis of Logan and Sadler (1996) (see Section 2 3 4 1 1) with Carlson-Radvansky and Irwin's (1994) multiple frame activation hypothesis (see Section 5 3 1 1) By combining these, a strategy for computationally selecting a frame of reference can be defined in Algorithm 8-1

228

If the frames of reference in a scene are dissociated Then

    Construct the spatial template for the preposition along the appropriate axes

    in both of the competing reference frames

    For each candidate trajector, sum its applicability ratings for each of the

    spatial templates

    Select the candidate trajector which has the highest overall applicability

End If

Algorithm 8-1: Frame of Reference Competition Resolution Algorithm 1

This algorithm provides a method for the selection between competing reference frames in terms of a combination of their individual contributions From a computational perspective, the ease of implementation of this approach is a major advantage arguing for its adoption However, while Algorithm 8-1 provides an easy method for the resolution of competing reference frames, at the core of this hypothesis is the assumption that the competition between reference frames is independent of the preposition used or the deviation of the axes within a particular frame from their canonical orientation

In Section 5 3 1 2, the results of experiments by Carlson-Radvansky and Irwin (1993) and Taylor *et al* (2000) were reviewed These experiments revealed that the competition between frames of reference does not solely depend on the position of the candidate trajectors in their respective reference frame The canonical orientation of the preposition biases which frame of reference is used Carlson-Radvansky and Irwin (1993) illustrated that the vertical axis is dominated by the viewer-centred frame of reference Indeed, my analysis of (Carlson-Radvansky and Irwin 1993) revealed a 2 1 bias toward the viewer-centred frame along the vertical axis Taylor *et al* 's (2000) experiments focused on the horizontal axes Their results indicated "strategic processing, with priority selection of the intrinsic frame" (Taylor *et al* 2000 pg 11) the authors qualify their analysis with the caveat that their findings may indicate a task requirement influence on spatial frame processing, rather than a plane based bias Carlson-Radvansky and Irwin's (1993) results are compelling and have the advantage from a computational perspective of allowing a threshold to be set Taylor *et al* 's (2000) results, while pointing to a bias on

229

the horizontal axes, are less conclusive Refining Algorithm 8-1 to account for these findings results in Algorithm 8-2

If the frames of reference in a scene are dissociated Then

    Construct the spatial template for the preposition along the appropriate axes in both of the competing reference frames

    If the preposition used is canonically aligned with the vertical axis Then

        Select the candidate trajector with the highest applicability in the absolute/viewer-centred reference frame, unless the intrinsic reference frame's primary candidate has an applicability rating over twice that of the absolute/viewer-centred reference frame

    Else If the preposition used is canonically aligned with the horizontal axis Then

        Select the candidate trajector with the highest applicability, with a bias towards the intrinsic frame in the event of a tie

    End If

End If

Algorithm 8-2· Frame of Reference Competition Resolution Algorithm 2

In Section 5 3 1 3, Carlson-Radvansky's (1996) and Carlson-Radvansky and Logan's (1997) experiments were reviewed These experiments examined what effect the competition between multiple activated frames of reference has on the construction of a preposition's spatial template The results of these experiments indicated that if there is a competition between reference frames, the construction of a preposition's spatial template in one frame of reference interferes with the construction of the spatial templates in the other frame of reference This interference between reference frames results in an amalgamated spatial template which extends over the areas covered by both of the individual spatial templates Furthermore, the constituency of this amalgamated spatial template differs from the viewer-centred or intrinsic spatial templates there is no good region, the acceptable regions are bigger and the bad regions are smaller The

regions that are rated as acceptable in both the viewer-centred and intrinsic frames of reference have a higher acceptability rating in the amalgamated frame of reference than those in the regions which are acceptable in only one of the individual spatial templates

The simplest way to merge the competing spatial templates is to sum the applicability ratings for a point in each of the competing templates and then normalise these results by dividing the summed values by the maximum value in the resulting template. However, this approach ignores the biases towards particular reference frames along the vertical or horizontal axes (see Section 5 3 1 2 and the Algorithm 8-2 above) One way to incorporate these biases into the amalgamated spatial template is to multiply the applicability ratings in a spatial template in the reference frame towards which there is a bias for a particular preposition by a value representing the bias before the spatial templates are summed. This process results in higher applicability ratings in the amalgamated template in the regions covered by the spatial template constructed in the preposition's preferred frame of reference. Based on the analysis of (Carlson-Radvansky and Irwin 1993) in Section 5 3 1 2, a ratio of 2 1 in favour of the viewer-centred frame of reference for prepositions that are canonically aligned with the vertical axis can be set. The work of Taylor et al indicates a slight bias towards the intrinsic frame of reference for prepositions canonically aligned with the horizontal axes. Although this bias has not been quantified, for the sake of computational ease, a ratio for the bias at 1 1 1 in favour of the intrinsic frame of reference for the spatial templates of prepositions canonically aligned with the horizontal axes is assumed. While there is a marginal difference across this ratio, it is sufficient to prefer the intrinsic frame of reference in the event of a tie. Adapting Algorithm 8-2 for reference frame selection to include these findings results in Algorithm 8-3.

If the frames of reference in a scene are dissociated Then

Construct the spatial template for the preposition along the appropriate axes in both of the competing reference frames

If the preposition used is canonically aligned with the vertical axis Then

Multiply the applicability ratings in the spatial template constructed in the viewer-centred frame of reference by 2

Assign each point an applicability rating equal to the sum of its applicability ratings in both spatial templates

Select the candidate trajector with the highest applicability in the amalgamated frame of reference as the referent

Else If the preposition used is canonically aligned with one of the horizontal axes Then

Multiply the applicability ratings in the spatial template constructed in the intrinsic frame of reference by 1 1

Assign each point an applicability rating equal to the sum of its applicability ratings in both spatial templates

Select the candidate trajector with the highest applicability in the amalgamated frame of reference as the referent

End If

End If

Algorithm 8-3   Frame of Reference Competition Resolution Algorithm 3.

It is important to note that the above competition resolution algorithm is congruent with Carlson-Radvansky and Logan's (1997) analysis (see Section 5 3 1 3) It should also be highlighted that this competition resolution algorithm does not claim to represent a cognitive processes used by humans in selecting a frame of reference but aims to model human preferences A key element in this algorithm is the construction of the spatial templates In the following section, a computational model of the spatial template of projective prepositions is developed

232

8.4 Modelling Projective Prepositions

The third stage in the general algorithm for interpreting a locative projective expression is to define the area of search for the trajector as defined by the spatial template associated with the preposition This requires a computational algorithm that defines the spatial template for a projective preposition To define a spatial template that can model the gradation of a preposition's applicability across a region and also avoid the paradoxical parsing of space that previous models were susceptible to, one must integrate a scruffy topological model with perceptually based precepts

In Section 2 3 4 2, the main issues defining the area of a projective preposition were introduced

1   the orientation of the canonical direction of a preposition,

2   the origin of the spatial template,

3   the constituency of the spatial template,

4   the scale dependency of the spatial template on the extension of the landmark

After examining the psycholinguistic literature pertaining to these issues (Section 2 3 4 2 1), a set of criteria that a semantic model of prepositions should accommodate were defined

1   There are three areas within the spatial template good, acceptable, and bad

2   These areas are symmetrical around the search axis

3   The good and acceptable regions blend into one another

4   There is a sharp boundary between bad and acceptable regions

5   The acceptability of a projective preposition decreases linearly as the angular deviation increases

6   Acceptability approached 0 as the angular deviation approaches $90°$

7   The extension of the landmark orthogonal to the preposition's canonical direction effects the scale of the angular deviation and consequently a relation's degree of applicability

8   The distance between the landmark and trajector impacts on the acceptability rating of the trajector by virtue of the fact that if there are two trajectors located at the same angular deviation from the search axis the trajector closer to the landmark will have a higher acceptability rating

9   There is a distinction between the angular dependence of the three main directions *in front of-behind*, *right-left* and *above-below* with the *in front of-below* direction rated highest followed by the *above-below* direction It is conjectured that perceptual cues are the basis of this variance

In Section 5 3 2, previous approaches to modelling prepositions were reviewed The neat models (Cooper 1968, Leech 1969, Bennett 1975, Miller and Johnson-Laird 1976, Herskovits 1986) were rejected because of their reliance on simple logical definitions that cannot account for the specificity or vagueness that may occur within a particular use of a preposition The scruffy models (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994, Fuhr *et al* 1998, Mukerjee *et al* 2000) were rejected because (a) all of these models use purely topological approaches, and, consequently, ignore perceptual cues such as object occlusion, (b) none of these models propose a solution to the issue attending to the location of a spatial template's origin, (c) none of these models account for the impact of the selection of a frame of reference on spatial template construction

In this section, a semantic model for the projective prepositions *in front of*, *behind*, *to the left of*, and *to the right of* that avoids the deficiencies identified in its predecessors and fulfils the criteria defined in Section 2 3 4 2 is described The basis of this model is a parameterised continuum function that works in three dimensions and integrates perceptual precepts

234

To begin, the two assumptions that this model is based on are examined

1   The perceptual phenomenon of object occlusion impacts on the spatial templates of projective prepositions *in front of* and *behind* in the viewer-centred frame of reference Consequently, each of these prepositions has similar but different spatial templates associated with it in each frame of reference

2   A projective preposition's spatial template is defined through a process of mutual exclusion with the spatial template of its antonym, i e , if an object's location cannot be described as being *in front of* a landmark it may be described as being *behind* the landmark and vice versa

Both of these assumptions are novel Indeed, they directly contradict the approaches adopted by previous computational models of projective prepositions which define purely topological definitions and furthermore assume that each preposition has only one spatial template associated with it hence they require some substantiation

Assumption 1 is not without precedent Claude Vandeloise (1991) posits that prepositions *devant/derriere* are bisemic[51], because the relationships they describe between the trajector and the landmark in the intrinsic frame of reference are different from the ones they describe in the contextual orientation[52] (viewer-centred frame of reference) In the intrinsic frame of reference "*a est devant/derriere b* if the target[53] is located on the positive/negative[54] side of the landmark's general orientation" (Vandeloise 1991 pg 100), while in the viewer-centred frame of reference "*a est devant/derriere b* if

---

[51] A bisemic word is a word which has two meanings

[52] Vandeloise uses the term contextual orientation to describe the frame of reference provided by the speaker This orientation is identical to viewer-centred frame of reference

[53] The term *target* in Vandeloise's terminology corresponds to the term *trajector* in the terminology of this dissertation

[54] Vandeloise's defines the positive and negative sides of the landmark's general orientation along the front back axis in an identical manner to (Clark 1973), i e , front is positive and back is negative (see Section 2 2 1)

the target/landmark is (potentially) the first (partial) obstacle to the perception of the landmark/target" (Vandeloise 1991 pg 131) Essentially, Vandeloise (1991) argues that the primary factor in the viewer-centred usages of *devant/derriere* is the perceptual cue object occlusion In this thesis, it is not posited that object occlusion is the primary factor in the semantics of the prepositions *in front of-behind*, rather the approach is more aligned with that of Jackendoff and Landau, who argue that while object occlusion impacts on the semantics of these prepositions, it plays "a secondary role, possibly forming a preference rule system with the directional criteria" (1992 pg 114) What is important to note here is that both Vandeloise's (1991) and Jackendoff and Landau's (1992) theories are congruent with the first assumption that perceptual cues impact on the semantics of projective prepositions Unfortunately, the results of the experiments that are pertinent to this thesis are not consistent some results (Carlson-Radvansky 1996) indicate that the shape of a preposition's spatial template is independent of the frame of reference it was constructed in This would seem to falsify the proposition, while other results (Gapp 1995b, Logan 1995) indicate that perceptual cues do impact on the semantics of horizontally aligned prepositions

In Section 5 3 1 3 the results of Carlson-Radvansky's (1996) psycholinguistic experiments was reviewed Of primary importance were the results pertaining to the issue of whether or not multiple reference frames are activated when the reference frames are dissociated However, it is the results of the other line of enquiry within Carlson-Radvansky's (1996) experiments that are of relevance here Apart from investigating the activation of reference frames, Carlson-Radvansky also attempted to ascertain whether the spatial template associated with a preposition was independent of the type of reference frame that was used to align it The results indicate that this was indeed the case, the spatial template for a preposition describes a similar shaped area regardless of whether it is constructed in an absolute, a viewer-centred, or an intrinsic frame of reference Clearly, these results contravene the first premise However, Carlson-Radvansky's (1996) experiments used a similar procedure to that used in (Logan and Sadler 1996) and consequently suffered from the same flaws (see Section 2 3 4 2 3) This means that visual cues such as object occlusion did not occur in the test Moreover, these experiments focused on the spatial template for the preposition *above* which is

236

canonically aligned with the vertical axis. Object occlusion would have little impact on this preposition and as a result may not have been a major consideration in the design of the experiments. Considering this it is not surprising that the experimental results indicated that the preposition's spatial template was consistent across frames of reference. However, for the current discussion what is important to note is that the results of these experiments do not falsify the first premise, because the preposition the experiments focused on does not fall within the set of prepositions defined by the premise. Moreover, if the experiments had focused on one of the horizontally aligned prepositions the results would be not be applicable to this discussion because the experimental procedures used precludes the occurrence of the perceptual cue that the premise is based upon.

In Section 2.3.4.2, the results of psycholinguistic experiments that investigated the constituency of the spatial templates for projective prepositions were reviewed. One of these experiments (Gapp 1995b) found that there was a difference between spatial templates of prepositions aligned with different spatial axes; regions in the spatial templates for *in front of-behind* and *above-below* were rated slightly higher than regions with the same angular deviation in the *right-left* spatial template. This finding echoes the results of earlier psycholinguistic work (Logan 1995; Logan and Sadler 1996) which found that: "Subjects were faster with *above* and *below* than with *front* and *back*, and faster with *front* and *back* than with *left* and *right*" (Logan and Sadler 1996 pg. 505). Concluding that section, it was noted that the higher ratings could be a result of the ease of perceiving the asymmetry along the *in front of-behind* and *above-below* axes. This analysis is congruent with Vandeloise's (1991) analysis and further strengthens the case for adopting assumption 1 above.

Assumption 2 claims that a process of mutual exclusion between the spatial templates associated with a projective preposition and its antonym plays a central role in the definition of the spatial template for each of the projective prepositions in the complementary pair. This assumption is inspired by the work of Terry Regier (1996). Regier's work attempted to characterise the semantic universals underpinning language by investigating the learnability of particular linguistically expressed concepts. This work is particularly relevant to this thesis as it focused on the acquisition of spatial language in a neural architecture that incorporated perceptual structures similar to those in the visual

237

system Regier proposes mutual exclusion as a solution to the problem of how children learn language almost entirely without the benefit of negative evidence "how can the child generalize from the input without overgeneralizing to include inappropriate usages, if these usages have never been explicitly flagged as infelicitous?" (Regier 1996 pg 59) The idea of mutual exclusion is "to take every positive instance of one spatial concept to be an implicit negative instance for all other spatial concepts being learned" (Regier 1996 pg 62) The results of the experiments indicated that mutual exclusion does provide a means for learning in the absence of explicit negative evidence Furthermore, giving the system a prior knowledge of antonyms facilitates the learning The results pertaining to learning with prior knowledge of antonyms is particularly revealing as these results indicate that the learning and, by inference, the semantics of antonymic prepositions are directly linked Of course, the plausibility of the idea that antonymic prepositions directly effect the semantics of each other hinges on the knowledge of which prepositions are antonymically paired being acquired prior to the acquisition of the semantics of the individual spatial terms There is evidence supporting this notion Regier notes that

"the opposite of *above* is *below* and not *under*, even though *under* is roughly synonymous with *below* This is of possible relevance since it indicates that it is the words themselves, rather than the meanings that are antonymically paired If this lexical pairing were known to children before the word meanings were, the knowledge could be used in acquiring the word meanings" (1996 pg 68)

Moreover, Tomasello's (1987) psycholinguistic work indicates that English prepositions which are members of antonym pairs (e g , *in-out, over-under*) are learned earlier than prepositions that are not (e g , *by, at*) Furthermore, the results of (Gapp 1995b) and (Logan 1995) which were discussed above indicated that the spatial templates associated with and the speed of comprehension of prepositions differed between prepositions aligned with different spatial axes In other words, a projective preposition and its antonym were similar to each other in terms of their spatial templates and people's speed at comprehending them

To summarise, assumption 1 is underpinned by Vandeloise's (1991) and Jackendoff and Landau's (Jackendoff and Landau 1992) theoretic analysis and the psycholinguistic work of Gapp (1995b) and Logan (1995) Assumption 2 is affirmed by Regier's (1996) findings which illustrate the plausibility of mutual exclusion as part of the semantic acquisition process for projective prepositions and furthermore indicates that a preposition's antonymic partner affects the definition of its semantics The semantic link between antonymic prepositions is also attested to in the results of (Gapp 1995b, Logan 1995) Based on this evidence, it is posited here that assumption 1 and 2 above are cognitively plausible as a basis for a semantic model of projective prepositions A consequence of adopting these assumptions is that a spatial template for each preposition in both the viewer-centred and intrinsic frames of reference must be defined However, while the spatial templates of projective prepositions differ from preposition to preposition and across different frames of reference, all the psycholinguistic (Gapp 1995b, Carlson-Radvansky 1996, Logan and Sadler 1996) evidence indicates a family resemblance across all the spatial templates for projective prepositions This resemblance emerges from their unilateral dependence of applicability on the angular deviation of the trajector position from their canonical direction and the distance of the trajector from the landmark To model this family resemblance, a continuum function is defined which models these topological factors of angular deviation and distance This function is the basis for the semantics of all the projective prepositions However, depending on the preposition and the frame of reference, this function is combined with perceptual cues which refine the spatial template it creates

In Section 8 4 1, the topological component of the spatial template model developed in this thesis is defined, henceforth, the SLI spatial template model The topological component of the SLI spatial template model consists of a continuum function similar to the potential field models reviewed in Section 5 3 2 2 The novel aspects of the SLI spatial template model's topological component are

1 It uses a new algorithm, developed in this thesis, which defines the spatial templates point of origin in the viewer-centred frame of reference based on the user's view of the landmark This algorithm for locating the spatial

template's origin permits the SLI spatial template model to avoid the problems of schematising the landmark by either its bounding box or its centroid in the viewer-centred frame of reference.

2. The SLI spatial template's potential field model is constructed using a <u>parameterised</u> continuum that works in 3-D. The parameterisation of the model means that the model is scalable and, consequently, it can be applied to different sized landmarks. Furthermore, the potential field model describes both the angular deviation of a point from the canonical direction of the preposition and the distance of the point from the spatial template's origin for every point in the spatial template. In Section 5.4.2.3, it was noted that the CSR-3-D system (Gapp 1994a) is the only system, prior to the SLI system developed here, that defines a scalable 3-D spatial template that accommodates a measure of the trajector's angular deviation from the canonical direction of the projective preposition's search axis and a measure of the distance of the trajector from the spatial template's origin. However, it was also noted that the potential field model proposed in (Gapp 1994a) uses a local coordinate system which is centred on the landmark's BRP. Although this local coordinate system ensures the scaling of the trajector's angular deviation and distance scores relative to the size of the landmark, it also <u>forces</u> the CSR-3-D system to use the landmark's centroid to represent the landmark. Importantly, the potential field model developed in this thesis avoids using a scaled local coordinate system to scale the trajector's angular deviation and distance scores by using parameterised continuum functions. It is this parameterised approach that allows the SLI potential field model to be combined with a ray casting algorithm for locating the spatial templates origin.

While the SLI spatial template model's topological component has several advantages over its predecessors, it suffers from a weakness that affects all purely topological models: when applied to a complex landmark in a viewer-centred frame of reference, there is a possibility that regions which are occluded by the landmark will be

240

topologically defined as *in front of* the landmark As a response to this issue, it is proposed that these areas can be accommodated by integrating perceptual cues into the topological component of the SLI spatial template model Accordingly, in Section 8 4 2, a set of the perceptual definitions for the prepositions aligned along the front-back axis in the viewer-centred frame of reference are defined, it is also illustrated why these perceptual definitions are not sufficient by themselves to model the spatial templates of the projective prepositions they apply to These perceptual definitions comprise the perceptual component of the SLI spatial template model Finally, Section 8 4 3 describes how the topological and perceptual components of the SLI spatial template model are integrated and how this integrated model overcomes the shortcomings of its component elements, i e , the topological and perceptual components It should be noted that none of the previous models of preposition's spatial templates integrate perceptual axioms into their framework

8.4 1 SLI Spatial Template Model: Topological Component

The SLI spatial template model's topological definitions are scruffy models similar to the potential field models, reviewed in Section 5 3 2 2 Recall that these potential field models attempt to model the gradation of a preposition's spatial template using a potential energy function which returns an applicability value for each location in the spatial template The SLI spatial template model's potential field model developed here avoids the deficiencies noted in its predecessors the SLI potential field model avoids the problems of schematising the landmark by either its bounding box or its centroid in the viewer-centred frame of reference, it works in three dimensions, it models both the angular deviation of a point from the canonical direction of the preposition and the distance of the point from the spatial template's origin for every point in the spatial template, and finally, it is parameterised to accommodate a scaling of the template relative to the size of the landmark

The origin of a potential field function / spatial template model is usually defined by abstracting the landmark to a point While there are an infinite number of points that may be selected to represent the landmark object, the previous potential field models (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994) use the centre of the landmark's bounding box or the bounding box itself However, these representations can be problematic (see Section 2 3 4 2 4)

Here, a new method is proposed that dynamically defines the spatial template's point of origin in the viewer-centred frame of reference based on the user's view of the landmark However, it is conjectured that the origin of the spatial templates in the intrinsic frame of reference is known by the user in a similar manner to the orientation of the prepositions around the landmark, i e , through prior knowledge The motivation for this treatment of the intrinsic frame of reference is that if a person associates an intrinsic frame of reference with an object, they must have learnt this intrinsic orientation based on prior experience with the object or objects of that type (for a review of the different strategies for defining an object's intrinsic horizontal axes see Section 2 3 3 2 1) Therefore, they are familiar with the object and should be able to gauge its probable dimensions and its centre Following this, the location of the spatial template origin for the intrinsic frame of reference is independent of the user's perception of the landmark and is known by the system through a priori knowledge In contrast with the intrinsic frame of reference, the view-centred frame of reference may be applied to an object without prior knowledge of the object, its dimensions or its centre From this, it is argued that it is cognitively implausible to assume that a person uses a point in space whose location they do not know (i e , the centre of an unfamiliar landmark) as the origin for their spatial orientation Importantly, this assumption is inherent in the previous potential field models that schematise the landmark by its centroid in the viewer-centred frame of reference (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994)

In Section 5 2 2 the graphics technique called ray casting was described ray casting can be functionally described as casting a ray (i e , drawing an invisible line) from one point in the 3-D world in a certain direction and then reporting back all the objects this

line intersected Ray casting was rejected as a viable technique for modelling visual perception because of the computational expense of utilising this function a large number of times However, it is now proposed that ray casting be used to locate a point on the landmark that can act as a suitable point of origin for the spatial templates, because this algorithm only requires one ray casting per iteration in comparison to the tens of thousands needed in order to model visual perception

What is proposed is to cast a ray from the user location though the point at the centre of the landmark's 3-D bounding box, and use the first point of intersection of the ray with the landmark's 3-D mesh as the spatial template's origin Once the spatial template's origin has been located, two planes are constructed both of which pass through the origin plane 1 runs orthogonal to the cast ray, plane 2 runs parallel to the ray Plane 1 bisects the space around the landmark into two regions, the region containing the user's location is defined as being *in front of* the landmark, while the other region is defined as being *behind* the landmark Plane 2 also bisects the space around the landmark into two regions These regions are aligned with the user's *left* and *right* The diagrams in Figure 8-1 illustrate the different steps in the SLI ray casting algorithm These diagrams use a bird's eye view of a spatial configuration containing a U-shaped landmark, the landmark's bounding box, the landmark's bounding box centre, and the user's location and field of view Figure 8-1(a) illustrates the path of the ray cast from the user's location through the centre of the landmark's bounding box and then on through the landmark itself, the point where the ray first intersects with the object mesh is also highlighted Figure 8-1(b)[55] shows how this approach parses the space around an object into the two regions *in front of-behind* once the intersection point has been located Figure 8-1(c) shows the parsing of space by the second plane into the regions *to the right of* and *to the left of* Note that these illustrations use the same user-landmark configuration that was shown in Section 2 3 4 2 4 to be problematic for approaches adopting the object's centroid as the spatial templates origin Indeed, the major advantage of the our approach

---

[55] It is apparent in diagram (b) of Figure 8-1 that there are regions of space which are perceptually *behind* the landmark but are topologically defined as *in front of* the landmark even after the shifting of the spatial template These problem areas will be accommodated in the SLI model by the integration perceptual cues which will be discussed in Section 8 4 3

is that it avoids the paradoxical parsing of space that can arise when the landmark is represented by its centroid: in Figure 8-1(d), the parsing of the space around the landmark is done using the centroid of the landmark's bounding box; the area highlighted in full red is defined as *behind the building* from the viewer's perspective. This is clearly wrong!

(a)

(b)

(c)

(d)
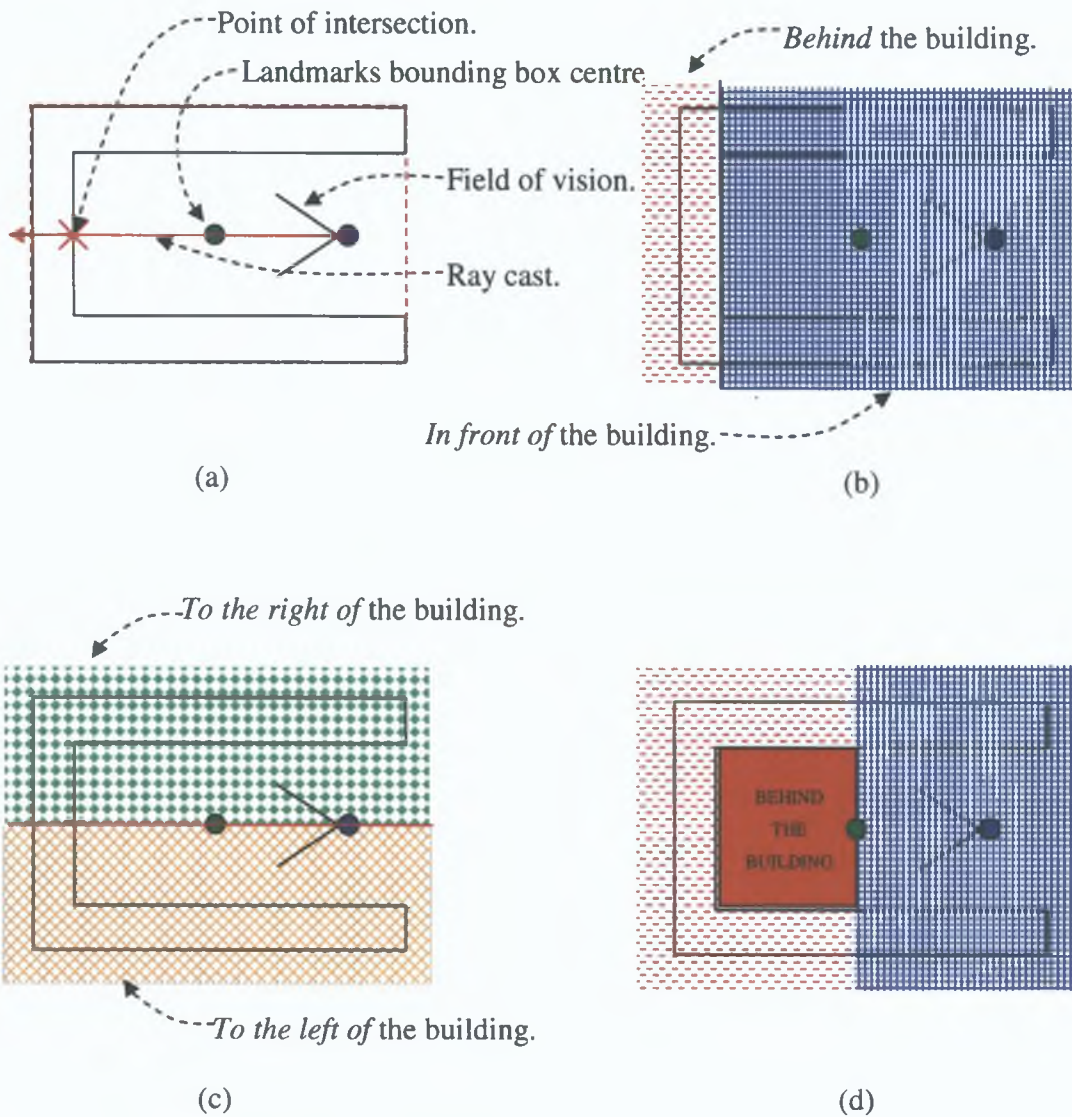
Figure 8-1: Diagrams illustrating the ray casting method for defining a spatial template's origin. These diagrams use a bird's eye view perspective. Diagram (a) illustrates the path of the ray from the user's location through the landmark's bounding box centre and then on through the landmark. The point where the ray

initially intersects the object's mesh is highlighted Diagram (b) illustrates the parsing of space into the *in front of* and *behind* regions once the spatial template origin has been defined. Diagram (c) illustrates the parsing of space into the regions *to the right of* and *to the left of* once the spatial template origin has been defined. Diagram (d) illustrates the parsing of space around the object using the object's centroid The area coloured in full red is defined as behind the building from the viewer's perspective. This is clearly wrong.

There are, however, situations where the cast ray might not intersect with the object Figure 8-2 illustrates one situation where this might occur here the user is located close to the end of a convex shaped landmark A ray sent from that location through the centre of the landmark's bounding box passes through the open end of the landmark, without intersecting the landmark's mesh It is important to note that the ray is not sent in the direction the user is looking, but rather, in the direction of the landmark's bounding box centre The reason for not correlating the direction of the ray with the direction of the user's view is that, if the landmark is not in the centre of the user's view and the ray is sent in the direction that the user is looking, the ray may well miss the landmark Clearly, the ray can also miss the landmark when it is sent in the direction of the landmark's bounding box centre However, sending the ray in the direction of the landmark's bounding box centre, rather than in the direction that the user is looking, reduces the possibility of the ray missing the landmark's mesh Moreover, an alternate strategy is developed below which allows the SLI framework to accommodate many of the situations where the first ray misses the landmark's mesh This alternate strategy, however, is dependent on the direction of the first ray Consequently, the alternate strategy would not be applicable if the first ray was sent in the direction of the user's gaze
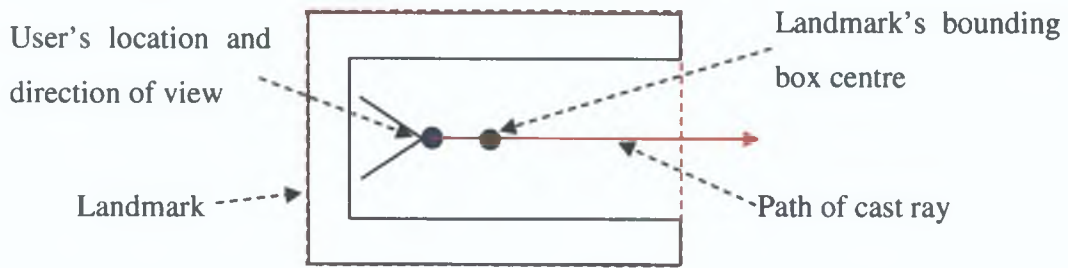
Figure 8-2: A situation where a cast ray does not intersect with the landmark's mesh.

We propose that in situations where the cast ray does not intersect with the intended object, a second ray should be sent in the opposite direction to the first. If the second ray intersects with the landmark's mesh, this point of intersection is taken as the spatial templates origin. Figure 8-3 illustrates this process.
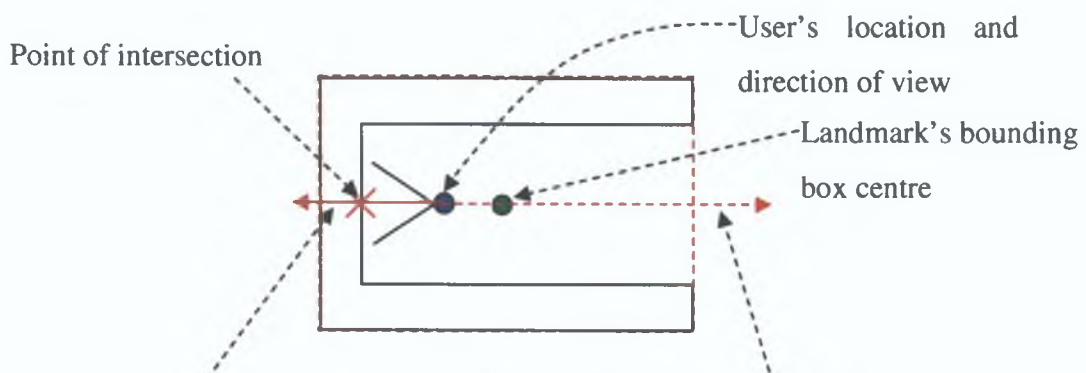
Path of 2nd ray. This ray is sent in the opposite direction to the first ray.

Path of 1st ray. The direction of this ray is not dependent on the direction the user is looking. Rather, this ray is sent from the user's location in the direction of the landmark's bounding box centre. In this example, this ray does not intersect with the landmark's mesh.

Figure 8-3: An illustration of the path of the second ray cast if the first ray through the object's bounding box centre fails to intersect with the object. The point of intersection with the second ray is highlighted; this point is taken to be the origin of the spatial template.

246

The casting of a second ray will allow us to define a suitable spatial template origin for the majority of instances where the first ray doesn't intersect with the object's mesh. However, there is still the possibility that neither ray will intersect with the object: for example, the user may be standing underneath a large arch. Figure 8-4 depicts this situation. This figure uses a bird's eye view of an arch with the arch top removed. The two boxes on either side of the user's viewpoint represent the supporting columns of the arch.

User's location and direction of view

Landmark's bounding box centre

Path of 2nd ray sent in the opposite direction to the first. In this example, this ray does not intersect with the landmark's mesh.

Path of 1st ray. This ray is sent from the user's location in the direction of the landmark's bounding box centre. In this example, this ray does not intersect with the landmark's mesh.
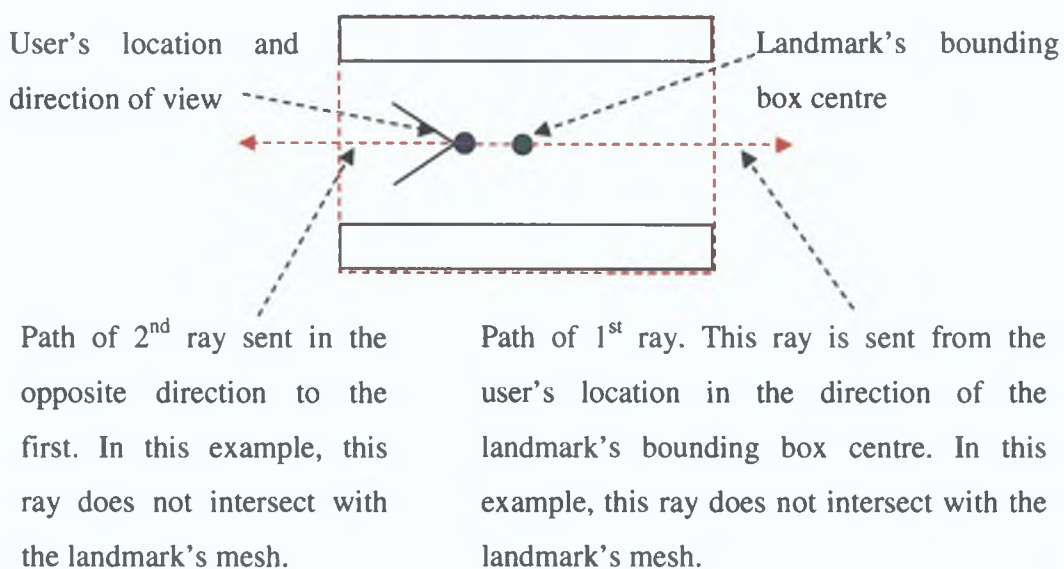
Figure 8-4: An illustration of the paths of the two cast rays in a situation where neither of the rays intersects with the object. This diagram represents a bird's eye view of an arch with its top removed. The user is standing under the arch with the supporting columns of the arch on either side.

In situations where neither of the rays intersect with the landmark's mesh, the framework uses the centre of the landmark's bounding box to represent the object. While this is not ideal, the number of situations where the centre of the landmark's bounding box is used is reduced by this process. Algorithm 8-4 lists the different steps in the process used to locate the spatial template origin in the viewer-centred frame of reference.

1	Cast a ray from the user's location through the point at landmark's bounding box centre If this ray intersects with the landmark's mesh, the point of intersection defines the spatial template origin
2	If the first ray does not intersect with the landmark's mesh, cast a ray in the opposite direction to the first ray If the second ray intersects with the landmark's mesh, the point of intersection defines the spatial template origin
3	If neither of the cast rays have intersected with the landmark's mesh, the landmark's bounding box centre is taken to define the spatial templates origin

Algorithm 8-4· The algorithm for locating the spatial template origin in the viewer-centred frame or reference.

As described above, once the origin has been located, the area around the landmark is parsed into different regions Having defined these regions, the next task is to model the gradation in each preposition's applicability across their respective region

8 4 1 2 Modelling the Gradation of a Preposition's Applicability

Within the criteria defined for the spatial templates of a projective preposition (defined in Section 2 3 4 2), there are two factors which are identified as impacting on the applicability of a preposition at a given point These are the angular deviation of the point from the canonical direction of the preposition and the distance of the point from the origin of the spatial template To model the directional constraint of projective prepositions using a spatial template, the modelling process must define a method for calculating the deviation of a point from a preposition's canonical direction The first stage of this process is to assign a canonical direction to each of the prepositions A

248

priori, the vector originating from the spatial template's origin to the user's location describes the canonical direction for *in front of* in the viewer-centred frame of reference Following this, the canonical direction for *behind* can be defined by rotating this front vector by $180°$ around the origin, and the directions for left and right can be defined by rotating the front vector $90°$ and $270°$ around the origin on the horizontal plane Given this, one way of computing the direction vector for *in front of* is to convert the world coordinate point representing the user's location in the world into a point in a local coordinate system centred on the spatial template's origin While there are several methods for achieving this, the one used here is to translate the spatial template origin to the world origin and then apply a translation with the same values to the point representing the user's location in the world Translating the spatial templates origin to the world origin is simply a matter of subtracting the spatial template's origin coordinates from themselves, in effect setting them to zero Applying the same translation to the user's location is as simple a procedure subtract the original spatial template origin coordinates from the coordinates of the user's location After these translations, the translated coordinates of the user's location are equivalent to the vector originating at the spatial template's origin in the direction of the user's location, i e , the direction vector for *in front of* the landmark in the viewer-centred frame of reference The direction vector describing the canonical direction for the preposition *behind* can be computed by rotating the front vector by $180°$ around the spatial template's origin The vectors for left and right can be defined by rotating the front vector by $90°$ and $270°$ on the horizontal plane around the translated spatial template origin Algorithm 8-5 formally defines the different stages in this algorithm and Figure 8-5 graphically illustrates the stages in this process Diagram (a) illustrates the world coordinates of the user, the spatial template origin, and a trajector Diagram (b) illustrates the translation of the spatial template origin to the world origin and the translation of the user's location world coordinates by the same translation as was applied to the spatial template origin Diagram (c) illustrates the vector defining *in front of* in the viewer-centred frame of reference after the translation of the spatial template's world coordinates and the user's location world coordinates

```
1   OLCS = WCSTO – WCSTO = [0 0, 0 0, 0 0]

2   LCUL = OLCS – WCSTO

3   VCFront = LCUL
```

Algorithm 8-5. The steps in calculating the front vector in the viewer-centred frame of reference. In this algorithm, the following acronyms are used WCSTO represents the vector describing the world coordinates of the spatial template origin, ULWC represents the vector describing the world coordinates of the user's location in the simulation, OLCS represents the vector describing the origin of the local coordinate system centred on the spatial template's origin, LCUL represents the vector describing the coordinates of the user's location in the local coordinate system centred on the spatial template's origin, and VCFront represents the direction vector for *in front of* the landmark in the viewer-centred frame of reference.

(a)

World origin

● World coordinates of point selected as spatial template origin

● Point representing trajector's world coordinates

● User's location in world coordinates

(b)

Translated spatial template origin

● World coordinates of point selected as spatial template origin

● Point representing trajector's world coordinates

● User's location in world coordinates

● The translated coordinates of the user's world coordinates

(c)

Translated spatial template origin

● Point representing trajector's world coordinates

Direction vector for *in front of* in viewer-centred frame of reference

● The translated coordinates of the user's world coordinates

Figure 8-5 Diagrams illustrating the different stages in defining the vector that defines the canonical direction for *in front of* in the viewer-centred frame of reference. Diagram (a) illustrates the world coordinates of the user, the spatial template origin, and a trajector. Diagram (b) illustrates the translation of the spatial template origin to the world origin and the translation of the user's location world coordinates by the same translation as was applied to the spatial template origin. Diagram (c) illustrates the vector defining *in front of* in the viewer-centred frame of reference after the translation of the spatial template's world coordinates and the user's location world coordinates Note that in these diagrams, full lines with arrow head endings are used to represent vectors and dashed lines are used to connect labels to objects.

Having assigned a direction to each preposition, the next step in the modelling process is to devise a method for calculating the angular deviation of a point from the direction vector of the preposition The angle between two vectors θ is given by the equation

$$\theta = \cos^{-1}\left(\frac{v \bullet w}{\|v\|\|w\|}\right)$$

Equation 8 The equation for the angle between two vectors.

252

Here v and w are two n-dimensional vectors with:

$$v = [x_1, \ldots, x_n]$$
$$w = [y_1, \ldots, y_n]$$
$$v \bullet w = x_1 y_1 + \ldots + x_n y_n$$
$$\|v\| = \sqrt{x_1^2 + \ldots + x_n^2}$$
$$\|w\| = \sqrt{y_1^2 + \ldots + y_n^2}$$

Equation 9: The equations for the dot product of two vectors and the length of two vectors.

The function $v \bullet w$ is known as the dot product and the functions ‖v‖ and ‖w‖ give the distance from the point represented by the vector to the coordinate system's origin. This function, however, only works for two vectors that have a common origin. As such, in order to use this function to measure the angular deviation of a point from the vector describing the canonical direction of a preposition, the point must be converted into a vector that shares a common origin with the preposition's direction vector. This conversion is done by applying the same translation to the point that was applied to the user's location: i.e., subtract the original spatial template's coordinates from the point. This converts the point into the local coordinate system centred on the spatial template's origin, which is the same coordinate system as the vectors describing the canonical direction of the projective prepositions. Consequently, after this translation, the translated coordinates of the point are equivalent to a vector whose origin is the same as the vector describing the direction of the preposition from the spatial template origin. Algorithm 8-6 formally defines the conversion of a world coordinate into a vector that shares a common origin with the direction vectors calculated using Algorithm 8-5. Figure 8-6 graphically illustrates the different stages in this conversion process. Diagram (a) illustrates the situation after the definition of the direction vector for *in front of* in the viewer-centred frame of reference. This diagram is equivalent to diagram (c) in Figure 8-5. Diagram (b) illustrates the translation of the point. The coordinates defined by this translation represent the location of the point in the local coordinate system defined around the

spatial template origin Diagram (c) illustrates the vector from the spatial template origin to the translated point It is this vector that is used in computing the angular deviation of the point from the vectors describing the canonical direction of the projective prepositions in the viewer-centred frame of reference

$$LCP = WCP - WCSTO$$

Algorithm 8-6: The calculation of the vector describing the location of a point in the world, which is to be rated in the spatial template, in the local coordinate system centred on the spatial template's origin In this algorithm, the following acronyms are used: WCP represents the vector describing the world coordinates of the point that is to be rated in the spatial template origin, WCSTO represents the vector describing the world coordinates of the spatial template origin, and LCP represents the vector describing the coordinates of the point to be rated in the local coordinate system centred on the spatial template's origin. Note that the vector LCP has the same origin as the direction vectors defined using Algorithm 8-5.

(a)

Translated spatial
template origin

● Point representing trajector's
world coordinates

Direction vector for *in front of* in viewer-centred frame
of reference

● The translated coordinates of the user's world coordinates

(b)

Translated spatial
template origin

in front of vector

● Point representing trajector's
world coordinates

● Coordinates of translated trajector's world coordinates

● The translated coordinates of the user's world coordinates

(c)

Translated spatial
template origin

The vector describing the location of the trajector
in the spatial template's local coordinate system.
This vector has a common origin with the vectors
describing the canonical direction of the
prepositions in the viewer-centred frame of
reference.

● Coordinates of translated trajector's world coordinates

● The translated coordinates of the user's world coordinates

Figure 8-6 Diagrams illustrating the different stages in converting the point that represents the trajector's location in world coordinates into a vector that shares a common origin with the vectors that describe the canonical direction of the projective prepositions in the viewer-centred frame of reference. Diagram (a) illustrates the situation after the definition of the direction vector for *in front of* in the viewer-centred frame of reference. This diagram is equivalent to diagram (c) in Figure 8-5. Diagram (b) illustrates the translation of the trajector's world coordinates The coordinates defined by this translation represent the location of the trajector in the local coordinate system defined around the spatial template origin Diagram (c) illustrates the vector from the spatial template origin to the coordinates of the trajector in the spatial template origin local coordinate system It is this vector that is used in computing the angular deviation of the trajector's position from the vectors describing the canonical direction of the projective prepositions in the viewer-centred frame of reference. Note that in these diagrams, full lines with arrow head endings represent vectors and dashed lines connect labels to objects

Let w equal the vector representing the preposition's direction, calculated using Algorithm 8-5, and v equal the vector that represents the translated point's coordinates, calculated using Algorithm 8-6, then the angular deviation of the point from the direction of the preposition can be calculated using Equation 8 Applying this process to all the points in an area surrounding a landmark assigns each point in the area an angular deviation from the preposition's canonical direction These angular deviations can be normalised by defining a maximum angle for the potential field let this be β Once β has been set, all points with an angular deviation greater than β are assigned an angular applicability rating of 0 and the remaining points are assigned an angular applicability rating equal to 1 minus their initial value divided by β Algorithm 8-7 formally defines the process used to normalise the angular deviation of points within the spatial template

256

```
If ı > β Then

        J = 0

Else

        J = 1 – ( ı / β )

End If
```

Algorithm 8-7 The algorithm used to normalise the angular deviation scores of points in the spatial template of a projective preposition. In this algorithm, ı represents the angular deviation of a point from the vector describing the canonical direction of a projective preposition, J represents the normalised angular applicability of a point within the spatial template, and β represents the maximum allowable angle in the spatial template

Figure 8-7 illustrates a cross section of the angular applicability ratings assigned to points in an area centred on a landmark anchoring an idealised meaning of *above* For the purposes of this illustration, the vector *[0, 1, 0]* was used to represent the canonical direction for the preposition The vertical red line in the figure illustrates this direction For this example, β was set to $90°$, the maximum angle of acceptability as evidenced in (Logan and Sadler 1996) The darker the colour, the higher the angular applicability (1 is the maximum applicability value)
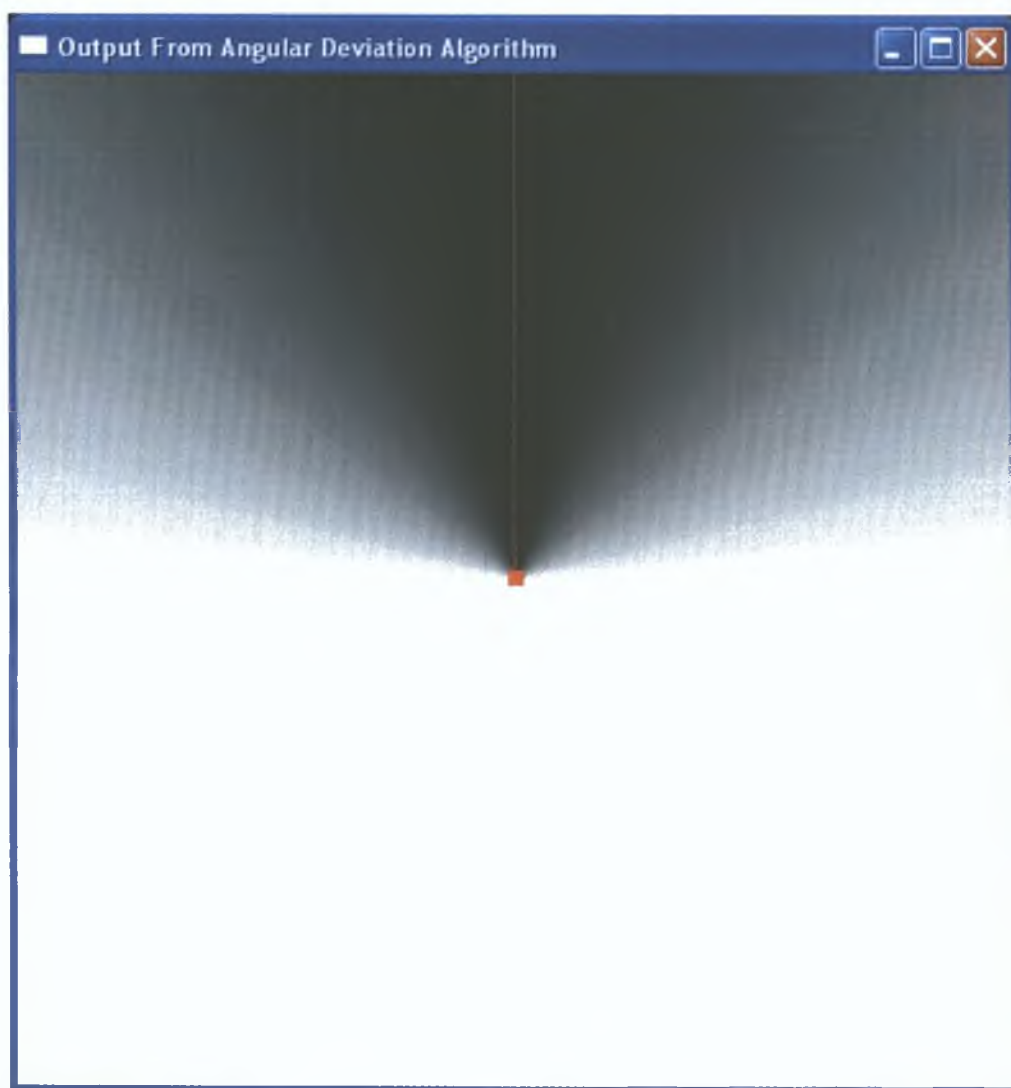
Figure 8-7: Graphical representation of the angular applicability ratings assigned to the points in an area using the equation of 1 − (angular deviation / β). The red square indicates the position of the landmark and the red line delineates the directional constraint of the preposition *above*.

Algorithm 8-5 assigns a vector to each preposition. This vector models the preposition's canonical direction. Moreover, using Algorithm 8-6, Equation 8, and Algorithm 8-7, a normalised rating of a point's angular deviation from a preposition's direction vectors can be calculated. In summary, this process models one of the topological factors affecting the applicability of a projective preposition. However, to

complete the topological definition for the projective prepositions, the distance of a point from the spatial template's origin must be modelled. Moreover, this distance applicability must be integrated with the angular applicability computed using the above process.

The distance applicability of a point can be computed using the standard coordinate geometry distance formula to calculate the distance between two points $[x_1, y_1, z_1]$ and $[x_2, y_2, z_2]$:

$$Dist. = \sqrt{\left((x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2\right)}$$

Equation 10: The equation for the distance between two points [x1, y1, z1], [x2, y2, z2].

By setting $[x_1, y_1, z_1]$ to the coordinates of the spatial template's origin and $[x_2, y_2, z_2]$ to the coordinates of each point that is to be rated, the distance of the point from the spatial template's origin is computed. These distances are then normalised using a process similar to the one used to normalise the angular deviations. First, a maximum range for the spatial template is defined; let this be Υ. Next, all points with a distance greater than Υ are assigned a distance applicability equal to 0 and the remaining points are assigned a distance applicability equal to 1 minus their distance divided by Υ. Algorithm 8-8 formally defines the steps in the process used to calculate the normalised distance applicability of points in a projective preposition's spatial template.

```
If i > Υ Then
        j = 0
Else
        j = 1 - ( i / Υ )
End If
```

Algorithm 8-8: The algorithm used to normalise the distance scores of points in the spatial template of a projective preposition. In this algorithm, i represents the distance of a point from the origin of the spatial template, j represents the

259

normalised distance applicability of the point in the spatial template, and Y represents the maximum distance allowed in the spatial template

It is important to note that although the psycholinguistic evidence (Gapp 1995b, Logan and Sadler 1996) indicates a maximum allowable angular deviation, no ratio of maximum distance to landmark size has been identified in these works Consequently, it is proposed here that Y should be set to the distance of the candidate trajector farthest from the spatial template origin but within the maximum allowable angular deviation This means that the distance from the spatial template does not preclude a candidate trajector being selected as the expression's referent, however, it does affect its rating within the process for selecting the referent Figure 8-8 illustrates a cross section of the distance applicability ratings assigned to points in an area centred on a landmark as in Figure 8-7, the darker the colour, the higher the applicability For this example, the range of the spatial template Y was arbitrarily set to 250 units As a result, all the points with a distance greater than 250 were assigned a distance applicability of 0, the other points were assigned a distance applicability of 1 minus their distance divided by 250 This procedure defines a range of distance applicability ratings between 0 and 1 Within this range, 1 represents a high applicability or nearness to the spatial template's origin and 0 represents a low applicability or distance from the spatial template's origin
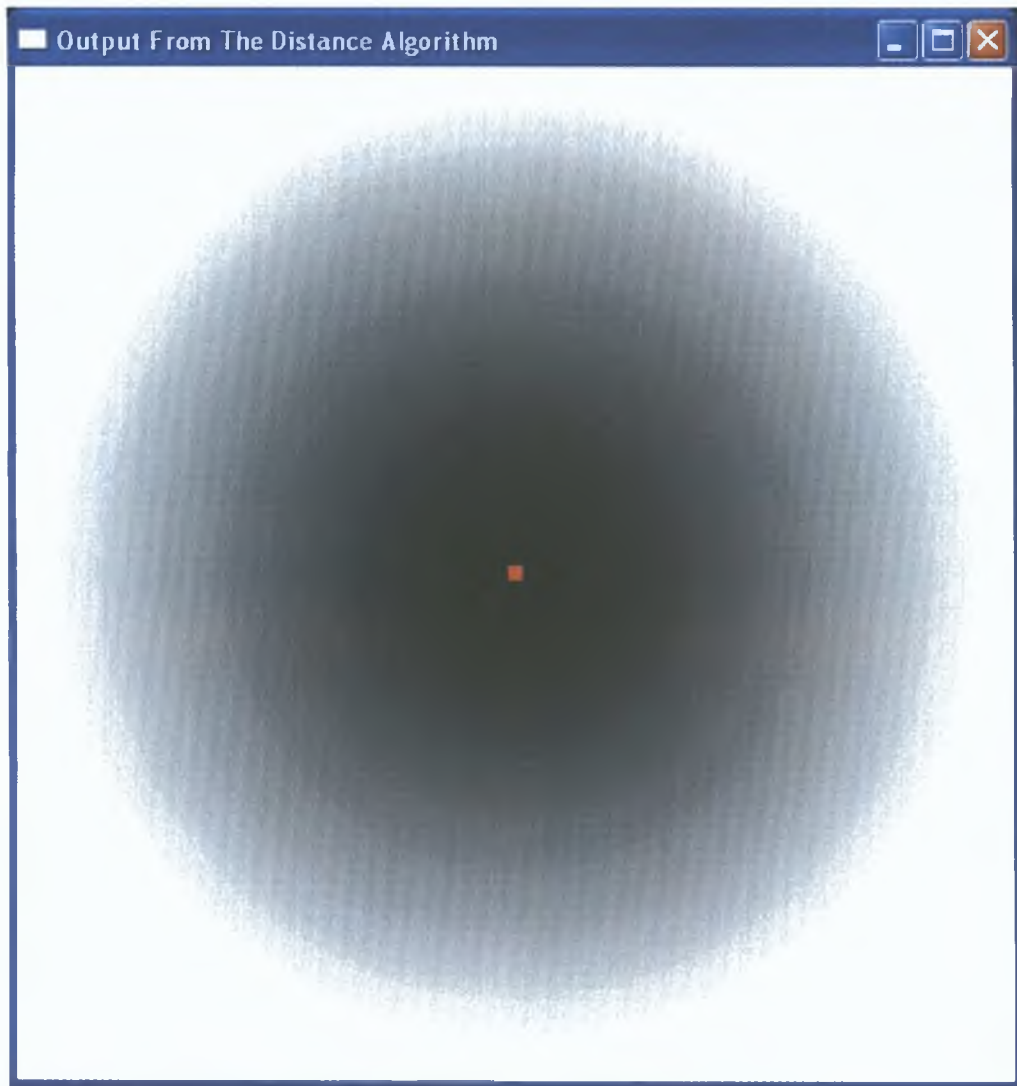
Figure 8-8: Graphical representation of the distance applicability ratings assigned to the points in an area using the equation of 1 – (distance / γ). The red square indicates the position of the landmark.

The above discussion has defined methods for quantifying both of the topological factors impacting on the applicability of a preposition relative to a given landmark at a point in space. However, to create the topological spatial template for a preposition, the angular applicability ratings must be combined with the distance applicability ratings.

This is done by first multiplying[56] the angular applicability rating for each point by its distance applicability The resulting values are then normalised by dividing all the points in the spatial template by the maximum value The normalised values are in the range between *[0 1]* Each of these values is taken to represent the applicability of the spatial template, the higher the value, the more applicable the preposition is at that point relative to the landmark

```
i = 1
While i <= The_Number_Of_Points_Being_Rated
        STRating[i] = AngleApp[i] * DistApp[i]
        If STRating[i] > MaxRating Then
                MaxRating = STRating[i]
        End If
        i = i + 1
End While
i = 1
While i <= The_Number_Of_Points_Being_Rated
        STRating[i] = STRating[i] / MaxRating
        i = i + 1
End While
```

Algorithm 8-9 The algorithm for combining the angular and distance applicability ratings in the spatial template In this algorithm, STRating is the array containing the overall ratings of points in the spatial template, AngleApp is the array containing the calculated angular applicability of the points being rated in the spatial template, and DistApp is the array containing the calculated distance

---

[56] The multiplication function is used, in preference to summation, to combine the angular and distance constraints in the spatial template because multiplication causes all points that have a rating of zero for either of these constraints to have an overall rating of zero In effect it is an AND function that requires all the points that rate > zero in the spatial template to be within both the maximum angle parameter β and maximum distance parameter γ

90° with the distance applicability ratings with Y set to 250 units. The values in this spatial template are normalised to the range of [0...1] – the higher the value assigned to a point, the darker the colour in the image The red square indicates the position of the landmark and the red line delineates the directional constraint of the preposition *above*.

The advantages of the SLI potential field model over previous potential field models are

1. It avoids the problems associated with using the landmark's bounding box centre as the spatial template origin in the viewer-centred frame of reference
2. It models the gradation of the preposition's applicability across a 3-D volume
3. It is able to accommodate different size landmarks by adjusting the maximum angle of deviation β and the maximum range Y of the spatial template

However, as noted in Section 8 4 1 1 footnote 55, there are problems with this model as it currently stands, in particular, there are regions which are occluded by the landmark but are topologically defined as *in front of* the landmark It is proposed here, that these areas can be accommodated by integrating perceptual cues into the framework

8.4 2 SLI Spatial Template Model Viewer-Centred Perceptual Component

Assumption 1 (Section 8 4 Page 235 above), states that perceptual cues impact on the semantics of projective prepositions *in front of* and *behind* in the viewer-centred frame of reference Accordingly, this section describes the perceptual cues which affect the semantics of these prepositions The basic perceptual definitions proposed for *in front of* and *behind* are

A is *in front of* B if object A occludes part of object B
B is *behind* A if object B is partly occluded by object A

264

applicability of the points being rated in the spatial template. Note that STRating[i], AngleApp[i], and DistApp[i] all describe the same point in space.

Figure 8-9 illustrates the shape of the spatial template that Algorithm 8-9 creates. It is important to note that the spatial template can be scaled to accommodate different sizes of landmarks by adjusting the maximum angle for the spatial template β and or the maximum range of the spatial template γ.
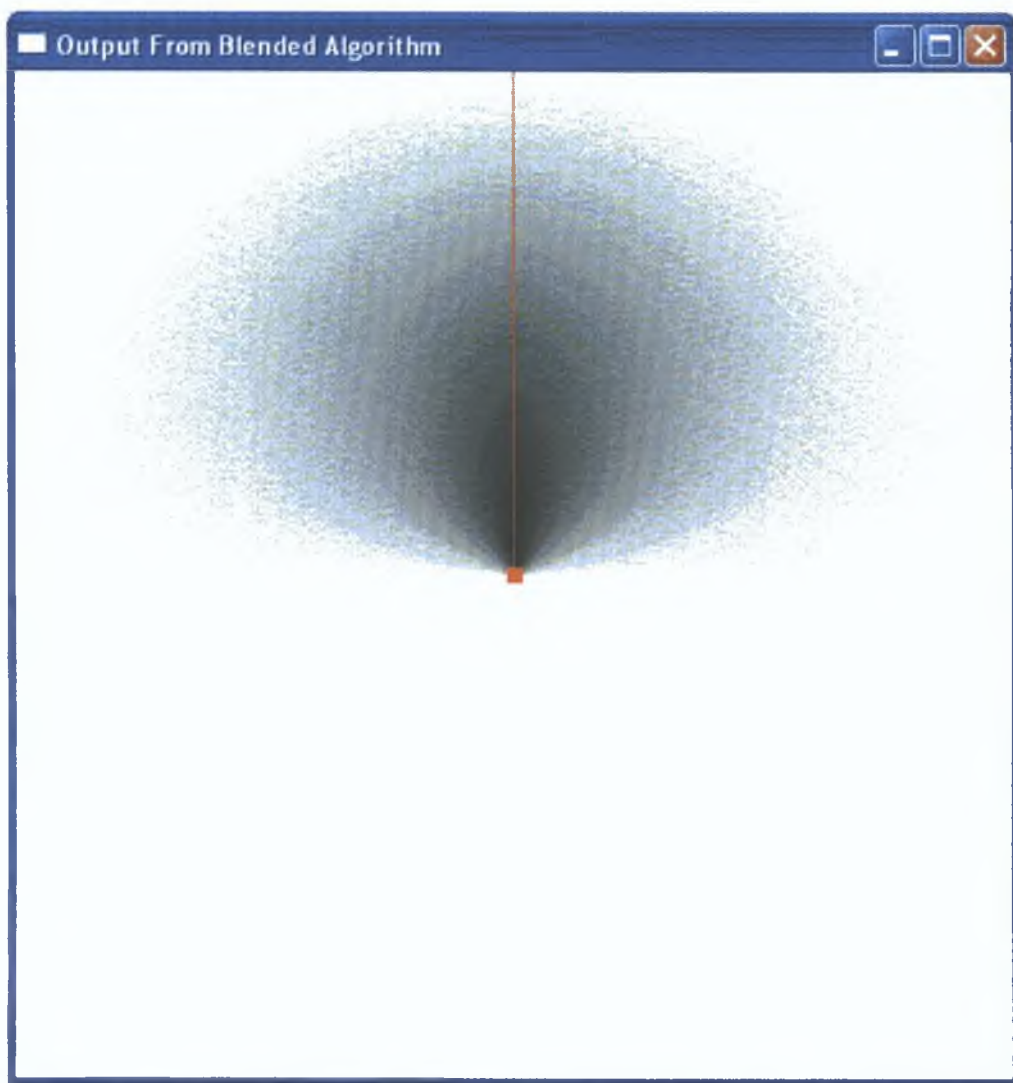
Figure 8-9: Graphical representation of the spatial template that results by combining the angular applicability ratings for the preposition *above* with β set to

Although these definitions allow the possibility of an object being both *in front of* and *behind* another object, this doesn't invalidate them as the object configurations where this occurs are inherently ambiguous. For example, a long horizontal object in a square defined by a building may partially occlude the building and be partially occluded by the building (see Figure 8-10).
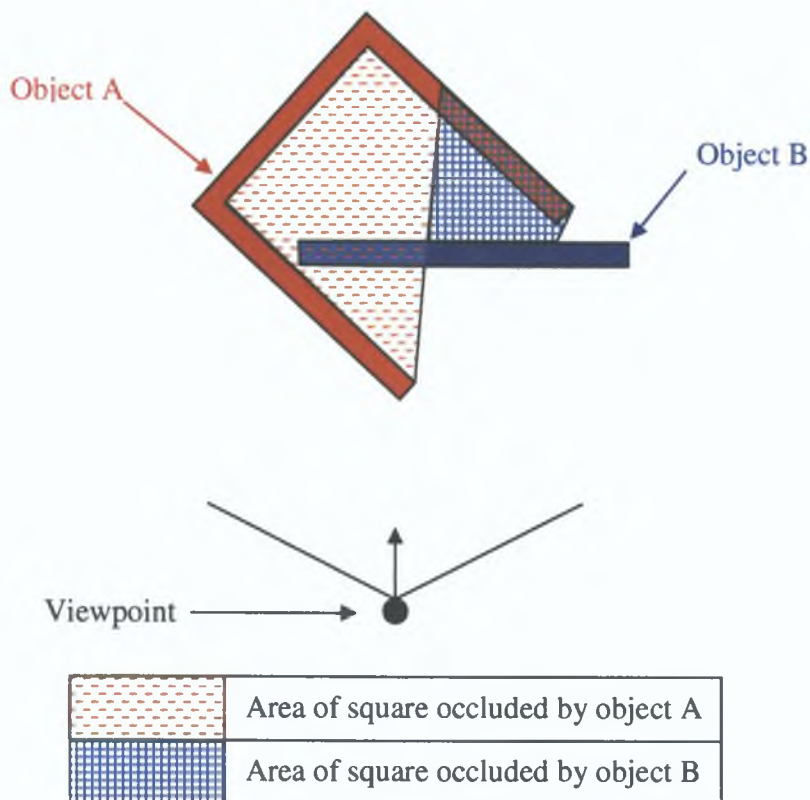
Figure 8-10: A spatial configuration of two objects, both of which occlude and are occluded by the other object.

There are, however, object configurations that can be correctly described using *in front of* or *behind* and which are not covered by these definitions. Figure 8-11 illustrates a situation where object B is *behind* object A even though none of object B is occluded by object A. Conversely, Figure 8-11 also illustrates a situation where object A *is in front of* object B even though object B is not occluded by object A. However, these configurations are accommodated by the topological model developed in Section 8.4.1

above. It is posited here that a model of projective prepositional semantics which integrates the SLI topological spatial template and the SLI perceptual definitions solves the problems that each of these component models are unable to accommodate. In the following section, we show how these topological and perceptual models are integrated to construct the SLI spatial template model in the viewer-centred frame of reference.
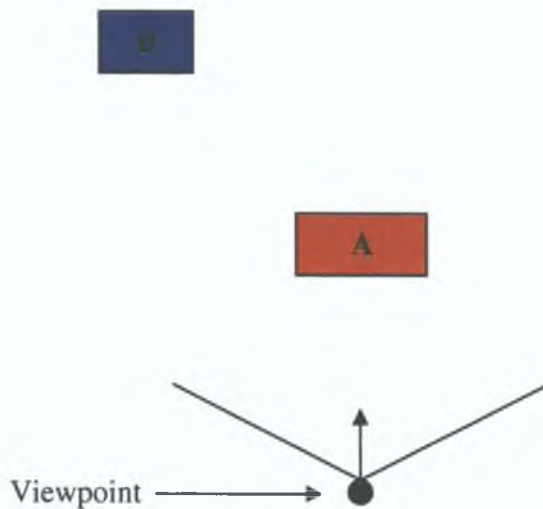
Figure 8-11: A spatial configuration where object B is behind object A without being occluded by object A and conversely object A is in front of object B without occluding any of object B.

8.4.3   SLI Spatial Template Model: Viewer-Centred Integrated Model

In Section 8.4.1 the topological model that forms the basis of the SLI semantic framework was defined. Concluding, it was noted that although the proposed model avoids many of the paradoxical definitions that can occur in systems that schematise the landmark by its bounding box or its centroid, certain configurations of user position and landmark shape can result in regions that are occluded by the landmark being topologically defined as *in front of* the landmark. Figure 8-12 is a modified version of diagram (b) in Figure 8-1. In this figure, the problematic regions are highlighted. In

266

Section 8 4 2, a set of perceptual definitions for the prepositions aligned along the front-back axis in the viewer-centred frame of reference were described However, it was shown that these perceptual definitions were unable to recognise certain valid configurations of objects within the set of relationships described by the prepositions they defined In summary, currently there are two approaches to defining the area of the projective prepositions along the front-back axis in the viewer-centred frame of reference However, both are flawed

1   The possibility of SLI spatial templates model's topological component defining regions that are occluded by the landmark as *in front of* the landmark was highlighted in Section 8 4 1 1 Footnote 55 and Section 8 4 1 2 above (see also Figure 8-12) It should be noted that this weakness in the SLI spatial templates topological component applies to all the previous potential field models (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994)

2   The inability of the SLI spatial template model's perceptual component to recognise valid front-back relationships between a set of objects, where none of the objects occlude other objects, was highlighted as the weakness of this component in Section 8 4 2 above

It is proposed here that, by integrating these two approaches, the topological and perceptual components defined in Sections 8 4 1 and 8 4 2 a semantic model which overcomes the shortcomings of each of these models can be defined This integrated model

1   avoids defining regions that are occluded by the landmark as *in front of* the landmark

2   is able to recognise the valid front-back object configurations which do not involve object occlusion
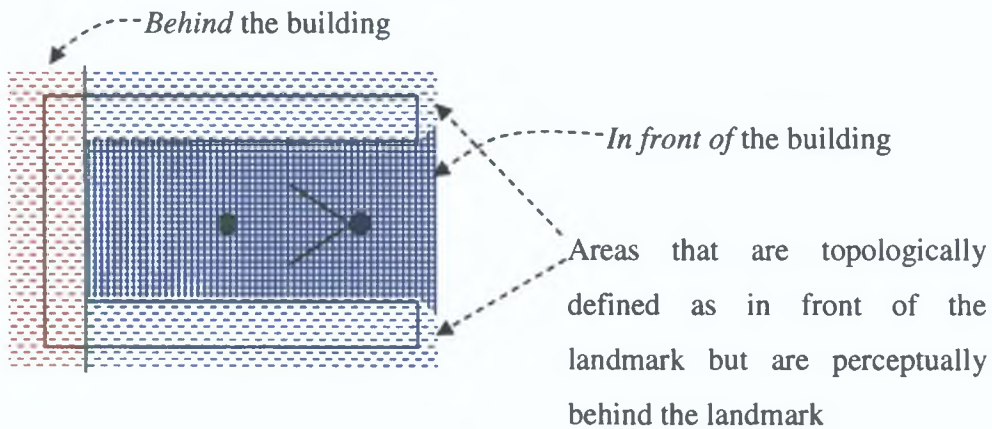
Figure 8-12: A diagram highlighting the regions which are topologically defined as being *in front of* the landmark but are perceptually occluded by the building.

The SLI topological model is integrated with the perceptual definitions by treating the perceptual definitions as privileged rules which override the topological definitions; in short:

1. An object which is topologically *in front of* the landmark but is partly occluded by the landmark is treated as *behind* the landmark.
2. An object which is topologically *behind* the landmark but which partially occludes the landmark is treated as *in front of* the landmark.

Moreover, an object which *occludes / is occluded by* the landmark is deemed to be in the good region and is assigned the maximum applicability within the *in front of-behind* spatial template. Other objects are rated based on the topological definitions. Figure 8-13 illustrates the parsing of space around a convex landmark using the integrated model.
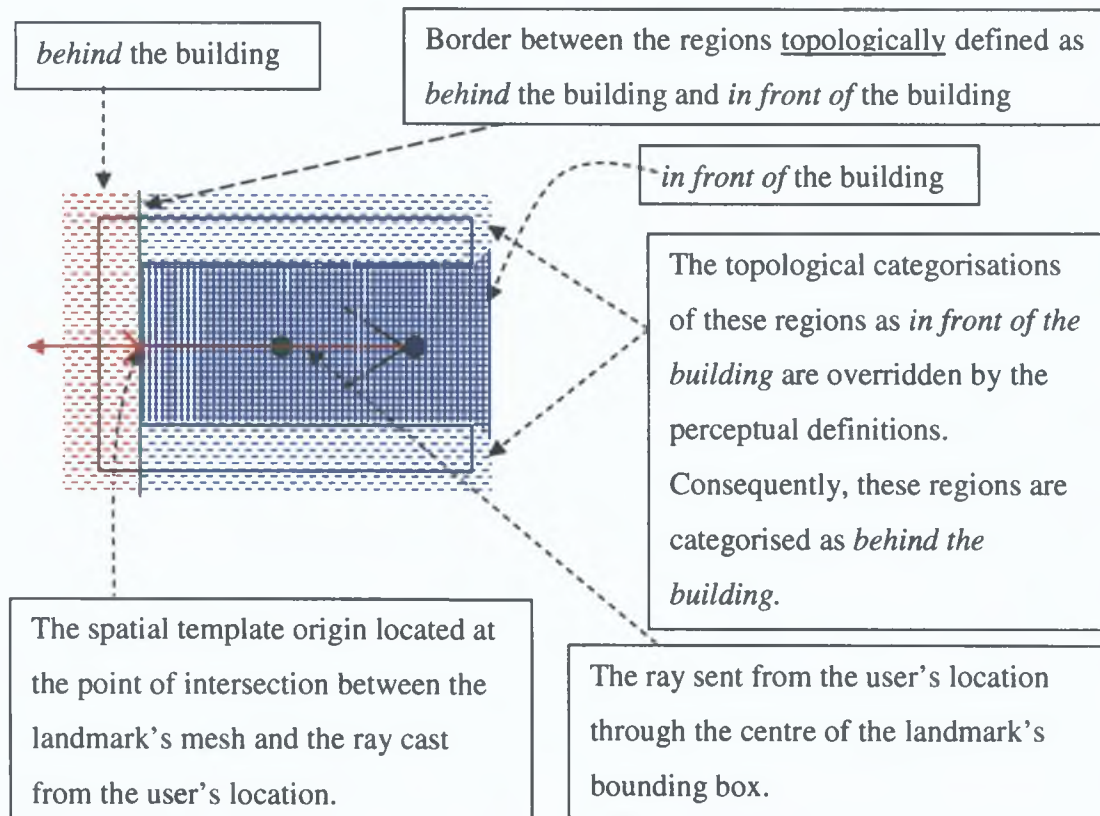
Figure 8-13: Figure illustrating the parsing of space around a landmark along the front/back axis in viewer-centred frame of reference using the integrated semantic model.

On first appearance, it could be argued that integrating the topological and perceptual definitions in this manner solves the problems associated with the schematisation of the landmark by its bounding box centre as the perceptual definitions would override the paradoxically defined regions in the topological model. An implication of this is that the SLI algorithm for locating the spatial template's origin would be redundant. This, however, is not the case. The diagrams in Figure 8-14 illustrate the differences in the parsing of space between a topological model centred on the landmark's bounding box centre that is integrated with the SLI perceptual definitions and the integrated SLI spatial template model that uses the ray casting algorithm to locate the spatial template's origin. Diagram (a) of Figure 8-14 names the different components in the example spatial configuration. For this example, assume that the landmark is already

in the system's model of the user's visual memory, i e , the user has seen the landmark at some point in their interaction with the environment Examples of the type of real-world or simulated world scenarios that could result in the particular configuration of objects in these diagrams and where the user is aware of the landmark are

1   The landmark is a building which the user has just exited from
2   The landmark is a building which the user has just walked up to and then turned away from

Diagram (b), in Figure 8-14, illustrates the topological parsing of space using the landmark's bounding box centre note that the candidate trajector object is wrongly defined as being *behind the landmark* in this topological model Diagram (c) depicts the parsing of space that results from integrating the SLI perceptual definitions into the topological model illustrated in diagram (b) note that the trajector is still wrongly defined as being *behind the landmark* Diagrams (d) and (e) illustrate the different stages in the ray casting algorithm in diagram (d) a ray is cast from the user's location in the direction of the landmark's bounding box centre This ray does not intersect the landmark's mesh As a result, a second ray is sent in the opposite direction to the first, the path of the second ray is illustrated in diagram (e) The second ray intersects the landmark's mesh This intersection point is marked by the red X in diagram (e) Moreover, it is taken as the location of the spatial template origin for the topological model that is depicted in diagram (f) However, the topological model of diagram (f) defines some regions that are occluded by the landmark as being *in front of the landmark* The final diagram in Figure 8-14, diagram (g), illustrates the parsing of space around the landmark as defined by the integrated SLI spatial template model It is worth noting that parsing of space in diagram (b) is equivalent to the parsing of space by the previous potential field models that schematised the landmark by its bounding box centroid and ignored perceptual factors (Yamada 1993, Gapp 1994a, Olivier and Tsuji 1994) Comparing diagram (b) and diagram (g) highlights the differences between these models and the SLI spatial template model

270

Landmark

User's location

Landmark's bounding box

User's view volume

Centre of landmark's bounding box

Candidate trajector

(a)

(b)

(c)

(d)

(e)

(f)

(g)

	in front of the landmark
	behind the landmark

271
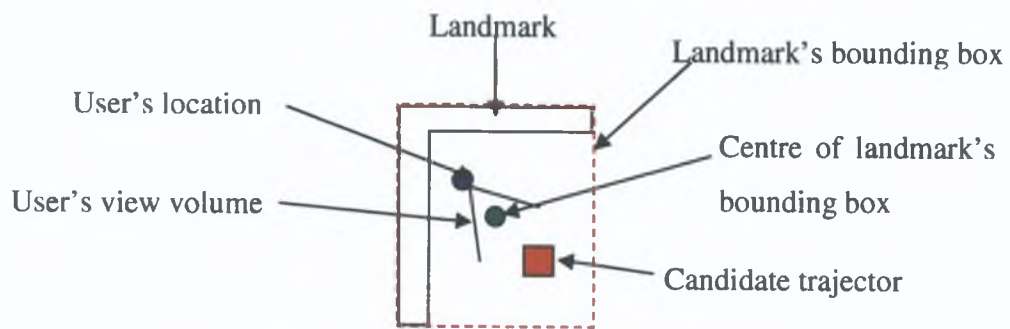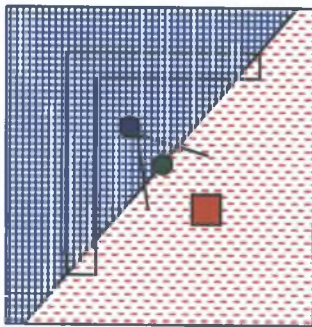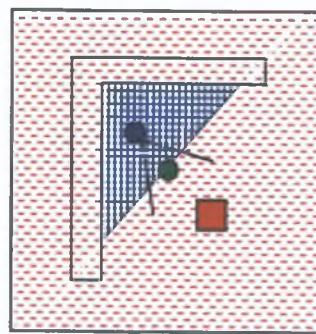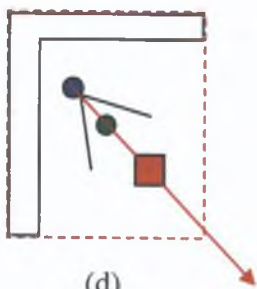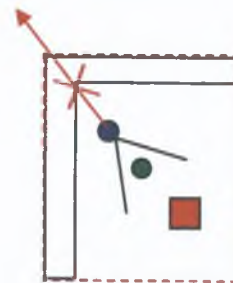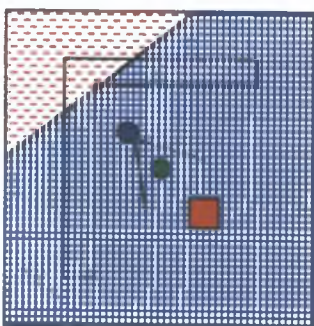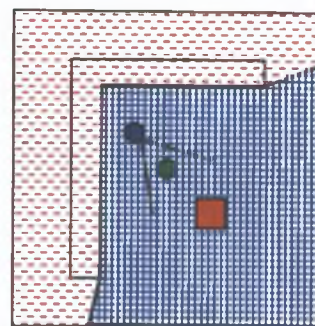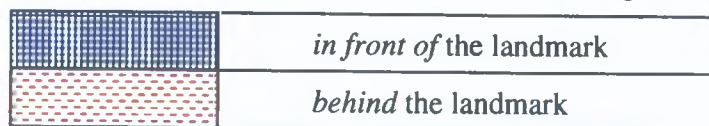
Figure 8-14 Diagrams illustrating the differences in the parsing of space between a topological model centred on the landmark's bounding box centre that is integrated with the SLI perceptual definitions and the integrated SLI spatial template model which uses the ray casting algorithm to locate the spatial template's origin Diagram (a) names the different components in the example spatial configuration Diagram (b) illustrates the topological parsing of space using the landmark's bounding box centre. Diagram (c) depicts the parsing of space that results from integrating the SLI perceptual definitions into the topological model in diagram (b). Diagrams (d) and (e) illustrate the different stages in the SLI ray casting algorithm The red X in diagram (e) is taken as the location of the spatial template's origin for the topological model which is depicted in diagram (f). Diagram (g) illustrates the parsing of space around the landmark as defined using the integrated SLI spatial template model

8 4.4   SLI Spatial Template Model Summary

In Section 8 4 1 2, a set of algorithms and equations were proposed which address the issues that impact on modelling the topological considerations of the semantics of projective prepositions Algorithm 8-10 defines how these different components are combined to construct a projective preposition's spatial template in the intrinsic frame of reference

To construct the potential field model of a projective preposition in the viewer-centred frame of reference, Algorithm 8-10 is combined with Algorithm 8-4 (the SLI algorithm for defining the spatial template origin), Algorithm 8-5 (the SLI algorithm for calculating the vectors describing the canonical direction of the horizontally aligned projective prepositions), and with the perceptual definitions developed in Section 8 4 3 Algorithm 8-11 defines how these algorithms are combined to construct the spatial template of a projective preposition in the viewer-centred frame of reference Integrating these algorithms results in a topological framework for the viewer-centred frame of

272

reference that avoids the problems inherent in systems that use the landmark's bounding box or its centroid to represent the landmark in this frame of reference

Finally, following Algorithm 8-3, if the intrinsic and viewer-centred frames of reference are dissociated, the spatial templates for each frame of reference should be adjusted to accommodate the bias in frame of reference use along the horizontal and vertical planes and then be amalgamated Algorithm 8-12 defines the amalgamation process and Figure 8-15 gives a graphical representation of the spatial template that results from the amalgamation process

Points[] = the array containing the set of points to be rated

Spatial_Template_Origin = Predefined

Canonical_Direction_Of_Preposition = Predefined

Let i = 1

While i < The_Number_Of_Points_To_Be_Rated_In_Spatial_Template

      Use Algorithm 8-6 to convert Point[i] into the local coordinate system centred on the spatial template origin Let LocPoint[i] be the result of this process

      Use Equation 8 to compute the angular deviation of LocPoint[i] from the vector Canonical_Direction_Of_Preposition Let AngleApp[i] be the result of this process

      Use Equation 10 to compute the distance of LocPoint[i] from the point Spatial_Template_Origin Let DistApp[i] be the result of this process

      Let i = i + 1

End While

Set β = 90°

Use Algorithm 8-7 to normalise values in the AngleApp[] array

Set Y = the maximum value in the DistApp[] array

Use Algorithm 8-8 to normalise values in the DistApp[] array

Use Algorithm 8-9 to combine the values in AngleApp[] and DistApp[]

Algorithm 8-10 The algorithm for calculating the ratings of a set of points in a preposition's spatial template in the intrinsic frame of reference.

```
Points[] = the array containing the set of points to be rated

Occludes() = function that returns True if parameter 1 occludes parameter 2

PrepStr = the string containing the preposition used in the locative expression


Use Algorithm 8-4 to locate the Spatial_Template_Origin

Use Algorithm 8-1 to calculate the Canonical_Direction_Of_Preposition

Let i = 1

While i < The_Number_Of_Points_To_Be_Rated_In_Spatial_Template

        If PrepStr == "in front of" Then

                If Occludes(Point[i], Landmark) Then

                        AngleApp[i] = 1 And DistApp[i] = 1

                End If

        Else If PrepStr == "behind" Then

                If Occludes(Landmark, Point[i]) Then

                        AngleApp[i] = 1 And DistApp[i] = 1

                End If

        Else

                Use Algorithm 8-6 to convert Point[i] into the local coordinate system centred on

                the spatial template origin Let LocPoint[i] be the result of this process

                Use Equation 8 to compute the angular deviation of LocPoint[i] from the vector

                Canonical_Direction_Of_Preposition Let AngleApp[i] be the result of this process

                Use Equation 10 to compute the distance of LocPoint[i] from the point

                Spatial_Template_Origin Let DistApp[i] be the result of this process

        End If

        Let i = i + 1

End While

Set β = 90°

Use Algorithm 8-7 to normalise values in the AngleApp[] array

Set Y = the maximum value in the DistApp[] array

Use Algorithm 8-8 to normalise values in the DistApp[] array

Use Algorithm 8-9 to combine the values in AngleApp[] and DistApp[]
```

Algorithm 8-11. The algorithm for calculating the ratings of a set of points in a projective preposition's spatial template in the viewer-centred frame of reference.

Points[] = the array containing the set of points to be rated

Use Algorithm 8-10 to rate each of the points in Points[] in the intrinsic spatial template. Let IntrinsicRatings[] contain the result of this process.

Use Algorithm 8-11 to rate each of the points in Points[] in the viewer-centred spatial template. Let ViewerCentredRatings[] contain the result of this process.

Adjust the rating that resulted from step 1 and 2 to reflect the bias in frame of reference use.

//Amalgamate the spatial templates

Let y = 1

Let MaxRating = 0

While y <= |Points[]|

    AmalgamatedRatings[y] = IntrinsicRatings[y] + ViewerCentredRatings[y]

    If AmalgamatedRatings[y] > MaxRating Then

        MaxRating = AmalgamatedRatings[y]

    End If

    y = y + 1

End While

//Normalise the results of the amalgamation process

Let y = 1

While y <= |Points[]|

    AmalgamatedRatings[y] = AmalgamatedRatings[y] / MaxRating

    y = y + 1

End While

Algorithm 8-12: The algorithm used to amalgamate the ratings of a set of points in a projective preposition's spatial template potential field model that are constructed when the intrinsic and viewer-centred frames of reference are dissociated.
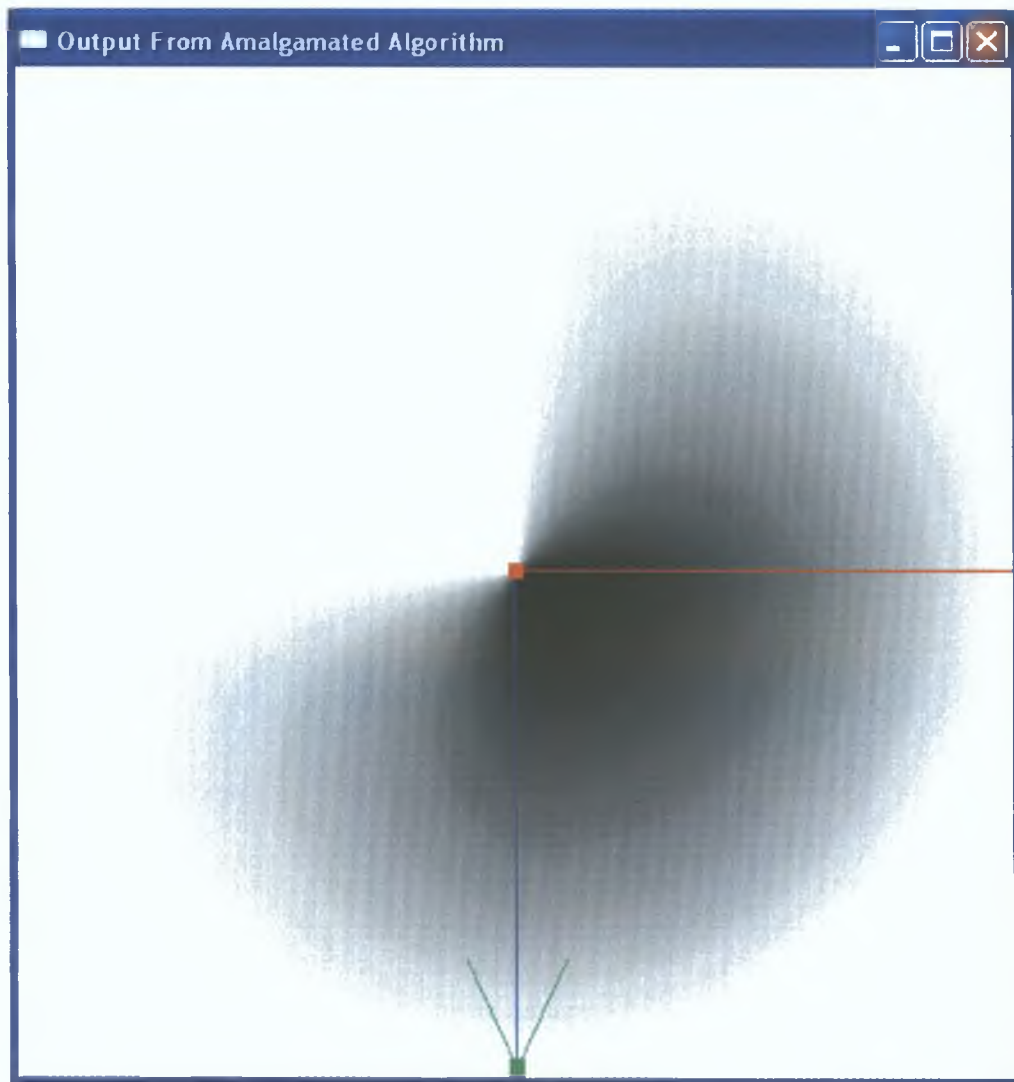
Figure 8-15: A graphical representation of the spatial template model that is constructed by amalgamating an intrinsic spatial template and a viewer-centred spatial template. This image represents a bird's eye view of a spatial configuration. The red box in the image represents the landmark and the red line delineates the canonical direction of the preposition *in front of* in the intrinsic frame of reference. The green box represents the viewer's location and the green lines extending away from the viewer delineate the view volume. The blue line running from the landmark to the viewer delineates the canonical direction of the preposition *in front of* in the viewer-centred frame of reference. The maximum angle parameter β was set to 90° and the maximum distance ratings γ was set to 250 units in both of the component spatial templates. The values in the amalgamated spatial template are

277

normalised to the range of [0.. 1] – the higher the value assigned to a point, the darker the colour in the image.

In summary, the advantages of the SLI spatial template model over previous models are that

1   It avoids the problems associated with using the landmark's bounding box centre as the spatial template origin in the viewer-centred frame of reference

2   It models the gradation of the preposition's applicability across a 3-D volume

3   It is able to accommodate different size landmarks by adjusting the maximum angle of deviation β and the maximum range γ of the spatial template

4   It accommodates the impact of reference frame selection on the construction of a spatial template model

5   It accommodates the perceptual cue of object occlusion

8.5   Selecting the Trajector

The semantic framework proposed in Section 8 4 above allows one to model the applicability of a preposition across a region Using this model, a projective locative expression can be resolved by selecting a referent from the set of candidate trajectors based on their fitness within the model However, there are several issues that impact on this process

Firstly, the list of candidate trajectors should not contain objects which the user has not seen (see Section 2 3 5 2) Secondly, the representation of the candidate objects in the system impacts on their fitness within the model (see Section 2 3 5 1) Thirdly, how does the system handle situations where two candidate objects are equally rated in the model?

The first of these issues is particularly relevant for systems that allow their interpretive module full knowledge of the world SHRDLU (Winograd 1973), CITYTOUR (Andre *et al* 1986, Andre *et al* 1987), CSR-3-D (Gapp 1994a), SPRINT (Yamada 1993) WIP (Olivier *et al* 1994, Olivier and Tsuji 1994), Situated Artificial

278

Communicator (Socher and Naeve 1996, Socher *et al* 1996, Vorwerg *et al* 1997, Fuhr *et al* 1998), Virtual Director (Mukerjee *et al* 2000), and CommandTalk (Dowding *et al* 1999, Stent *et al* 1999, Goldwater *et al* 2000) In these systems, the set of candidate trajectors will contain all the objects which are topologically valid irrespective of whether or not the user is aware of their existence However, in the SLI framework the set of candidate trajectors is drawn from the context model (which will be developed in Chapter 9) that is built from the output of the perceptual saliency model developed in Chapter 7 and the discourse history This architecture ensures that only objects that the user has seen or which are currently in the view volume will be included in the list of candidate trajectors Moreover, the integration of perceptual cues into the SLI semantic models of prepositions allows the system to correctly process candidate trajectors which are rated by the topological model as having a high applicability even though perceptual cues indicate that they should not be considered as candidate referents

The second issue pertains to how the system represents the candidate trajectors This is important because it impacts on the applicability ratings assigned to candidate trajectors by the model Section 2 4 5 1 illustrated the problems with using the object's centroid to represent its location in the spatial template Another approach is to take the value of the average applicability for all the vertices in the candidate trajector's geometric mesh However, the distribution of vertices across an object's mesh may not be uniform across its area this can result in the average value being skewed Here a third alternative is suggested use the vertex in the object's mesh which has the highest applicability rating to represent the object This ensures that the candidate trajector with a point at the highest applicability will be selected It is this approach that is implemented in the SLI system

Finally, how does the framework adjudicate between two candidate trajectors which have the same applicability ratings? In situations where there are two or more candidate trajectors which have the same applicability ratings in the spatial template model the candidate with the highest visual salience is selected If the visual salience ratings do not distinguish between the candidates, the system asks the user for clarification Algorithm 8-13 defines the process used to select a referent for a locative expression This algorithm depends on the other stages in the process of resolving a

locative Algorithm 8-14 defines how Algorithm 8-13 is used in the general algorithm for resolving locatives

Let x = 1 And MaxRating = 0

While x <= |Trajectors[]|

    If MaxRating > Trajectors[x] Rating Then

        Delete Trajectors[x] From Trajectors[] And x = x - 1

    Else If MaxRating < Trajectors[x] Rating Then

        Delete Trajectors[1   x-1] From Trajectors[]

        MaxRating = Trajectors[x] Rating And x = 1

    End If

    x = x + 1

End While

If |Trajectors[]| == 1 Then

    Referent = Trajectors[1]

Else If |Trajector[]| > 1 Then

    Select the element in Trajectors[] which has the highest visual salience on condition that the difference between its salience and the salience ascribed to the other elements remaining in Trajectors[] is greater than a predefined confidence interval

    If the saliency requirement is not met treat the reference as ambiguous and ask the user for clarification

End If

Algorithm 8-13 The algorithm for selecting a referent from the set of candidate trajectors In this algorithm, Trajectors[] represents the array of objects which fulfil the linguistic restrictions of the referring expression on the trajector This array is supplied by the SLI discourse model, which will be developed in Chapter 9 and Trajectors[x].Rating = the maximum rating in the preposition's spatial template assigned to a vertex in Trajectors[x]'s 3-D mesh.

280

8.6   Chapter Summary

This chapter began by listing the four main stages to interpreting a projective preposition:

1. Identify the landmark.
2. Select a frame of reference and superimpose it on the landmark.
3. Define the area of search for the trajector as defined by the spatial template associated with the preposition.
4. Identify the primary trajector within the search area.

In each stage, solutions were proposed. Section 8.2 focused on resolving the landmark reference as a case of general reference (see Chapter 9).

In Section 8.3 an algorithm was proposed to handle the issue of frames of reference: Algorithm 8-3. This algorithm draws together the results of several psycholinguistic experiments (Carlson-Radvansky and Irwin 1994; Logan and Sadler 1996; Carlson-Radvansky and Logan 1997; Taylor *et al.* 2000). The major point of note in this algorithm is the construction of an amalgamated spatial template by a weighted combination of competing spatial templates. It is important to note that this approach impacts on the construction of the spatial templates. How Algorithm 8-3 is integrated with Algorithm 8-10 and Algorithm 8-11 is illustrated below.

In Section 8.4, a novel spatial template model of projective prepositions that defines prepositions in terms of perceptual and topological axioms was proposed. One of the most important aspects of this model is the dynamic location of the spatial template's origin in the viewer-centred frame of reference based on the user's location relative to the landmark (see Section 8.4.1.1 Algorithm 8-4). This approach allows the algorithm to take into account the user's perception of the landmark which in turn avoids the paradoxical parsing of space that can arise by using a predefined spatial template origin in the viewer-centred frame of reference irrespective of the user's position. In the intrinsic frame of reference, the location of the spatial template origin for a given landmark is known through a priori knowledge. Another important feature of the SLI spatial template model

is that the model uses a scalable potential field model that works in three dimensions and models both the angular deviation of a point from the canonical direction of the preposition and the distance of a point from the spatial template's origin This potential field model defines the semantics of the prepositions in the intrinsic frame of reference However, unlike previous models, it was assumed that the spatial template associated with the prepositions *in front of* and *behind* in the viewer-centred frame of reference differ from their intrinsic counterparts due to the impact of object occlusion Accordingly, a model of spatial semantics for these prepositions was proposed for the viewer-centred frame of reference which integrates this perceptual phenomenon with the topological model used for the intrinsic frame of reference Algorithm 8-10 and Algorithm 8-11 define the construction of a potential field model of a projective preposition's spatial template in the intrinsic and viewer-centred frames of reference, respectively Section 8 4 3 illustrates how the SLI semantic model is able to define the regions surrounding landmarks with complex geometries in a consistent manner Finally, in Section 8 4 4, a solution to the issues attending the representation of the candidate trajectors and the selection of a referent from the set of candidate trajectors was proposed

At the beginning of this chapter, it was noted that the interdependence of the different stages in the interpretation of a locative expression complicated the development of a unified general approach to the interpretation process However, having developed independent algorithmic solutions for each of the main stages in the process, an algorithm for resolving locative expressions will now be given, Algorithm 8-14 A prerequisite of using this interpretive algorithm for projective locative expressions is a discourse model that will resolve the landmark reference and supply a list of candidate trajectors for the algorithm to rate In Chapter 9, the SLI discourse framework is developed

Let Landmark = the landmark reference resolved using the SLI general model of reference resolution, which will be developed in Chapter 9

Let Trajectors[] = the array of objects which fulfil the linguistic restrictions of the referring expression This array is supplied by the SLI discourse model, which will be developed in Chapter 9

Let Mesh[x] = the set of points in Trajectors[x]'s 3-D mesh

If (Landmark Has Intrinsic Frame Of Reference) AND (Viewer-Centred And Intrinsic Frames of Reference Are Dissociated) Then

      Let x = 1

      While x <= |Trajectors[]|

            Use Algorithm 8-12 to rate the points in Mesh[x]

            Trajector[x] Ratings = the max amalgamated rating template assigned to a point in Mesh[x]

            x = x + 1

      End While

Else

      Let x = 1

      While x <= |Trajectors[]|

            Use Algorithm 8-11 to rate the points in Mesh[x]

            Trajector[x] Rating = the maximum rating assigned to a point in

      Mesh[x]

            x = x + 1

      End While

End If

Use Algorithm 8-13 to select a referent from the set Trajectors[]

Algorithm 8-14: The SLI algorithm for interpreting a locative expression

9 Integrating Visual and Linguistic Discourse Context For Reference Resolution in Simulated 3-D Environments

9.1 Introduction

In Section 2 4, it was noted that a linguistic context model is needed because often a user's commands can only be understood by considering them as part of an ongoing dialogue In this section, the SLI discourse framework is developed The approach adopted in designing this model was inspired by Langacker's (1987, 1991b, 1994) Cognitive Grammar (see Section 3 2), and Salmon-Alt and Romary's (2001) reference resolution framework (see Section 4 4)

Following Langacker (1987, 1991b, 1994) and Salmon-Alt and Romary (2001), it is assumed that the process of resolving a referential expression is achieved by accessing and restructuring cognitive domains or reference domains Moreover, these reference domains are not limited linguistic constructs but rather conceptual entities based on different knowledge sources linguistic, encyclopaedic, and visual Furthermore, many reference domains may figure in the semantics of a given expression In this approach, the role of grammar is to define patterns for combining simpler reference domains into more complex ones, which may then be used to interpret complex expressions Finally, a linguistic expression evokes one of a set of possible images on the conceptual domain it draws upon These are the fundamental assumptions underpinning the interpretive approach of this thesis Adopting these as guiding principles results in the division of the SLI discourse framework into three sections

1 the context model which supplies the reference domains,

2 an interpretive process which accesses and restructures these domains,

3 a constructional schema which integrates existing domains to create more complex domains

284

The following sections contain a description of the SLI discourse framework, beginning with a description of the structure of the SLI reference domains (see Section 9 2), the basic unit of the SLI context model They function as local context structure, and are analogous to the concept of a cognitive domain in Cognitive Grammar (Langacker 1987, Langacker 1991b, Langacker 1994) In Section 9 3, the structure of the SLI context model is described Section 9 3 3 summarises the description of the SLI context model structure and presents an overview of how its components are created and interact with the other components in the SLI framework Section 9 4 describes the processes used to interpret nominal expressions Section 9 5 focuses on how the framework models the semantics of relational expressions How the discourse framework functions is illustrated in Section 9 6 Section 9 7 describes the overlap and differences between the SLI framework and its predecessors

9.2   The SLI Reference Domains

The basic units of the SLI context model are **reference domains** These reference domains function as local context structures, similar to the concept of a cognitive domain in Cognitive Grammar (Langacker 1987, Langacker 1991b, Langacker 1994) Internally, each reference domain consists of a domain name, a profiled elements list, a TYPE partition, and zero or more basic partitions Figure 9-1 illustrates the internal structure of a reference domain in the SLI discourse model

```
┌─────────────────────────────────┐
│  ┌──────────────────────────┐   │
│  │ Domain Name              │   │
│  └──────────────────────────┘   │
│  ┌──────────────────────────┐   │
│  │ Profiled Element List    │   │
│  └──────────────────────────┘   │
│  ┌──────────────────────────┐   │
│  │ TYPE Partition           │   │
│  └──────────────────────────┘   │
│  ┌──────────────────────────┐   │
│  │ Partition 1              │   │
│  └──────────────────────────┘   │
│  ┌──────────────────────────┐   │
│  │ Partition 2              │   │
│  └──────────────────────────┘   │
│  ┌──────────────────────────┐   │
│  │ ...                      │   │
│  └──────────────────────────┘   │
│  ┌──────────────────────────┐   │
│  │ Partition n              │   │
│  └──────────────────────────┘   │
└─────────────────────────────────┘
```
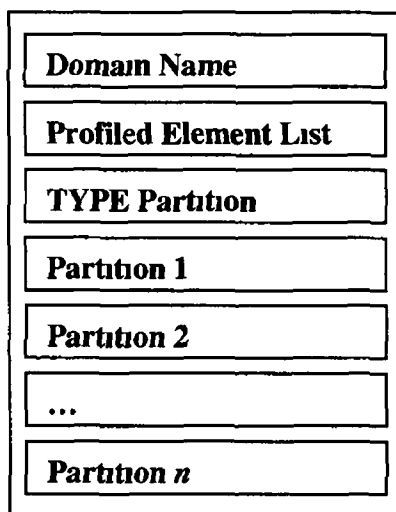
Figure 9-1: The internal structure of a reference domain in the SLI discourse framework.

9 2.1   Domain Name

The SLI reference domains are named after the type of object they contain, in the current implementation of the SLI framework the type information is extracted from the world model For example, if a domain describes a set of houses it is named *house*, likewise if it describes a set of trees it is named *tree* and if it describes a set of objects of different types it is set to a generic type *thing*

9.2 2   Partitions

The SLI reference domains contain one TYPE partition and a set of zero or more basic partitions The function of these partitions, both TYPE and basic, is to predict the different ways that a user may refer to an object in the domain These predictions may be based on previous discourse information, the physical attributes of the objects currently in the domain, or conceptual knowledge of these objects The partitions are comprised of a differentiation criterion and an element list The **differentiation criterion** of a partition

286

is the object attribute that distinguishes the elements of a partition from the elements of the domain that are excluded from the partition Examples of typical differentiation criterion values in the SLI scenarios are *house, tree, thing* for TYPE partitions, and *red, green, blue, tall, wide,* etc for basic partitions The partition's element list is the data structure where the partition elements are stored The **elements** of a partition in a particular reference domain represent objects in the 3-D simulation that are of the correct type for the reference domain and that have the property specified by the partition's differentiation criterion For example, the set elements in a partition, whose differentiation criterion is *red* and whose domain is called *house,* would consist of references to *red houses* in the 3-D simulation Figure 9-2 illustrates the internal structure of a partition in an SLI reference domain
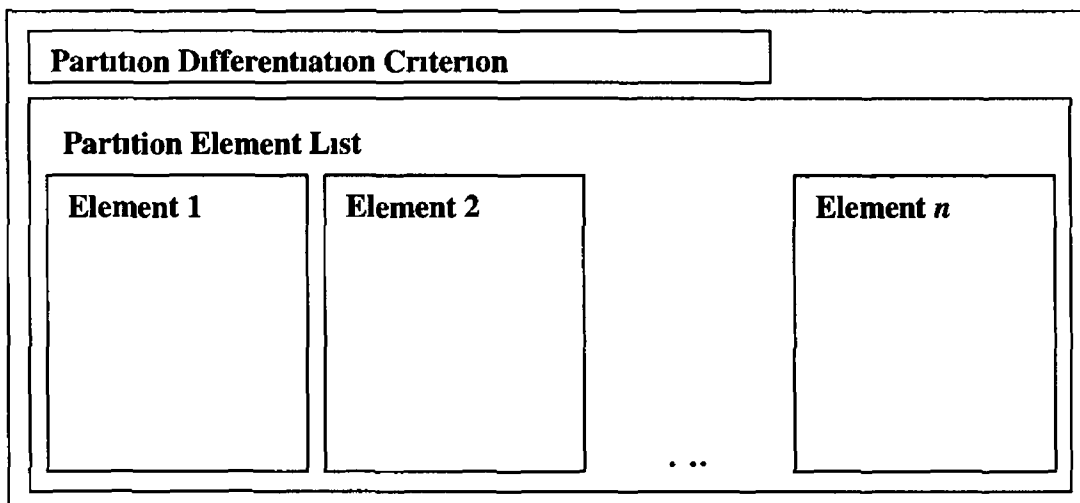
Figure 9-2. The internal structure of a partition in an SLI reference domain

Each of a partition's elements has two components a pointer to the description of an object in the simulated environment and the visual saliency rating of the object that the pointer describes Figure 9-3 illustrates the internal structure of a partition's element and how an element's components relate to the other modules in the SLI system

287
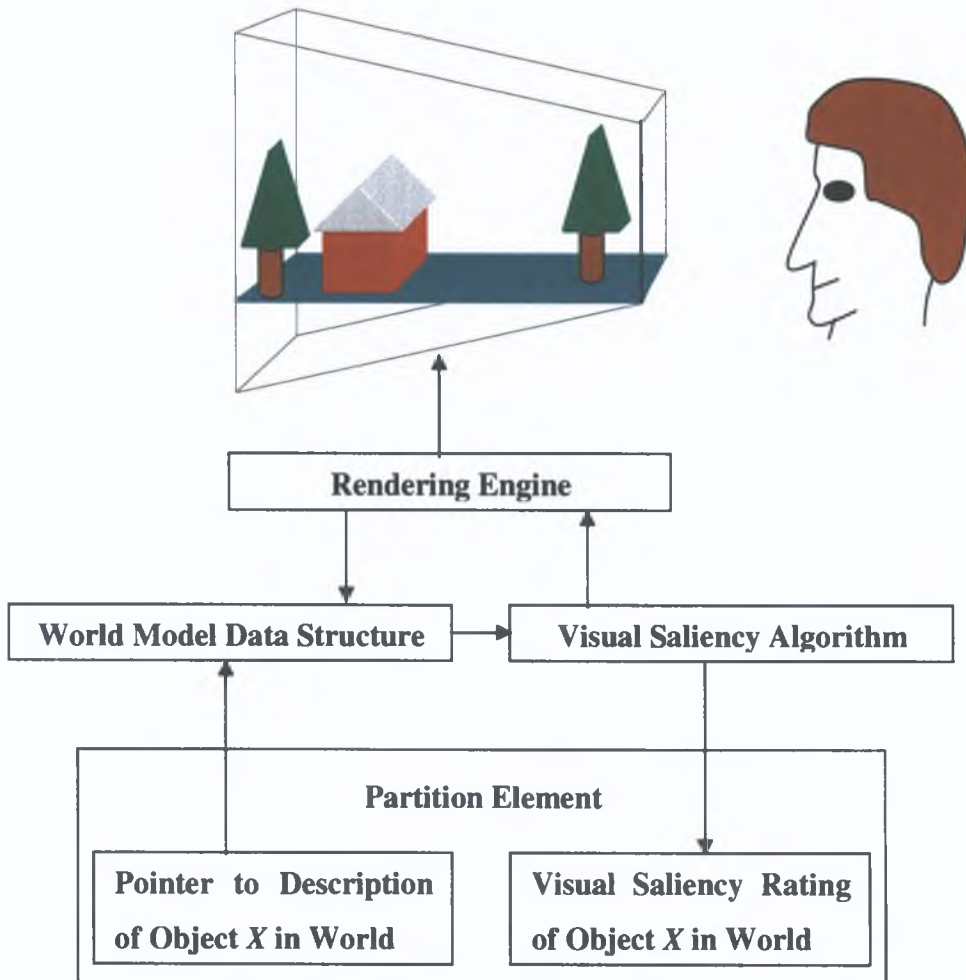
Figure 9-3: The internal structure of a partition's element in the SLI discourse framework and how it relates to the other components in the SLI system.

The element's saliency component is used to order the elements within the partition. This ordering of the elements is a key component within the proposed model. An important point in this context is that the partitions use a last-in-first-out-access policy; i.e., the partitions are implemented using stacks. The default ordering process is to insert elements into a partition in an ascending order based on their salience. This results in the element with the highest salience being inserted at the head of the list; i.e., the first access location within the partition. This organisation reflects one of the fundamental assumptions underlying the interpretive approach of this work; that is, all other factors being equal, objects which have a higher visual salience are more likely to be the

referents of a referring expression than objects which have a lower visual salience Figure 9-4 illustrates the ordering of elements in a partition's element list based on salience
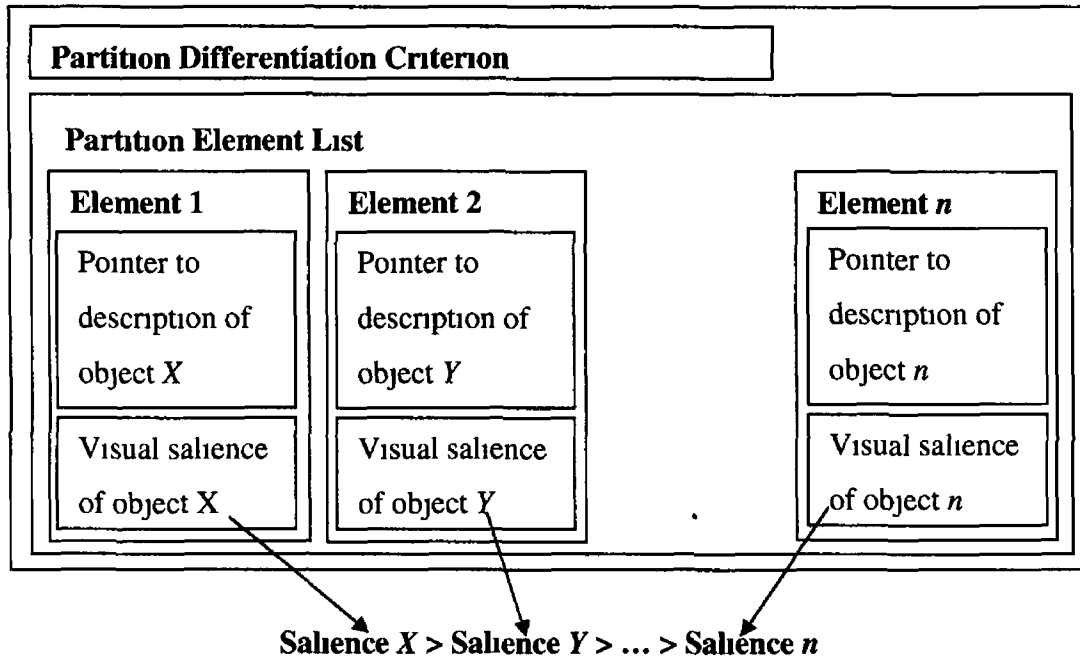
Figure 9-4 Figure illustrating the ordering of elements based on salience in a Partition's Element List.

The saliency based insertion ordering is used for partitions which describe object type and colour For other partitions that describe things such as object size or location, elements are inserted in an ascending order based on their fitness with respect to the partition's criterion In situations where two elements within a partition are equal with respect to the differentiation criterion, the element with the lower saliency rating is inserted first Figure 9-5 illustrates the ordering of elements in a partition modelling a quantifiable property of the elements in the domain In this example, the property being modelled is height Consequently, the taller the element the lower its index in the list Where two or more elements are equal in height they are ordered based on their visual salience, the element with the highest salience is inserted first

289

height X > **height** Y > ... > **height** n

or

if **height** *Element i* = **height** *Element i+1 then*

Salience *Element i* > **Salience** *Element i+1*
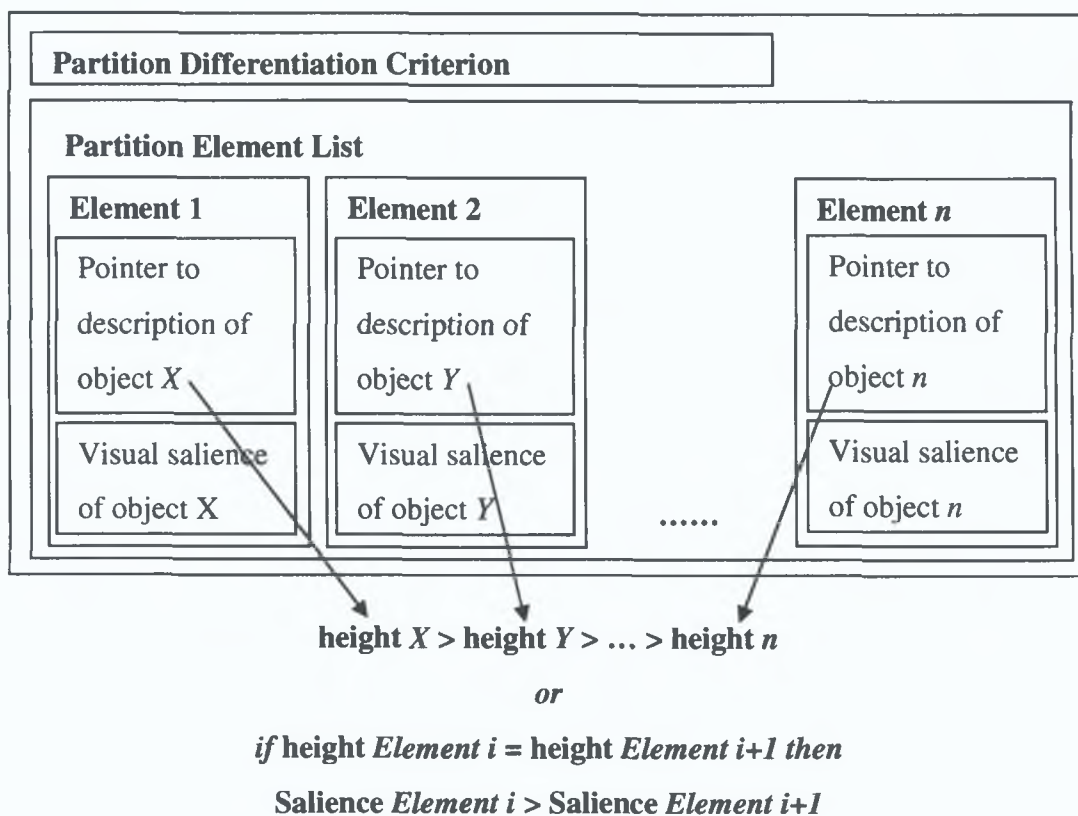
Figure 9-5: Figure illustrating the ordering of elements in a partition, modelling a quantifiable property; in this instance height. Elements are ordered firstly by their fitness with respect to the property specified in the partition's differentiation criterion; in this instance the taller an element is the lower its index in the list. Where two or more elements have an equal fitness with respect to the partition's differentiation criterion, they are inserted into the list based on their visual salience scores. The higher the element's visual salience, the lower its index in the list.

Each domain has at least one partition, called the **TYPE partition**, whose differentiation criterion is set to the domain's type. This partition lists all the elements in the domain apart from the profiled element(s); other partitions – **basic partitions** – may be added to a domain based on the physical characteristics of the objects in the domain or discourse information.

Figure 9-6 illustrates the structure of a reference domain in the SLI system. This domain models the three houses in the accompanying scene. Accordingly, the domain

name is *house* The differentiation criterion of the TYPE partition is also set to *house* As there are no profiled elements in this domain, the domain's TYPE partition lists all the elements in the domain *house8, house6* and *house7* These elements are ordered in the TYPE partition based on their visual salience *house8* has the maximum normalised visual salience (1 0000) As a result, *house8* is at the head of the TYPE partition's element list Of the remaining two elements, *house6* has the higher visual salience (0 5348) It is next in the list *house7* has the lowest visual saliency (0 3282) Consequently, it is stored in the last position in the TYPE partition's element list There are three basic partitions that model qualitative attributes These are the *red, blue,* and *green* partitions The default ordering in these partitions is to order their elements based on their visual salience However, in this instance each of these partitions has only one element, because for each of these partitions there is only one object in the scene that fulfils the partition's differentiation criterion There are six partitions that model quantitative attributes *tall, short, wide, narrow, deep, shallow* These partitions order their elements primarily based on the fitness relative to the element attribute the partition is modelling, i e , the partition's differentiation criterion In instances where two or more of the partition's elements score equally relative to the differentiation criterion, they are ordered based on salience The *tall* partition illustrates the ordering of elements in these quantitative partitions In the *tall* partition, *house6* is inserted before *house8*, even though *house8* has a higher visual salience This is because *house6* is taller and, consequently, scores higher relative to the partition's differentiation criterion *tall house8* and *house7* are the same height However, *house8* has a higher visual salience and, consequently, is inserted ahead of *house7* in the tall partition's element list
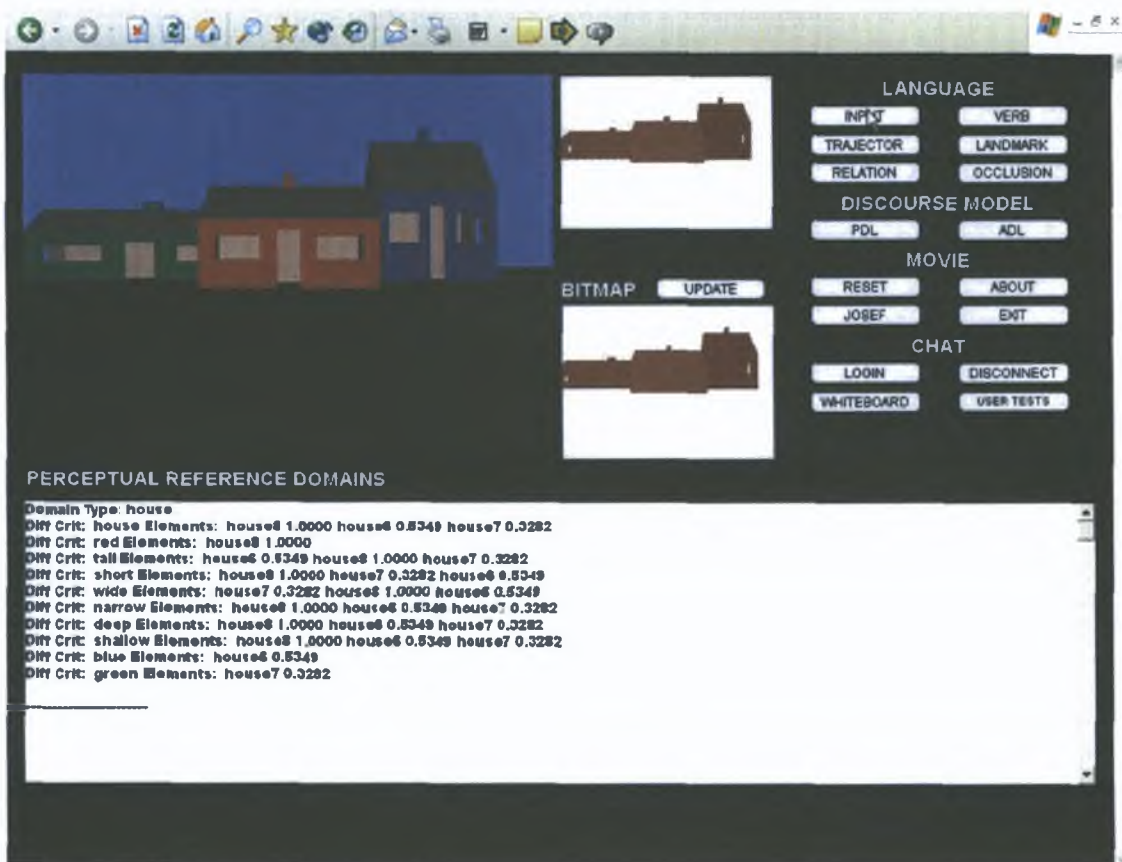
Figure 9-6: Screen shot illustrating the structure of a reference domain in the SLI system.

9.2.3 Profiled Elements List

The **profiled elements list** stores references to entities that are currently profiled in the domain. Profiling an element designates it as being prominent within a domain. More than one element may be profiled within a domain. For example, taking the scene in Figure 9-6 as the visual context and the accompanying reference domain as a local discourse context, if the user inputs the command *make the red house taller,* the reference domain element *house8* would be profiled as it represents the red house in the scene.

When an element is profiled it is removed from all the partitions in the domain and a reference to it is stored in the profiled elements list. Moreover, the partition which modelled the decomposition of the domain that was used to intend on the profiled object

292

is also profiled The motivation for this is that if, for example, a user accesses an element using the expression *the red house*, not only is the prominence of the intended element increased but also the prominence of the other elements within the domain which fit the description but are not selected However, if the removal of a profiled element from a partition empties the partition, the partition is deleted from the domain Furthermore, if the deleted partition was profiled, the domain type partition is profiled by default Figure 9-7 illustrates a reference domain that has a profiled element This domain was created by the SLI interpretive module in response to the user's command *make the red house taller* The local context used during the interpretation of this command was the reference domain illustrated in Figure 9-6 The restructuring of the original domain during the profiling process is evident in Figure 9-7 The restructured domain contains a profiled element list with one element, *house8* A consequence of this profiling process is that the red partition which was used to access the element was also profiled However, another stage in the profiling process is the removal of the profiled element, here *house8,* from the domain's partitions As a result of this removal, the red partition was emptied and deleted The deletion of the profiled red partition resulted in the domain's TYPE partition being profiled The other partitions in the domain have also been updated to reflect the removal of the *house8* element
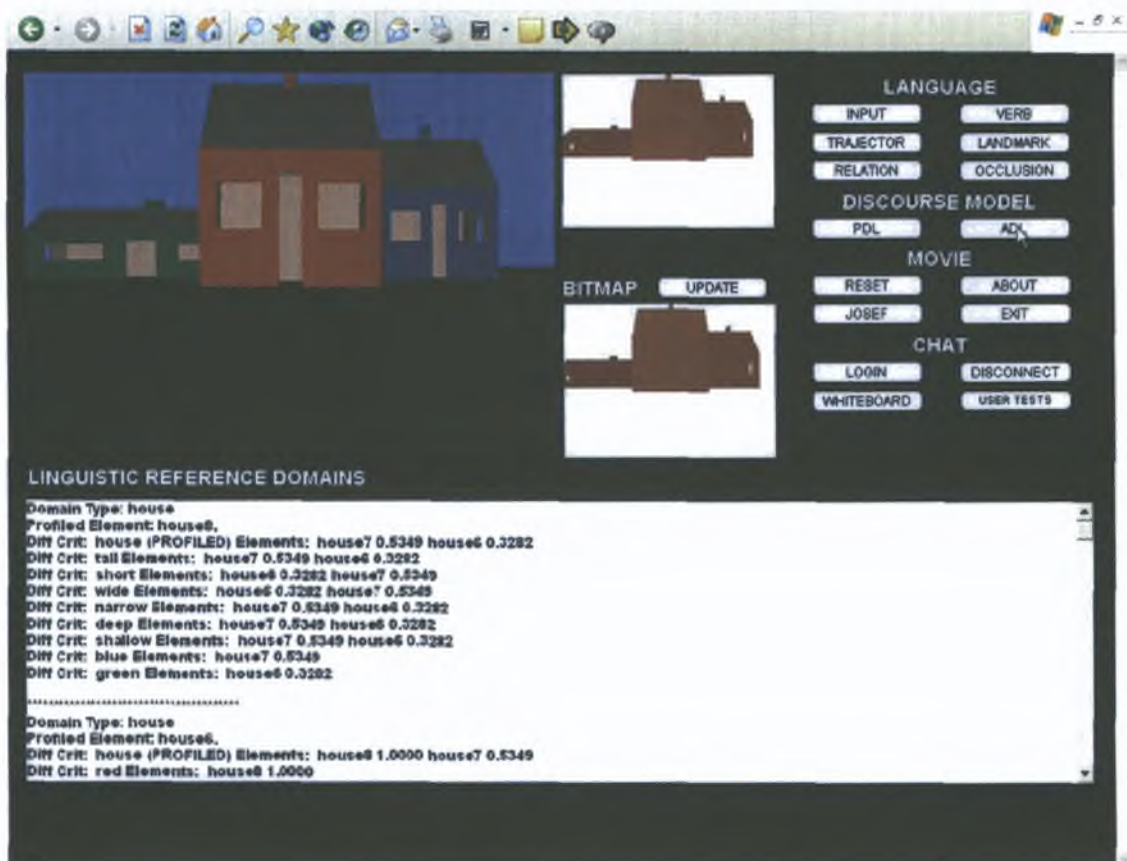
Figure 9-7: Figure illustrating a reference domain which has a profiled element and a profiled partition. This reference domain was created by the SLI interpretive module as a result of processing the command *make the red house taller*.

9.3   The Structure of the SLI Context Model

There are two types of reference domain: those that are created by perceptual cues and those that are created by linguistic cues. Based on this distinction, the context model is divided into two modules: the Visual Domains List (VDL) and the Linguistic Domains List (LDL). Although both of these modules comprise lists of reference domains, how they are created, what they represent, and how they function is quite different. Figure 9-8 illustrates the structure of the SLI context model. The context model is split into two lists of reference domains. The reference domains in the VDL are created using the output of the visual saliency model described in Chapter 7. The function of the VDL is to model

the user's visual perception of the simulated world The reference domains in the LDL are created by the SLI's interpretive module The function of the LDL is to model the linguistic component of the user-system interaction dialogue

Context Model

Visual Domains List	Linguistic Domains List
Domain Name	Domain Name
Profiled Element List	Profiled Element List
TYPE Partition	TYPE Partition
Partition 1	Partition 1
Partition 2	Partition 2
.	. .
Partition *n*	Partition *n*
Domain Name	Domain Name
Profiled Element List	Profiled Element List
TYPE Partition	TYPE Partition
Partition 1	Partition 1
Partition 2	Partition 2
...	...
Partition *n*	Partition *n*

Figure 9-8 The structure of the SLI context model The overall model is divided to two lists of reference domains. The reference domains in the Visual Domains List are created as a result of visual perceptual events. The reference domains in the Linguistic Domains List are created in response to utterances in the discourse. Note the reference domains in both lists have the same structure

9.3 1 Visual Domains List

The VDL represents a model of the user's visual memory of the environment and is used as a referent source when new entities are introduced into the linguistic discourse The reference domains in this module are called Visual Perceived Domains (VPDs) VPDs are created continually and are constructed based on the information supplied by the visual salience algorithm Recall that the visual salience algorithm runs each time a frame is rendered and creates a list of visible elements which have an associated saliency rating The visual context module takes this list and restructures the information to form reference domains The reference domain in Figure 9-6 is an example of a VPD Once a VPD has been created, it is inserted in the VDL This list functions like a stack, i e , it uses a last-in-first-out policy As new domains are created, they are added to the top of the stack – the domains that have been in the list longest are discarded once the list is full and new domains are added This structure was adopted as it mimics human memory and allows us to restrict the size of the perceptual domain The length of the system's perceptual memory is given by the equation

$$\textit{Length of system memory in seconds} = N / (F * T)$$

Equation 11: The equation defining the length of the system's perceptual memory. N = length of the list; F = frame rate of the system; T = average number of types of elements in each frame

In the current system, this list can contain up to 3000 VPDs, each representing an observation of all the objects of a particular type in the view volume for one frame Taking an average rendering speed of 30 frames per second and 3 different types of objects in the view volume per frame, this results in the system having a visual memory of 3000 / (30 * 3) = 33 33 seconds It should be noted that the number of a particular type of object in the view volume does not impact on the size of the visual memory since all the objects of one type will be stored within the reference domain modelling that type of object Consequently, although the above calculation for the length of the system's

visual memory is based on the assumption of three different types of objects in the view volume per frame, the calculation length of the system's memory would not alter even if there was dozens or hundreds of each type of object in each frame

9 3 2 Linguistic Domains List

The LDL represents the linguistic context and is used as the linguistic information source when the system is resolving anaphoric references The domains in this module are called Linguistic Domains (LDs) and are added to this module each time a user inputs a command to the system The LDs are created as a result of the interpretive process which restructures and extracts referents from domains in the context model The reference domain in Figure 9-7 is an example of an LD The LDL's structure is similar to the VDL's

1   It is a list that uses a last-in-first-out policy

2   New domains are added to the top of the stack

3   The domains that have been in the list the longest are discarded when a new domain is added and the list is full

9 3 3   The SLI Context Model Summary

In summary, the SLI context model is comprised of two stacks of reference domains These reference domains function as local context structure The reference domains in both stacks have a similar internal structure They each contain a domain name, a profiled element list, a TYPE partition, and a set of zero or more basic partitions Each partition in a domain has two components a differentiation criterion and a partition element list The differentiation criterion defines an object attribute that the objects, represented by the elements in the partition, possess The partition element list holds a set of elements ordered by visual salience or by fitness with respect to the partition's differentiation criterion Each element in a partition also has two components a pointer to

298

an object in the world model and the visual saliency rating ascribed to the object in the world that the element's pointer intends on

One of the stacks of reference domains is called the Visual Domains List or VDL The VDL represents a model of the user's visual memory of the environment and is used as a referent source for deictic references Reference domains are added to this stack in response to visual perceptual cues captured by the visual saliency algorithm developed in Chapter 7 Figure 9-9 illustrates the relationships between the components of a VDL reference domain and the visual salience and world model modules in the SLI framework The visual salience values used in this diagram are taken from Figure 9-6 above Indeed, this schematic is representative of the data flow that occurs in the SLI system during the processing of that scene The green parts of the figure represent the information stored in and flowing from the world model The red parts of the figure represent the creation and use of the visual saliency information

Rendering Engine

Visual Salience Module

World Model

Context Model

Visual Domains List

Linguistic Domains List

VPD1

Domain Name: *house*

Profiled Element List: *Null*

TYPE Partition:

Differentiation Criterion: *house*

Partition Element List:

Element 1:

Object Pointer: *house8*

Visual Salience: *1.0000*

Element 2:

Object Pointer: *house6*

Visual Salience: *0.5349*

Element 3:

Object Pointer: *house7*

Visual Salience: *0.3287*

Basic Partitions 1 ... *n*

VPD2

Domain Name: *X*

Profiled Element List: *Null*

Partitions: *TYPE & Basic*

...
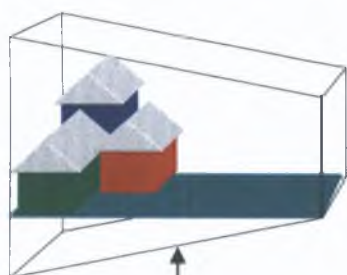
Figure 9-9· A figure illustrating the relationships between the components of a VDL reference domain and the visual salience and world model modules of the SLI framework. The actual saliency values used in this figure are taken from the system processing of Figure 9-6 The green parts of the figure represent the information stored in and flowing from the world model The red parts of the figure represent the creation and use of the visual saliency information

The second stack of reference domains in the SLI context model is called the Linguistic Domains List or LDL The LDL models the linguistic context of the discourse It is used as the referent source when the system is resolving an anaphoric reference New reference domains are added to this stack by the interpretive module The process of adding reference domains to this stack is triggered by the user inputting a command that contains a referring expression Figure 9-10 below illustrates the construction of an LD by the interpretive module and is indicative of the data flow triggered in the SLI system after the user's input *make the red house taller* (see Figure 9-7 above for the results of this process) The parts of the diagram that are drawn in blue represent the flow of information during the interpretation process, i e , the creation and insertion of the LD and the updating of the world model

Comparing LD1 and VPD1 in *Figure 9-10 illustrates some of the reference domain* restructuring that occurs during the SLI interpretive process In particular, LD1 has an element in the profiled element list but VPD1 does not Furthermore, the profiled element in LD1 has been removed from the TYPE partition It is also worth noting the changes in the LD element's visual salience scores, reflecting the fact that the scene has changed, i e , less of *house6* is visible because *house8* has been made taller However, the restructuring of a reference domain during the interpretation of a referring expression is not restricted to these changes In Section 9 4, the algorithms used by the SLI framework during the interpretation of a referring expression are developed, and the impact of these algorithms on the structure of the data in the context model is illustrated
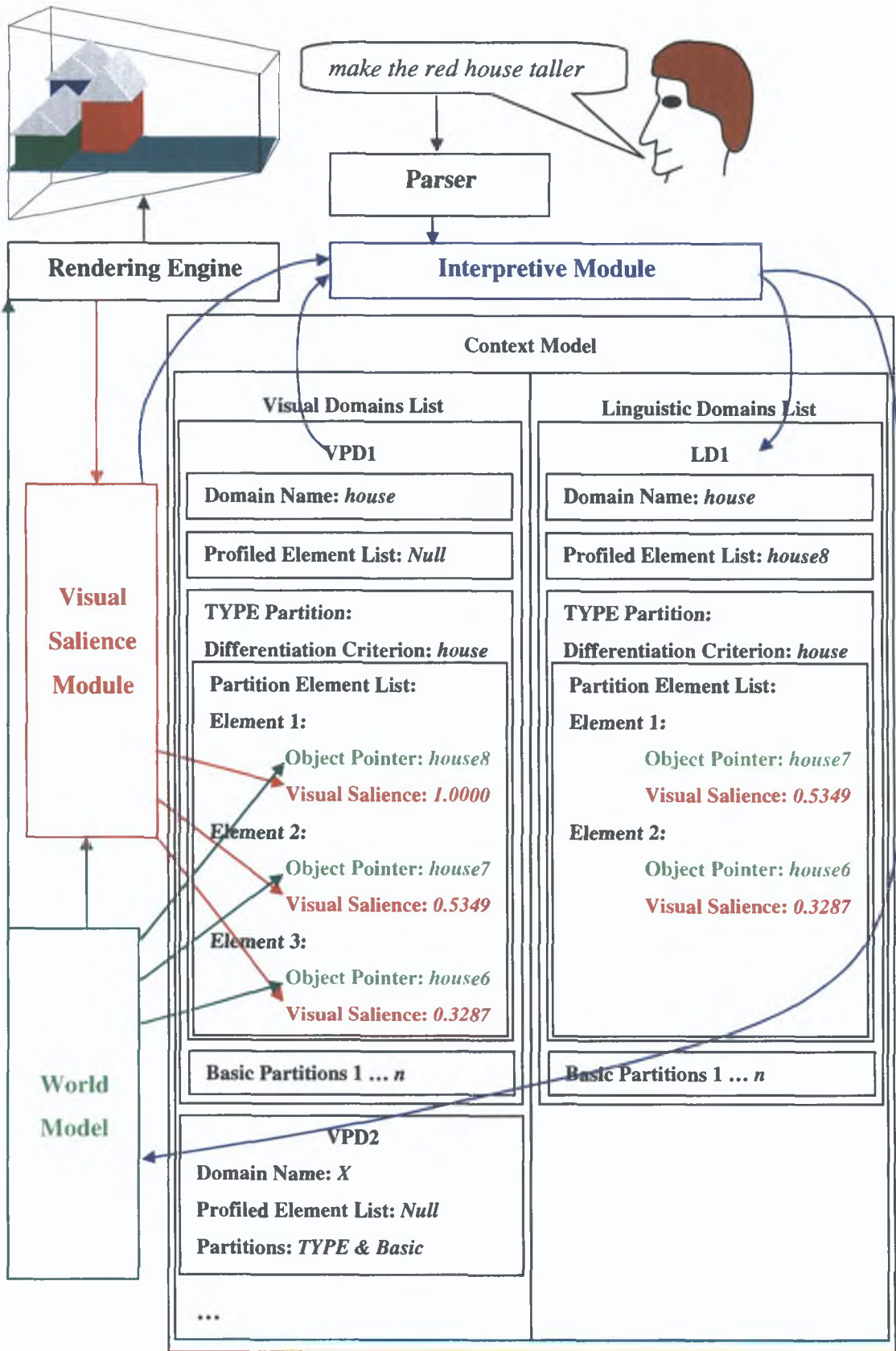
Figure 9-10: Diagram illustrating the creation of an LD and its insertion at the head of the LDL stack.

9.4   Interpretation Process

Sections 9 2, 9 3, and 9 3 3 described data structures comprising the SLI context model and illustrated how these data structures related to the other components in the SLI framework Moreover, the SLI interpretation module was introduced In this section, the algorithms that the SLI interpretive module uses to manipulate the context model's data structures are developed

First, however, it should be noted that there are several grammatical classes defined within cognitive grammar, the most fundamental distinction is between a nominal and a relational expression A **nominal expression** designates a thing[57]. This class of expression includes such traditional classes as noun, pronoun, and noun phrase A **relational expression** describes a relationship between things This class of expression includes adjectives, prepositions, adverbs, infinitives, participles, verbs, clauses, and full sentences The interpretation process described in this section focuses on how to resolve nominal expressions in particular, definite and indefinite descriptions, other-anaphoric expressions, one-anaphoric expressions, the unmarked pronoun *it*, and the singular primary demonstratives *this* and *that* when they are accompanied by a deictic gesture The interpretation of relational expressions is treated as a grouping operation applied to the domains created by the nominal expressions within the relational expression (see Section 9 5)

The context model presented in Sections 9 2, 9 3, and 9 3 3 provides the interpretive process with two distinct sources of information VDL and LDL Each of these information sources are considered as separate dialogues within the user-computer

---

[57] "A thing is characterised schematically as a 'region in some domain,' where a region can be established from any set of entities (e g , the stars in a constellation) just by conceiving of them in relation to one another" (Langacker 1994 pg 592)

303

discourse the VDL represents the visual perceptual dialogue and the LDL the linguistic dialogue Moreover, each dialogue is comprised of a set of chronologically ordered local context models Given this, the process of interpreting a referring expression may be defined as selecting a referent from a local context within a dialogue A three step algorithm was developed to achieve this goal Algorithm 9-1

---

1 Select the relevant dialogue VDL versus LDL

2 Select the local context of the utterance select a reference domain from within the relevant dialogue

3 Select and profile the expression's referent

Algorithm 9-1 The SLI interpretive algorithm

9.4 1 Selecting the Dialogue: VDL or LDL.

The first stage in interpreting a referring expression is to select which dialogue, visual versus linguistic, is appropriate as a general context for a given referring expression The distinction between these two dialogues is equivalent to the distinction between anaphoric and deictic references Anaphoric references refer to the referent of an antecedent noun phrase introduced by previous discourse, deictic references refer to an object which is physically present in the situation of the utterance where the identification is often supported by a demonstrative gesture (Pinkal 1986) Following this, deciding which dialogue is the appropriate context for a given referring expression is equivalent to determining whether the expression is anaphoric or deictic However, determining whether an expression is anaphoric or deictic is difficult because most forms of referring expression can be used in both an anaphoric and deictic manner

304

9 4 1 1   Definite Descriptions

It is generally held that the term definite description describes the set of noun phrases which are introduced by the definite article, *the* However, although this definition captures the widespread understanding of the term, it is neither universal nor unequivocal One issue with this definition is whether definite descriptions can be plural as well as singular A second issue is whether phrases which have possessive NPs for their determiners should be included as a definite description, e g , *his house* or *John's house* A third issue concerns the categorisation of phrases which begin with *the*, but are proper names, e g , *the Sun*, *the Grand Canyon*, etc In this thesis the term **definite description** is taken to denote singular noun phrases which are introduced by the definite article and cannot function as a proper name Moreover, following Kamp and Reyle, NPS with possessive determiners are treated " as a distinct semantical category" (1993 pg 249)

It is not surprising, when one considers the difficulties in describing what definite descriptions are, that there are also difficulties in defining what they do Most discussions about definite descriptions begin with Russell's (1905) Theory of Descriptions (the name he gave to his account of the logical function of descriptive phrases) One of the keynotes of Russell's theory is uniqueness of the referent of a definite description However, if a definite description, such as *the man*, is analysed, the problems with the uniqueness presupposition become apparent obviously, there is more than one man in the world Clearly, when someone uses such a phrase, the addressee is able to resolve the reference to one referent, consequently, it is evident that there is some form of uniqueness involved in identifying the referent However, this uniqueness criteria only works relative to a local context set Following this, Lyons' (1977) contextualised definition of Russell's theory of the function of definite descriptions is adopted they "identify a referent, not only by naming it, but also by providing the hearer or reader with a description of it, sufficiently detailed, in the particular context of utterance, to distinguish it from all other individuals in the universe of discourse" (1977 pg 179)

While Lyons' definition loosens the uniqueness presupposition to a degree which is more reconcilable with the actual usage of definite descriptions, his use of the term *universe of discourse* to describe the referential context of a definite description highlights the fact that a definite description may be used within several different kinds of universe of discourse: for a deictic definite description the universe of discourse is the physical environment of the discourse, for an anaphoric definite description the universe of discourse is the linguistic context of the discourse. As Pinkal states, definite descriptions "apply freely to object introduced in discourse, present in the physical environments, or available through the common background of the discourse participants" (Pinkal 1986 pg. 371). For this work, of particular importance, is the fact that definite descriptions can refer to objects already introduced to the discourse (anaphoric) or to objects present in the physical environment that are new to the discourse (deictic). Indeed, Poesio notes that "the two most common cases of definite descriptions in the TRAINS[58] conversations are anaphoric definites and definites interpreted with respect to the visual situation" (Poesio 1994 pg. 214). Given this, how can a distinction between a deictic and anaphoric use of a definite description be motivated?

In some instances, the categorisation of a definite description can be based on syntactic information. For example, "some non-anaphoric definite descriptions can be identified by looking for syntactic clues like attached prepositional phrases or restrictive relative clauses" (Bean and Riloff 1999 pg. 373). When processing these types of definite descriptions, the VDL is selected as the general context. In the SLI context, the presence of the modifier *other* or the use of the pronoun *one* as the head noun in the noun phrase is taken as syntactic indication that the definite description is an anaphoric reference. Accordingly, when interpreting these phrases, the LDL is selected as the general context. Indeed, a computational system that uses syntactic cues as a method of categorising definite descriptions has been developed by Poesio and Vieira (2000). While this system was successful in many instances, it could not process all of the definite descriptions in the test data it was given. Moreover, the test data it used was extracted from the Penn

---

[58] The TRAINS corpus is a multimodal corpus created at the University of Rochester. See
http://www.cs.rochester.edu/research/trains/

Treebank I corpus, a collection of newspaper articles from the Wall Street Journal This means that deictic uses of definite descriptions were not included in the test data (Poesio and Vieira 1998 pg 185) As it is often the case that anaphoric and deictic definite descriptions are not syntactically distinguishable, syntactic approaches will have difficulties in distinguishing between anaphoric and deictic definite descriptions that do not contain syntactic cues

In the absence of a syntactic cue, one is forced to use heuristic rules that utilise perceptual cues Given that, in the general context of these discourses, a user interacting with a simulated 3-D environment, the main information source is the visual simulation, it is expected that in the majority of cases, definite descriptions will be deictic references In other words, their referent is drawn from the visual perceptual dialogue While acknowledging that this assumption is a simplification of the issue, there is empirical evidence that suggests it is the correcat approach for the majority of cases Unfortunatly, the statistics describing the occurrences of anaphoric and deictic definites in the TRAINS corpus are not available (Poesio 2003) However, Poesio and Vieira's (1998) work on the Penn Treebank I corpus did quantify the anaphoric and deictic definite descriptions The results indicated that "about 50% of the definites in the collection were classified as discourse-new, 30% as anaphoric, and 18% as associative/bridging" (Poesio and Vieira 1998 pg 1) Another corpus-based investigation into definite descriptions found that non-anaphoric NPs "account for 63% of all definite NPs" (Bean and Riloff 1999 pg 374) in the 1600 MUC-4 corpus It should be noted that, both of the above corpora (Penn Treeback I and 1600 MUC-4) are purely textual, as opposed to transcribed spoken, resources and as such are not directly relevant to multimodal discourse However, in the context of a visually grounded discourse we conjecture that deictic uses of definite descriptions are at least as frequent as deictic uses attested in the text based corpora and that it is reasonable to take the deictic interpretation to be the default and treat the anaphoric interpretation as an exception that occurs under certain conditions

How can these conditions be defined? Pinkal's (1986) description of the interpretation of definite descriptions can be used as a starting point "the most salient object meeting the description is selected as the referent, independently of its offspring" (1986 pg 371) Recall from cognitive grammar, that profiling an element marks it as

307

being prominent Consequently, at the time an utterance occurs, if the preceding utterance in a discourse was referential, the most salient element in the discourse is the element that the preceding utterance has profiled, i e , the referent of the previous utterance Following Pinkal's (1986) interpretation, if the referent of the previous utterance is available to be selected as the referent for a definite description, it should be selected as such Here, this is taken as the anaphoric interpretation of a definite description Two conditions that the referent of the previous utterance must fulfil for selection, are

1 The referent of the previous utterance matches the type specification and adjectival descriptions provided by the definite description

2 The referent of the previous utterance should be currently visible in the view volume

The motivation for condition (1) is that an object cannot be selected as the referent for an expression if it does not match the description of the referent provided by the expression The function of condition (2) is to catch situations where a user has referred to an object and subsequently relocated in the world If the user then referred to an object of a similar type to their previous referential utterance and the system did not check that the previous referent was still visible, the interpretation of the user's new utterance would be applied to an object off screen

If either of these conditions is not met by the referent of the previous utterance, the framework resorts to the default, i e , the deictic interpretation of definite descriptions That is, the most salient element in the VDL that matches the linguistic description provided by the expression is selected as the referent It should be noted that, as the deictic interpretation is only triggered when the anaphoric interpretation is not available, the deictic interpretation is also congruent with Pinkal's (1986) interpretation

Based on the above discussion, a strategy for selecting whether or not a definite description is anaphoric or deictic and, consequently, whether to adopt the LDL or VDL as the general context for the interpretation process can be defined This strategy is defined in Algorithm 9-2

308

```
If ∧((NPStr.Det == 'the'), (NPStr.Modifiers[] ∩ {'other'} == ∅ ), (NPStr.Head ≠
{'one'})) Then

        If {x : ∧ (x ∈ LDL[1].Profiled[]), (x.Visible = TRUE), (fulfils(x))}≠ ∅

        Then

                IntExp.Dialogue = LDL

        Else

                IntExp.Dialogue = VDL

        End If

End If
```

Algorithm 9-2: The interpretive algorithm for selecting the general context for definite descriptions. The conditions containing the terms *other* and *one* indicate that other-anaphora and one-anaphora are treated as special classes of definite descriptions for which different strategies are used. For a definition of the terms used in the algorithm see Appendix A.

The following two examples from an SLI user-system dialogue illustrate how Algorithm 9-2 impacts on the interpretation of a definite description. In both of these examples, the user inputs a command that intends on an object using the referential expression *the house*. In the first example, *the house* is interpreted deictically. In the second example, *the house* is interpreted anaphorically.

Taking Figure 9-11 as a visual context, and assuming that none of the preceding user inputs have referred to any of the objects in the scene, if the user inputs the command *make the house brown*, the SLI system will treat the definite description *the house* as a deictic reference and interpret it as referring to the most salient element in the scene that matches the linguistic description provided by the expression[59]. Figure 9-12

---

[59] Note that where more than one object in the scene matches the description provided by the referring expression, the selection of the most salient object in this set as the referent is subject to the condition that the difference in saliency ratings ascribed to the primary candidate and each of the other candidate objects exceeds a predefined confidence interval, see Section 7.4.

309

illustrates the change in the visual scene after the interpretation of the command *make the house* brown.
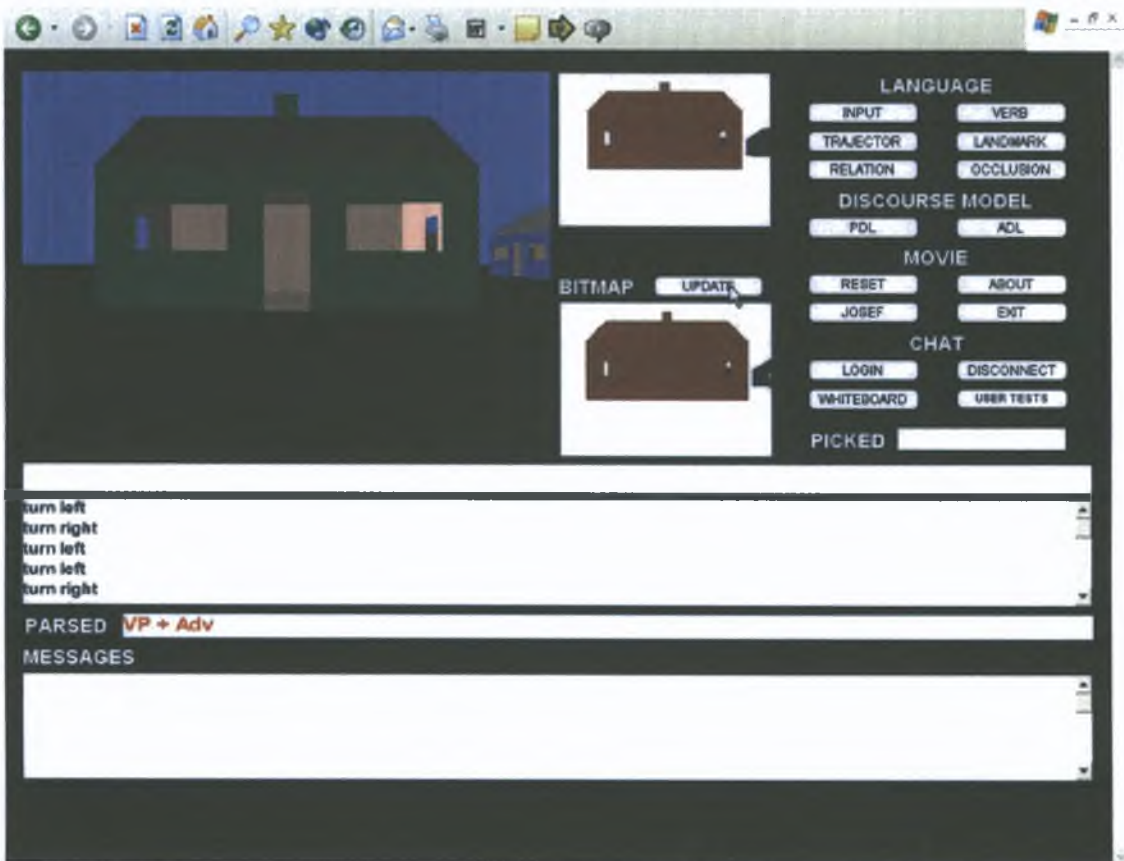
Figure 9-11: The initial visual context for an example illustrating a deictic interpretation of the definite description *the house*.
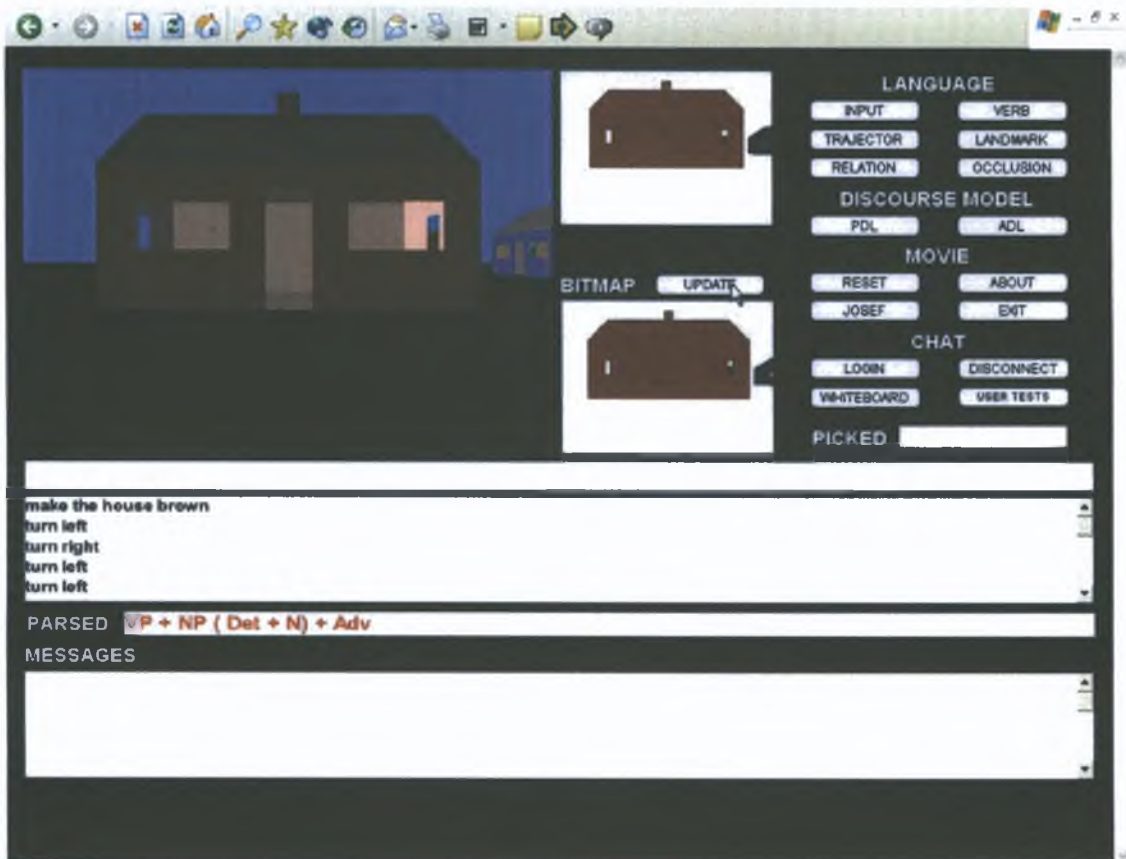
Figure 9-12: The state of the simulation after the system has interpreted the command *make the house brown*. **Note that in this instance the expression** *the house* **was treated as a deictic reference.**

An example of an anaphoric interpretation of the definite description *the house* can be given if the linguistic context is changed so that the command preceding the current command, containing the definite description *the house*, refers to an entity in the scene. For example, taking Figure 9-13[60] as the visual context, if the user inputs the command sequence[61]:

---

[60] It should be noted that Figure 9-13 is identical to Figure 9-11 above. Indeed, the state of the SLI system at the point that Figure 9-13 was captured was equivalent to the state of the system when Figure 9-11 was captured; i.e., the previous discourse was identical.

[61] Co-indexing (i.e., assigning identical subscripts) is used in this command sequence to indicate anaphoric relations

(17a) *Make the blue house$_i$ red.*

(17b) *Make the house$_i$ brown.*

– assuming that the referent of the definite description in (17a) is available for selection as the referent of the expression *the house*; i.e., it fulfils the description of the referent provided by the expression and is also currently visible in the view volume – the SLI system will treat the expression *the house* as an anaphoric rather than deictic reference. Figure 9-14 illustrates the state of the visual context after the system has interpreted *make the blue house red.* Figure 9-15 illustrates the state of the visual context after the system has interpreted *make the house brown.*
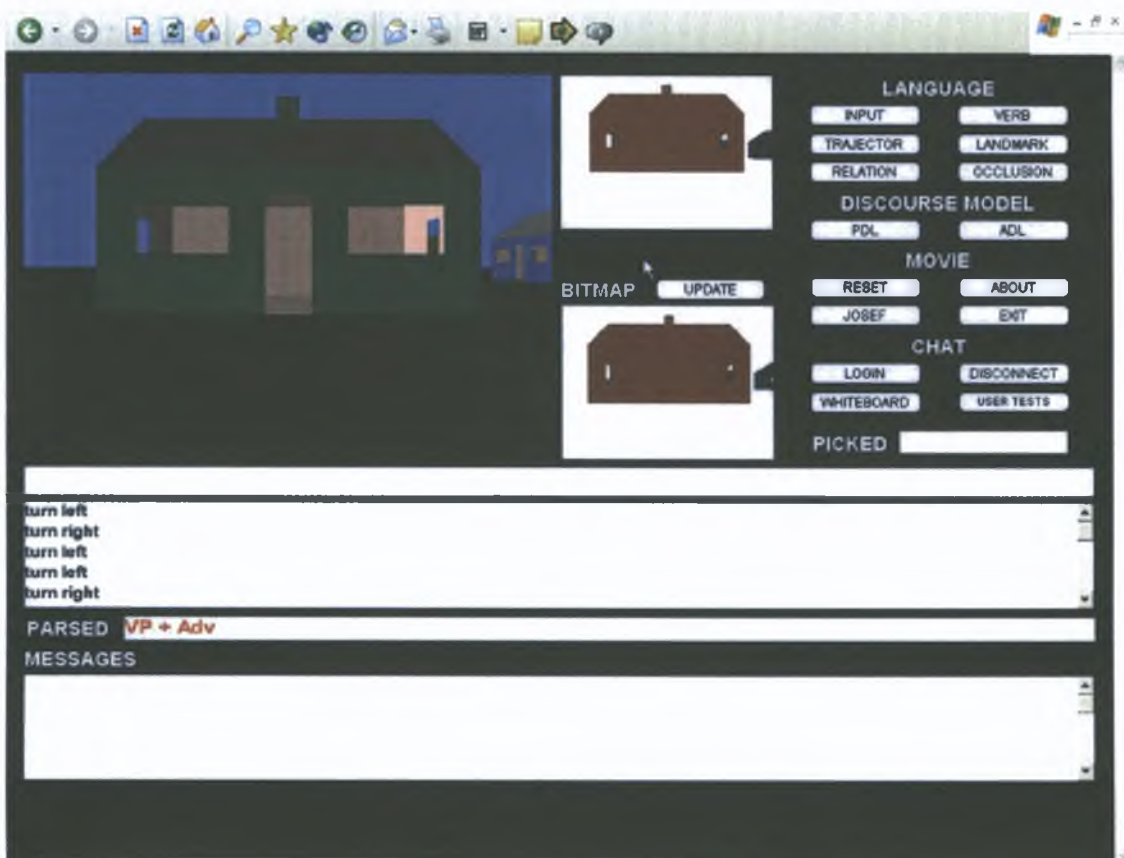
Figure 9-13: The initial visual context for an example illustrating an anaphoric interpretation of a definite description.
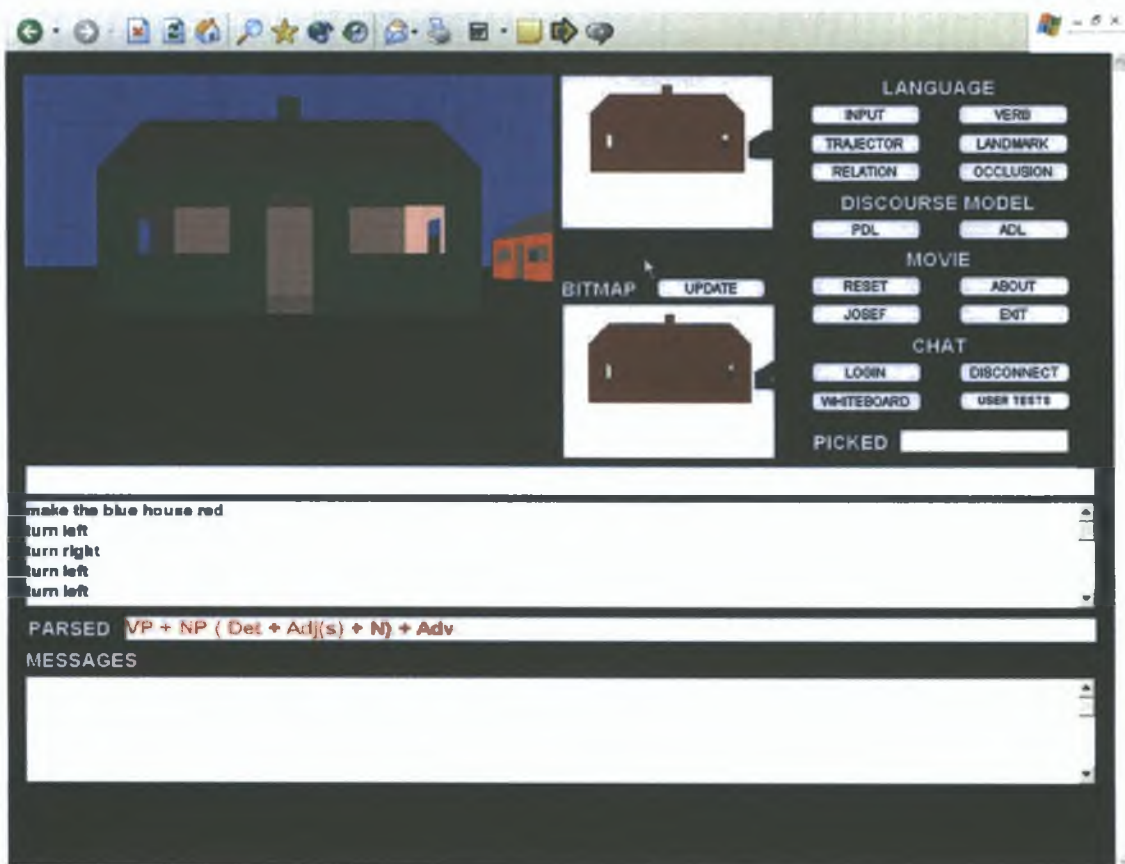
Figure 9-14: The visual context after the interpretation of the command *make the blue house red*. Note that the definite description in this command *the blue house* was interpreted as a deictic reference. Consequently, it introduces a new referent into the linguistic discourse.
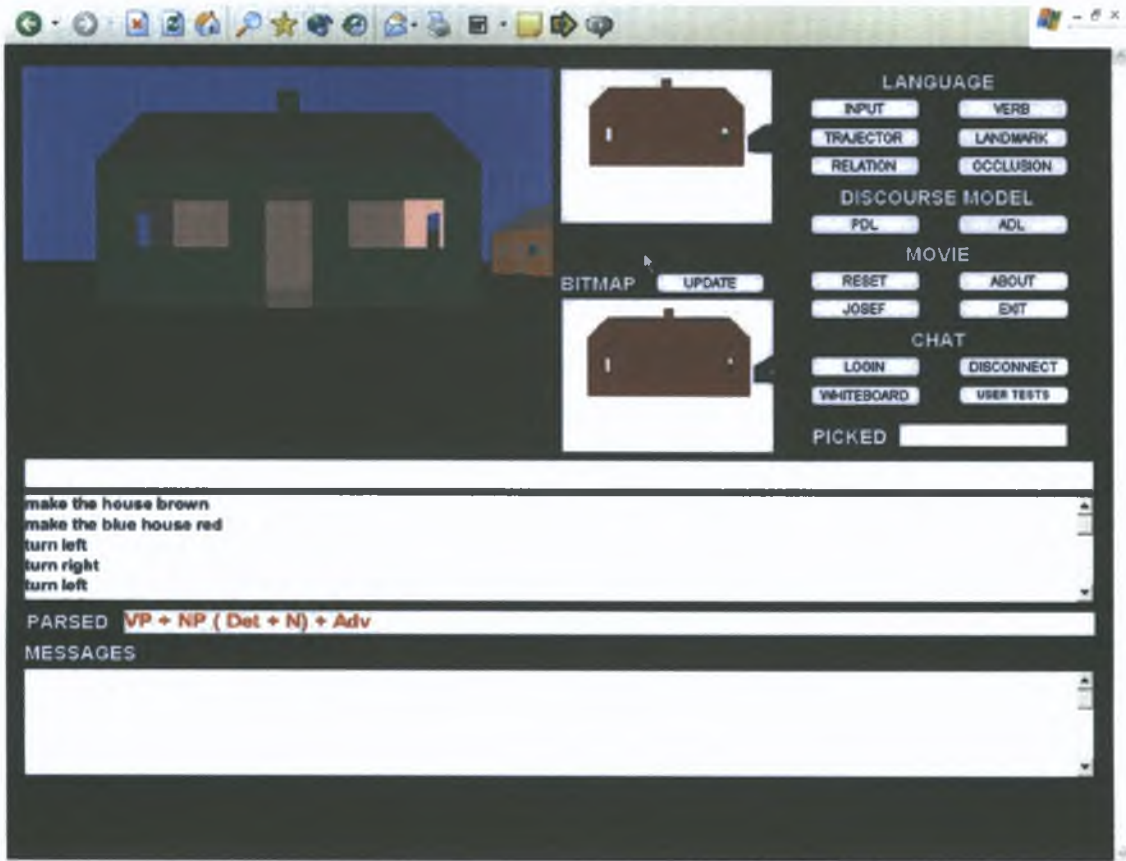
Figure 9-15: The visual context after the anaphoric interpretation of the referring expression *the house* in the user command *make the house red*.

Comparing the results of a deictic interpretation of *make the house brown*, Figure 9-12, and the anaphoric interpretation of *make the house brown*, Figure 9-15, illustrates the impact that Algorithm 9-2 has on the interpretation of definite descriptions.

9 4 1 2   One-Anaphora

In dialogue, the token *one* can be used as a generic pronoun, a numeral, and as a substitute pronoun (Greenbaum 1996) As a pronoun, *one* may substitute for an indefinite noun phrase (18) or for the head of a noun phrase and, perhaps, one or more of its modifiers (19)

(18a) *Well I could have a party*

(18b) *She's planning <u>one</u>*

(19a) *Which car is yours?*

(19b) *It's the blue <u>one</u>*

In (18b), *one* substitutes for *a party* and in (19b), *one* substitutes for *car* It is the uses of *one* as a substitute for the head of a noun phrase, as exemplified in (19b), that this thesis concerns itself with Semantically, this use of *one* seems to be situated between classical referential anaphora and ellipsis

"One anaphora seems to occupy a position that is half-way between the classical cases of referential anaphora (in particular pronoun anaphora) on the one side and the paradigmatic cases discussed elsewhere in this section [ellipsis] on the other side " (Cooper *et al* 1994 pg 130)

Furthermore, in these uses, the token *one* picks up some property of an antecedent noun phrase's referent

"Tokens of one that are used in this way might be described as 'second order anaphors' For what they do is to pick up (or stand in for) some property " (Cooper *et al* 1994 pg 131)

315

Because the properties that *one* can access are restricted to those introduced by simple or common noun phrases, this use of the token *one* has much in common with ellipsis; i.e., gapping or VP-deletion (Cooper *et al.* 1994).

Taking Figure 9-16 as the visual context and (20a) as the most recent utterance in the linguistic context, a typical example of these uses of *one* in a user-3-D system dialogue is (20b):

(20a) *Make the yellow house taller.*
(20b) *Make the green one shorter.*
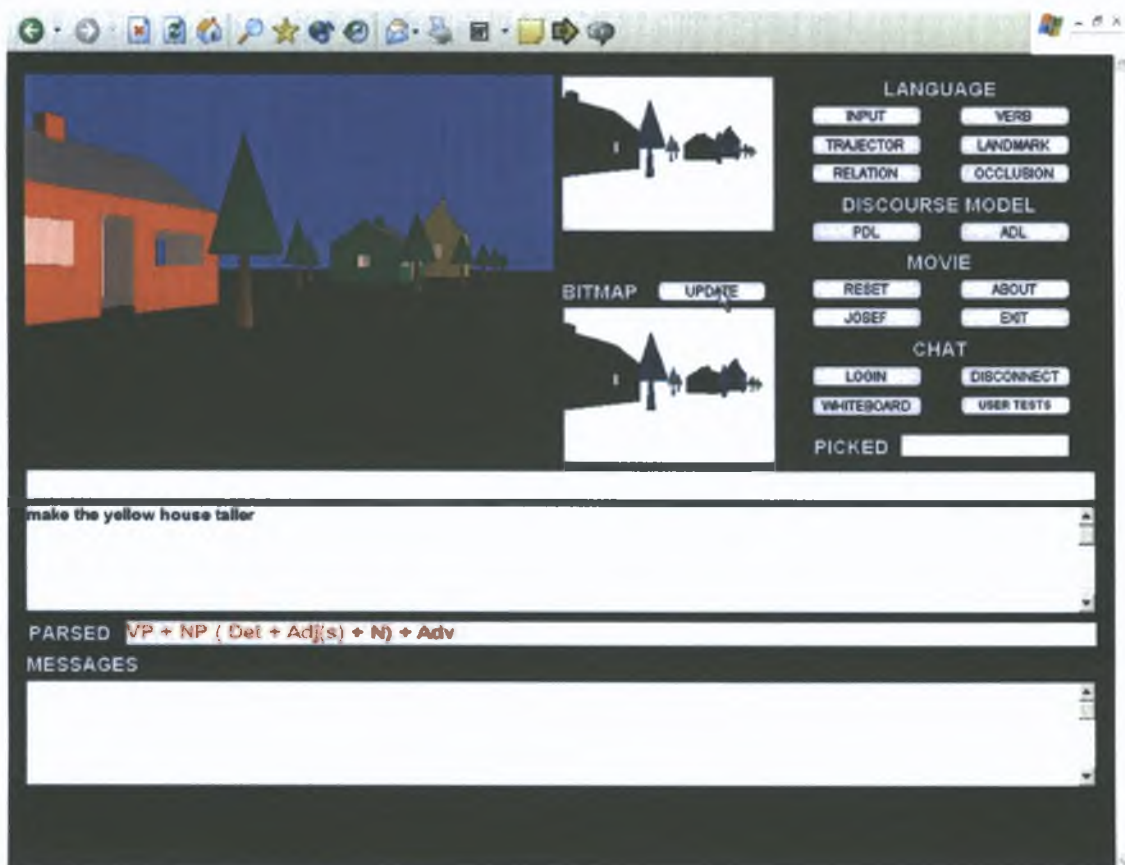
Figure 9-16: The visual context after the interpretation of *make the yellow house taller.*

Figure 9-17 illustrates the visual context after the interpretation of (20b). Importantly, in the context of an interaction dialogue between a user and a 3-D system,

the type information of an expression's referent is normally given by the head noun of a referring expression; i.e., the token *one* has picked up the type information from the previous utterance. Consequently, in order to interpret a referring expression in which *one* has been used to substitute for the head noun, the referent's type information must be extracted from the preceding linguistic utterance. In the SLI discourse model, this information is stored in the LDL. Accordingly, the LDL is selected as the general context for the interpretation process. Algorithm 9-3 gives a formal definition of this strategy.
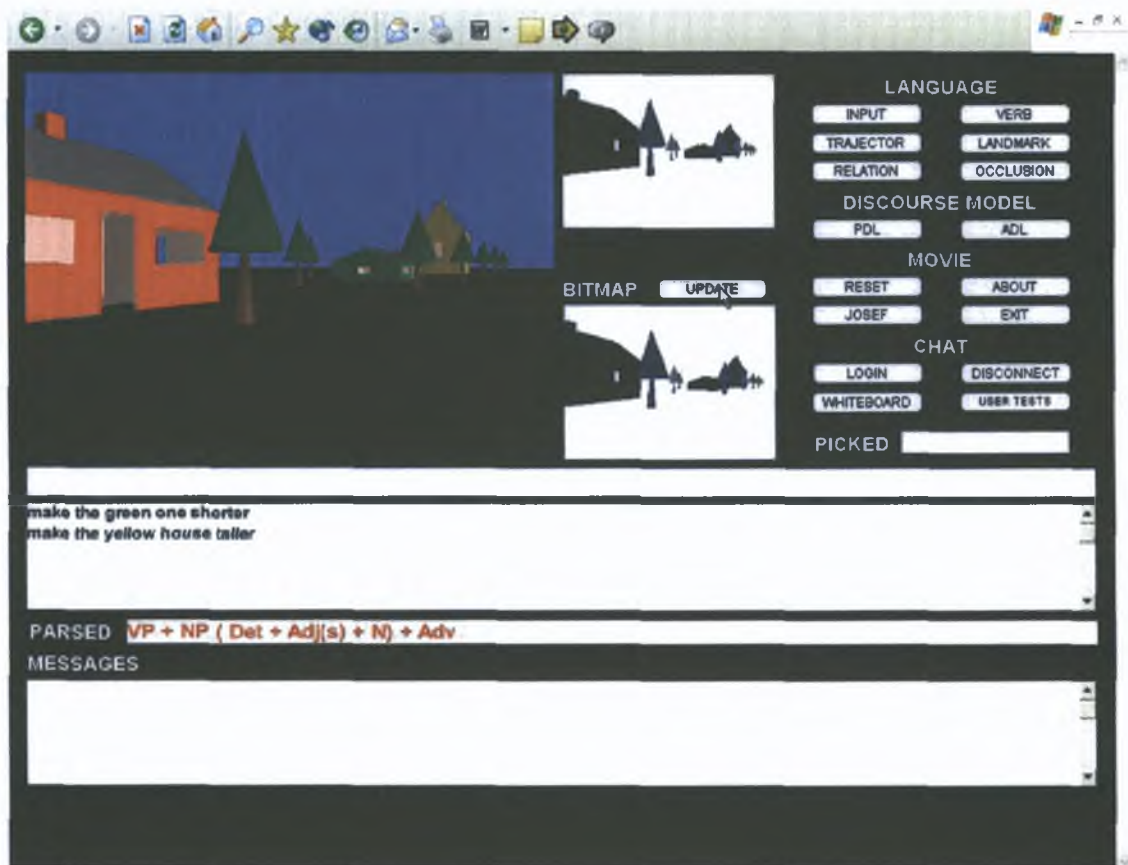
Figure 9-17: The visual context after the interpretation of *make the green one shorter*.

```
If ∧ ((NPStr Det == 'the'), (NPStr Modifiers[] ∩  {'other'} == ∅ ), (NPStr Head ∈
{'one'})) Then
                IntExp Dialogue = LDL
End If
```

Algorithm 9-3: The interpretive algorithm for selecting the dialogue for one-anaphoric definite descriptions. The precondition that the noun phrase does not contain the modifier *other* indicates that a different strategy is used for other-anaphoric definite descriptions. For a definitions of the terms used in this algorithm see Appendix A.

9 4 1 3   Other-Anaphora

Other-anaphora occurs when a definite description contains the modifier *other* In predicate logic, McCawley (1993) defines the semantics of a clauses containing *other* as a manifestation of ~ and = (where ~ is the negation operator and = is the identity relation), i e , the referent of a clause containing *other* is an entity that is not equal to some specified entity Following this, the modifier *other* designates an object that has been excluded from a specified or implied group Accordingly, the first step in interpreting an other-anaphoric expression is to define the grouping that the referent of the expression has been excluded from

In DRT, "other must be represented by a discourse referent that is presented as distinct form some discourse referent already introduced in the DRS" (Kamp and Reyle 1993 pg 463) Recall from Section 4 2 that the DRS or Discourse Representation Structure in DRT represents a global context model comprising all potential referents introduced into the discourse In effect, the referent of an other-anaphora expression must be distinct from a previously mentioned referent Importantly, the SLI discourse model's profiling mechanism (see Section 9 4 4) carries all the visual perceptual information in the local context (modelled by the selected reference domain) at the time an utterance was interpreted forward into the restructured LD Consequently, each LD contains a list

of the elements representing the referents of the expression in its profiled element list and a list of the elements representing the objects not selected as the expression's referent in their TYPE partition and basic partitions As a result, the LDL contains all the information that is required to interpret an other-anaphora expression Accordingly, it is selected as the general context for the interpretation of these expressions, as in Algorithm 9-4

If $\bigwedge$ ((NPStr Det == '*the*'), (NPStr Modifiers[] $\cap$ {'*other*'} $\neq \varnothing$)) Then

        IntExp Dialogue = LDL

End If

Algorithm 9-4 The interpretive algorithm for selecting the general context of an other-anaphoric definite description. For a definition of the terms used in this algorithm see Appendix A.

9 4 1 4 Indefinites

Although there are several types of indefinite referring expressions, this thesis focuses on singular noun phrases introduced by the indefinite article, e g , *a house, a man*, etc While there are instances where an indefinite noun phrase may be used anaphorically (21), they are paradigmatically viewed as introducing new, non-specific entities into the discourse (Kamp and Reyle 1993)

(21) "*Dr Smith, told me that exercise helps Since I heard it from a doctor,, I'm inclined to believe it*" (Byron 1998 pg 21)

Following the standard analysis of indefinites in this thesis, it is assumed that indefinite references are not anaphoric However, this assumption does not indicate that they are exclusively deictic In the context of a simulated environment an indefinite noun phrase *a N* may be used to arbitrarily refer to one of the elements of type *N* in the spatio-

319

temporal context (22a), or to the generic type N (22b) in commands that create new objects in the world

(22a) *Make a house taller*

(22b) *Add a tree to the right of the red house*

To adjudicate between these options, the framework first checks the verb used In the SLI context, some verbs (e g , *add, create*, etc) can only be used in commands that create new objects in the simulation Consequently, indefinites that complement these verbs are interpreted as referring to the generic type N In contrast, some other verbs in the SLI context (e g , *go to, look,* etc) can only be complemented by references to objects that already exist in the simulation Indefinites that complement these verbs are interpreted as arbitrarily referring to an element within the set of objects in the spatio-temporal context that fulfil the selection restrictions specified by the reference Finally, some verbs (e g , *make*) can be used in both types of commands If the verb used does not allow the categorisation of the indefinite, the framework then checks whether there is an adjective in a post-verbal or predicative position in the input If there is an adjective in a predicative position in the input, the indefinite is interpreted as a deictic reference, i e , the indefinite refers to an arbitrarily selected element from the set of objects in the spatio-temporal context that fulfil the selectional restrictions of the reference

Examples (23a) and (23b) illustrate inputs where the categorisation of the indefinite is based on the verb In (23a) the verb *go to* is used In the SLI context, *go to* can only be complemented by indefinite noun phrases that arbitrarily refer to an element in the spatio-temporal context In (23b) the verb *add* is used Following this, the indefinite in (23b), *a green house*, would be interpreted as referring to the generic type house Examples (23c) and (23d) illustrate indefinites that are categorised based on the presence or absence of a predicative adjective In both these examples the verb *make* is used In the SLI context the verb *make* can be used with an indefinite, *a N*, to arbitrarily refer to one of the elements of type *N* in the spatio-temporal context or to refer to the generic type *N* In (23c) the adjective *green* occurs in a predicative position in the input Consequently, the indefinite *a house* is interpreted as arbitrarily referring to one of the houses in the spatio-

temporal context Conversely, in (23d) there is no adjective used in a predicative position[62], and, as a result, the indefinite in (23d), *a green house*, is interpreted as referring to the generic type house

(23a) <u>Go to</u> a green house

(23b) <u>Add</u> a green house

(23c) <u>Make</u> a house <u>green</u>

(23d) <u>Make</u> a green house

The referents of commands that insert a new element into the simulated environment (e g , <u>*add*</u> *a green house*, <u>*make*</u> *a green house*, etc) are extracted from the system's hierarchy of object models This knowledge is assumed to be resident in the system before the discourse begins It is akin to the conceptual or encyclopaedic world knowledge that is assumed by a human interlocutor to be shared with their audience on the basis of common knowledge, the knowledge source that humans use to interpret references to abstract entities or entities which are not in the immediate environment The first stage in the interpretation of these commands is to instantiate their referent in the discourse context Following this, the most recent VPD that contains an element which represents the object created as the expression's referent is restructured This creates an LD whose profiled element represents the newly instantiated object, and this LD is inserted at the top of the LDL stack

All other indefinite phrases (e g , *<u>go to</u> a green house*, *make a house <u>bigger</u>*, *move a house <u>forward</u>*, etc) are treated as arbitrarily selecting an element from an appropriate reference domain within the VDL This form of command takes as its referent an object in the perceptual dialogue

Algorithm 9-5 gives a formal description of the strategy for selecting a general interpretive context for indefinite descriptions It should be noted that the first path through this decision tree (i e , If ($\vee$(Verb $\in$ Creation_Verbs[], $\wedge$(Verb $\in$ Either_Verbs[], Predicative_Adjectives[] == $\varnothing$))) Then) defines how inputs interpreted

---

[62] In (23d) the adjective *green* occurs within the noun phrase in an attributive position

as commands to insert a new object into the simulation are processed The strategy defined in this section of the algorithm defines the entire interpretive process used for this type of input The second path through this decision tree captures user input that arbitrarily refers to an object in the scene, e g , *make a house red* For these commands, the VDL is selected as the general interpretive context

```
If (NPStr det == 'a') Then
        If (∨(Verb ∈ Creation_Verbs[] ,
        ∧ ( Verb ∈ Either_Verbs[], Predicative_Adjectives[] == ∅ ))) Then
                RefPtr = createReferent(NPStr Head, NPStr Adjectives[])
                i = MinIndex(VDL, {x    ∧(( x ∈VDL), (RefPtr ∈
                x TYPE Elements[] Object)) })
                LDL[1] = restructure(VDL[i])
        Else
                IntExp Dialogue = VDL
        End If
End If
```

Algorithm 9-5 The interpretive algorithm for selecting the general context for an indefinite referential expression. For a definition of the terms used in this algorithm see Appendix A

9 4 1 5 Pronouns

Pronouns are in effect a closed set of nouns They carry very little information about their antecedent Besides gender markings on third person pronouns and number constraints they have very few clues that help to resolve their reference This thesis focuses on interpreting the unmarked pronoun *it*

Although pronouns are the paradigmatic example for anaphora, not all pronoun uses are anaphoric Examples of non-anaphoric uses of pronouns include the *it* in

322

mentions of time and weather, e g , *it's raining*, and the *it* introducing a cleft sentence[63] (Hirst 1994, Byron 1998) However, these non-anaphoric examples do not occur in the SLI scenarios

Salmon-Alt and Romary (2001) treat pronouns as expressions that designate an element which has been previously profiled In the SLI discourse framework, the domains containing profiled elements are exclusive to the linguistic dialogue This means that when interpreting a pronoun the linguistic dialogue is selected as the general context This approach is in agreement with Byron's (1998) analysis of pronouns Byron notes that due to their minimal marking, the determination of a pronoun's referents is difficult As a result, "pronouns should only be used in cases where the referent is obvious to the listener because of the discourse context" (1998 pg 12) Algorithm 9-6 lists a formal definition of the dialogue context selection

```
If NPStr Head =='it' Then
        IntExp Dialogue = LDL
End IF
```

Algorithm 9-6: The algorithm for selecting dialogue context for the pronoun *it*. For a definition of the terms used in the algorithm see Appendix A

---

[63] A cleft sentence is a sentence that is split so as to put focus on one part of it A cleft sentence is introduced by *it*, which is followed by a verb phrase whose main verb is generally *be*, e g , *It was yesterday that the movie was on TV*

323

9 4 1 6 Demonstratives

In English there are four primary demonstratives *this, that, these, those* They can be used as pronouns (24a) or as determiners (24b)

(24a) *Look at this*

(24b) *I'd like some of that apple*

Primary demonstratives present two types of contrast number and proximity Numerically *this* and *that* are singular and *these* and *those* are plural With respect to proximity, *this* and *these* indicate relative nearness while *that* and *those* indicate relative remoteness This proximity may be in space or in time

Often a demonstrative is accompanied by a physical gesture that identifies the intended referent These gestures usually occur when there is more than one element within the perceptual context that could function as a referent If there is only one element of the type specified by the noun in instances where the demonstrative is used as a determiner or only one element in the perceptual context that is suitable to act as a referent, the gesture is not really necessary to identify it The gesture is used to make the intended referent more salient, or to profile the referent

This thesis focuses on instances where singular demonstratives are used as pronouns and are accompanied by a gesture Following (Rieser 1999), deictic gestures are treated as semiotic objects Furthermore, a deictic gesture is interpreted as being equivalent to a deictic definite description and the demonstrative as an anaphoric noun phrase which takes the gesture as its antecedent and refers to the gesture's referent As the interpretation of the antecedent gesture is equivalent to an antecedent definite description, the general context for the interpretation of a demonstrative is the linguistic dialogue Consequently, the linguistic dialogue is selected as the general context when interpreting the utterance containing the demonstrative Importantly, this dialog selection does not occur until the utterance is completed This means that

324

1. when the demonstrative is being interpreted, the most recent reference domain in the linguistic context (i.e., the reference domain that will be used as the local context for interpreting the demonstrative) will be the reference domain that was created by the interpretation of the deictic gesture that accompanied the demonstrative.

2. the linking between the demonstrative and the gesture may be cataphoric as well anaphoric.

Algorithm 9-7 gives a formal definition of the dialogue context selection.

If $\vee$ ((NPStr.Head == 'this'), (NPStr.Head == 'that')) Then

      IntExp.Dialogue = LDL

End If

Algorithm 9-7: The algorithm defining the selection of the dialogue context for the interpretation of the singular demonstratives *this* and *that*. For a definition of the terms used in the algorithm see Appendix A.

9.4.1.7 *Selecting the Dialogue: VDL or LDL Summary.*

Section 9.4.1 defines the strategies used during the first stage of the SLI interpretive process: the selection of which dialogue is appropriate, visual versus linguistic, to be used as the general context for a given referring expression. It was noted at the beginning of Section 9.4.1 that the distinction between these two dialogues is equivalent to the distinction between anaphoric and deictic references. However, it was also noted that determining whether an expression is anaphoric or deictic is difficult because most forms of referring expression can be used in both an anaphoric and deictic manner.

This framework explicitly handles: definite descriptions including both one-anaphora and other-anaphora, indefinite descriptions, pronominal reference specifically the pronoun *it*, and the singular demonstratives *this* and *that* when accompanied by a

deictic gesture (a mouse click) For each of these forms of reference, an introductory discussion explaining the motivation underpinning the approach and a formally defined algorithm to select the relevant dialogue was provided Algorithm 9-8 lists the complete algorithm for selecting the dialogue for all the types of reference that the SLI framework supports

```
If (NPStr.Det == 'the') Then          //Definite Description
        If ∧ ((NPStr.Modifiers[] ∩ {'other'} == ∅ ), (NPStr.Head ≠ {'one'})) Then
                If {x : ∧ (x ∈ LDL [1].Profiled[]), (x.Visible = TRUE), (fulfils(x))}≠ ∅   Then
                        //Anaphoric Interpretation
                        IntExp.Dialogue = LDL
                Else
                        //Deictic Interpretation
                        IntExp.Dialogue = VDL
                End If
        Else If ∧((NPStr.Modifiers[] ∩ {'other'} == ∅ ), (NPStr.Head ∈ {'one'})) Then
                //One-anaphora
                IntExp.Dialogue = LDL
        Else If (NPStr.Modifiers[] ∩ {'other'} ≠ ∅ ) Then
                //Other-anaphora
                IntExp.Dialogue = LDL
        End If
Else If (NPStr.det == 'a') Then       //Indefinite Description
        If ( ∨ (Verb ∈ Creation_Verbs[] ,
        ∧ ( Verb ∈ Either_Verbs[], Predicative_Adjectives[] == ∅ ))) Then
                //Reference to Generic Object
                RefPtr = createReferent(NPStr.Head, NPStr.Adjectives[])
                i = MinIndex(VDL, {x : ∧ (( x ∈ VDL), (RefPtr ∈  x.TYPE.Elements[].Object))})
                LDL[1] = restructure(VDL[i])
        Else
                //Arbitrary Deictic Reference
                IntExp.Dialogue = VDL
        End If
Else If NPStr.Head =='it' Then        //Pronominal Reference
        IntExp.Dialogue = LDL
Else If ∨ ((NPStr.Head == 'this'), (NPStr.Head == 'that')) Then        //Demonstrative
        IntExp.Dialogue = LDL
End If
```

Algorithm 9-8: The interpretive algorithm for selecting the dialogue context for the interpretation of an expression. All text in red font which is preceded by the symbol

'//' are explanatory comments and are not part of the algorithm. For a definition of the terms used in the algorithm see Appendix A.

9 4 2   Selecting a Reference Domain

The second stage in the interpretive process is the selection of the local context of the utterance Recall (see Section 9 4 1 1) that each of these local contexts is named based on the types of objects they contain Furthermore, they are temporally organised by their position in the VDL and LDL The process for selecting the reference domain uses both the temporal and lexical domain information The general approach is to select the most recent domain within the relevant dialogue that contains one or more elements which match the description of the object in the expression Where no description is given (*it*, *this*, etc), the selection process simply returns the most recent domain in the dialogue

For this stage of the interpretive process, all deictic references (e g , deictic definite descriptions or deictic indefinite descriptions), are treated as equivalent Selecting their reference domain involves searching for the most recent domain in the VDL which contains elements of type N If the linguistic description contains a colour adjective, a further restriction is that the domain elements which fulfil the type restriction must also fulfil the colour restrictions Algorithm 9-9 lists the algorithm for selecting the reference domain for a deictic reference

```
If IntExp.Dialogue = VDL Then

        If ({ y : y == colour adjective } ∩ NPStr.Adjectives[]) ≠ ∅ ) Then

                //check for elements matching type and colour

                i = MinIndex(VDL, { x : ⋀((x ∈ VDL), (∃ j :

                fulfils(x.TYPE.Elements[j].Object) ))})

                IntExp.Rd = VDL[i]

        Else

                //check for elements matching type

                i = MinIndex(VDL, { x : ⋀((x ∈ VDL), ( ∃ j :

                fulfils(x.TYPE.Elements[j].Object) ))} )

                IntExp.Rd = VDL[i]

        End If

End If
```

Algorithm 9-9: The interpretive algorithm for selecting a reference domain for a deictic reference. For a definition of the terms used in the algorithm see Appendix A.

Selecting the reference domain for an anaphoric expression is more complicated than the process used for deictic expressions. This is because different forms of anaphoric expressions make different presuppositions about the structure of their local context.

Recall from Section 9.4.1.1 that a precondition of a definite description being anaphorically interpreted is that the referent of the preceding discourse utterance is eligible and available as the referent of the current expression being interpreted. Consequently, the most recent reference domain in the LDL is selected as the local context for anaphoric-definite descriptions.

In Section 9.4.1.5, pronouns were described as expressions that designate elements which have been previously profiled. In the case of singular pronouns, this description can be refined to: singular pronouns are expressions that designate a single previously profiled element. Therefore an a priori condition for selecting a reference domain as a

local context for a singular pronoun is that the domain should contain exactly one profiled element.

When used as pronouns, the singular demonstratives *this* and *that* also designate a single previously profiled element. In many instances, these demonstratives are accompanied by a deictic gesture which profiles the intended referent of the demonstrative. Indeed, this is the reason why deictic gestures are treated as semiotic objects and the demonstratives as anaphoric expressions which take the accompanying pointing gesture as their antecedent and the object the gesture intends on as their referent. It is important to note that, in this framework, deictic gestures are interpreted as they occur, and demonstratives are interpreted when the utterance they are in is completed. Consequently, the most recent reference domain in the LDL at the time a demonstrative is being interpreted will have been created by the interpretation of the deictic pointing gesture that accompanied the demonstrative. Following this, demonstrative expressions take the most recent reference domain in the LDL that has exactly one profiled element as their reference domain.

One-anaphora occurs when the pronoun *one* substitutes for the head of a definite noun phrase. The pronoun *one* is generic and can be used for any singular noun. Because the pronoun *one* carries very little information that helps in resolving its referent, it is usual for a one-anaphora expression to contain an adjectival description that restricts the possible referents of the expression; e.g., *the blue one*, *the tall one*, *the tall blue one*. Consequently, in contrast with pronouns and demonstratives, one-anaphoric expressions may be interpreted in a context that does not contain a profiled entity once it contains an element matching the adjectival description of the object in the expression. Therefore, they are assigned the most recent reference domain in the LDL that contains an element matching any supplied adjectival description as their local context.

Finally, other-anaphora occurs when a definite description contains the modifier *other*. The modifier *other* designates an object that has been excluded from a specified or implied group. Following this, the primary consideration in selecting a local context for an other-anaphoric expression is that the domain of reference should contain both a specified or implied grouping and an element that has been excluded from that group. Selecting a suitable context for these types of expression is further complicated by the

330

possibility of an adjectival description being included within the referring expression, e g , *the other blue house* In the parlance of this framework, these considerations translate into the requirement that the reference domain contain one or more profiled elements and a non-profiled element that fulfils the adjectival and type descriptions of the referent Following this, the reference domain selected as the local context for other-anaphora references is the most recent LD that fulfils these requirements

Algorithm 9-10 lists the strategies used to select a reference domain for the different types of anaphoric reference accommodated by the framework

```
If IntExp.Dialogue = LDL Then

        If NPStr.Det = 'the' Then

                If (NPStr.Modifiers[] ∩ {'other'} == ∅ ) ∧ (NPStr.Head ∉ {'one'})

                Then

                        //Anaphoric Definite Description

                        IntExp.RD = LDL[1]

                Else If (NPStr.Modifiers[] ∩ {'other'} == ∅ ) ∧ (NPStr.Head ∈

                {'one'}) Then

                        //One-anaphora

                        i = MinIndex(LDL, {x : ∧((x ∈ LDL), ( ∃ j:

                        fulfils(x.TYPE.Elements[j].Object)))})

                        IntExp.Dialogue = LDL[i]

                Else If (NPStr.Modifiers[] ∩ {'other'} ≠ ∅ ) Then

                        //Other-anaphora

                        i = MinIndex(LDL, {x : ∧((x ∈ LDL), (|x.Profiled| > 0),

                        ( ∃ j: fulfils(x.TYPE.Elements[j].Object)))})

                End If

        Else If (NPStr.Head ∈ {'it', 'this', 'that'}) Then

                //Pronominal or demonstrative references

                i = MinIndex(LDL, { x: ∧((x ∈ LDL), ( |x.Profiled| == 1))})

                IntExp.RD = LDL[i]

        End If

End IF
```

Algorithm 9-10: The algorithm used to select the reference domain for anaphoric references. For a definition of the terms used in the algorithm see Appendix A.

9.4.3 Selecting the Expression's Referent

Once a suitable domain has been selected from the relevant dialog, the referent of the expression is extracted from the domain Some referring expressions, such as the pronoun *it* and the singular primary demonstratives *this* and *that*, refer to entities which are already profiled in the selected reference domain For these expressions, no restructuring of the reference domain occurs However, for the other types of referring expression, the extraction process comprises a restructuring of the domain that results in the profiling of the element in the domain that represents the expression's referent This restructuring models the imposition of a particular construal on the domain content by the expression There are two parts to this stage of the interpretation process The first part is the selection of the element in the domain that represents the expression's referent, this stage of the process is different for each type of expression The second part of the process is the profiling of the selected element, the profiling mechanism differs between domains in the VDL and domains in the LDL This section describes the processes used to select the referent for each type of referring expression Following this, Section 9 4 4 describes how elements are profiled within a domain

The process of selecting the referent of the expression under interpretation attempts to select the most salient element in the domain that matches the description of the referent in the expression A crucial factor in this process is the internal structure of the reference domains Recall that each domain is divided into partitions which consist of a differentiation criterion that specifies an attribute or a list of attributes that the partition elements have and a list of elements sorted by attribute fitness and then by salience Each domain has at least one partition whose differentiation criterion is set to the domain type, this partition lists all the elements in the domain apart from the profiled elements These partitions are attempts to predict the different ways that a user may refer to an object in the domain Furthermore, each reference domain has a profiled element list that contains a list of the elements of the domain that are currently profiled It is important to note that when an element is profiled, it is removed from the domain's TYPE partition and all the domain's basic partitions

333

In general, the selection of a referent for a deictic definite description consists of selecting the most salient element in the reference domain that fulfils the description of the referent in the expression Consequently, if no adjectives are used to describe the referent, the most salient element in the domain's TYPE partition is selected For example, the deictic interpretation of the expression *the house* would ascribe the most salient element in the partition *house* within the most recent perceptual domain of type *house* as the referent for the expression For deictic definite descriptions which contain an adjectival description of the referent, the selection procedure consists of searching the domain for a basic partition whose differentiation criterion matches the adjectival description and selecting the most salient element in that partition For instance, under a deictic interpretation, the referent ascribed to the expression *the red house* would be the most salient element in the *red* partition within the most recent perceptual domain of type *house* If, however, the domain does not have a partition whose differentiation criterion matches the adjectival description, a partition of this type is created For example, if the system is deictically interpreting the expression *the tall blue house*, and the chosen reference domain does not contain a partition with the differentiation criterion *tall, blue,* the system creates a partition of this type within the domain by copying the partition *blue*[64] and reorganising the copied partition's elements based on their height and then salience Once this partition has been created, the first element in this domain will be the tallest blue house[65] and the selection procedure will take this element to represent the referent of the expression

A final point about the structure of the reference domains and function of the partitions in these domains if there is more than one element in the partition that the referent of the expression is being extracted from, then the expression is ambiguous, i e ,

---

[64] The existence of a partition within the domain that matches the colour adjective within the object description is guaranteed by reference domain selection process which has a condition that if the linguistic description contains a colour adjective the selected domain must contain a partition differentiation criterion that matches the adjective (see Section 9 4 2)

[65] If there are two or more candidates for the primary position in a partition they are sorted by salience

there is more than one object in the visual context that fulfils the linguistic description of the referent However, recall in Section 7 4 how the system framework used the visual saliency associated with the elements in the referents domain and a predefined saliency confidence interval to disambiguate references Following this, if there is more than one element in the partition that the referent of the expression is being selected from, the saliency of the primary element in the partition is compared relative to the other elements in the partition If the difference between the primary element's saliency rating and the candidate elements is not greater than or equal to a predefined confidence interval, the system informs the user that it is unable to disambiguate the reference Algorithm 9-11 gives the procedure for selecting the referent of a deictic definite description from a reference domain

```
If ∧ ((IntExp.Dialogue = VDL), (NPStr.Det = 'the')) Then

        If NPStr.Adjectives[] == ∅   Then

                If checkSalience(IntExp.RD.TYPE) Then

                        IntExp.Referent = IntExp.RD.TYPE.Elements[1]

                Else

                        //Ambiguous Reference – Output Message to User

                End if

        Else

                If ∧(( ∀i : NPStr.Modifiers[i] ∈ IntExp.RD.Partitions[j].Criterion),

                (|NPStr.Modifiers[i]| == |IntExp.RD.Partitions[j].Criterion|)) Then

                        If checkSalience(IntExp.RD.Partitions[j]) Then

                                IntExp.Referent =

                                IntExp.RD.Partitions[j].Elements[1]

                        Else

                                //Ambiguous Reference – Output Message to User

                        End If

                Else

                        i = createPartition(IntExp.RD, NPStr)

                        If checkSalience(IntExp.RD.Partitions[i]) Then

                                IntExp.Referent =

                                IntExp.RD.Partitions[i].Elements[1]

                        Else

                                //Ambiguous Reference – Output Message to User

                        End If

                End If

        End If

End If
```

Algorithm 9-11: The algorithm for selecting the referent of a deictic definite description from a reference domain. For a definition of the terms used in the algorithm see Appendix A.

336

9 4 3 2 Anaphoric Definite Descriptions

A definite description is only interpreted anaphorically if there is a profiled element in the LDL[1] reference domain whose object fulfils the linguistic description of the referent in the expression and is still visible in the view volume. Accordingly, the referent ascribed to an anaphoric definite description is the object in the view volume which matches the description of the expression's referent and whose element is currently profiled. Algorithm 9-12 lists the algorithm for selecting a referent for an anaphoric definite description.

If $\wedge$ (((IntExp Dialogue = LDL), (NPStr Det = '*the*'), (NPStr Modifiers[] $\cap$

{'*other*'} == $\varnothing$), (NPStr Head $\neq$ {'*one*'})) Then

      {x $\wedge$ ((x ϵ IntExp RD Profiled), (x Visible = TRUE), (fulfils(x)))}

      IntExp Referent = x

End If

Algorithm 9-12: The algorithm for selecting the referent of an anaphoric definite description. For a definition of the terms used in the algorithm see Appendix A.

9 4 3 3 Indefinites

The selection process for deictic indefinite descriptions is similar to that for deictic definite description. The major difference between the two processes is that for indefinite descriptions, the referent is arbitrarily selected from the relevant partition. As a consequence, the requirement that the salience score of the selected element should be greater than the other candidates by more than a predefined confidence interval is dropped. Algorithm 9-13 lists the formal definition for this algorithm.

```
If ∧ ((IntExp Dialogue = VDL), (NPStr Det = 'a')) Then

        If NPStr Adjectives[] == ∅   Then

                IntExp Referent = random(IntExp RD TYPE)

        Else

                If ∧(( ∀ı   NPStr Adjectives[ı] ∈ IntExp RD Partitions[j] Criterion),

                (|NPStr Adjectives[ı]| == |IntExp RD Partitions[j] Criterion|)) Then

                        IntExp Referent = random(IntExp RD Partitions[j])

                Else

                        ı = createPartition(IntExp RD, NPStr)

                        IntExp Referent = random(IntExp RD Partitions[ı])

                End If

        End If

End If
```

Algorithm 9-13· The algorithm for selecting the referent of an indefinite expression.
For a definition of the terms used in the algorithm see Appendix A.

9 4 3 4   Pronouns and Demonstratives

This thesis focuses on analysing the pronoun *ıt* and the singular primary demonstratives *this* and *that* Clearly, there are semantic differences between these three lexemes However, they share several important characteristics Firstly, they are all singular Secondly this and that are not marked with respect to gender Thirdly, none of them change the attentional focus of the discourse, ı e , they all refer to entities which are already in focus or profiled This profiling may be the result of linguistic reference, accompanying gesture, etc (see Section 9 4 1 4)

Based on these similarities, the extraction of a referent from a reference domain for pronouns and demonstratives is defined as equivalent in this thesis Indeed, the selection and profiling process for these lexemes is empty they do not change the value of the

338

profiled element in the domain or the partitions in the domain An a priori condition for this interpretation is that these expressions must be interpreted within a profiled domain of reference however, this condition is an explicit element in the selection of reference domains for these expressions The referent ascribed to them is the reference domain's profiled element

If NPStr Head ∈ {'it', 'this', 'that'} Then

    IntExp Referent = IntExp RD Profiled[1]
End If

Algorithm 9-14: The algorithm for selecting the referent for the pronoun *it* or either of the singular demonstratives. *this*, *that*. For a definition of the terms used in the algorithm see Appendix A.

9 4 3 5   One-Anaphora

The genericness of the pronoun *one* means that it carries a minimal amount of information that can be used to resolve its referent Consequently, if the expression contains an adjectival description of the referent, this description is the primary source of information designating the referent In these cases, the referent of these types of expressions is the most salient element in the reference domain whose attributes matches the supplied adjectival description As profiled elements are by definition more prominent than unprofiled elements, if the domain contains a profiled element that matches the adjectival description in the expression, it is selected as the referent If more than one profiled element matches the linguistic description of the referent, then the element within this set with the highest saliency is selected as the referent with the condition that the difference between its saliency and the saliencies of the other profiled elements that fulfil the linguistic description is equal to or greater than the predefined confidence interval

339

If, on the other hand, the domain does not contain any profiled elements or none of the domain's profiled elements match the description, then the referent is the most salient element in the domain that matches the adjectival description This element is found by searching for a partition whose differentiation criterion matches the adjectives in the expression If a partition is found, the most salient element in the partition is selected as the referent with the condition that the difference between this element's saliency and the saliencies of the other elements in the partition equals or exceeds the predefined confidence interval If, however, the domain does not have a partition whose differentiation criterion matches the adjectival description, a partition of this type is created, populated with the elements of the domain that fulfil its differentiation criterion, and sorted Once this partition has been created, the first element in this domain will be the fittest element in the domain with respect to the adjectival description in the expression The selection procedure takes this element to represent the referent of the expression if the difference between its saliency and the saliencies of the other elements in the partition exceeds the predefined confidence interval If this condition is not met the reference is deemed to be ambiguous

If no adjectival description is provided in the expression, the referent is the most salient element in the reference domain Again, profiled elements are taken as more prominent than unprofiled elements Hence, the referent of a one-anaphora reference which does not contain an adjectival description of the object it denotes is the reference domain's profiled element which has the highest salience, or if there is no profiled element in the domain, the most salient element in the reference domain's TYPE partition Again, the selection of these referents is subject to the condition that the difference in saliency ascribed to the element selected as the referent and the other elements which fulfil the linguistic restrictions on the referent exceeds the predefined confidence interval

```
If ∧ ((NPStr.Det = 'the'), (NPStr.Modifiers[] ∩ {'other'} == ∅ ), (NPStr.Head == 'one')) Then

        { x : ∧ ((x ∈ IntExp.RD.Profiled[]), (fulfils(x.Object)) }

        Let x[1].Object.Salience >= x[2].Object.Salience …>= x[n].Object.Salience

        If x ≠ ∅  Then

                If checkSalience(x) Then

                        IntExp.Referent = x[1]

                Else

                        //Ambiguous reference – output message to user

                End If

        Else

                If NPStr.Adjectives[] = ∅  Then

                        If checkSalience(IntExp.RD.TYPE) Then

                                IntExp.Referent = IntExp.RD.TYPE.Elements[1]

                        Else

                                //Ambiguous reference – output message to user

                        End If

                Else

                        If ∧(( ∀i : NPStr.Adjectives[i] ∈ IntExp.RD.Partitions[j].Criterion),

                        (|NPStr.Adjectives[i]| == |IntExp.RD.Partitions[j].Criterion|)) Then

                                If checkSalience(IntExp.RD.Partitions[i]) Then

                                        IntExp.Referent = IntExp.RD.Partitions[i].Elements[1]

                                Else

                                        //Ambiguous reference – output message to user

                                End If

                        Else

                                i = createPartition(IntExp.RD, NPStr)

                                If checkSalience(IntExp.RD.Partitions[i]) Then

                                        IntExp.Referent = random(IntExp.RD.Partitions[1])

                                Else

                                        //Ambiguous reference – output message to user

                                End If

                        End If

                End If

        End If

End If
```

Algorithm 9-15: The algorithm for selecting the referent from the reference domain for a one-anaphora referring expression. For a definition of the terms used in the algorithm see Appendix A

9 4 3 6   Other-Anaphora

As noted in Section 9 4 2, the definite description modifier *other* designates an entity that has been excluded from a specified grouping Following this, in the SLI framework, the term *other* is interpreted as designating an element in a domain that has been excluded from a set of one or more profiled elements The specification of the exclusive grouping often occurs in the referential expression preceding the other-anaphora expression Figure 9-18 illustrates the creation of a profiled grouping by a definite description *the N* Note that by creating this grouping, a second set of objects, the elements of which have been excluded from the grouping, ~ *the N*, is also created Importantly, the set, ~ *the N*, can be divided into objects of type N, which were not profiled due to their relative salience within the set of objects of type N, and objects of type ~N

$\sim N = \{\ e1\ \}$

Universe of Discourse

e1

e4

$N == \{\ e2,\ e3,\ e4,\ e5\}$

Exclusive grouping '*the N*' == {e4}

$\sim the\ N == \{\ e1,\ e2,\ e3,\ e5\ \}$

$\sim the\ N \cap N == \{\ e2,\ e3,\ e5\ \}$
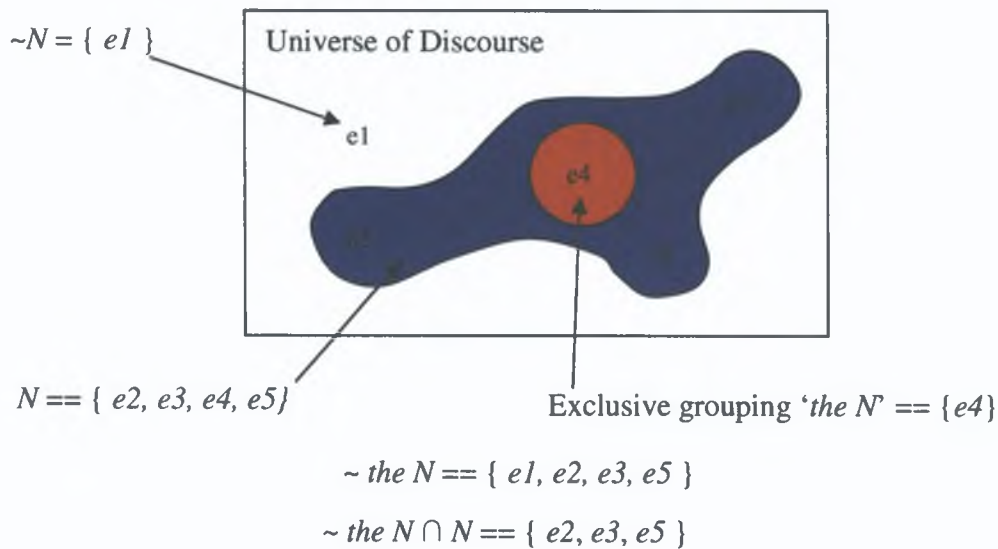
Figure 9-18: The sets created by processing the expression *the N*. Element e4 was selected as the referent for the expression.

Following McCawley's (1993) definition of *other* as the function $\lambda y.\ \sim(y = x)$, for a given x (see Section 9.4.1.3), the referent of an other-anaphoric expression such as *the other* [*N* / '*one*'] should be extracted from the set ~ *the N*; i.e., { *e1, e2, e3, e5* }. However, it is evident that *e1* is not a suitable referent as it does not fulfil the type restriction N[66]. Accordingly, the referent for an other-anaphoric expression such as *the other* [*N* / '*one*'], should be extracted from the set ~ *the N* $\cap$ *N*; i.e., { *e2, e3, e5* }. The process of selecting the referent from this set is driven by the saliencies associated with each of the set's elements. As with previous forms of reference, a requirement for this selection process is that the saliency of the primary element should exceed the saliency of each of the other elements by a predefined confidence interval (see Section 7.4). Algorithm 9-16 lists the initial algorithm for the selection of a referent for an other-anaphora expression from a reference domain.

---

[66]In the context of interpreting linguistic input to 3-D simulations the token *one* picks up the type information from the preceding utterance, see Section 9.4.1.2. Consequently, the type restriction *N* applies to the referent of a referring expression *the other one* where *one* has substituted for *N*.

343

```
If ∧ ((NPStr.Det = 'the'), (NPStr.Modifiers[] ∩ {'other'} ≠ ∅ )) Then

        Let N = NPStr.head

        Let IntExp.RD.Partitions[i] == ~ the N ∩ N

        If checkSalience(IntExp.RD.Partitions[i]) Then

                IntExp.Referent = IntExp.RD.Partitions[i].Elements[1]

        Else

                //Ambiguous reference – output message to user

        End If

End If
```

Algorithm 9-16: The initial algorithm for the selection of a referent from a reference domain for an other-anaphoric expression. This algorithm assumes that the sets *the N*, ~ *the N*, *N*, and ~*N* are defined analogously to Figure 9-18. The existence of a partition equivalent to ~ *the N* ∩ *N* within the reference domain IntExp.RD is guaranteed as the reference domain selection algorithm – Algorithm 9-10 – requires that the reference domain has one or more profiled elements and that there is at least one element in the domain's TYPE partition that fulfils the linguistic restrictions of the utterance. Note that for an other-anaphoric expression, it is assumed that the type restrictions defined by NPStr.head are equivalent to those defined by NPStr-1.head. For a definition of the terms used in the algorithm see Appendix A.

The impact of Algorithm 9-16 on the interpretation process can be illustrated by using Figure 9-20 as the visual context and by assuming that the user entered the following sequence of commands:

(25a) *Make the red house taller.*
(25b) *Make the other house blue.*

The referring expression in (25a), *the red house*, creates a profiled grouping which contains one element which is the object selected as the referent for the expression; i.e.,

344

the red house. However, in creating this profiled grouping, it also implicitly creates a grouping of the objects which have not been profiled; i.e., the set of houses in the context that are not the red house. In this instance, this unprofiled grouping contains one element, the green house, which is of type house.

$\sim N == \varnothing$

Universe of Discourse

{the red house}

$N == \{\text{ the green house, the red house }\}$ the $N == \{\text{ the red house }\}$

$\sim$ the $N = \{\text{ the green house }\}$

$\sim$ the $N \cap N = \{\text{ the green house }\}$
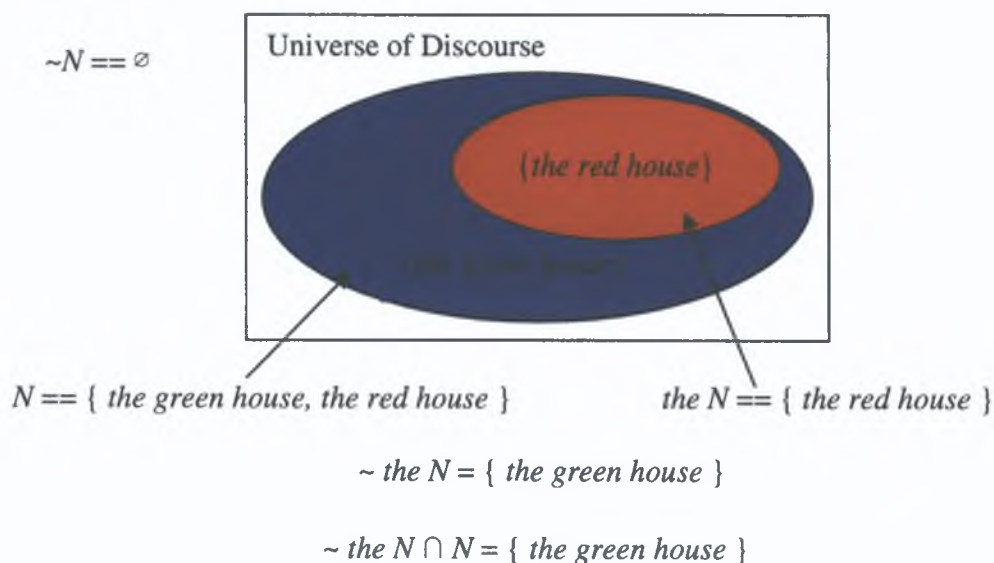
Figure 9-19: The sets created by interpreting the referring expression *the red house* **in the context supplied by Figure 9-20.**

The use of the nominal modifier *other* in the definite description in (25b), *the other house*, specifies that the referent of this expression should be extracted from the set of objects which have not been profiled by the preceding utterance; i.e., the set ∼ *the N*. However, as a result of the type restriction *N*, the set of candidate referents can be restricted to ∼ *the N* ∩ *N*. In this instance, the sets ∼ *the N* and N are equivalent; however, as Figure 9-18 illustrates this is not always the case. In this example | ∼ *the N* ∩ *N* | == 1; consequently, the selection of the referent from this set is trivial. Figure 9-21 illustrates the state of the visual scene after the sequence of commands (25a) and (25b) have been interpreted.

345
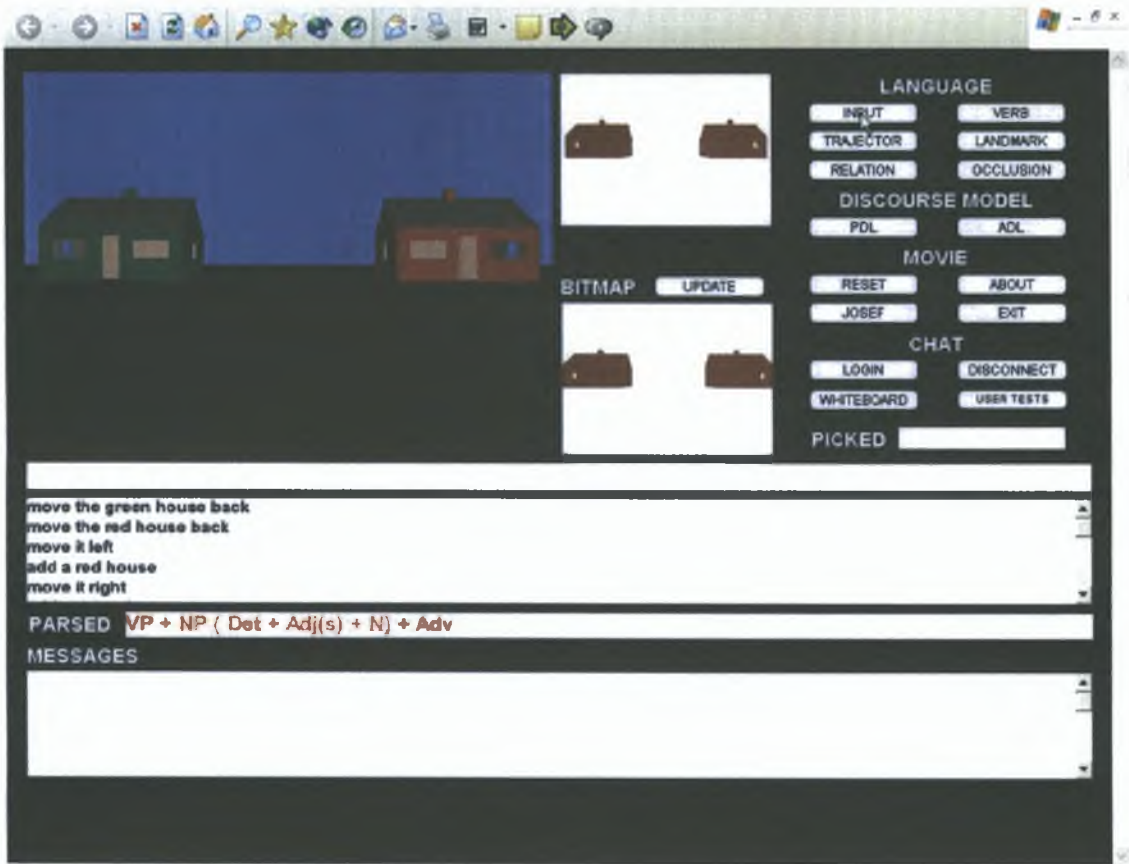
Figure 9-20: The initial visual context for a simple other-anaphora resolution example.
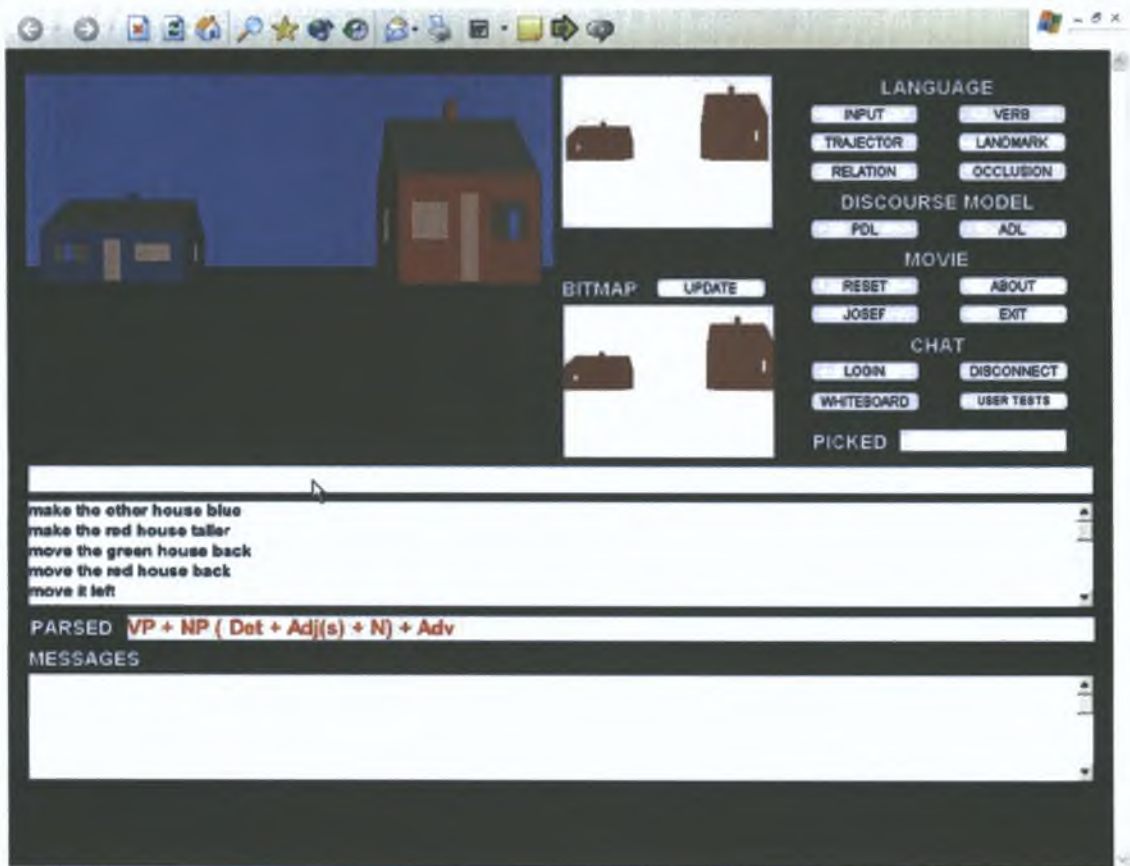
Figure 9-21: The state of the simulation after the system has interpreted the command *make the other house blue*.

In the preceding example, it was noted that the type restriction specified by the head of the nominal, *N*, in an other-anaphoric expression *the other N*, or in the utterance preceding an other-anaphoric expression *the other one*, impacts on the selection of the referent by requiring that it should be of type *N*. A consequence of this was that the set of elements excluded from a profiled grouping could be subdivided into those that meet the linguistic restrictions of the preceding utterance and those that did not. Moreover, the elements that meet the linguistic restrictions of the preceding utterance were more suitable as referents for an other-anaphoric expression than those that did not. However, so far, only the type restrictions that impact on the selection of the referent have been considered. The question now addressed is: what impact does an adjectival description in the referential expression preceding an other-anaphoric expression have on the interpretation of the other-anaphora?

Let *the X N* symbolise a definite description, where *X* represents an adjectival description, and *N* the head noun of the expression, which defines the type restriction on the referent. Figure 9-22 illustrates the sets created by the utterance *the X N*.

$\sim N == \{\ e1\ \}$

Universe of Discourse

the X N == { *e4* }

e1

e4

$N == \{\ e2, e3, e4, e5\ \}$

$X N = \{\ e3, e4, e5\ \}$

$\sim the\ X\ N = \{\ e1, e2, e3, e5\ \}$

$\sim the\ X\ N \cap X\ N = \{\ e3, e5\ \}$
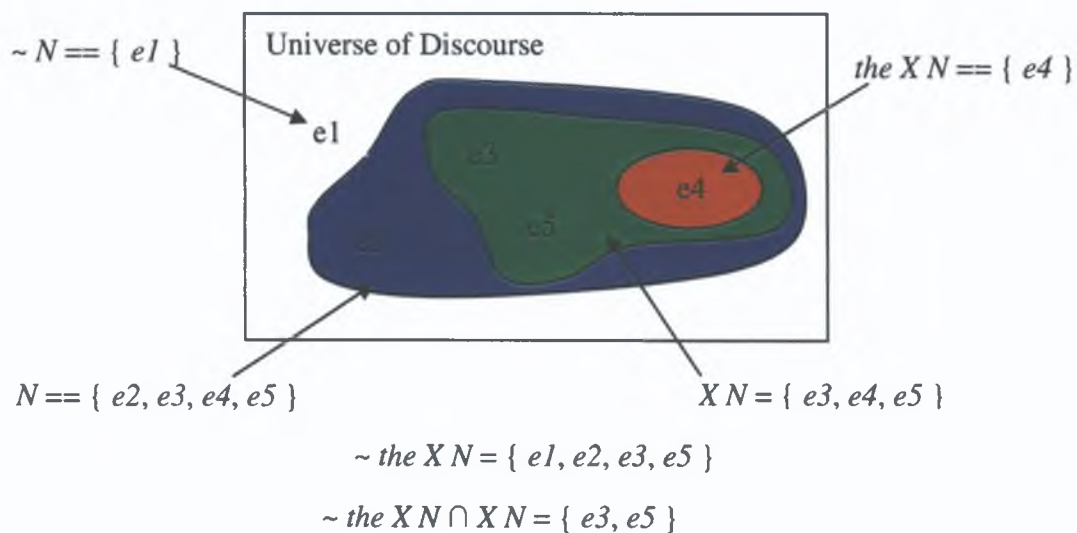
Figure 9-22: The set created by the referring expression *the X N*, where X is an adjectival description and *N* symbolises the head noun of the expression. In this figure, e4 represents the object that was selected as the referent for the expression, e3 and e5 represent objects that fulfilled both the type restriction specified by *N* and the adjectival restrictions specified by *X*. The selection of e4 as the referent in preference to e3 or e5 would have been driven by the saliency ratings associated with these elements. The element e2 represents an object that fulfils the type restriction but not the adjectival restrictions and the element e1 represents an object that does not fulfil the type restriction.

Given the context illustrated in Figure 9-22, how should an other-anaphoric expression such as *the other* [*N* / *'one'*] be interpreted? As in the preceding example, and following McCawley's (1993) definition of the *other* as the function $\sim= x$ (see Section 9.4.1.3) the referent of an other-anaphoric expression such as *the other* [*N* / *'one'*] should be extracted from the set ~ *the N*; i.e., { *e1, e2, e3, e5* }. Again, however, *e1* is

not a suitable referent as it does not fulfil the type restriction N. Crucially, it should be noted that, just as $e1$ does not fulfil the type restrictions specified by N and is therefore excluded from being considered as a referent, e2 does not fulfil the adjectival restrictions specified by X. However, in the preceding example, a green house was a suitable referent for an other-anaphoric expression even though it did not fulfil the adjectival restriction *red* which was specified in the preceding utterance. This illustrates that objects which do not fulfil the adjectival restrictions of the preceding referring expression but do fulfil the type restrictions are suitable as referents for a subsequent other-anaphoric expression of type *the other* [N / *'one'*]. Nonetheless, it is evident that the adjectival restrictions X decomposes the set ~ *the X N* $\cap$ N into two sets: ~ *the X N* $\cap$ $X N$ and (~ *the X N* $\cap$ N) – $X N$. It is posited in this thesis that the objects in the set ~ *the X N* $\cap$ $X N$ are more suitable as candidate referents for an other-anaphoric expression, *the other* [N / *'one'*], than objects in the set (~ *the X N* $\cap$ N)- $X N$; i.e., in the context defined in Figure 9-22, *e3* and *e5* are more suitable as referents than *e2*.

From this, following a referring expression of the form *the X̱ N* (where X is an adjectival description and N is a noun), if there are unprofiled elements of type $X N$ in the context, the referent of an other-anaphoric expression *the other* [N / *'one'*] is selected from the set ~ *the X N* $\cap$ $X N$. However, if there are no unprofiled elements of type $X N$ in the context, ~ *the X N* $\cap$ $X N$ == $\varnothing$, the expression *the other* [N / *'one'*] is interpreted as *the other N*; i.e., the referent is selected from the set (~ *the X N* $\cap$ N) – $X N$. In either case, the process of selecting the referent from the set of candidate referents is driven by the saliency associated with each of the set's elements. As with previous forms of reference, a requirement of the saliency driven selection process is that the saliency of the primary element should exceed the saliency of each of the other elements by a predefined confidence interval (see Section 7.4). However, for other-anaphora expressions of type *the other* [N / *'one'*] where the referent is being selected from the set ~*the X N* $\cap$ $X N$ there is a further stipulation on the selection of the referent: the saliency ratings ascribed to the elements of the set (~*the X N* $\cap$ N) – $X N$ should not exceed the salience of the primary element within the set ~ *the X N* $\cap$ $X N$. This later requirement safeguards against situations where an object that fulfils the adjectival and type selection restrictions specified by the preceding utterance in a sequence of commands but has a very low visual

saliency rating might be selected as a referent. If this condition is not fulfilled, the reference is deemed to be ambiguous and the user is notified of this. Algorithm 9-17 formally defines an algorithm, which accommodates the information provided by the adjectival description supplied in the preceding utterance, into the selection process for a referent for an other-anaphoric expression from a reference domain.

```
If ∧ ((NPStr.Det = 'the'), (NPStr.Modifiers[] ∩ { 'other'} ≠ ∅ )) Then
        Let X = NPStr-1.Adjectives[] And N = NPStr-1.head
        If ~ the X N ∩ X N ≠ ∅  Then
                Let IntExp.RD.Partitions[i] == ~ the X N ∩ X N
                Let j == ( ~ theX N ∩ N ) − X N
                If ∧ ( (checkSalience(IntExp.RD.Partitions[i])),
                (∀ k ∈ j : k.Salience −
                IntExp.RD.Partitions[i].Elements[1].Object.Salience < C_Int)) Then
                        IntExp.Referent = IntExp.RD.Partitions[i].Elements[1]
                Else
                        //Ambiguous reference − output message to user
                End If
        Else
                Let IntExp.RD.Partitions[i] == ( ~ the X N ∩ N ) − X N
                If checkSalience(IntExp.RD.Partitions[i]) Then
                        IntExp.Referent = IntExp.RD.Partitions[i].Elements[1]
                Else
                        //Ambiguous reference − output message to user
                End If
        End If
End If
```

Algorithm 9-17: A refined algorithm for the selection of the referent of an other-anaphoric expression from a reference domain. This algorithm accommodates the impact of an adjectival description in the referential expression in the utterance

350

preceding the utterance containing the other-anaphoric expression, on the selection of the referent for the other-anaphoric expression This algorithm assumes that the sets *the X N*, ~ *the X N*, *X N*, *N*, and ~ *N* are defined analogously to Figure 9-22 Note that for other-anaphoric expressions the type restrictions stipulated by NPStr-1.head are assumed to be equivalent to those stipulated by NPStr.head. For a definition of the terms used in the algorithm see Appendix A

To ground the above discussion on the impact of an adjectival description in a referential expression in the utterance preceding an other-anaphoric expression, an example dialogue between a user and the SLI system is given For this example, Figure 9-23 is taken as the initial visual context and the user enters the following sequence of commands

(26a) *Make the red house taller*
(26b) *Make the other one wider*
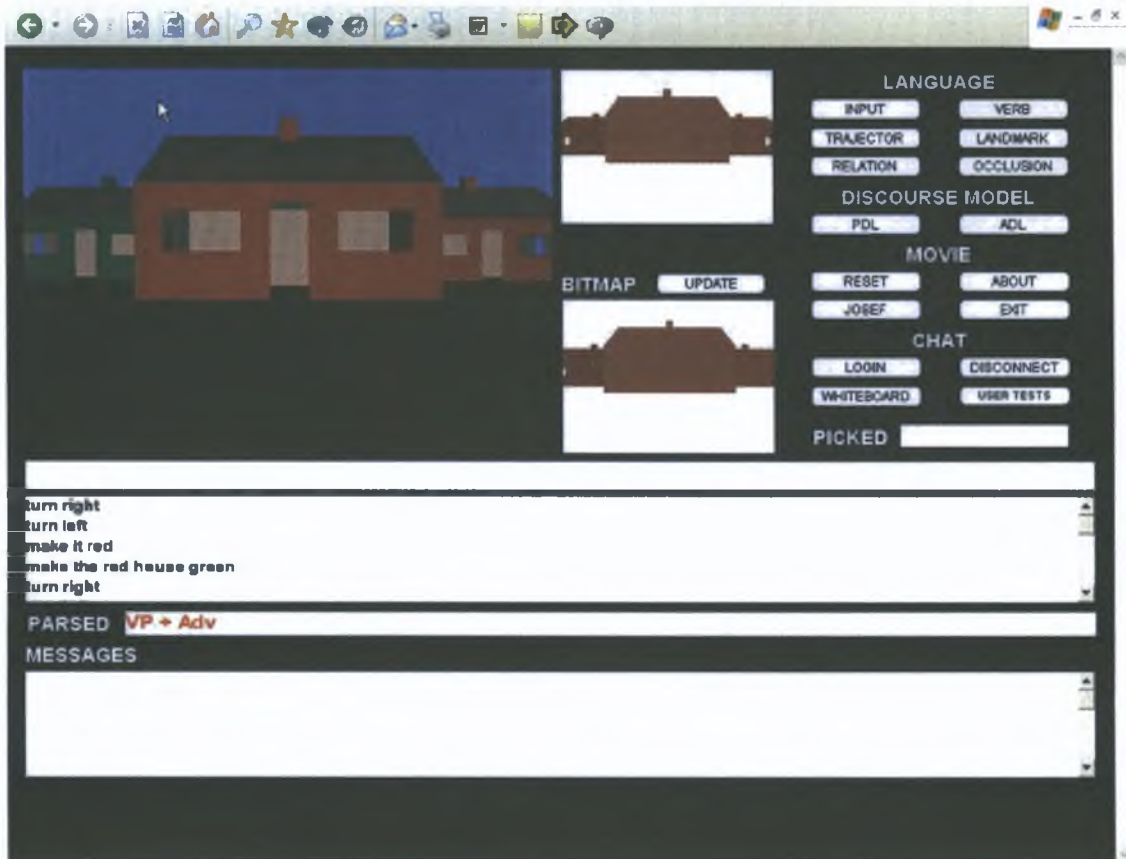
Figure 9-23: The initial visual context for a complex other-anaphora resolution example.

Note that the referring expression in (26a), *the red house*, contains an adjectival description: *red*. The use of the adjective *red* in the nominal in (26a) defines a selectional restriction on the referent of the expression and, in so doing, creates a set of possible candidates: $X N == \{ red\ house\ 1, red\ house\ 2 \}$. It is from this set that the expression's referent is selected based on its relative salience within this set[67]. For this example, it assumed that *red house 1* was the element selected as the referent for the expression. The selection of the referent creates an exclusive grouping consisting of the referent of the

---

[67] Note that the resolution of the reference *the red house* was dependent on saliency of the primary candidate referent exceeding the saliency of the other candidate referents by a predefined confidence interval (see Section 7.4). If this criterion had not been met, the system would have treated the input as ambiguous and would have informed the user of this.

expression: this is the set *the X N* == { *red house 1* }. This selection process also implicitly creates a set of objects which have not been profiled: ~ *the X N*. The number of elements in this set is two: ~ *the X N* == { *the green house, red house 2* }. This set can be further divided into the sets ~ *the X N* ∩ *X N* == { red house 2 } and (~ *the X N* ∩ *N*) − *X N*. Figure 9-24 illustrates the state of the visual context after the system has interpreted (26a).
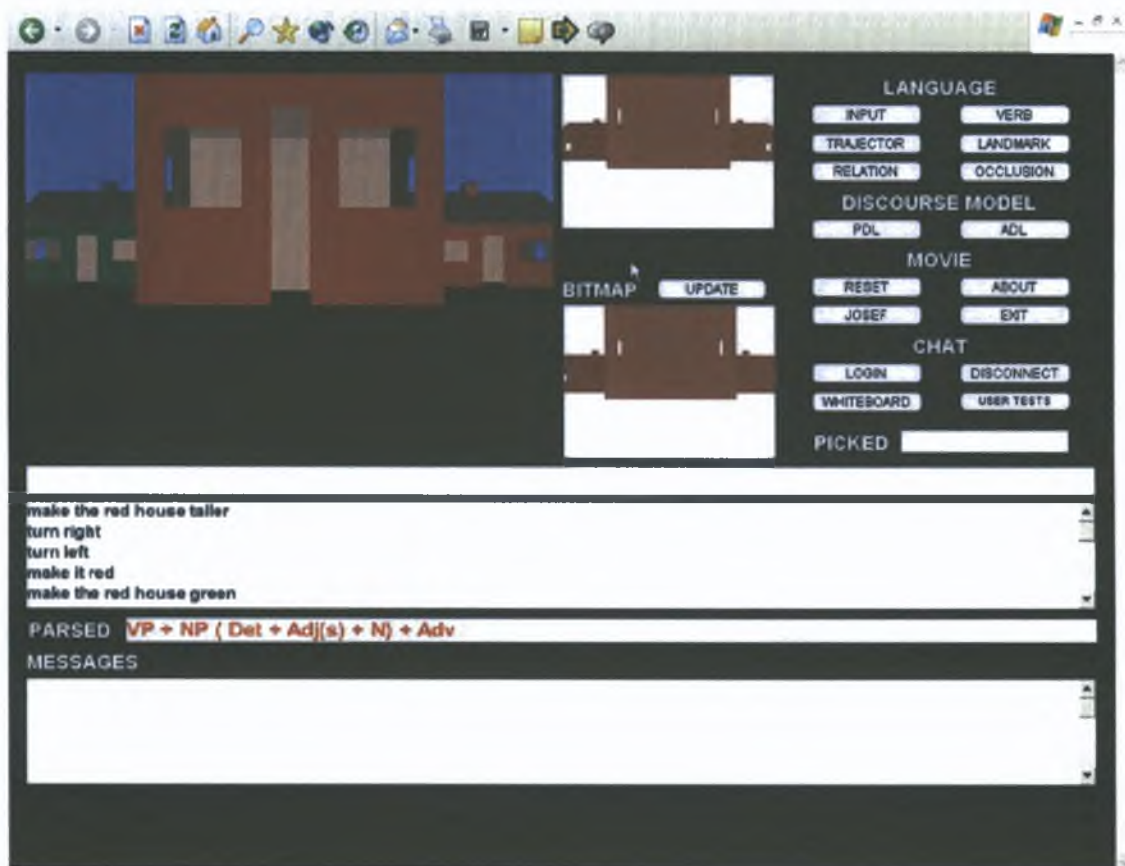
Figure 9-24: The state of the visual context after the system interpreted the command *make the red house taller*.

The utterance (26b) *Make the other one wider* denotes its referent using an other-anaphoric expression: *the other one*. It has already been noted in this section (Footnote 66 Page 343), that in the SLI scenarios the token *one* picks up the type information of the referent from the preceding utterance. Consequently, this referent for the utterance (26b) should fulfil the type restrictions defined in (26a); i.e., it should be a house. However,

353

Algorithm 9-17 states that where the set *~the X N* ∩ *X N* ≠ ∅ , the referent for an other-anaphoric expression of type *the other* [*N* / '*one*'] should be selected from this set Moreover, there are two requirements for the saliency ascribed to the primary element in this set for the reference to be successfully interpreted These were

1   The saliency of the primary element in the set *~the X N* ∩ *X N* should be exceed the saliency ascribed to each of the other elements in the set by greater than or equal to a predefined confidence interval

2   None of the saliency scores ascribed to the elements of the set (*~ the X N* ∩ *N*) – *X N* should exceed the saliency score of the primary element in the set *~ the X N* ∩ *X N* by the predefined confidence interval

In this example, the set *~the X N* ∩ *X N* == { *red house 2* } As the set *~ the X N* ∩ *X N* ≠ ∅ , the referent for the expression (26b) should be selected from this set Furthermore, as | *~ the X N* ∩ *X N* | = 1, the first saliency requirement is not applicable This leaves the requirement that the saliency of the elements in the set (*~ the X N* ∩ *N*) – *X N* == { *the green house* } should not exceed the saliency of *red house* 2 by more than the predefined confidence interval It is evident from inspection of Figure 9-24 that the saliency of *the green house* and *red house* 2 are approximately equal Consequently, *red house* 2 fulfils both the saliency requirements of the selection process and, as the only element in the set *~ the X N* ∩ *X N*, is the default choice for the referent of expression (26b)

In summary, the red house which was not selected as the referent for expression (26a) is a more likely referent for (26b) than the green house, because it fulfilled both the adjectival and type restrictions specified in (26a) Figure 9-25 illustrates the state of the visual context after the interpretation of the command (26b) *Make the other one wider* Note that because of the impact of the preceding linguistic utterance (26a), *Make the red house taller*, the command (26b) *Make the other one wider* was interpreted as *Make the other red house wider*

354
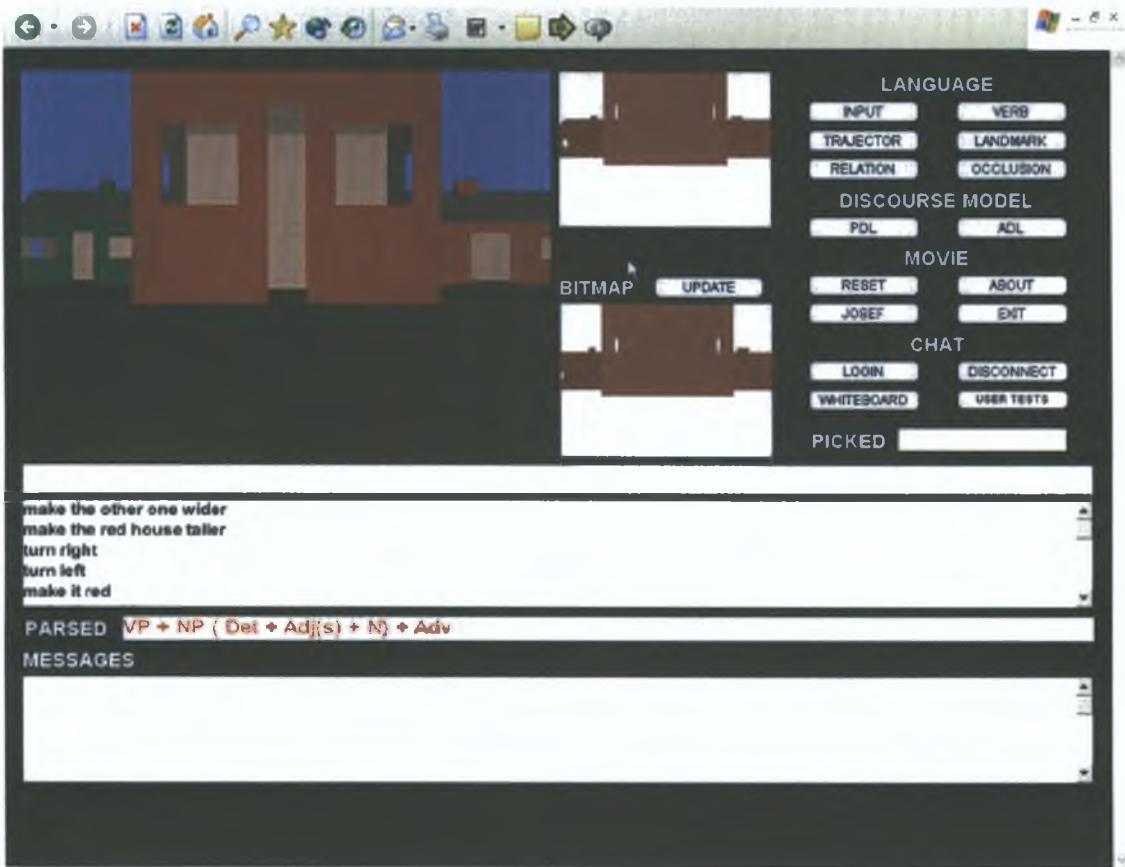
Figure 9-25: The state of the visual context after the system has processed a complex (i.e., more than one candidate referent) other-anaphora reference.

To model the impact of an adjectival description in the preceding linguistic utterance on the selection of a referent for an other-anaphoric expression, the SLI discourse framework marks the grouping used to select a referent of an expression by profiling the partition that models the domain decomposition expressed in a referring expression. This profiling mechanism will be described in greater detail in Section 9.4.4; for the current discussion, it is sufficient to be aware that the partition whose differentiation criterion matched the adjectival description supplied in the utterance whose interpretation created the reference domain is profiled. The purpose of profiling the partition is to mark it as defining the set ~ *the X N* ∩ *X N.*

The final issue considered with respect to the selection of a referent for an other-anaphoric expression is the impact that an adjectival description within the expression itself has on the selection of the referent. A priori, the referent of an expression should

355

match any adjectival description supplied in the expression Hence if the anaphoric expression contains an adjectival description Y, *the other Y* [N / *'one'*], that differs from the adjectival description X supplied in the preceding utterance, the referent of the expression should be selected from the set $\sim$ *the X N* $\cap$ $Y N$, where $Y N$ defines the set of elements of type N that fulfil the adjectival description Y If $| \sim the X N \cap Y N | > 1$, then the selection of the referent from this set is based on visual salience The only requirement attached to this selection process is that the saliency of the primary candidate should exceed the saliency of the other elements in the set by greater than or equal to a predefined confidence interval Algorithm 9-18 formally defines the SLI algorithm for the selection of a referent from a reference domain for an other-anaphoric expression This algorithm accommodates both the impact of an adjectival description in the referential expression in the utterance preceding the utterance containing the other-anaphoric expression and the impact of an adjectival description in the other-anaphoric expression, on the selection of the referent for the other-anaphoric expression

```
If ∧ ((NPStr.Det = 'the'), (NPStr.Modifiers[] ∩ {'other'} ≠ ∅ )) Then

        Let Y = NPStr.Adjectives[] And X = NPStr-1.Adjectives

        Let NPStr-1.head == NPStr.head And N =NPStr.head/NPStr-1.head

        If ∧ (( Y ≠ ∅ ), ( Y ≠ X ))Then

                Let IntExp.RD.Partitions[i] = ~ the X N ∩ Y N

                If checkSalience(IntExp.RD.Partitions[i]) Then

                        IntExp.Referent = IntExp.RD.Partitions[i].Elements[1]

                Else

                        //Ambiguous reference – output message to user

                End If

        Else

                If ~ the X N ∩ X N ≠ ∅  Then

                        Let IntExp.RD.Partitions[i] == ~ the X N ∩ X N

                        Let j == ( ~ the  X N ∩ N ) – X N

                        If ∧ ( checkSalience(IntExp.RD.Partitions[i]),

                        ( ∀ k ∈ j : k.Salience – IntExp.RD.Partitions[i].Elements[1] < C_Int) )

                        Then

                                IntExp.Referent = IntExp.RD.Partitions[i].Elements[1]

                        Else

                                //Ambiguous reference – output message to user

                        End If

                Else

                        Let IntExp.RD.Partitions[i] == ( ~ the X N ∩ N ) – X N

                        If checkSalience(IntExp.RD.Partitions[i]) Then

                                IntExp.Referent = IntExp.RD.Partitions[i].Elements[1]

                        Else

                                //Ambiguous reference – output message to user

                        End If

                End If

        End If

End If
```

Algorithm 9-18: The SLI algorithm for the selection of a referent from a reference domain for an other-anaphoric reference. This algorithm assumes that the set *the X N*, ~ *the X N*, *X N*, *N*, and ~ *N* are defined analogously to Figure 9-22. For a definition of the terms used in the algorithm see Appendix A.

9 4.4   Profiling an Element

Once the referent of an expression has been selected, the final stage of interpreting a nominal expression is to profile the referent This profiling of the referent restructures the reference domain Algorithm 9-19 lists the different steps in this process Once the referent has been profiled and the domain has been restructured, the final step in the interpretation process is the insertion of the new domain at the top of the LDL

---

1   A new reference domain is created which has the same partitions as the original reference domain that was used as the local context during the selection of the referent

2   Copies of the element(s) that represent the referent(s) in the original domain are stored in the profiled elements list in the new domain

3   Copies of all the other elements in the original domain are copied into the new domain's partitions

4   The partition that modelled the decomposition of the domain that was used to intend on the profiled object is also profiled Profiling the partition used in *the selection of the referent has the advantage of explicitly marking the grouping used to designate the profiled element(s)* As noted in Section 9 4 3 6, this information is particularly useful when interpreting other-anaphora expressions

---

Algorithm 9-19: The algorithm for profiling the referent of an expression in a reference domain.

9.5   Grammatical Constructions

The interpretation process described in Section 9 4 models the semantics of nominal expressions However, there are many types of referring expressions which are more complex than nominal expressions Cognitive grammar defines a grammatical class called relational expressions to describe these more complex expressions Although the class of relational expressions is divided into three types, simple atemporal relation, complex atemporal relation, and process[68], the analysis in this thesis is restricted to describing simple atemporal relations These relations describe a stative relation that is, they describe a consistent relationship between two or more conceived entities In particular, nominal expressions with a prepositional phrase modifier and coordinating expressions will be focused on here Example (27a) and (27b) illustrates the types of simple atemporal relation examined

(27a) *The tree to the right of the house*

(27b) *The tree and the blue house*

One of the fundamental observations of cognitive grammar is that the semantics of an expression may presuppose the semantics of another expression For example, the semantics of a complex expression such as *the lamp above the table* presupposes the semantics of the expressions *the lamp* and *above the table* This observation introduces the notion of a hierarchy of domains where some domains may be included as components of others

While admitting that semantics is not fully compositional, Langacker posits that there are "conventional patterns of composition that determine central aspects of a 'composite structure's' organisation" (1991b pg 25) Composite structures describe the semantics of complex expressions such as *the lamp above the table* They are created by the integration of simpler component structures, e g , *the lamp* and *above the table* The

---

[68] See Section 3 2 for definitions of complex atemporal relations and processes

359

conventional patterns of composition that guide this integration process are called constructional schemas

The role of constructional schemas is implemented in the SLI discourse framework by a **grouping operation** The function of the grouping operation is to create complex domains by integrating two or more existing domains and to make these complex domains available to the interpretation process The grouping operation may be triggered by relational expressions such as prepositions or coordination occurring in the discourse, or perceptual factors for example grouping objects using the principles of gestalt theory (see Section 2 2 2) Here, this thesis focuses on grouping that has been triggered by discursive factors, i e , when a linguistic input contains a prepositional phrase or a coordinating expression Algorithm 9-20 lists the steps in the SLI grouping algorithm

1 Create a new domain
2 Set the type of this domain to the type of elements described by the complex expression If the expression describes elements of different types, set the domain name to the generic marker *thing*
3 Create a partition whose differentiation criterion is set to the complex expression This partition is called the **complex expression partition**
4 Insert elements into this partition that match this criterion, order them by fitness and salience The set of elements that match this criterion is defined as the set of elements within the domains created by the component expressions of the complex expression that fulfil the differentiation criterion of the new composite domain's TYPE partition
5 Profile the referent or referents of the complex expression The referents of the complex expressions are either the profiled elements of the domains created by the processing of the complex expression's component expressions or, if one of the component expressions designates an area (rather than an object), the referent of the complex expression is the first element in the complex expression partition
6 If, after the profiling of the complex expression's referents, there are elements remaining in the complex expression partition, profile the complex expression partition Otherwise, delete the complex expression partition and profile the domain's TYPE partition

Algorithm 9-20: The SLI grouping algorithm.

9.6 Updating and Integrating VDL and LDL in Discourse: A Worked Example

To illustrate how the SLI discourse framework functions, an example discourse is given As the discourse model is designed to be used in applications which contain both a visual and a linguistic context, it is necessary to supply both of these contextual elements as inputs to the framework The example includes an initial visual context for the discourse, the resulting context model and examines how the framework represents the evolution of the discourse after each utterance

(28a) *Add a blue house*

(28b) *Make the red house green*

(28c) *Make the other house yellow*

(28d) *Make it blue*

(28e) *Make the blue house and the tree red*

(28f) *Make the house to the left of the tree red*

The analysis of each input includes a diagram illustrating the most recent domains in the context model at that point in the discourse For the sake of clarity, the scope of the example will be restricted to presenting only the most relevant parts of the context model The major impacts of this restriction are that only the most recent domains in the VDL and LDL will be presented and within these domains only the partitions that are directly relevant to the analysis will be presented Also, it is assumed the system can update the visual domain in response to the linguistic utterances once the referent has been profiled

9.6.1   The Initial Context

Figure 9-26 illustrates the initial visual context.

Figure 9-26: The initial visual context of the example.

Supplying this visual context as an input to the SLI discourse framework triggers the creation of two perceptual domains which are added to the VDL; since there has been no linguistic dialogue, the LDL is empty. The domains in the VDL are called *House* and *Tree*; neither domain has a profiled element. However, each contains at least one partition, their type partition with differentiation criterion set to *House* and *Tree* respectively. There is one element in each domain: *H1* in the domain *House* and *T1* in the domain *Tree*. Figure 9-27 illustrates the state of the context model after the creation and insertion of these domains. Note that, although each of the domains in Figure 9-27 contains only one partition, they may contain many more: for example, the *House* domain could contain a partition for each type of linguistic access that could be used to refer to its elements (e.g., *tall*, *wide*, etc.).

VDL	LDL
Domain· *House* **Profiled Element List.** *Null,* **Type Partition** **Differentiation Criteria·** *House* **Elements.** *H1* **Partition** **Differentiation Criteria:** *Red* **Elements·** *H1* **Domain·** *Tree* **Profiled Element List:** *Null,* **Type Partition** **Differentiation Criteria·** *Tree* **Elements** *T1* **Partition** **Differentiation Criteria:** *Green* **Elements.** *T1*	**EMPTY**
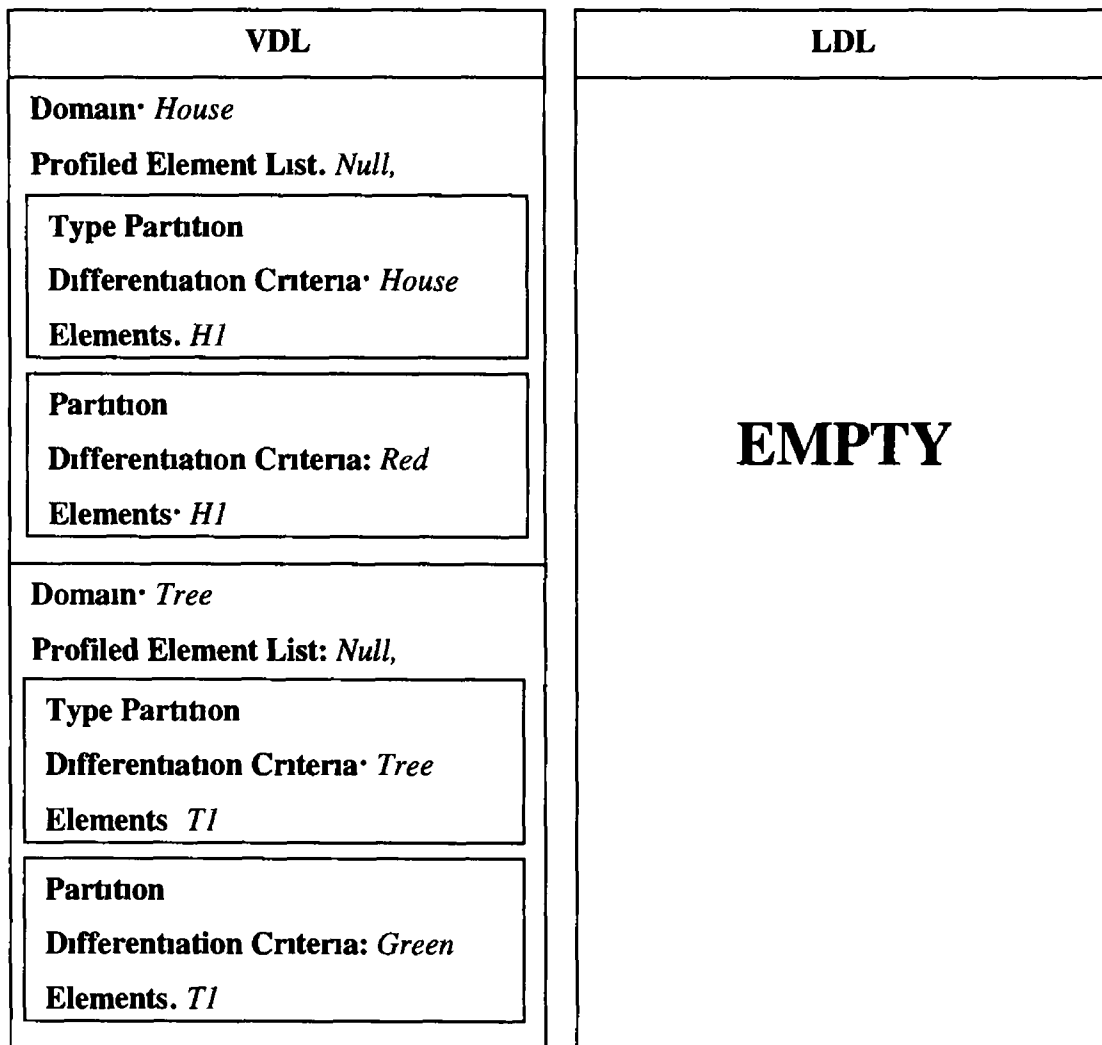
Figure 9-27 **The state of the context model after the creation and insertion of domains triggered by the visual context in Figure 9-26.**

9 6 2   Interpreting an Indefinite

The first linguistic input to the system is expression (28a) *Add a blue house* This input contains an indefinite expression *a blue house* The algorithms used in interpreting indefinite expressions are Algorithm 9-5, Algorithm 9-9, Algorithm 9-13, and Algorithm 9-19 Recall that in the context of a simulated environment an indefinite expression *a N*

may be used to arbitrarily designate one of the entities of type N in the spatio-temporal context, or to refer to the generic type N in commands that insert new objects into the simulation. It is evident from examining the input that, in this instance, the indefinite expression refers to the generic type: this deduction is based on the use of the verb *add* in the input. These types of commands are special cases as their referent must first be instantiated within the discourse domain using conceptual encyclopaedic knowledge as a blueprint before it can be processed in the discourse context model. Instantiating the expression's referent is achieved by creating an object using a model extracted from the system's model database and inserting the created object in the simulated environment. When this is done, the visual context will change to reflect the new object's presence (see Figure 9-28).

Figure 9-28: The visual context after the addition of *a blue house*.

After the new object has been inserted, the VDL automatically changes to reflect the new perceptual state of the discourse. Since the new entity is a house, the changes in the visual perceptual context will affect the structure of all perceptual *House* domains that are created after the insertion of the new square into the visual context. The *House*

365

domain will now contain two elements instead of one and the number of partitions within the domain may increase if the new element has any attributes that distinguishes it from the other elements in the domain Figure 9-29 illustrates the context model once the new element has been inserted into the visual context

VDL	LDL
Domain. *House* **Profiled Element List** *Null,* **Type Partition** **Differentiation Criteria:** *House* **Elements.** *H1, H2* **Partition** **Differentiation Criteria:** *Red* **Elements·** *H1* **Partition** **Differentiation Criteria:** *Blue* **Elements·** *H2* **Domain·** *Tree* **Profiled Element List:** *Null,* **Type Partition** **Differentiation Criteria:** *Tree* **Elements** *T1* **Partition** **Differentiation Criteria:** *Green* **Elements.** *T1*	**EMPTY**

Figure 9-29: The state of the context model after the creation and insertion of a blue house into the visual context.

366

Once the expression referent has been instantiated in the discourse, the context model must be updated to reflect the expression's occurrence This is done by inserting a restructured copy of the most recent perceptual domain matching the expression's referent type into the LDL This restructuring of the domain reflects the hearer's construal of the expression, as in Langacker's cognitive grammar (see Section 3 2), it consists of profiling the expression's referent within the domain

The first stage in this process is to select the relevant reference domain from the VDL This domain is the most recent domain in the VDL whose name matches the type description of the referent in the expression, i e , domain *House* in Figure 9-29 Once a reference domain has been selected, a copy of the reference domain is created It is this copy that will be restructured and inserted into the LDL

Having selected and copied the reference domain, the next stage in the process is to select the domain element that represents the expression's referent This is achieved by searching the domain for the partition whose differentiation criterion matches the adjectival description of the object in the expression, this partition will contain an element that represents the newly created object Once the partition has been found, the partition's element that represents the newly created object is selected as the expression's referent In this example, the adjectival description used in the expression is *blue* There is one partition in the domain whose differentiation criterion matches this description and within this partition there is only one element This element represents the newly created object, *H2* This element is selected as representing the referent of the expression

The final stage of the restructuring process is to profile the selected element and insert the restructured domain into the LDL This is done by

1   copying the selected element into the profiled element list
2   removing the profiled element from the domain's partitions (Note that any partitions that are empty after the removal of the profiled element are deleted)
3   marking the partition used to identify the selected element as profiled If the partition has been deleted, the domain's TYPE partition is profiled instead

367

Once the element is profiled, the restructured domain is inserted at the top of the LDL. Figure 9-30 illustrates the state of the context model after expression (28a) has been fully processed. The red shading of a partition in the LDL *House* domain indicates that this partition has been profiled by the interpretation process.

VDL	LDL
Domain: *House*	**Domain:** *House*
Profiled Element List: *Null, …*	**Profiled Element List:** *H2*
Type Partition **Differentiation Criteria:** *House* **Elements:** *H1, H2*	**Type Partition** **Differentiation Criteria:** *House* **Elements:** *H1*
Partition **Differentiation Criteria:** *Red* **Elements:** *H1*	**Partition** **Differentiation Criteria:** *Red* **Elements:** *H1*
Partition **Differentiation Criteria:** *Blue* **Elements:** *H2*	
Domain: *Tree*	
Profiled Element List: *Null, …*	
Type Partition **Differentiation Criteria:** *Tree* **Elements:** *T1*	
Partition **Differentiation Criteria:** *Green* **Elements:** *T1*	

Figure 9-30: The state of the context model after expression (28a) *Add a blue house* **has been fully processed.**

9 6.3   Interpreting a Definite Description

The second linguistic input to the system is (28b) *Make the red house green* This input contains a definite description, *the red house* The algorithms used in interpreting a definite description are Algorithm 9-2, Algorithm 9-9, Algorithm 9-10, Algorithm 9-11, Algorithm 9-12, and Algorithm 9-19 As there are no linguistic cues indicating that this definite description is anaphoric and the profiled element in LDL[1] does not fulfil the adjectival linguistic restrictions in this definite description, it is assumed that it is deictic Consequently, the VDL is selected as the general context for interpreting this expression

The second stage in the interpretive process involves selecting the reference domain that will act as the local context for the expression The reference domain for a deictic definite description which contains a colour adjective is the most recent domain in the VDL which contains elements that match the type specified by the head noun in the expression and consequently has a partition whose differentiation criterion matches the adjective The *House* domain in the VDL section of Figure 9-30 contains a partition with a differentiation criterion set to *red* As a result, it is a possible candidate as the reference domain for this expression For this example, it is assumed that it is the most recent domain in the VDL and hence is assigned as the reference domain for expression (28b)

One the reference domain has been selected, it is copied The referent of the expression is extracted and profiled within the copied domain For definite descriptions the process of selecting a referent within a reference domain consists of searching the domain for a partition whose differentiation criterion matches the adjectival description of the object in the expression and selecting the most salient element in that partition The domain contains one partition which fulfils the adjectival restrictions of the description, the *red* partition This partition has only one element *H1* in it, so this is selected as the expression's referent To profile the *H1* element, it is stored in the profiled elements list and extracted from the domain's partitions, and finally the partition used in the selection process is profiled Figure 9-31 gives the visual context after this command has been interpreted Figure 9-32 illustrates the context model after the interpretation

Note in Figure 9-32 that (a) the *House* domain in the VDL has changed to reflect the changes in the visual context (there is no longer a *red* partition in the domain and a

369

green partition has been added), and (b) the domain at the top of the LDL represents the linguistic context after expression (28b) has been processed. As a result, the referent of this expression is profiled. Thirdly, the domain at the bottom of the LDL represents the linguistic context before expression (28b) was uttered.

Figure 9-31: The visual context after the processing of expression (28b) *Make the red house green.*

VDL	LDL
Domain: *House*	**Domain:** *House*
Profiled Element List: *Null, …*	**Profiled Element List:** *H1*
Type Partition **Differentiation Criteria:** *House* **Elements:** *H1, H2*	**Type Partition** **Differentiation Criteria:** *House* **Elements:** *H2*
Partition **Differentiation Criteria:** *Green* **Elements:** *H1*	**Partition** **Differentiation Criteria:** *Blue* **Elements:** *H2*
Partition **Differentiation Criteria:** *Blue* **Elements:** *H2*	**Domain:** *House*
Domain: *Tree*	**Profiled Element List:** *H2*
Profiled Element List: *Null, …*	**Type Partition** **Differentiation Criteria:** *House* **Elements:** *H1*
Type Partition **Differentiation Criteria:** *Tree* **Elements:** *T1*	**Partition** **Differentiation Criteria:** *Red* **Elements:** *H1*
Partition **Differentiation Criteria:** *Green* **Elements:** *T1*	

Figure 9-32: The state of the context model after expression (28b) *Make the red house green* **has been fully processed.**

371

9 6.4 Interpreting Other-Anaphora

The third linguistic input in the example uses other-anaphora to designate its referent (28c) *Make the other house yellow* The algorithms used in interpreting an other-anaphoric expression are Algorithm 9-4, Algorithm 9-10, Algorithm 9-18, and Algorithm 9-19 In the SLI discourse framework, the discourse domain takes all the information in the perceptual domain at the time of an utterance and carries it forward in a structured manner Consequently, any input that is dependent on prior discourse, even if the expression suggests a visual context from which its referent is extracted (e g, expressions such as one-anaphora, other-anaphora (Salmon-Alt and Romary 2001)), is treated as discourse anaphoric As a result, other-anaphoric expressions are treated as discourse anaphoric in the SLI discourse framework Following this, the LDL is selected as the general context for other-anaphora expressions

As other-anaphora expressions designate an entity that has been excluded from a specified or implied grouping, the local context assigned to the expression must contain both a grouping and an excluded element Therefore, the referenced domain selected to act as the local context for the expression must contain at least one profiled element which fulfils the role of the exclusive grouping and at least one element that is not profiled and fulfils the linguistic type and attribute restrictions specified in the expression In this instance, the only restriction placed on the referent is that it must be of type *House* This means that for a reference domain to be a candidate context for this expression it must have at least one profiled element and at least one non-profiled element that represents an entity of type *House* The domain at the top of the LDL in Figure 9-32 has a profiled element Moreover, it has a non-profiled element that matches the description of the referent in the expression Consequently, this domain fulfils the conditions associated with the selection of a reference domain for other-anaphora expressions As it is the most recent domain in the LDL that fulfils these conditions, it will be selected as the expression's local context

The final part of the interpretation process may now begin the selection and profiling of the expression's referent This is achieved by creating a new domain and

372

copying the elements in the expression's local context into the new domain in a structured manner This results in a new reference domain that is a restructured version of the reference domain that was selected as the expression's local context The first stage of this process is the creation of a new blank domain In the second stage of the process, the partitions in the expression's local context are copied over into the new domain This results in a new domain that has the same partition structure as the expression's local context reference domain, but has no elements in it In the third stage of the process, the referent for the other-anaphora expression is profiled in the new domain The referent for other-anaphora expressions is the most salient element in the expression's local context reference domain's profiled partition that matches the description of the referent in the referring expression, i e , *H2* Profiling this element in the new domain involves copying it into the new domain's profiled element list In the fourth stage of the process, all the other elements in the expression's local context reference domain are copied into the new domain's partitions In the fifth stage of the process, the partition in the new domain that mirrors the partition in the expression's local context reference domain from which the element that is now profiled in the new domain was selected is profiled Finally, all the empty partitions in the new domain are deleted and the new domain is inserted at the top of the LDL Figure 9-33 gives the visual context after this command has been interpreted Figure 9-34 illustrates the context model after the interpretation It is again important to note that in Figure 9-34 the *House* domain in the VDL has changed to reflect the changes in the visual context there is no longer a *blue* partition in the domain and a *yellow* partition has been added The domain at the top of the LDL represents the linguistic context after expression (28c) has been processed As a result, the referent of this expression is profiled Finally, the domain at the bottom of the LDL represents the linguistic context before expression (28c) was uttered This domain functioned as the local context of the expression during the interpretation process

373

Figure 9-33: The visual context after the processing of expression (28c) *Make the other house yellow*.

VDL	LDL
Domain: *House* **Profiled Element List:** *Null, ...* **Type Partition** **Differentiation Criteria:** *House* **Elements:** *H1, H2* **Partition** **Differentiation Criteria:** *Green* **Elements:** *H1* **Partition** **Differentiation Criteria:** *Yellow* **Elements:** *H2* **Domain:** *Tree* **Profiled Element List:** *Null, ...* **Type Partition** **Differentiation Criteria:** *Tree* **Elements:** *T1* **Partition** **Differentiation Criteria:** *Green* **Elements:** *T1*	**Domain:** *House* **Profiled Element List:** *H2* **Type Partition** **Differentiation Criteria:** *House* **Elements:** *H1* **Partition** **Differentiation Criteria:** *Green* **Elements:** *H1* **Domain:** *House* **Profiled Element List:** *H1* **Type Partition** **Differentiation Criteria:** *House* **Elements:** *H2* **Partition** **Differentiation Criteria:** *Blue* **Elements:** *H2*

Figure 9-34: The state of the context model after expression (28c) *Make the other house yellow* **has been fully processed.**

375

9 6 5   Interpreting Pronouns

The fourth utterance in the example discourse (28d) *Make it blue* uses the pronoun *it* to designate its referent The algorithms used in interpreting the pronoun *it* are Algorithm 9-6, Algorithm 9-10, Algorithm 9-14, and Algorithm 9-19 As the pronoun *it* normally designates a single previously profiled element, the LDL is selected as the general context for the expression For a reference domain to act as a local context for this pronoun, it should contain exactly one profiled element The most recent domain in the current LDL, Figure 9-34 above, fulfils this condition Furthermore, as the referent of the pronoun *it* is usually the element already profiled in local context, the referent ascribed to it is its reference domain's profiled element, *H2* Moreover, the selection and profiling process for this pronoun consists of copying the expression's local context and inserting the copy at the top of the LDL

Figure 9-35 gives the visual context after this command has been interpreted Figure 9-36 illustrates the context model after the interpretation As with the previous utterance, the *House* domain in the VDL in Figure 9-36 has changed to reflect the changes in the visual context There is no longer a *yellow* partition in the domain and a *blue* partition has been added Focusing on the LDL, the domain at the top of the stack represents the linguistic context after expression (28d) has been processed, while the domain at the bottom of the LDL represents the linguistic context before expression (28d) was uttered As the referring expression in (28d) is the pronoun *it*, the process of selecting and profiling a referent for this expression consists of copying the expression's local context and inserting the copy into the LDL Consequently, the two LDs in Figure 9-36 – representing the expression's local context and the system's construal of the expression – are identical

376

Figure 9-35: The visual context after the processing of expression (28d) *Make it blue.*

VDL	LDL
Domain: *House*	**Domain:** *House*
Profiled Element List: *Null, …*	**Profiled Element List:** *H2*
Type Partition **Differentiation Criteria:** *House* **Elements:** *H1, H2*	**Type Partition** **Differentiation Criteria:** *House* **Elements:** *H1*
Partition **Differentiation Criteria:** *Green* **Elements:** *H1*	**Partition** **Differentiation Criteria:** *Green* **Elements:** *H1*
Partition **Differentiation Criteria:** *Blue* **Elements:** *H2*	**Domain:** *House* **Profiled Element List:** *H2*
Domain: *Tree* **Profiled Element List:** *Null, …*	**Type Partition** **Differentiation Criteria:** *House* **Elements:** *H1*
Type Partition **Differentiation Criteria:** *Tree* **Elements:** *T1*	**Partition** **Differentiation Criteria:** *Green* **Elements:** *H1*
Partition **Differentiation Criteria:** *Green* **Elements:** *T1*	

Figure 9-36: The state of the context model after expression (28d) *Make it blue* has been fully processed.

378

9 6.6   Interpreting a Coordinating Expression.

The fifth utterance in the example is (28e) *Make the blue house and the tree red*
This expression contains the coordinating conjunction *and* Conjunctions trigger the grouping operation, Algorithm 9-20 In this example, the grouping operation groups the domain formed by interpreting the definite descriptions *the blue house* and *the tree* As Section 9 6 3 has already illustrated how the SLI discourse framework processes definite descriptions, for the purposes of this example, it is assumed that the domains for the nominal expressions *the blue house* and *the tree* have already been created These domains are illustrated in the LDL stack in Figure 9-37

VDL
Domain: *House*
Profiled Element List: *Null, ...*
Type Partition   **Differentiation Criteria:** *House*   **Elements:** *H1, H2*
Partition   **Differentiation Criteria:** *Green*   **Elements:** *H1*
Partition   **Differentiation Criteria:** *Blue*   **Elements:** *H2*
Domain: *Tree*
Profiled Element List: *Null, ...*
Type Partition   **Differentiation Criteria:** *Tree*   **Elements:** *T1*
Partition   **Differentiation Criteria:** *Green*   **Elements:** *T1*

LDL
Domain: *Tree*
Profiled Element List: *T1*
Type Partition   **Differentiation Criteria:** *Tree*   **Elements:** *Null*
Domain: *House*
Profiled Element List: *H2*
Type Partition   **Differentiation Criteria:** *House*   **Elements:** *H1*
Partition   **Differentiation Criteria:** *Green*   **Elements:** *H1*

Figure 9-37: The state of the context model after the component nominal expressions *the blue house* and *the tree* of relational expression (28e) *Make the blue house and the tree red* have been processed. But before these domains have been grouped.

The first step in the grouping operation is to create a new domain. The name of this domain is set to the type of elements described by the complex expression. As this

380

expression describes elements of different types, the domain name is set to the generic marker *thing* Note that the differentiation criterion of the domain's type partition is also set to the generic marker Consequently, all the elements in the current discourse are eligible to be inserted into this partition Once the new domain has been created, the complex expression partition is created This partition's differentiation criterion is set to the complex expression itself The elements that fulfil this criterion are the set of elements within the domains created by the component expressions of the complex expression that fulfil the differentiation criterion of the new composite domain's type partition In this example, the type partition's differentiation criterion is set to the generic *thing*, therefore, all the elements within the component expression's domains are inserted into this partition Next, the referents of the complex expression are profiled As all the component expressions of the complex expression profile an object, the referents of the complex expression are the profiled elements in the component expression's domains, i e , *H2* and *T1* After profiling these elements, the complex expression partition still contains one element *H1* As it is not empty, the complex expression partition is profiled

Figure 9-38 gives the visual context after command (28e) has been interpreted Figure 9-39 illustrates the context model after the interpretation Again, the domains in the VDL have been automatically updated to reflect the changes in the visual context The domain in the LDL represents the linguistic context after the processing of the full complex expression

381

Figure 9-38: The visual context after the processing of expression (28e) *Make the blue house and the tree red.*

VDL	LDL
Domain: *House*	**Domain:** *Thing*
Profiled Element List: *Null, ...*	**Profiled Element List:** *H2, T1*
Type Partition **Differentiation Criteria:** *House* **Elements:** *H1, H2*	**Type Partition** **Differentiation Criteria:** *Thing* **Elements:** *H1*
Partition **Differentiation Criteria:** *Green* **Elements:** *H1*	**Partition** **Differentiation Criteria:** *The blue house and the tree.* **Elements:** *H1*
Partition **Differentiation Criteria:** *Red* **Elements:** *H2*	**Domain:** *Tree* **Profiled Element List:** *T1*
Domain: *Tree* **Profiled Element List:** *Null, ...*	**Type Partition** **Differentiation Criteria:** *Tree* **Elements:** *Null*
Type Partition **Differentiation Criteria:** *Tree* **Elements:** *T1*	**Domain:** *House* **Profiled Element List:** *H2*
Partition **Differentiation Criteria:** *Red* **Elements:** *T1*	**Type Partition** **Differentiation Criteria:** *House* **Elements:** *H1*
	Partition **Differentiation Criteria:** *Green* **Elements:** *H1*

Figure 9-39: The state of the context model after expression (28e) *Make the blue house and the tree red* **has been fully processed.**

9.6 7 Interpreting a Locative Expression

The final utterance in the example is (28f) *Make the house to the left of the tree red*
As with the last example, this utterance contains a complex expression, in this instance,
the locative expression *the house to the left of the tree* The two nominal expressions in
this complex relational expression, *the house* and *the tree,* are definite descriptions and
will be processed in a similar manner to example (28b) (see Section 9 6 3) For the sake
of clarity, it is assumed that the reference domains representing the interpretation of these
nominals have already been created This will allow the description of the interpretation
process to focus on the grouping operation (Algorithm 9-20) when it is applied to these
nominal reference domains

The prepositional phrase *to the left of the tree* is a complex expression whose
reference domain would be created by grouping the semantics of *the tree* and *to the left
of* A prepositional phrase *to the left of the tree* describes an area relative to the profiled
element in the reference domain created by processing *the tree* Since more than one type
of object can be in this area, the reference domain of a prepositional phrase is named
thing This admits references to all elements within the described area into the domain
Since the prepositional phrase is a complex expression, its reference domain will contain
a complex expression partition whose differentiation criterion matches the expression,
i e , *to the left of the tree* All the elements whose referents fulfil this criterion are inserted
into this partition They are sorted based on their fitness with respect to the criterion and
in the case of draws by salience Recall, that in Chapter 8, an algorithm for interpreting
projective locative expressions was developed This interpretive algorithm provides the
SLI framework with a mechanism for rating the fitness of candidate trajectors within the
semantics of a given preposition Using this algorithm, the framework can decide whether
or not an element's referent fulfils the differentiation criterion of the complex expression
partition within the domain created by the grouping process Moreover, the elements that
do fulfil the criterion can be ordered based on their fitness with respect to the criterion
This ordering of the elements allows the discourse framework's interpretive process to
extract the correct referent from the reference domain

384

There is no profiled element in this domain since the expression it represents describes an area rather than an object With respect to the nominal expression, *the house*, *H1* was selected as the referent in this instance Figure 9-40 illustrates the context model after the component expressions have been processed, but before their domains are grouped The domain at the top of the LDL represents the construal of the expression *the house*, the second domain in the LDL represents the construal of *to the left of the tree* and the domain at the bottom of the LDL represents the construal of *the tree*

VDL	LDL
Domain: *House*	**Domain:** *House*
Profiled Element List: *Null, ...*	**Profiled Element List:** *H1*
Type Partition **Differentiation Criteria:** *House* **Elements:** *H1, H2*	**Type Partition** **Differentiation Criteria:** *House* **Elements:** *H2*
Partition **Differentiation Criteria:** *Green* **Elements:** *H1*	**Partition** **Differentiation Criteria:** *Blue* **Elements:** *H1*
Partition **Differentiation Criteria:** *Blue* **Elements:** *H2*	**Domain:** *Thing* **Profiled Element List:** *Null*
Domain: *Tree* **Profiled Element List:** *Null, ...*	**Type Partition** **Differentiation Criteria:** *Thing* **Elements:** *T1, H1, H2*
Type Partition **Differentiation Criteria:** *Tree* **Elements:** *T1*	**Partition** **Differentiation Criteria:** *To the left of T1.* **Elements:** *H1*
Partition **Differentiation Criteria:** *Red* **Elements:** *T1*	**Domain:** *Tree* **Profiled Element List:** *T1*
	Type Partition **Differentiation Criteria:** *Tree* **Elements:** *Null*

Figure 9-40: The state of the context model after the component expressions *to the left of the tree* and *the house* of expression (28f) *Make the house to the left of the tree red* have been processed, and before the domains are grouped.

The first step in grouping the domains, *the house* and *to the left of the tree*, is the creation of a blank domain Recall from Section 3 2 that in sentences where a head combines with a modifier, the head's conceptual substructure is designated within the cognitive domain of the expression Inspecting the grammatical structure of (28f) reveals that the head *the house* is combining with a modifier, the prepositional phrase *to the left of* This results in the expression designating the conceptual structure of the head *the house* Consequently, this expression's domain describes elements of type *house* and the domain is named after these elements *house* This results in the composite domain's TYPE partition's differentiation criterion being set to *house* Next, the complex expression partition is created the differentiation criterion for this partition is *the house to the right of the tree* The set of elements that match this criterion is the set of elements within the domains created by the component expressions of the complex expression that fulfil the differentiation criterion of the new composite domain's TYPE partition In this example, there are two elements in this set *H1* and *H2* These elements are inserted into the partition based on their fitness with respect to the differentiation criterion In Chapter 8, a computational model for interpreting locative expressions was defined Using this model, the candidate referents *H1* and *H2* can be graded on their fitness with respect to the differentiation criterion Since the element *H2* is not to the left of *T1*, it would score zero in this model and, consequently, would not be inserted into the partition Consequently, *H1* is the only element in the partition Since one of the component domains of the expression *to the left of the tree* designates an area, the referent of the complex expression is the first element in the complex expression partition *H1* Once the referent has been selected, it is profiled It should be noted that since *H1* is the only element in the complex expression partition, the profiling of *H1* empties this partition and, consequently, the partition is deleted and the domain's TYPE partition is profiled (see step 6 Algorithm 9-20)

Figure 9-41 gives the visual context after (28f) has been interpreted Figure 9-42 illustrates the context model after the interpretation of (28f) Again, the domains in the VDL have been automatically updated to reflect the changes in the visual context The domain in the LDL represents the linguistic context after the processing of the full complex expression

387

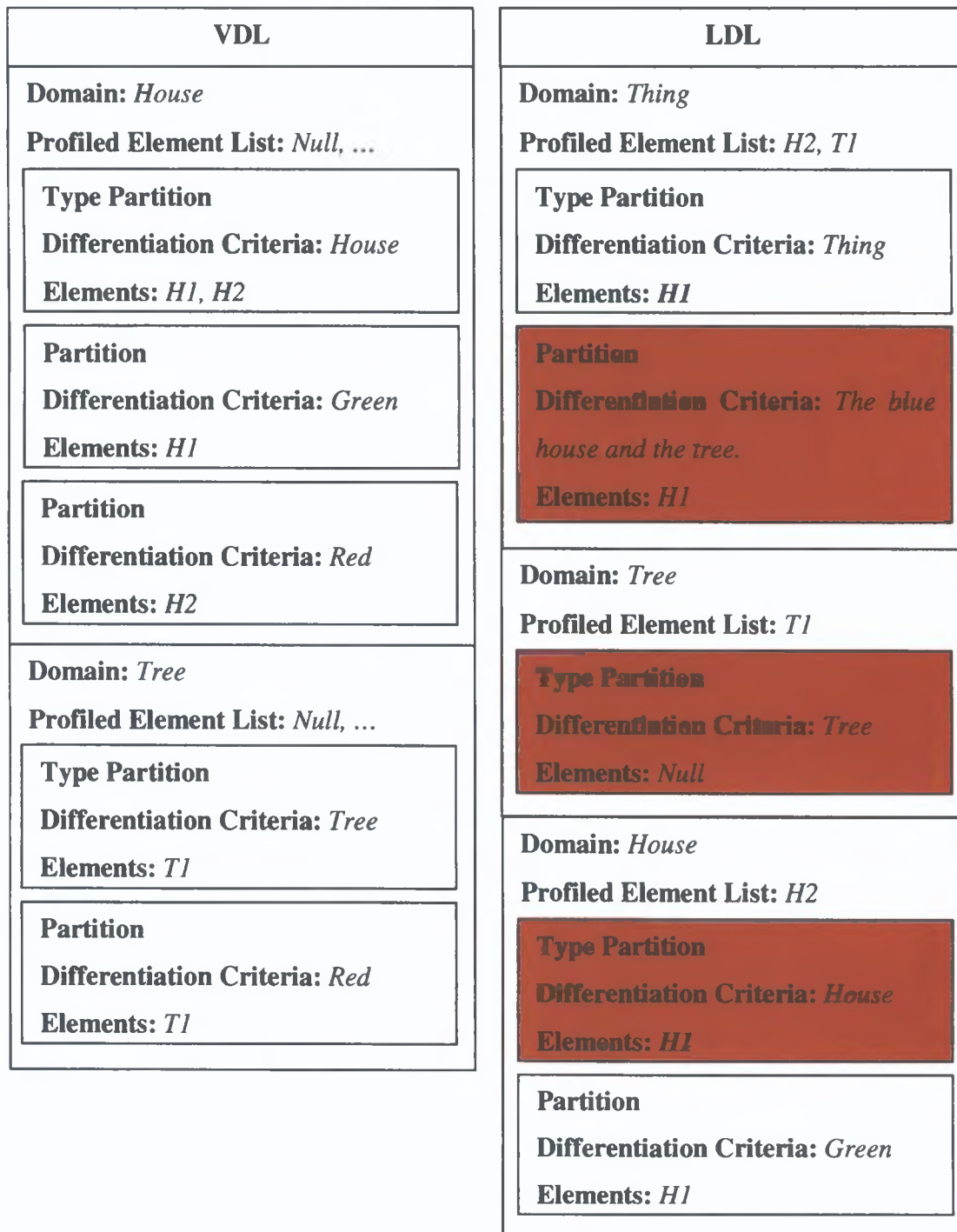**Figure 9-41: The visual context after the processing of expression (28f)** *Make the house to the left of the tree red.*

VDL
Domain: *House*
Profiled Element List: *Null, ...*
Type Partition **Differentiation Criteria:** *House* **Elements:** *H1, H2*
Partition **Differentiation Criteria:** *Red* **Elements:** *H1, H2*
Domain: *Tree*
Profiled Element List: *Null, ...*
Type Partition **Differentiation Criteria:** *Tree* **Elements:** *T1*
Partition **Differentiation Criteria:** *Red* **Elements:** *T1*

LDL
Domain: *House*
Profiled Element List: *H1*
Type Partition **Differentiation Criteria:** *House* **Elements:** *H2*

Figure 9-42: The state of the context model after expression (28f) *Make the house to the left of the tree red* **has been fully processed.**

9.7 Similarities, Differences, and Advantages

The approach adopted in designing the SLI discourse framework is inspired by the reference resolution framework described in (Salmon-Alt and Romary 2001) Below, the similarities and differences between the two frameworks are discussed

Following the work of Salmon-Alt and Romary (2001), it is assumed that the process of resolving a referential expression is achieved by accessing and restructuring domains of reference Further similarities between the two frameworks can be seen in the structure of the reference domains In both frameworks, the reference domains are divided into one or more partitions to model the possible linguistic decompositions of the domains Another similarity between the two frameworks is the association with each partition of a "differentiation criterion, which represents a particular point of view on the domain and therefore predicts a particular referential access to its elements" (Salmon-Alt and Romary 2001) Finally, both frameworks use a grouping operation to model the role of constructional schema within cognitive grammar Having noted the similarities within the context models of the two frameworks the areas where the SLI discourse framework differs from its predecessor are now highlighted

The major difference between the context models proposed by the two frameworks is that Salmon-Alt and Romary's (2001) context model has a monolithic architecture, while the SLI discourse framework divides the context model into two interacting but distinct dialogues (represented by the VDL and LDL data structures) By splitting the discourse into separate dialogues, the SLI discourse framework can explicitly incorporate visual perceptual information into the discourse Furthermore, the SLI discourse framework gives an account of when and how this perceptual information is used in resolving references and updating the context model Although the Salmon-Alt and Romary (2001) framework admits perceptual information by allowing it to trigger certain events (i e , domain creation, the grouping operation, and partitioning), their model gives no description of how to computationally gather visual perceptual information or how this visual perceptual information is to be combined with the linguistic information when resolving references

A second difference is the organisation of the domains. In the SLI discourse framework the domains are organised chronologically within the dialogues. This ordering obviates the need for a unique domain identifier. Instead, the domain can be named after the type of objects that it contains; in the case of a domain with mixed element types, the domain name is set to a generic marker *thing*. This approach has the advantage of being more cognitively plausible and less computationally complex since there no need to create unique domain names.

A third difference is the type of decomposition that the frameworks utilise. Both frameworks use partitions to model the possible decompositions of a domain. Moreover, they both associate a differentiation criterion with each partition. However, the role of these criteria differs in the two frameworks. In Salmon-Alt and Romary (2001), a partition's differentiation criterion is used to distinguish between the different elements within the partition. In contrast, the differentiation criterion of a partition in the SLI discourse framework distinguishes the elements of a partition from the elements of the domain that are excluded from the partition. The difference between these two uses can be illustrated by extending an example given by Salmon-Alt and Romary. The example is a domain that describes a group of two marbles, one red and one green. The domain for this group contains a partition with a differentiation criterion of *colour*; this partition contained pointers to both marbles. In the SLI discourse framework, the decomposition of the domain contains two partitions: one with a differentiation criterion equal to *red* and one with a differentiation criterion equal to *green*. Each of these partitions contains one element: the *red* partition would have an element representing the red marble and the green partition would have an element representing the *green* marble. The approach used by the SLI discourse framework results in a greater number of partitions within a domain relative to the model proposed by Salmon-Alt and Romary (2001). One of the advantages of this is a finer decomposition of the domain, which results in a broader coverage of the possible forms of referential access to the domain elements.

A fourth difference between the two frameworks is the representation of the elements within a partition. In Salmon-Alt and Romary's (2001) framework, the elements of a partition are pointers to representations of the sub-components of the domain. In the SLI discourse framework, the elements of a partition are comprised of a pointer similar to

391

those found in Salmon-Alt and Romary and the visual saliency of the object that the pointer describes. Associating a saliency with the elements of a domain allows the SLI framework to differentiate between the elements based on deictic visual perceptual factors as well by physical attributes. An important consequence of this is that the SLI discourse framework – in contrast to Salmon-Alt and Romary's (2001) framework – is able to interpret a definite description without requiring that the referent of the expression be uniquely identifiable by a differentiation criterion within the reference domain used as a local context for the expression (see Section 7 4)

A fifth difference between the two frameworks is the ordering of elements within the partitions. Salmon-Alt and Romary's (2001) framework gives no specification of how the elements within a partition are ordered. However, the ordering of the elements is a key component within the SLI discourse framework. An important point in this context is that in the SLI discourse framework the partitions use a last-in-first-out access policy, i e , the partitions are implemented using stacks. The default ordering process is to insert elements into a partition in ascending order based on their salience. This results in the element with the highest salience being inserted at the head of the list, i e , the first access location within the partition. This organisation reflects one of the fundamental assumptions underlying the interpretive approach of this thesis all other factors being equal, objects which have a higher visual salience are more likely to be the referents of a referring expression than objects which have a lower visual salience. This saliency-based insertion ordering is used for partitions which describe object type and colour. For other partitions which describe object size or location, elements are inserted in an ascending order based on their fitness with respect to the partition's differentiation criterion. In situations where two elements within a domain score equal with respect to the criterion of a partition, the element with the lower saliency is inserted first

The final major difference between the two models is the profiling mechanisms used by the two frameworks. Both frameworks profile elements to designate them as prominent within a domain. However, in Salmon-Alt and Romary's (2001) framework, at most one element can be profiled in a partition, while in the SLI discourse framework more than one element can be profiled within a domain. Furthermore, the partition which models the decomposition of the domain that the referring expression accessed to profile

392

the object is also profiled This profiling mechanism allows the proposed model to explicitly mark the implicit specification of a profiled group within the composite domain created by a coordinating expression The advantage of noting this information is particularly evident when processing a subsequent other-anaphora expression (see Section 9 4 4)

9.8 Chapter Summary

The function of a discourse model is to create a context which can be used to interpret language The novelty of the SLI discourse model is its integration of visual perceptual information into its context model and an explicit description of how this perceptual information is combined with the linguistic information to resolve references There are three components within this discourse model the context model, the interpretive process, and the grouping operation The context model and the interpretive process are sufficient to resolve nominal expressions, however, for more complex grammatical constructions the grouping operation is necessary For locative expressions, the grouping operation is augmented by a semantic model for prepositions This model gives the framework the ability to sort the elements representing the candidate trajectors within the grouped domain complex expression partition based on their fitness with respect to the prepositional phrase which in turn allows the discourse framework's interpretive process to profile the correct element as the referent of the expression

10 Testing the Framework

10.1 Introduction

Three computer based experiments were developed to test different aspects of the model Experiment 1 examined the impact of size on visual salience Experiment 2 examined the spatial template of the prepositions *in front of* and *behind* Experiment 3 examined whether subjects found the interpretation of anaphoric and deictic references by the SLI's discourse model reasonable

10.2 Experiment 1

There are various factors of the SLI visual salience model which are undoubtedly psychologically realistic Previous psychological research indicates that centrality and distance both impact on salience Consequently, these factors are accepted a priori However, the model also assumes that the size of an object impacts on its salience This has not been verified as a psychologically reasonable assumption The task of experiment 1 was to examine this assumption

10 2.1 Method

10 2 1 1 Subjects

14 subjects took part in this experiment 11 men and 3 women

10 2 1 2 Materials

The materials were based on four images rendered using the SLI system Each image contained two objects of the same type and colour but of different sizes Two of the images contained houses, the other two images contained trees

The effect of distance was controlled by rendering the objects at the same distance from the viewpoint of the image Furthermore, a chequered floor was used in the images to aid in depth perception A pre-test, which checked the apparent distance of the objects, was carried out on the images During this pre-test, each image was shown to 4 subjects who were asked whether they thought that either of the objects in the image was closer than the other one, or whether both objects were at the same distance In all the pre-test trials, the subjects responded that they thought both objects were at the same distance

In each image one object was drawn on the right of the image and the other object was drawn on the left of the image To control for position effects, the larger house was located on the left of one of the house images and on the right of the other house image The location of the larger tree was alternated in a similar manner For a listing of the images used in this experiment see Appendix B

Each house image was paired with each of the tree images This resulted in four sets of images Each subject was shown one set Figure 10-1 illustrates one of the images used in this experiment
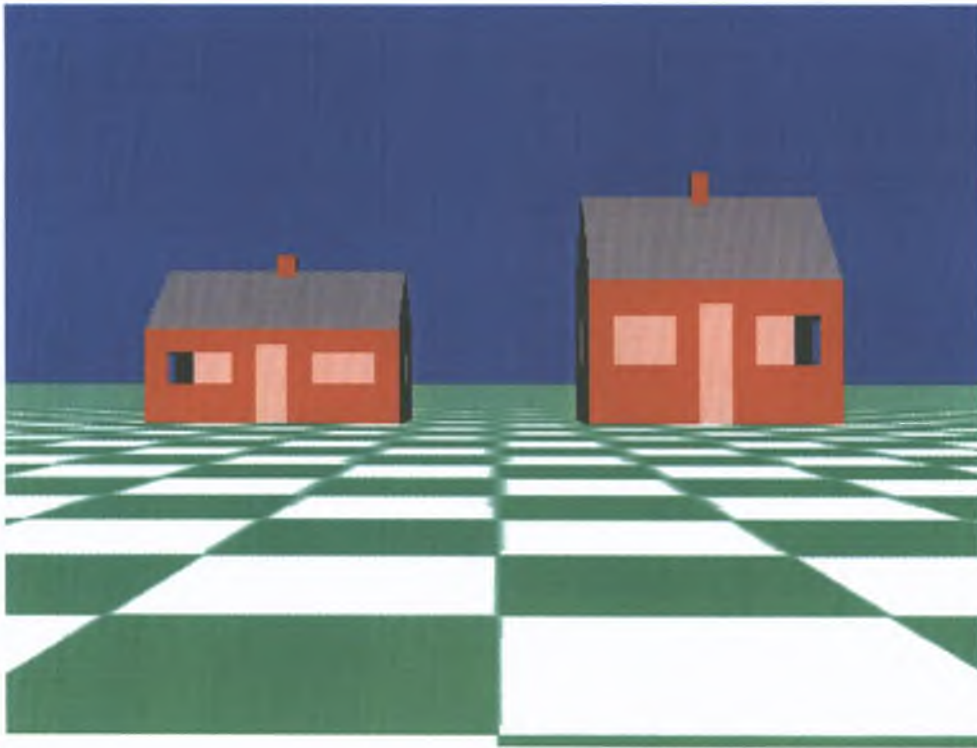
Figure 10-1: Sample image used in Experiment 1.

10.2.1.3 Procedure

Subjects were tested individually. Before each trial the subjects were presented
with a set of instructions:

1. The subjects were told that when they clicked on a button they would be
   shown a picture containing two ___. The blank was filled in with the noun
   houses, or *trees*.

2. If the image for the trial contained two houses the subjects were presented
   with the question: *If someone told you to: 'Meet me outside the house.'*
   Which house would you stand outside? However, if the trial used an image
   containing two trees the question was of the form: *If someone told you to:*
   'Meet me beside the tree.' Which tree would you stand beside?

3. Finally, the subjects were instructed to click the option button under the object in the image they felt most suitable, or click the option button labelled *Ambiguous* if they had no preference.

Figure 10-2 contains a sample of the instructions used in the experiment.
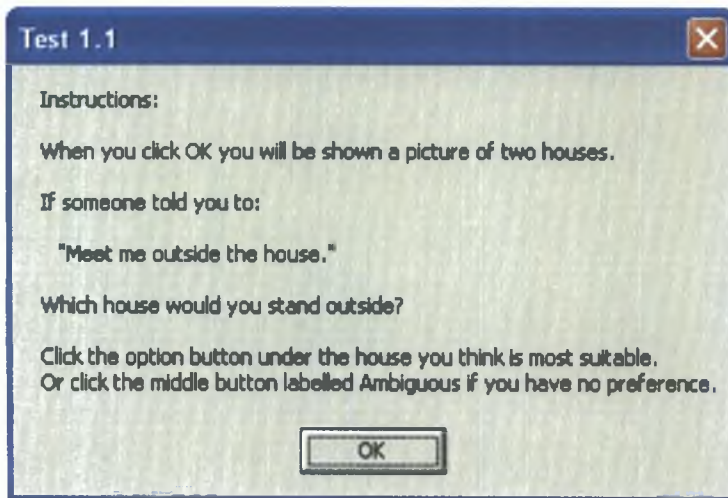
Figure 10-2: Sample set of instructions used in Experiment 1.

After the subject clicked on the button they were shown the image. Under the image were three option buttons. The middle button was labelled *Ambiguous*. Each of the other two buttons was underneath one of the objects displayed in the scene. The subject input their selection by clicking on one of the option buttons and then pressing the OK button to move on to the next trial. Figure 10-3 illustrates the form the user was presented with. In this figure the user has selected the option button under the larger house.
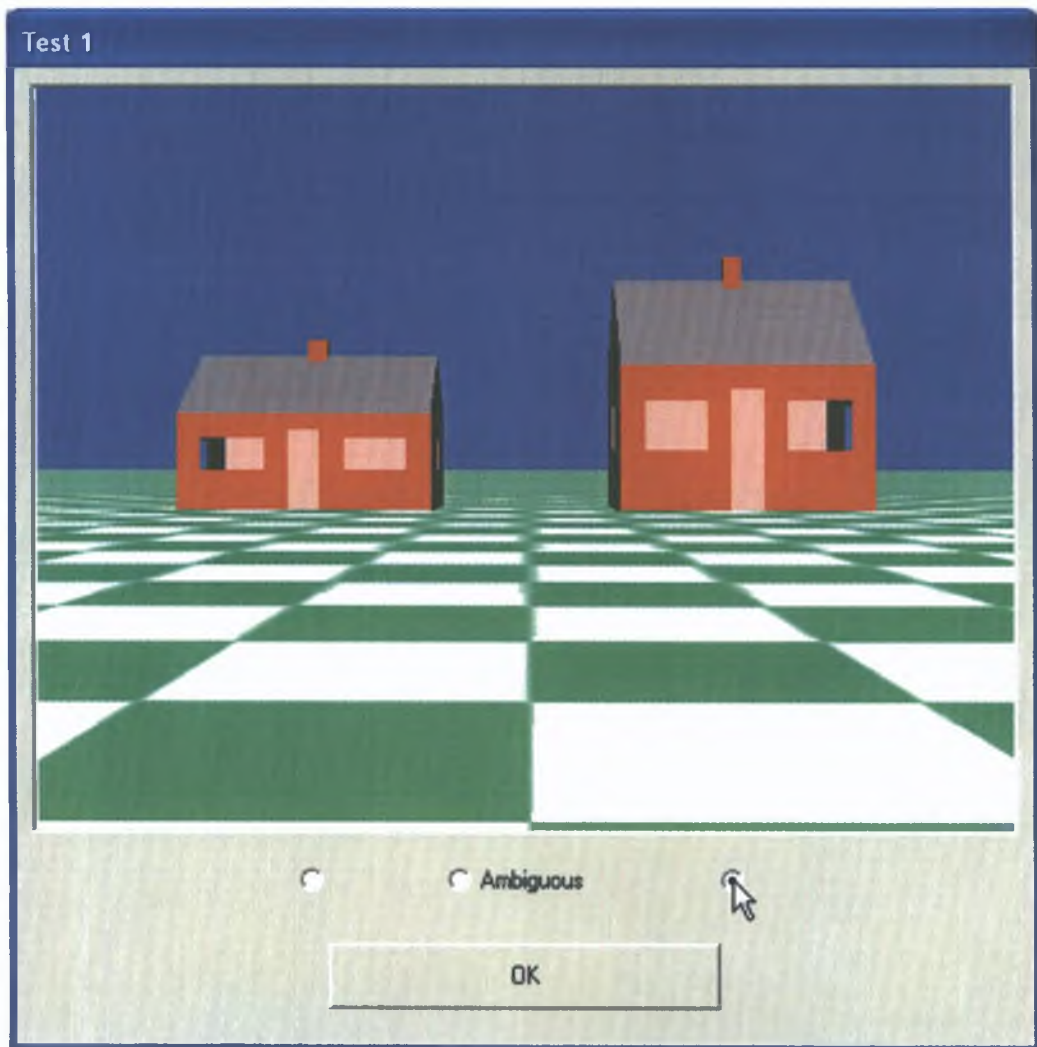
Figure 10-3: Sample of the form used to present images during Experiment 1.

10.2.2 Results and Discussion

There were 28 trials in total, 2 per subject. The data was broken down into three categories: positive, negative, and ambiguous. A positive response was one where the subject selected the larger object in the image. A negative response was one where the subject selected the smaller object in the image. An ambiguous response was one where the subject had selected the ambiguous option. 71.43% of the responses were positive, 3.57% were negative, and 25% were ambiguous. The high percentage of positive results

supports the assumption that the larger and object is the greater its salience and the more likely it is to be selected as the referent of an underspecified referring expression

10.3 Experiment 2

The psycholinguistic experiments of Carlson-Radvansky and Logan (1997) revealed that the process of selecting a frame of reference impacts on the construction of a preposition's spatial template It is important to note that, in contrast to the framework proposed here, previous NLVR systems that interpreted locative expressions (CITYTOUR (Andre *et al* 1986, Andre *et al* 1987), CSR-3-D (Gapp 1994a), SPRINT (Yamada 1993), WIP (Olivier *et al* 1994, Olivier and Tsuji 1994), Situated Artificial Communicator (Socher and Naeve 1996, Socher *et al* 1996, Vorwerg *et al* 1997, Fuhr *et al* 1998), Virtual Director (Mukerjee *et al* 2000)) neglected to account for this phenomenon

However, the work of (Carlson-Radvansky and Logan 1997) focused on the prepositions, *above*, *below*, *left*, and *right* Experiment 2 was designed to examine whether the process of selecting a frame of reference had a similar impact of the construction of the spatial templates for the prepositions *in front of* and *behind* and to examine whether there is a bias in frame of reference use for the prepositions aligned with the horizontal plane

10 3.1 Method

10 3 1 1 Subjects

13 subjects took part in this experiment 10 men and 3 women

10.3.1.2 Materials

The images used in this experiment contained a central landmark, an upright man, on a tiled surface. The tiling of the surface was designed to help depth perception in the images. The landmark was always placed in the middle of a seven by seven grid (row four, column four). In 25% of the images the landmark was facing the subject's viewpoint; i.e., the man was in a canonical orientation. In 25% of the images the landmark was facing away from the subject; i.e., the man's back was facing the subject's viewpoint. In the rest of the images, the landmark was rotated 90° around the vertical axis into a noncanonical orientation. In half of these images the rotation was clockwise – i.e., the man was facing the user's left –, and in the other half the rotation was anti-clockwise – i.e., the man was facing the user's right. The images also contained a trajector object, a red square. For each of the landmark's four orientations the trajector was placed at one of the 48 locations in the seven by seven grid surrounding the landmark. During the experiment, each image was displayed with a sentence of the form *The box is _____ the man*. The blank was filled with one of spatial relations *in front of* or *behind*. The sentence was presented under the image. The appropriate sentence was also spoken by the testing system at the beginning of each trial.

10.3.1.3 Procedure

There were 384 trials, constructed from the following variables: 4 orientations, 2 spatial terms, and 48 trajector locations. To avoid sequence effects the landmark's orientation was changed for each trial and the spatial term was alternated. Furthermore, the location of the trajector was randomly selected for each trial. Consequently, the trials were presented in a different random order to each subject.

Subjects were instructed that they would be shown sentence-picture pairs and were be asked to rate the acceptability of the sentence as a description of the picture using a 10-point scale, with zero denoting not acceptable at all; four or five denoting moderately acceptable; and nine perfectly acceptable. Trials were self-paced, and the experiments

lasted about 25-30 minutes. Figure 10-4 illustrates the instructions given to the subjects for the experiment. Figure 10-5 illustrates how the trials were presented.
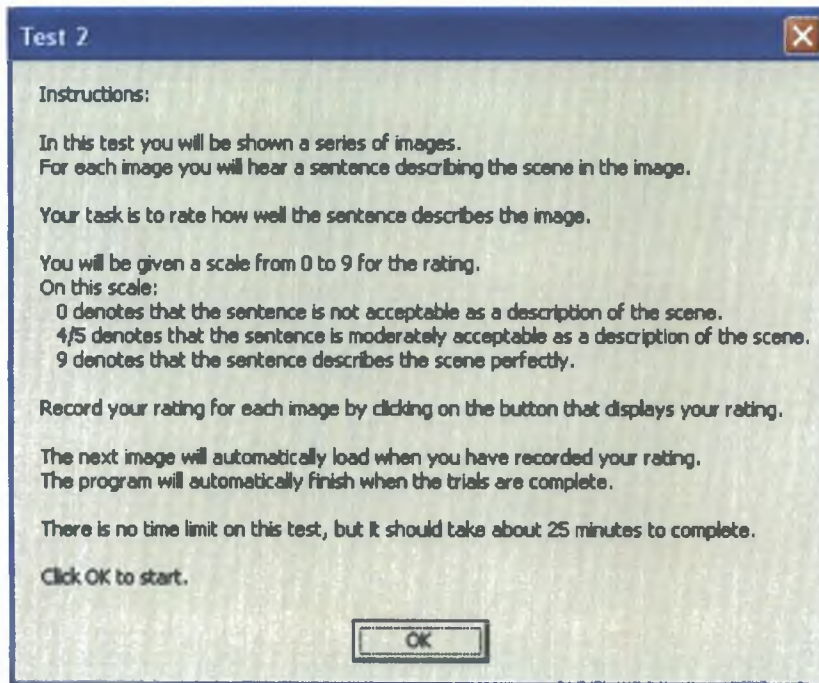
Figure 10-4: The instructions used in Experiment 2.
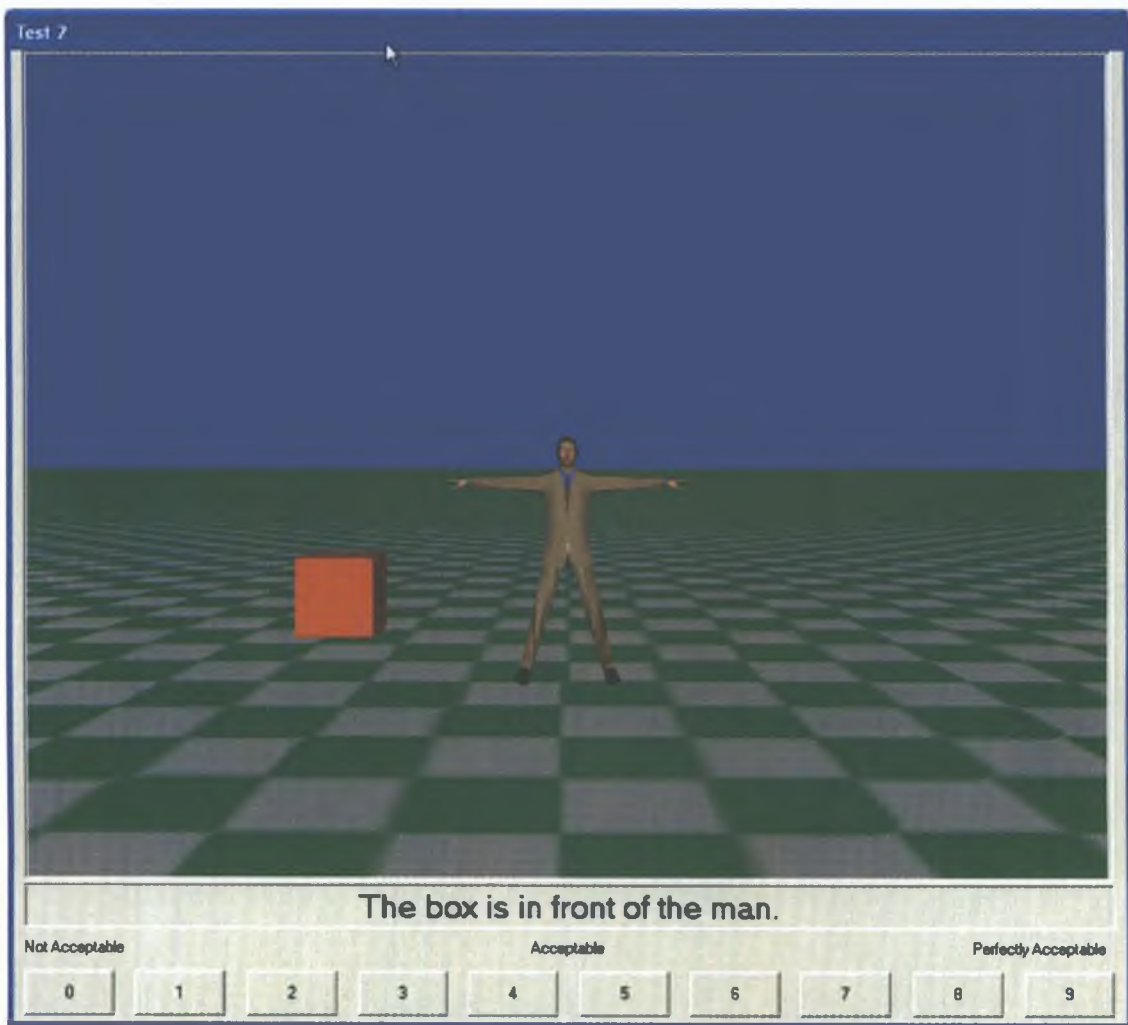
Figure 10-5: Sample presentation of a trial during Experiment 2.

10.3.2 Results and Discussion

Mean acceptability ratings broken down by orientation of the landmark (canonical and noncanonical) and spatial relation (*in front of* and *behind*) were calculated across subjects for each position of the trajector. A p-level of .05 was adopted for significance. Follow-up tests were based on critical differences required for significance. The critical differences were calculated on the basis of 95% confidence intervals using the error term from the interaction or main effect of the appropriate analysis of variance.

10 3 2 1 Canonical Trials

In the canonical trials the landmark object was facing towards the viewer The cross subject mean acceptability ratings for each position of the trajector for the canonical trials for each spatial relation are presented in Table 6 and Table 7

Table 6 A bird's eye view of the cross subject mean acceptability ratings for *in front of* a canonically oriented landmark by position in a 7 * 7 grid. The arrow indicates the direction of the landmark's intrinsic front The face symbol represents the position of the viewer

in front of						
0 0	0 0	0 8	0 2	0 1	0 1	0 1
0 1	0 5	1 0	0 0	0 5	0 2	0 2
0 2	0 2	0 1	0 0	0 3	0 5	0 1
3 9	5 1	5 4	↓	5 8	3 5	3 3
6 8	7 2	8 0	8 8	7 9	6 1	6 5
7 5	7 8	8 2	8 8	7 9	7 6	7 1
7 8	7 9	8 2	9 0	8 3	8 0	7 6
			☺			

Table 7: A bird's eye view of the cross subject mean acceptability ratings for *behind* a canonically oriented landmark by position in a 7 * 7 grid The arrow indicates the direction of the landmark's intrinsic front. The face symbol represents the position

behind						
7 2	7 6	8 5	8 9	8 4	8 1	7 0
6 9	6 1	7 2	9 0	8 1	7 3	6 8
2 7	4 2	7 2	9 0	6 7	5 7	3 2
0 2	0 4	0 5	↓	0 5	0 3	0 4
0 1	0 2	0 3	0 0	0 8	0 1	0 2
0 1	0 1	0 0	0 0	0 2	0 0	0 0
0 0	0 7	0 4	0 0	0 0	0 0	0 0
			☺			

403

The spatial template for each preposition defined by each subject's data was divided into good, acceptable, and bad regions following the designations made by Logan and Sadler (1996), see Figure 2-16. The good region contained the cells along the front axis of the coincident viewer-centred and intrinsic frames of reference; i.e., column 4 in rows 5 – 7 for *in front of* and column 4 rows 1-3 for *behind*. Acceptable regions consisted of the cells in the remainder of these rows; i.e., columns 1 – 3 and 5 – 7 in rows 5 – 7 for *in front of*, and columns 1 – 3 and 5 – 7 in rows 1 – 3 for *behind*. Finally the bad regions consisted of cells in the remaining rows; i.e., rows 1 – 4 for *in front of* and rows 4 – 7 for *behind*. Figure 10-6 illustrates the definition of the good, acceptable, and bad regions in the 7 * 7 grid for the prepositions *in front of* and *behind* in the canonical trials.
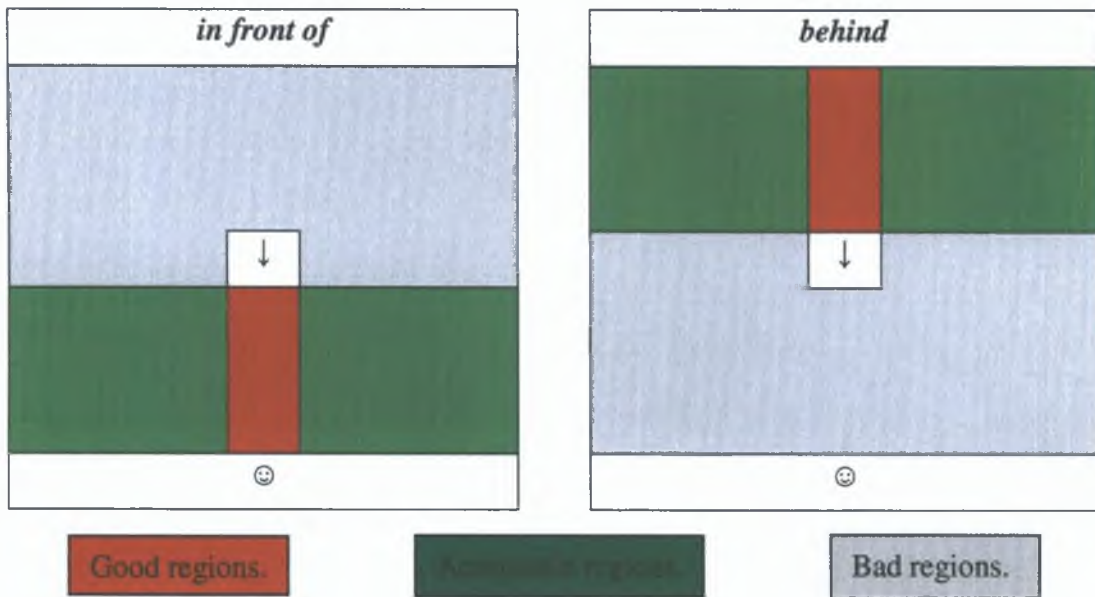
Figure 10-6: A bird's eye view of the definitions of the good, acceptable, and bad regions in the 7 * 7 grid for the prepositions *in front of* and *behind* when the landmark is in a canonical orientation. The arrow in the center of each grid indicates the direction of the front of the landmark, the face symbol at the bottom of each grid indicates the viewpoint of the subject.

Average acceptability ratings were calculated across each of these regions for each subject's data. A 3 (regions: good, acceptable, bad) * 2 (spatial relation: *in front of* and *behind*) repeated measure analysis of variance (**ANOVA**) performed on these mean ratings validated the region classification and replicated, for *in front of* and *behind*, Logan and Sadler's (1996) and Carlson-Radvansky and Logan's (1997) findings for *above* and *below*. There was a main effect of region ($F_{(2, 24)} = 762.000$, $p < .05$); a main effect of relation ($F_{(1, 12)} = 58.098$, $p < .05$); and a significant interaction ($F_{(2, 24)} = 32.336$, $p < .05$). This interaction is displayed on the graph in Figure 10-7, which illustrates the significant difference between *in front of* and *behind* in the acceptable and bad regions but not in the good regions.
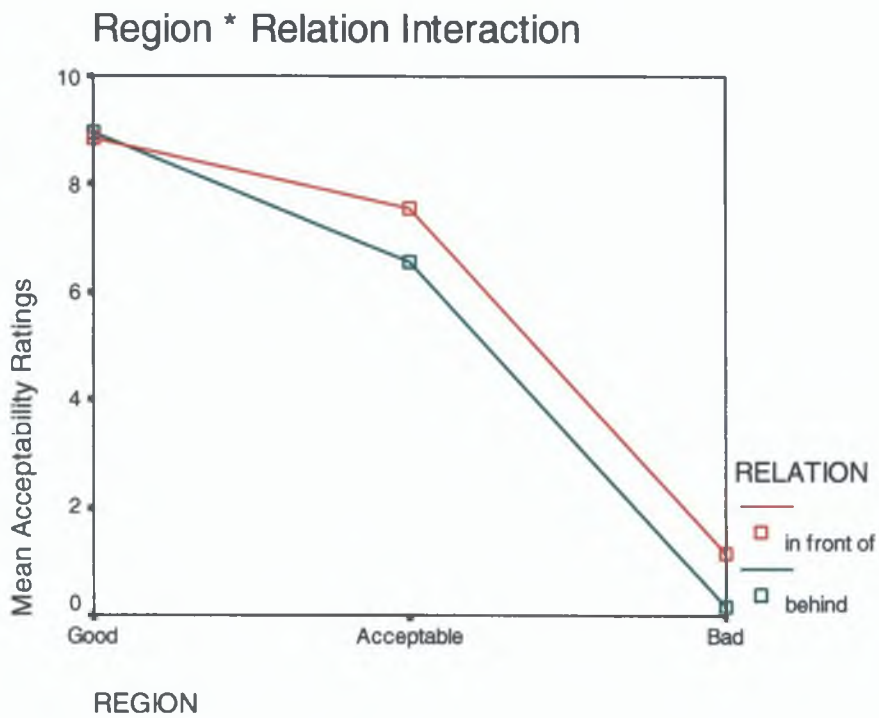
Figure 10-7: Graph plotting the interaction of region and relation acceptability during the canonical trials.

Figure 10-7 also illustrates that the acceptability ratings of both spatial relations dropped as the trajector moved from the good to the acceptable and finally to the bad regions The cross subjects mean acceptability ratings for each spatial relation in each of the regions is given in Table 8 Using a critical difference of 1 3, for both *in front of* and *behind*, the mean ratings for *in front of* and *behind* for the good regions (M = 8 88 and 8 98 respectively) were significantly higher than the mean ratings for the acceptable regions (M = 7 58 and 6 59 respectively), which in turn were significantly higher than the mean ratings for the bad regions (M = 1 19 and 19 respectively)

Table 8· Mean acceptability ratings for the canonical trials for all subjects broken down by region and spatial relation.

Region	Spatial Relation	
	in front of	*behind*
Good	8 88	8 97
Acceptable	7 58	6 59
Bad	1 19	19

Because the reference frames were aligned during the canonical trials, these spatial templates are consistent with either an exclusive use of one reference frame or the use of a combination of reference frames across trials In order to distinguish between these possibilities it is necessary to examine the spatial templates for the noncanonical trials

10 3 2 2 Noncanonical Trials

The noncanonical results were divided into two categories The first category contained the results of the trials where the landmark was rotated 90° to the left or the right The second category contained the results of trials where the landmark was rotated 180 degrees i e , the landmark object's intrinsic back was facing the subjects

The category 1 noncanomcal trials consisted of all the trials where the landmark was to rotated 90° to the left or right In this orientation the landmark's intrinsic spatial template overlaps with the viewer-centred frame of reference It was expected that an analysis of the data from the noncanonical category 1 trials would reveal

1 Whether there was an exclusive use of a viewer-centred or intrinsic frame of reference across trials, or whether the subjects used a mixture of the viewer-centred and intrinsic spatial templates across trials

2 Whether there was a bias towards either frame of reference along the horizontal plane

If the viewer-centred frame of reference was solely used across these noncanonical trials, then the good, acceptable and bad regions should correspond to the definitions in Figure 10-8 If an intrinsic frame of reference was solely used across these noncanomcal trials, the good, acceptable and bad regions should correspond to the definitions in Figure 10-9 However, if a mixture of spatial templates was used across noncanonical trials, the regions should correspond closer to the regions specified in Figure 10-10 The regions in Figure 10-10 are derived from overlaying the two independent spatial templates defined in Figure 10-8 and Figure 10-9 The major differences between the mixed spatial templates and the spatial templates in the viewer-centred or intrinsic frame of reference are

1 the relatively small bad region in the mixed spatial template,

2 the cells in mixed spatial template which have an acceptable rating in both the viewer-centred and intrinsic frames of reference have higher ratings than the cells in the acceptable regions of either the viewer-centred or intrinsic frame of reference

| Viewer-centred *in front of* | Viewer-centred *behind* |

Good regions.   Acceptable regions.   Bad regions.
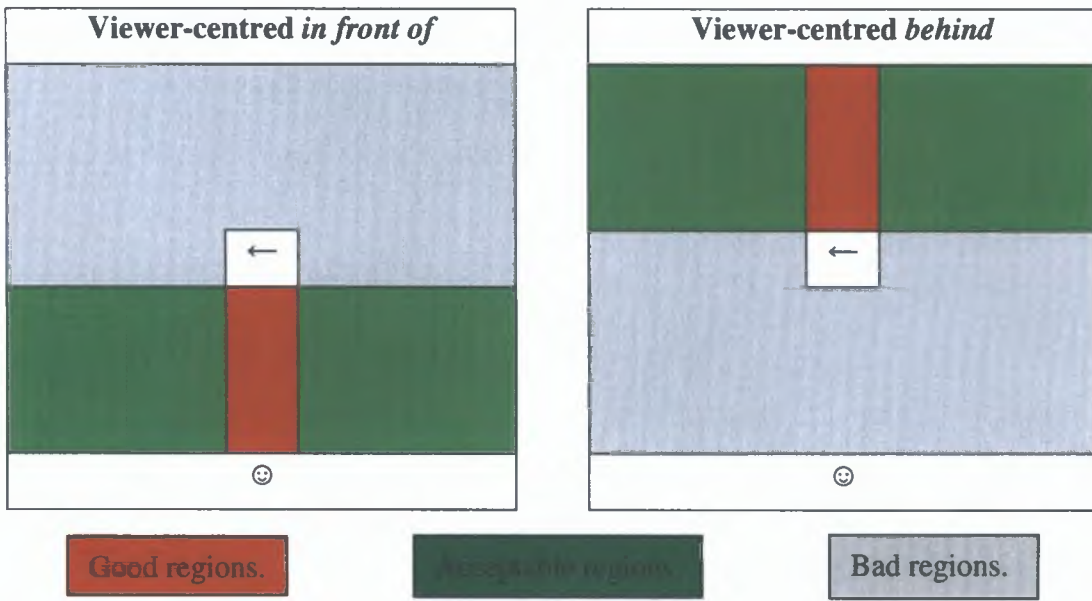
Figure 10-8: A bird's eye view of the definition of the good, acceptable, and bad regions in the 7 * 7 grid, modelling the spatial templates of the prepositions *in front of* and *behind*, in a viewer-centred frame of reference during a noncanical trial. The arrow at the center of each grid denotes the orientation of the landmark's front. In these figures the landmark is facing to the left, however the same regional definitions apply for the trials where the landmark was facing to the right. The face symbol at the bottom of each figure denotes the subject's view point.

408
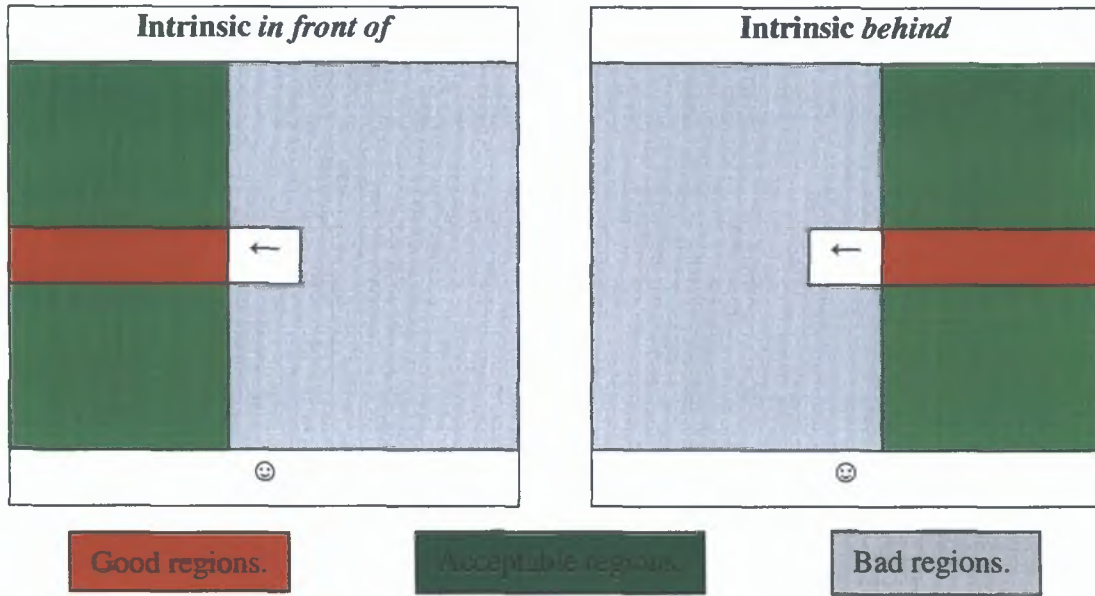
Figure 10-9: A bird's eye view of the definition of the good, acceptable, and bad regions in the 7 * 7 grid, modelling the spatial templates of the prepositions *in front of* and *behind*, in an intrinsic frame of reference during a noncanonical trial. The arrow at the center of the grid denotes the orientation of the landmark's front. In these figures the landmark is facing to the left. The regional definitions for the trials where the landmark was facing to the right are obtained by reflecting over the vertical midline. The face symbol at the bottom of each figure denotes the subject's view point.

| Mixture *in front of* | Mixture *behind* |

Cells with a good designation in either spatial template or with an acceptable designation in both.

Acceptable regions.
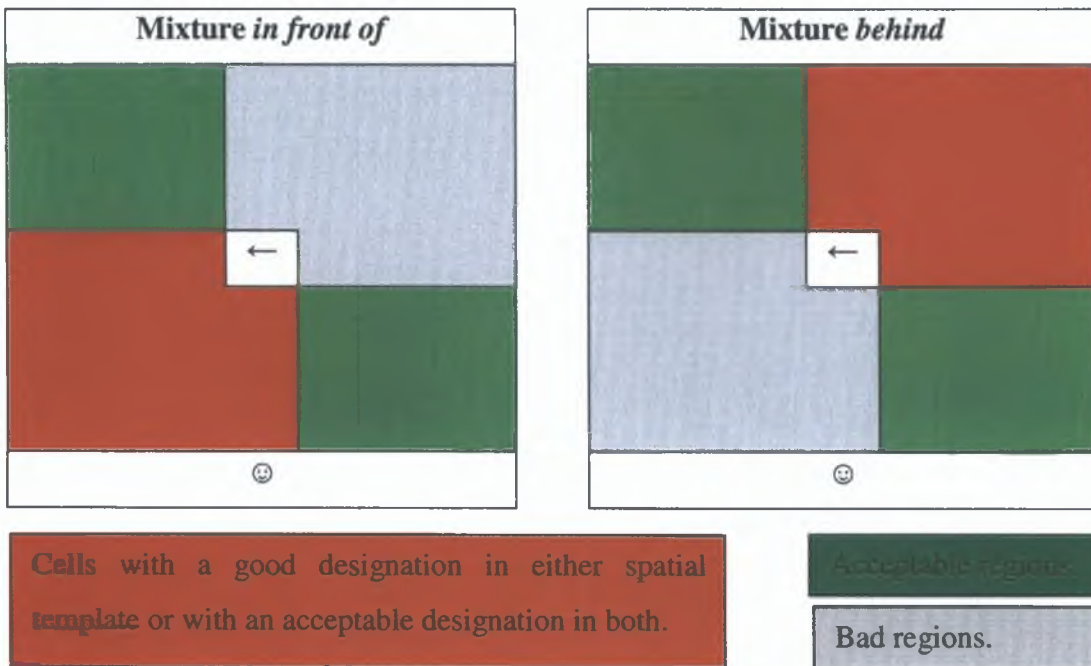
Bad regions.

Figure 10-10: A bird's eye view of the definition of the good, acceptable, and bad regions in the 7 * 7 grid, modelling the spatial templates of the prepositions *in front of* and *behind*, in a mixture frame of reference during a noncanonical trial. The arrow at the center of the grid denotes the orientation of the landmark's front. In these figures the landmark is facing to the left. The regional definitions for the trials where the landmark was facing to the right are obtained by reflecting over the vertical midline. The face symbol at the bottom of each figure denotes the subject's view point.

The cross subject mean acceptability ratings for each position for each spatial relation are presented in Table 9 and Table 10. The templates are based on a landmark that was rotated 90° to the left (consistent with Figure 10-8, Figure 10-9, and Figure 10-10). In computing these means, the data from the trials where the landmark was rotated 90° to the right was included by reflecting across the vertical midline.

Table 9: Cross subject mean acceptability ratings for each position in the 7 * 7 grid for the preposition *in front of* in noncanonical category 1 trials.

in front of						
6 9	5 9	5 3	1 1	0 4	0 1	0 1
6 9	6 7	6 3	1 0	0 0	0 3	0 1
7 4	7 5	7 4	0 8	0 2	0 1	0 2
7 9	7 6	8 1	←	0 2	0 2	0 4
7 7	7 6	7 5	4 7	2 1	1 7	1 7
7 2	7 4	7 2	4 9	2 6	2 5	1 8
7 0	7 1	6 4	4 7	3 6	3 5	2 4
			☺			

Table 10 Cross subject mean acceptability ratings for each position in the 7 * 7 grid for the preposition *behind* in noncanonical category 1 trials.

behind						
2 0	2 4	2 6	4 4	7 0	7 3	7 5
1 8	2 2	2 3	4 2	7 1	7 4	7 7
1 6	1 3	2 1	3 9	7 4	8 0	7 7
0 1	0 1	0 2	←	7 9	7 8	7 7
0 2	0 2	0 4	0 4	7 2	7 0	7 1
0 1	0 1	0 1	0 9	6 0	6 8	6 6
0 1	0 3	0 3	0 3	6 0	5 7	6 3
			☺			

Examining the data in Table 9 and Table 10 it is evident that the spatial template emerging from the category 1 noncanonical trials most closely corresponds to the designations in Figure 10-10 For the noncanonical spatial templates, the size of the bad region is diminished consisting of only 15 cells, rather than the 27 that made up the bad regions for the canonical trials, and the cells that have an acceptable designation in both the viewer-centred and intrinsic frames of reference have a higher rating than the cells in the acceptable area of only one of the spatial templates

To further evaluate whether the noncanonical templates reflect a mixture of viewer-centred and intrinsic spatial templates the three acceptable regions in the noncanonical templates (i e , the cells that are designated as acceptable in both the viewer-centred and intrinsic template, the cells that are designated as acceptable in only the intrinsic template, and the cells that have an acceptable designation in the viewer-centred frame of reference) were compared For each subject the mean acceptability rating for each relation in each of the acceptable regions was computed, see Table 11

Table 11 The noncanonical mean acceptability ratings for the three acceptable regions broken down by subject and spatial relation

Subject	Region					
	Viewer-centred & Intrinsic		Intrinsic		viewer-centred	
	in front of	*behind*	*in front of*	*behind*	*in front of*	*behind*
1	4 89	4 67	4 94	5 00	1 89	1 39
2	5 22	6 89	6 17	5 67	2 39	1 61
3	7 22	7 06	6 61	5 44	89	61
4	8 33	8 33	6 72	5 11	6 11	6 11
5	8 22	8 44	7 44	7 89	5 11	3 50
6	8 28	8 06	6 83	7 67	4 61	5 61
7	6 78	7 50	6 94	7 06	50	00
8	6 61	6 83	5 44	6 06	1 44	83
9	7 28	7 00	3 94	4 11	3 89	3 56
10	8 89	8 83	8 67	8 39	00	44
11	8 72	9 00	8 94	8 83	1 11	50
12	6 28	7 06	6 89	6 17	3 28	2 06
13	6 94	6 94	7 11	7 11	00	00

Using the data in Table 11 a 3 region (viewer-centred and intrinsic, intrinsic only, and viewer-centred only) * 2 spatial relation (*in front of* and *behind*) repeated measure ANOVA was performed. There was a main effect of region ($F_{(2, 24)} = 46.742.000$, p < .05); no main effect of relation ($F_{(1, 12)} = 1.674$, p = .220); and a significant interaction ($F_{(2, 24)} = 2.600$, p < .095). This interaction is plotted on the graph in Figure 10-11. The graph illustrates the difference between the acceptability of *in front of* and the acceptability of *behind* in the intrinsic and viewer-centred acceptable region, and in the viewer-centred only acceptable region, but not in the intrinsic only region.

Acceptable Region * Relation Interacton
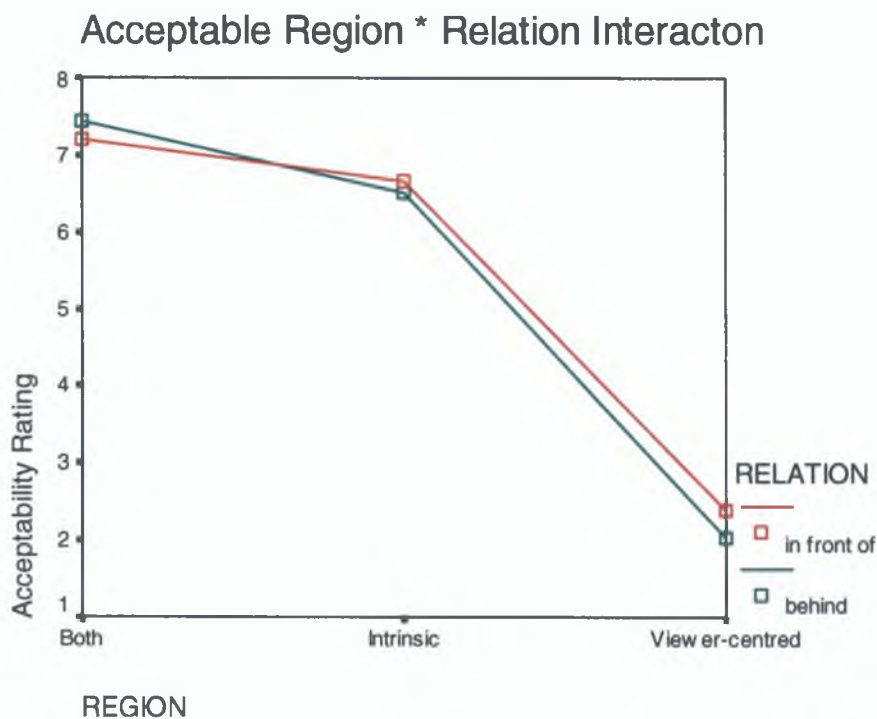
Figure 10-11: Graph plotting the interaction of region and relation during the noncanonical category one trials.

Figure 10-11 also illustrates that the acceptability of both spatial relations is highest in the region that is designated as acceptable in both the viewer-centred and intrinsic spatial template. Using a critical difference of .53, for both in front of and behind, the cross subject mean acceptability rating for the viewer-centred and intrinsic acceptable

413

regions (M = 7 205 and 7 432, respectively) were significantly higher than for either the intrinsic only acceptable regions (M = 6 667 and 6 5, respectively) or the viewer-centred acceptable regions (M = 2 402 and 2 017, respectively), see Table 12

Table 12 The noncanonical mean acceptability ratings for the three acceptable regions broken down by subject and spatial relation

Region					
Viewer-centred & Intrinsic		Intrinsic		Viewer-centred	
in front of	behind	in front of	behind	in front of	behind
7 205	7 432	6 667	6 500	2 402	2 017

It should be noted that the intrinsic only regions were rated significantly higher that the viewer-centred only regions This finding contrasts with Carlson-Radvansky and Logan's (1997) results which indicated that the viewer-centred frame of reference was rated higher than the intrinsic frame of reference for *above* and *below*, and points to a bias towards the intrinsic frame of reference along the horizontal plane

Excluding the difference with respect to reference frame bias, the results of this experiment replicated, for the preposition *in front of* and *behind*, the findings of Carlson-Radvansky and Logan's (1997) work, on the prepositions *above* and *below* As such, they support the hypothesis that the spatial template constructed across noncanonical trials reflect a combination of using the viewer-centred and intrinsic spatial templates

10 3 2 2 2 Noncanonical Trials – Category 2

The category 2 noncanomcal trials consisted of all the trials where the landmark was facing away from the subject In this orientation the landmark's intrinsic spatial template is completely dissociated from the viewer-centred frame of reference It should be noted that, in Carlson-Radvansky and Logan's work there were no trials which tested situations where the frame of reference were completely dissociated The focus of the

analysis of the noncanonical category 2 trials data was to examine the bias in reference frame use for prepositions canonically aligned with the horizontal plane

The cross subject mean acceptability ratings for each position of the trajector for the noncanonical category two trials for each spatial relation are presented in Table 13 and Table 14

Table 13: The cross subject mean acceptability ratings for the preposition _in front of_ for each position in the 7 * 7 grid in noncanonical category 2 trials.

in front of						
6 2	6 7	7 5	8 2	6 8	6 9	6 0
5 4	5 8	7 0	8 2	6 9	6 7	5 9
3 5	5 2	6 0	7 5	6 4	5 5	3 6
2 1	1 4	2 3	↑	2 4	0 6	0 8
2 2	2 5	2 0	3 4	2 8	1 7	2 7
2 4	3 8	3 2	3 5	2 8	3 0	3 1
2 9	2 3	2 7	3 2	3 7	3 4	2 5
			☺			

Table 14· The cross subject mean acceptability ratings for the preposition _behind_ for each position in the 7 * 7 grid in noncanonical category 2 trials

behind						
1 9	2 3	2 5	3 2	2 8	2 2	2 0
2 2	2 0	2 3	3 3	2 5	2 1	2 2
1 1	1 5	2 5	3 3	2 2	1 5	2 5
3 8	4 0	3 5	↑	4 2	2 5	4 2
5 9	6 2	6 9	7 6	6 9	6 2	5 0
6 4	6 4	7 4	7 6	7 0	6 2	6 6
6 5	6 7	7 0	7 7	7 0	6 9	6 6
			☺			

For each subject's data the spatial template of each spatial relation was broken down into four regions intrinsic good, intrinsic acceptable, viewer-centred good, and

415

viewer-centred acceptable For the preposition *in front of*, the intrinsic good region consisted of the cells in rows one, two, and three in column four, the intrinsic acceptable regions consisted of the other cells in rows one, two, and three, the viewer-centred good regions consisted of the cells in rows five, six, and seven in column four, the viewer-centred acceptable regions consisted of the other cells in rows five, six, and seven For the preposition *behind*, the intrinsic good region consisted of the cells in rows five, six, and seven in column four, the intrinsic acceptable region consisted of the other cells in rows five, six and seven, the viewer-centred good regions consisted of the cells in rows one, two, and three in column four, the viewer-centred acceptable regions consisted of the other cells in rows one, two, and three For each of these regions (*in front of* intrinsic good and acceptable, viewer-centred good and acceptable, *behind* intrinsic good and acceptable, viewer-centred good and acceptable) the mean acceptability values for each subject were calculated These means are presented in Table 15 and Table 16 Next, the mean acceptability for each subject for each region and frame of reference, independent of the relation used, was calculated These mean values are presented in Table 17

Table 15 The mean acceptability values broken down by subject for the noncanonical category 2 trials for the preposition *in front of* for each region (good and acceptable) and each frame of reference (intrinsic and viewer-centred).

	in front of			
	Frame of Reference			
	Intrinsic		*Viewer-centred*	
	Region			
Subject	*Good*	*Acceptable*	*Good*	*Acceptable*
1	5 33	4 94	3 00	1 78
2	8 67	5 28	3 67	2 78
3	8 33	7 39	0 00	0 83
4	7 00	6 44	5 67	6 44
5	9 00	5 17	5 67	5 72
6	8 67	6 22	9 00	6 17
7	9 00	6 56	0 00	0 72
8	8 33	5 50	3 67	1 61
9	3 33	3 44	4 67	3 67
10	9 00	7 28	0 00	0 94
11	9 00	8 22	3 33	1 22
12	9 00	5 17	4 67	4 11
13	9 00	6 50	0 00	0 00

Table 16· The mean acceptability values broken down by subject for the noncanonical category 2 trials for the preposition *behind* for each region (good and acceptable) and each frame of reference (intrinsic and viewer-centred)

	behind			
	Frame of Reference			
	Intrinsic		*Viewer-centred*	
	Region			
Subject	*Good*	*Acceptable*	*Good*	*Acceptable*
1	5 33	4 11	1 33	1 78
2	8 67	6 06	3 33	2 39
3	8 67	6 61	0 00	0 56
4	3 33	6 22	8 33	5 67
5	8 67	7 28	4 33	3 61
6	9 00	7 78	9 00	5 33
7	9 00	7 11	0 00	0 61
8	7 33	6 00	3 33	0 61
9	4 00	3 67	4 33	3 72
10	9 00	8 67	0 00	0 11
11	9 00	8 78	6 00	1 44
12	8 33	6 11	2 33	1 78
13	9 00	6 78	0 00	0 00

Table 17 The mean acceptability rating broken down by subject for each region (good and acceptable) and frame of reference (intrinsic and viewer-centred) in the noncanonical category two trials.

Subject	Frame of Reference			
	Intrinsic		Viewer-centred	
	Region			
	Good	Acceptable	Good	Acceptable
1	5 33	4 53	2 17	1 78
2	8 67	4 86	3 50	2 59
3	8 50	7 00	0 00	0 70
4	5 17	5 94	7 00	6 06
5	8 84	6 47	5 00	4 67
6	8 84	7 00	9 00	5 75
7	9 00	6 84	0 00	0 67
8	7 83	4 83	3 50	1 11
9	3 67	4 17	4 50	3 70
10	9 00	7 98	0 00	0 53
11	9 00	6 06	4 67	1 33
12	8 67	5 39	3 50	2 95
13	9 00	6 64	0 00	0 00

Using the data in Table 17 a 2 (frame of reference intrinsic, viewer-centred) * 2 (region good, acceptable) repeated measure ANOVA was performed There was a main effect of frame of reference ($F_{(1\ 12)} = 22\ 148$, p = 001), a main effect of region ($F_{(1\ 12)} = 20\ 106$, p = 001), and a significant interaction ($F(1, 12) = 4\ 264$, p = 061) This interaction is plotted on the graph in Figure 10-12, which shows the difference between the good and acceptable regions was greater in the intrinsic frame of reference than in the viewer-centred frame of reference The graph also illustrates that both regions were given a higher average rating in the intrinsic frame of reference than in the viewer-centred frame of reference

Frame of Reference * Region
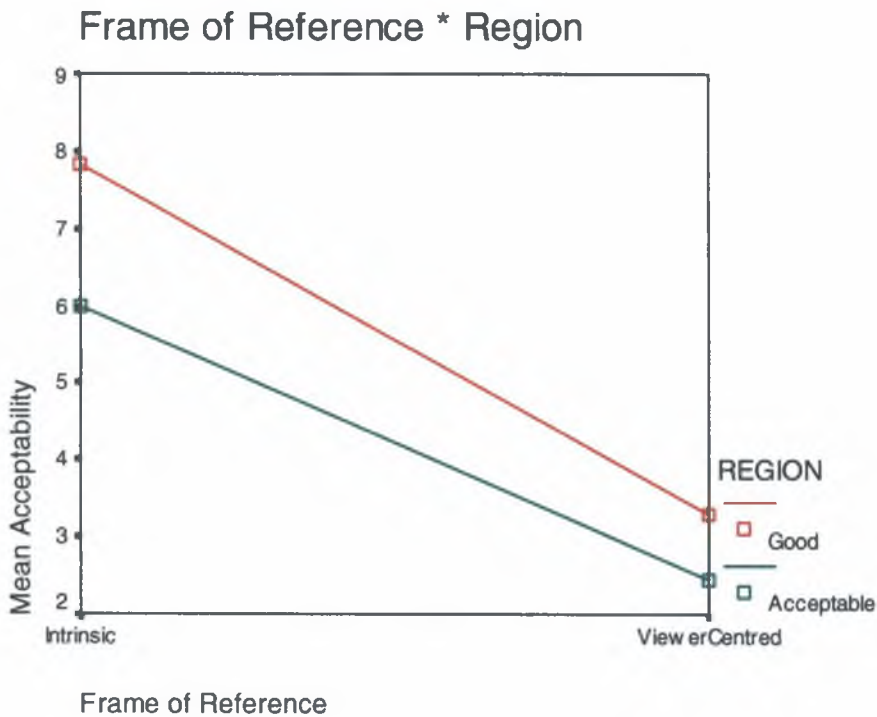
Figure 10-12: Graph plotting the interaction of frame of reference and region during the noncanonical category two trials.

Using a critical difference of 4.51 for both the good and acceptable regions, the mean ratings for the intrinsic frame or reference (M = 7.81 and 5.98, respectively) were significantly higher than the mean ratings for the viewer-centred frame of reference (M = 3.30 and 2.45, respectively). These results suggest a strong cross subject bias towards the use of an intrinsic frame of reference for horizontally aligned prepositions. This contrasts with the viewer-centred bias found by Carlson-Radvansky and Logan (1997) for the vertically aligned prepositions.

10.4 Experiment 3

The most distinctive attribute of the SLI discourse model is its integration of the visual context, modelled by visual salience, with a linguistic context model The integration of these information sources gives the discourse model the ability to distinguish between and resolve anaphoric and deictic references The focus of experiment 3 was to examine the cognitive plausibility of the algorithms used by the model to resolve deictic and anaphoric references Due to the difficulties in controlling the factors that impact on the resolution of a referring expression, it was decided to test the discourse model by showing each subject a video of the SLI system as it interacted with a user Following the system's interpretation of some of the more complicated user inputs, the video was paused and the subjects were asked whether the system had responded to the last user input as they had expected The advantage of this approach was that it allowed for the standardisation of the experiment between subjects

10.4.1 Method

10 4 1 1 Subjects

14 people took part in this experiment 11 men and 3 women

10 4 1 2 Materials

A 2 minute video containing the SLI 3-D world and an audio recording of a user interacting with the system was created The script of the video was designed to incorporate both anaphoric and deictic references This video was then edited into segments containing one user input followed by the system's interpretation of the input -- as illustrated by the system updating the 3-D simulation A Visual Basic application was

developed which played each of these segments in turn, and after each segment asked the subject whether the system had responded to the input as they had expected. Figure 10-13 shows the form used by the application to show the videos. Figure 10-14 shows the dialog box that appeared at the end of each segment asking the subject if the system had responded as they had expected. Table 18 lists

1   a set of sample images from the test video,

2   the accompanying video dialog,

3   an explanation of the system functions that each input was designed to test,

4   markers indicating the points in the video where the dialog box, illustrated in Figure 10-14, appeared asking the subjects for input

Text area: displays the previous user input.

Video display area.
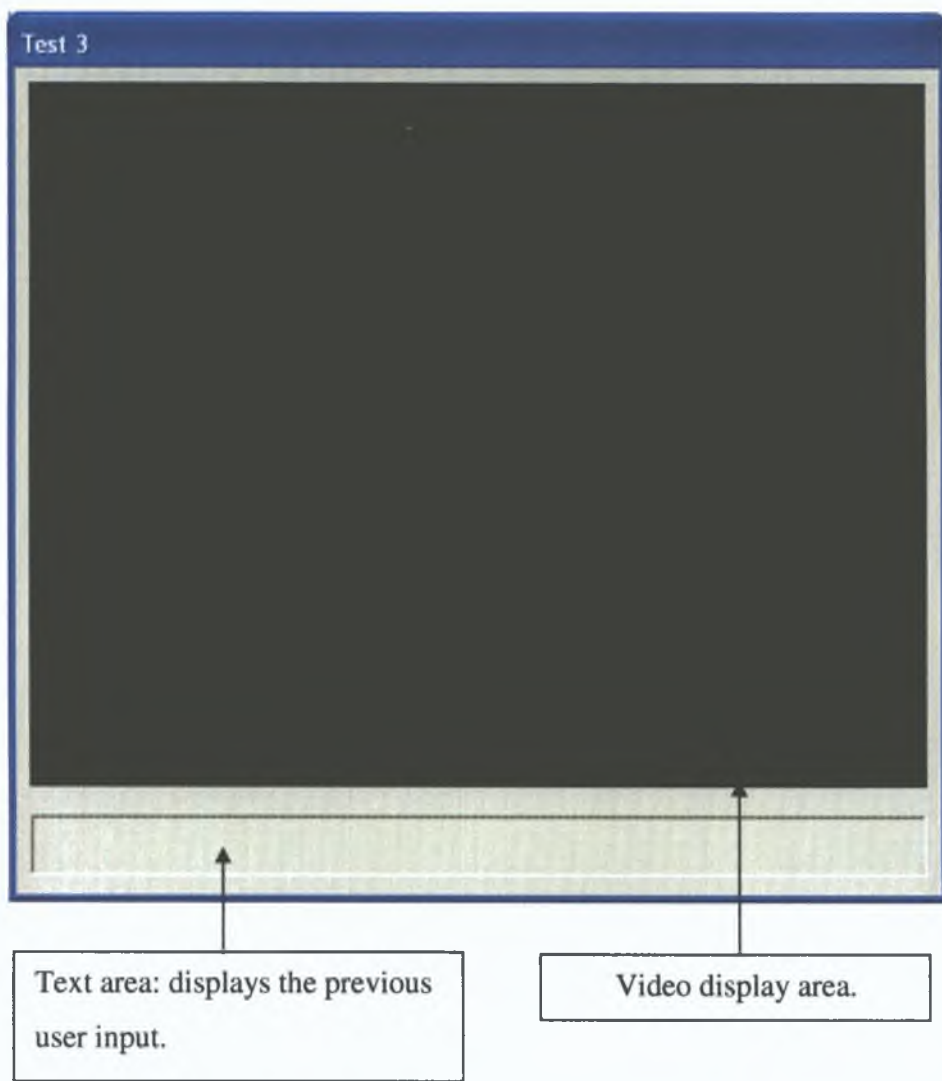
Figure 10-13: The form used to display the video of the SLI system during the experiment three trials.

Test3

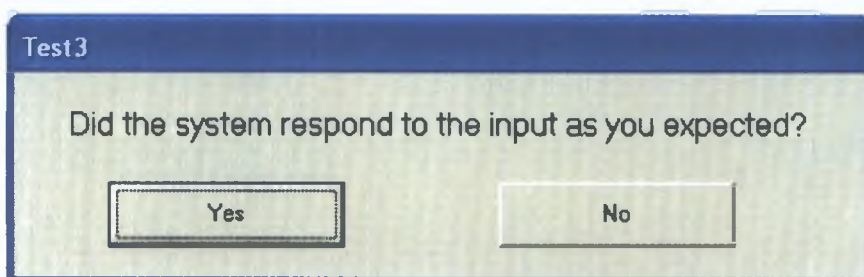Did the system respond to the input as you expected?

Yes   No

Figure 10-14: The dialog box that appeared at the end of each video segment in experiment three.

Table 18: This table lists a chronologically ordered sequence of sample images from the test video used in experiment 3. The accompanying video dialog is also listed along with markers indicating the points in the video where test subjects were asked for input. The inputs to the system are numbered and printed in a red italic font. An explanation of the system functions that each input tests along with a desciption of the approach adopted by the system to resolve each of these inputs is given. The locations in the video where the subjects were asked for input are indicated by the text "Question x: Did the system respond as you expected? Yes/No", where x is a number.

<u>Visual Context</u>	<u>Linguistic Context</u>
	(1) make the red house taller

Input (1) contains an underspecified deictic definite description, *the red house.* Resolving this reference tests the system's ability to resolve a deictic (visible situation use) referring expressions.

It should be noted that there are several red houses in the simulation, two of which are currently visible. In order to resolve this reference the system uses the information from the visual salience model to create a local context that restricts the number of possible referents in the world to the two houses that are currently visible. It then uses the saliencies associated with each candidate to make a graded judgement and selects a referent. Once a referent has been selected the simulation is updated.

424

(2) turn left

425

(3) stop

(4) look at the green house

Input (4) contains the deictic definite description *the green house*. Resolving this reference tests the system's approach to resolving a deictic (immediate situation use) referring expression.

There are no green houses currently visible in the view volume. However, at this point in the discourse three green houses have been seen -- none of these have been explicitly referred to in the linguistic dialogue.

426

In order to resolve this reference the system selects the most salient green house in the most recent VPD that contains an element that fulfils the linguistic selection restrictions of the referring expression. Once a referent has been selected the user's view volume is adjusted so that the referent is in the middle of the screen.

Question 2: Did the system respond as you expected? Yes/No

(5) make the tree in front of the house taller

Input (5) contains a locative expression that combines a deictic reference, *the tree*, with an anaphoric reference, *the house*. Interpreting this input tests the system's approach to: resolving anaphoric and deictic definite descriptions and modelling projective prepositions.

It should be noted that the reference *the house* could also be treated as an underspecified deictic definite description. However, because the referent of the preceding utterance fulfils the linguistic selectional restrictions of the reference and is currently visible in the view volume, the reference is treated as anaphoric and the green house is selected as the referent for the object noun phrase in the input; i.e., the green house (selected as the referent for the preceding input) is selected as the landmark. Once the landmark has been selected the reference to the trajector is resolved using the SLI spatial template model. Once the trajector has been   selected the simulation is updated.

427

(6) turn left

(7) stop

(8) make a blue house red

428

Input (8) contains an indefinite reference *a blue house*. In order to resolve this reference the system randomly selects one of the candidate referents from the local context that is created using the visual salience information.

(9) make the other one taller

Input (9) contains an other-anaphoric reference *the other one*. In order to resolve this reference the system uses both the type information from the preceding reference and the adjectival description in the preceding reference. It is important to note that if the system did not record the impact of the adjectival description in the preceding reference (i.e., by profiling the partition that the referent was extracted from), it would only have access to the type information of the preceding reference. This would result in the reference being underspecified and the most salient house in the scene that is not currently profiled being selected – in this instance the yellow house.

This input tested whether the system's use of the preceding adjectival description was cognitively plausible.

429

Question 5: Did the system respond as you expected? Yes/No

10.4.1.3 Procedure

Subjects were instructed that they would be shown a short video containing a sample interaction between a user and a 3D system. Furthermore, they were informed that the system would be given a set of instructions and that after receiving each instruction the system would respond by changing the visual scene. Finally they were told that at intervals during the video the video will pause and a dialog box will appear asking them whether they felt that the system had interpreted the last input as they expected. They were to respond to this question by clicking on the yes or no button on the dialog box. Figure 10-15 illustrates the instructions shown to each subject prior to the beginning of the test.
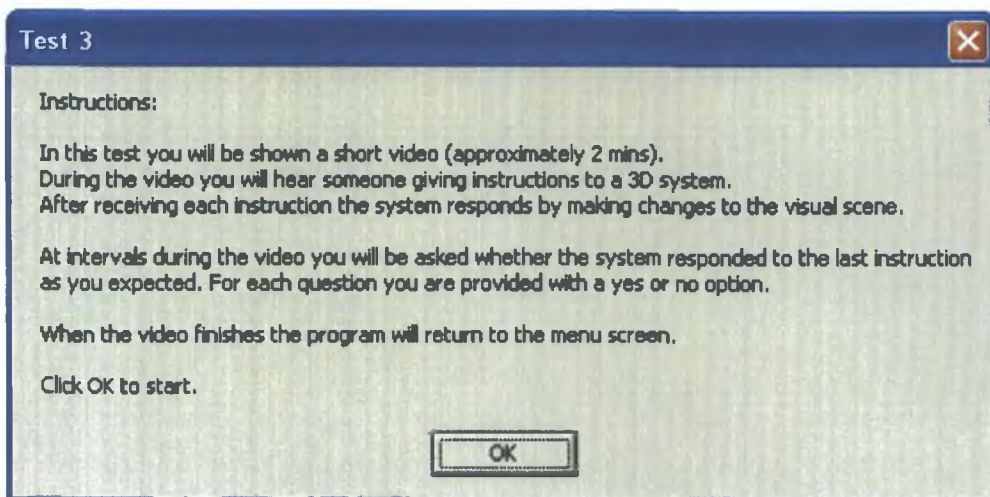
430

Figure 10-15: The instructions given to subjects before experiment 3.

10.4.2 Results and Discussion

The focus of experiment 3 was to examine whether the reference resolution algorithms developed for the SLI framework were cognitively plausible. In particular, the experiment focused on inputs that required the framework to:

1. distinguishing between anaphoric an deictic references,
2. switch between the perceptual and linguistic information sources when creating a local context for reference resolution,
3. make graded judgements, using saliency information, within the local context when resolving underspecified references.

Table 19 lists the data collected for experiment three. It is evident from these findings that in the vast majority of cases the subjects agreed with the results of the system's interpretation process; and, while these findings are not taken to suggest that the algorithms developed for the SLI framework are similar to those used by humans, the findings do indicate that the framework's interpretation algorithms are cognitively reasonable, at least with respect to the end product of the interpretation.

431

Table 19: The subject responses for experiment 3

Question Number	Yes Responses	No Responses
1	14	0
2	14	0
3	13	1
4	14	0
5	14	0

10.5 Chapter Summary

This chapter described three experiments that tested different aspects of the SLI framework Experiment 1 examined whether or not the absolute size of an object affected its visual salience, and consequently the probability of it being interpreted as the referent for a referring expression Of the 28 trials conducted, it was found that in 71 43% subjects selected the larger objects as the referent for the expression Notwithstanding the small sample size, the magnitude of the preference, within the results, for selecting the larger object as a referent supports the hypothesis that size matters[1]

Much of experiment 2 replicated Carlson-Radvansky and Logan's (1997) work on the prepositions *above* and *below*, for the prepositions *in front of* and *behind* The focus of this experiment was (1) to examine if the spatial template associated with a projective preposition is affected by the process of selecting a frame of reference, and (2) to examine the bias in frame of reference use for the prepositions canonically aligned with the horizontal plane The results of this experiment support the hypothesis that reference frame selection impacts on the construction of a spatial template, with the results indicating the use of a mixed spatial template to resolve locative references in situations where the frames of reference are dissociated This finding is in line with Carlson-Radvansky and Logan's (1997) results It should be noted that previous NLVR systems that interpreted locative expressions (CITYTOUR (Andre *et al* 1986, Andre *et al* 1987), CSR-3-D (Gapp 1994a), SPRINT (Yamada 1993), WIP (Olivier *et al* 1994, Olivier and Tsuji 1994), Situated Artificial Communicator (Socher and Naeve 1996, Socher *et al*

432

1996, Vorwerg *et al* 1997, Fuhr *et al* 1998), Virtual Director (Mukerjee *et al* 2000)) neglected to account for this phenomenon Regarding the bias in reference frame use for prepositions along the horizontal plane, it was found that there was a preference towards the intrinsic frame of reference This result differs from the viewer-centred bias, reported in (Carlson-Radvansky and Logan 1997), for preposition's aligned with the vertical frame of reference, and supports the biasing towards the intrinsic frame of reference for horizontally aligned prepositions in the frame of reference resolution algorithm developed in this thesis, Algorithm 8-3

Experiment 3 examined whether subjects found interpretations produced by the SLI discourse model's reference resolution algorithms acceptable The results indicated that in the vast majority of cases the subjects agreed with the system's interpretation of the inputs While it is not claimed that these algorithms mirror the cognitive processes in the human brain, these results do indicate that the saliency based approach to reference resolution developed in this thesis is reasonable as the basis for an NL interface

11 Conclusions

In this thesis, a perceptually based framework for interpreting spatial language in a real-time system has been developed There are three major components integrated in this framework a model of visual attention, a semantic model for projective prepositions, and a discourse framework

Modelling visual salience allows the system to incrementally create a model of what the user assumes as mutual knowledge shared between the user and the system This model aids in the resolution of references since it restricts the set of candidate referents to those objects that are in the mutual knowledge set Furthermore, by modelling visual salience, the system is able to rank candidate referents Using this ranking, the system can resolve underspecified references

The SLI model of visual attention is described in Chapter 7 This model is a novel application and extension of a synthetic model of vision that uses a graphics technique called false colouring (Noser *et al* 1995) The function of this visual attention model is to try to capture the perceptual information flowing from the environment to the user The output of this model feeds into the SLI discourse model, which uses it to model the visual perceptual aspects of the dialogue

In Chapter 8, the stages in the SLI algorithm for interpreting locative expressions were described There were several novel components of this model

1   In Section 8 3, a computational algorithm, based on psycholinguistic research, that attempts to resolve the issue of reference frame use was developed

2   In Section 8 4, a computational model of the semantics of projective prepositions that defines prepositions in terms of perceptual and topological axioms was developed The major components on this model were

    2 1 In Section 8 4 4, three algorithms that define how the frame of reference resolution algorithm developed in Section 8 3 and the novel components of the semantic models for projective prepositions developed in Sections 8 4 1 1, 8 4 1 2, and 8 4 3 are integrated The novel components developed were

2 1 1   In Section 8 4 1 1, an algorithm that dynamically locates the spatial template's origin in the viewer-centred frame of reference based on the user location relative to the landmark as developed Using this algorithm, the model avoided many of the paradoxical definitions that occur with models which default to using the landmark's bounding box centre as the origin

2 1 2   In Section 8 4 1 2, a parameterised potential field model of the spatial template of projective prepositions was developed The parameterisation of the model means that the model can be scaled to accommodate different sized landmarks Importantly, unlike the CSR-3-D system, this scaling does not depend on the scaling of a local coordinate system centred on the landmark, the major advantage of this parameterised approach is that the model is not forced to adopt the landmark's centre as the spatial template origin The model works in 3-D and measures both the angular deviation of a point from the canonical direction of the preposition and the distance of the point from the spatial template origin

2 1 3   In Section 8 4 2, a set of perceptually based definitions for the prepositions along the front-back axis in the viewer-centred frame of reference were defined Section 8 4 3 illustrated how these perceptual definitions can be combined with the topological potential model developed in Section 8 4 1 Moreover, it was shown how this integrated semantic model is able to define the regions surrounding landmarks with complex geometries in a consistent manner

3   Section 8 5 developed solutions for the issues pertaining to the representation of candidate trajectors in the framework

4   In Section 8 6, a novel general computational algorithm for interpreting locative expressions was developed

In Chapter 9, the SLI discourse model was developed The discourse model creates a context which can be used to interpret language The SLI discourse model adapts and extends the model proposed by Salmon-Alt and Romary (2001) The novelty of the SLI model is its integration of visual perceptual information into its context model and an

435

explicit description of how this perceptual information is combined with the linguistic information to resolve references There are three components within this discourse model the context model, the interpretive process, and the grouping operation The context model and the interpretive process are sufficient to resolve nominal expressions, however, for more complex grammatical constructions the grouping operation is necessary For locative expressions, the grouping operation must be augmented by a semantic model for prepositions This model gives the framework the ability to sort the elements representing the candidate trajectors within the grouped domain complex expression partition based on their fitness with respect to the prepositional phrase which in turn allows the discourse framework's interpretive process to profile the correct element as the referent of the expression

Chapter 10 describes a set of psycholinguistic experiments that examined different aspects of the SLI framework The results of these experiments indicate that

1   The assumption that an object's absolute size affects its visual salience, and consequently the probability of it being interpreted as the referent for and expression, is valid

2   The process of selecting a frame of reference impacts on the shape of the spatial template associated with the prepositions *in front of* and *behind*

3   There is a bias towards the use of the intrinsic frame of reference for the prepositions *in front of* and *behind*

4   The SLI reference resolution algorithms, which integrate both visual perceptual and linguistic information, are cognitively plausible

In conclusion, the SLI framework provides a computational model for interpreting some of major linguistic constructions in spatial language We argue that this has achieved the goal of this work to develop a semantic framework to underpin the development of NL interfaces that allow a user to navigate through and interact with a rendered 3-D environment Moreover, this framework illustrates the feasibility of using visual perceptual information as the foundation for an NL interface for simulated environments Finally, the ability of the SLI framework to resolve references that

previous models have found problematic reinforces the thesis of this work namely that the ability of computational models to interpret spatial language is greatly increased if they utilise information from a visual context that is shared with the user

12 Future Work and Open Questions

12.1 Introduction

There are many areas where the framework developed in this thesis can be extended and refined In this chapter, I would like to highlight where I think more work would be particularly fruitful I focus on each of the components in the framework in turn I discuss how they could be improved and outline some of the issues facing the work I conclude by discussing how the SLI framework can be extended to the generation of referring expressions

12.2 Visual Salience

Although the SLI false colouring visual salience algorithm has several advantages (speed and sensitivity), some of its weaknesses are

1 it assumes that the viewer's attention is focused on the centre of the scene,
2 it does not model all the factors affecting visual salience

The first of these could be addressed using eye tracking technology Using this technology a system could dynamically set the weightings assigned to the pixels in each scene relative to their distance from the user's center of gaze This approach would provide a more realistic model of the user's attention

In regard to point 2 above, Section 2 2 2 contains a discussion of some of the different factors that affect visual salience In concluding that section, it was noted that many of these factors were very difficult to quantify and model, and that in many cases the different factors compete Furthermore, it was argued that using the most basic determiners of visual salience as the input to a model had the benefits of simplifying the model and making it more generic While I feel that this is a valid and reasonable approach, I would like to extend the scope of the visual salience model to include some

438

other factors impacting on visual salience In particular, I think that the model should accommodate colour Many psychological tests have attested to the importance of colour in human visual perception Gapp (1995c) notes that colour is the easiest object feature to perceive and is directly responsible for the salience of an object However, there are many issues in modelling the impact of colour on salience For example, although red is well known to be an extremely salient colour, a red object against a red background would not be salient Consequently, I think that the major issue to be tackled in modelling the impact of an object's colour on its visual salience in a given scene is defining an algorithm that can accommodate the context dependency of a colour's salience One approach to this issue may be to calculate the average colour of a scene and to define an object's colour salience as the distance of the object's colour from the scene's average colour in the Munsell (Munsell 1905) colour space

12.3 Locative Expressions

The SLI algorithm for interpreting locative expressions containing projective prepositions proposes several novel components The one I would like to focus on in this discussion is the algorithm for locating the spatial template's origin in the viewer-centred frame of reference This algorithm is an improvement on previous work as, in many cases, it permits the system to avoid the problems associated with using the landmark's bounding box centre as the spatial template's origin However, in its current form, this algorithm requires the system to use the landmark object's bounding box centroid in the location process I think that this is psychologically implausible because in many instances where a viewer-centred frame of reference is used, a person will not have knowledge of the location of the object's bounding box A better approach would be to calculate the pixel at the center of the landmark object's 2D projection onto the viewport, and to use a ray cast through that pixel to locate the spatial template's origin The advantage of this approach is that it is grounded in the viewer perception and does not assume a priori knowledge of the landmark object's geometry

439

12.4 Discourse Model

An obvious extension to the SLI discourse model is the inclusion of plural references within its ambit. While there are several problems that need to be addressed to fulfil this goal, I think that the major issue in this area is quantifying the number of intended referents; e.g., given a visual context containing more than two red houses and assuming a deictic reference, does the reference *the red houses* refer to all the red houses in the scene or to a particularly salient subset of the set of red houses. To my knowledge, there is no obvious satisfactory resolution to this issue; however, a possible approach may be to search for discontinuities in the spectrum of visual saliencies assigned to the candidate referents and to take this discontinuity as the demarcation on the set of referents.

12.5 Natural Language Generation

This thesis has focused on the interpretation of spatial language. One area of future work is in the related field of natural language generation (NLG). Dale and Mellish define NLG as the "body of research that is concerned with the process of mapping from some underlying representation of information to a presentation of that information in linguistic form, whether textual or spoken" (1998 pg. 1). Reiter and Dale (1997) provides an overview of this field. They define six categories of work in NLG: content determination, document structuring, lexicalisation, aggregation, generation of referring expressions (GRE) and surface realisation. Of these, GRE is the area that this thesis is of most relevance too.

GRE focuses on the semantic questions involving the factual content of the description, and does not concern itself with the linguistic realisation of the description. There have been many GRE algorithms proposed (see among others Appelt 1985; Dale 1992; Dale and Reiter 1995; Krahmer and Theune 2002; van Deemter 2002). Most of these algorithms deal with the same problem definition: given a single target object, for which a description is to be generated, and a set of distractor objects, from which the

target object is to be distinguished, determine which set of properties is needed to single out the target object from the distracters. The term content determination is used to describe the task of determining which set of properties is needed to single out the target object from the distracters. On the basis of these properties a distinguishing description of the target object can be generated; i.e., a distinguishing description is a description of the target object that excludes all the elements of the distractors set.

The current state of the art in the area is the incremental algorithm (Dale and Reiter 1995), with most later algorithms extending this. The incremental algorithm "sequentially iterates through a (task-dependent) list of attributes, adding an attribute to the description being constructed if it rules out any distractors that have not already been ruled out, and terminating when a distinguishing description has been constructed" (Dale and Reiter 1995 pg. 247). If the end of the list of attributes is reached before a distinguishing description has been generated the algorithm fails. It should be noted that, in the incremental algorithm the target object's type is always included in the generated description even if it has no distinguishing value.

There are, however, several limitations to the incremental algorithm. van Deemter (van Deemter 2001; van Deemter *et al.* 2002) note the following simplifying assumptions made by the algorithm:

- the target object is always a single object,
- all objects in the domain are equally salient,
- an object has (or does not have) a property regardless of context,
- a property never consists of a relation to another object,
- referring expressions do not use negations or conjunctions.

In the context of integrating the SLI framework with a GRE algorithm, it is interesting to note the similarities between the data structure requirements of GRE algorithms and the type of information the SLI data structures can accommodate. GRE algorithms take as input a data structure containing a set of entities (target object and distractors) each with a set of attributes. The reference domains in the SLI framework model these requirements exactly: the entities are modelled by the elements in the

441

domain and the attributes and relationships between the entities are modelled by the differentiation criterion of the reference domain partitions Furthermore, many of the GRE concepts can be easily formulated within the SLI framework

- content determination can be defined as partition creation,
- a description can be defined as a differentiation criterion of a partition,
- a description refers to an object in the scene if the element representing the object in the reference domain is in the partition for which the description is the differentiation criterion,
- a description is a distinguishing description of an object in the scene if there is only one element in the partition for which the description is the differentiation criterion and that element represents the object in the scene

Using these definitions Dale and Reiter's (1995) incremental algorithm can be defined in the SLI framework as

Input a reference domain containing an element representing the target object in the scene, a set of elements representing the distracter objects in the scene
1   Sequentially iterate through the list of preferred attributes <type, colour, tall, short, wide, narrow, deep, shallow>
2   For each attribute create a partition whose differentiation criterion is set to the attribute plus all the previously accepted attributes
3   If the number of elements in the newly created partition is less than the number of elements in the partition created using the previously accepted attributes, add the current attribute to the list of accepted attributes
4   Terminate when the element representing the target object is the only element left in the created partition (success) or when the end of the preferred attribute list is reached (fail)
5   Always include the target object's type in the set of accepted attributes

Algorithm 12-1· A definition of Dale and Reiter's (1995) Incremental Algorithm within the SLI framework.

442

Algorithm 12-1 has been implemented in the SLI system

One of the simplifying assumptions made by the incremental algorithm is that all objects in the domain are equally salient In the SLI framework visual salience was exploited for reference resolution A symmetric process might be used in reference generation Krahmer and Theune (2002) present an algorithm which extends Dale and Reiter's (1995) incremental algorithm to allow for linguistic salience The underlying idea of their modification is to modify the definition of a distinguishing description to the following

"A definite description 'the N' is a suitable description of an object d in a state s iff d is the most salient object with the property expressed by N in state s " (Krahmer and Theune 2002 pg 176)

It should be noted that (Krahmer and Theune 2002) focus on linguistic salience Indeed, they propose a framework for modelling linguistic salience that is a synthesis of the hierarchical focusing constraints of Hajičova (1993) and the constraints of Centering Theory (Grosz et al 1995) In the SLI framework, each element in the context model has a visual salience associated with it Using these salience values the incremental algorithm can be extended to accommodate the generation of underspecified references where the visual salience of the target object is sufficient to allow the hearer to resolve the reference This can be accomplished by defining a distinguishing description as follows a description is distinguishing if it excludes all the distractors that have a visual salience greater than the target object's salience minus a predefined confidence interval A version of the incremental algorithm using this definition for distinguishing description as a terminating condition has been implemented in the SLI system While there has been no formal testing of this algorithm, preliminary results indicate that the underspecified references it generates are cognitively reasonable It should be noted that in some contexts this modified algorithm can generate reasonable underspecified references where the unmodified implementation of the incremental algorithm failed

Appendix A

This appendix contains definitions for the terms written in bold in the algorithms defined in Chapter 9 This appendix is organised in two parts The first part defines the terms used in the algorithms to refer to elements in the input, the context domain, or the interpretive process The second part defines the functions used in the algorithms The definitions in each section are ordered alphabetically and are separated by blank lines

Section 1: Terms referring to entities in the input, in the context model, or that are part of the interpretive process.

The term **C_Int** represents the predefined salience confidence interval

The term **Creation_Verbs[]** symbolises the set of verbs that in the SLI context can only be used in commands that create new objects in the simulation **Creation_Verbs[]** == { *add, create* }

The term **Deictic_Verbs[]** symbolises the set of verbs that in the SLI context can only be complemented by deictic references
Deictic_Verbs[] == { *look, go to, move* }

The term **Either_Verbs[]** symbolises the set of verbs that in the SLI context can be used in command that create new object in the simulation or can be complemented by deictic references
Either_Verbs[] == { *make* }

The term **IntExp** denotes symbolises the process of interpreting a referring expression The notations used to describe the values of the properties associated with this process are

1   **IntExp Dialogue** – denotes the general context selected for the interpretation process The range of values this property can take is defined by the set { VDL, LDL }

2   **IntExp.RD** – denotes the local context or reference domain selected for the interpretation process This property can be set to either an LD or a VPD

3   **IntExp.Referent** – denotes the element of the local context that is selected as the referent for an expression

LDL[] represents the set of LDs in the LDL LDL[x] represents the LD at index x in the LDL stack LDL[1] represents the most recently created LD in the LDL stack

The term **NPStr** denotes the string that represents the referring expression being processed The term **NPStr-1** denotes the string that represents the referring expression in the utterance preceding the utterance that is currently being interpreted The object NPStr has four properties These are

1   **NPStr.Det** – denotes the determiner of the referring expression The range of values NPStr Det can take is defined by the set { 'a', 'the', '' }

2   **NPStr Adjectives[]** – denotes the set { x $\wedge$((x is an attributive adjective), (x $\in$ NPStr)) }

3   **NPStr.Modifiers[]** – denotes the set { x $\wedge$((x is a nominal modifier), (x $\notin$ NPStr Adjectives[])) }

4   **NP head** – denotes the head noun of the referring expression

The term **Predicative_Adjectives[]** symbolises the set of adjectives used in a post-verbal or predicative position in the user input

445

The term **Verb** symbolises the verb used in the user input

VDL[] represents the set of VPDs in the VDL **VDL[x]** represents the VPD at index *x* in the VDL stack **VDL[1]** represents the most recently created VPD in the VDL stack

The term *RefPtr* symbolises a variable used to hold the object pointer returned by the function *createReferent()*

Let x be an element of LDL[] or VDL[], i e , x is a reference domain in the context model Then

x.Name denotes the name attribute of x

x.Profiled[] denotes the set of elements that are profiled in x

x Profiled[y].Visible is a Boolean value that is set to true if the object represented by element y of the set of profiled elements in x is visible in the current view volume

x.TYPE denotes the TYPE partition in x

x TYPE Elements[] denotes the set of elements in the TYPE partition in x

x.TYPE.Elements[] Object refers to the set of world objects that are represented by the set of elements in the TYPE partition in x

x.TYPE.Elements[y] denotes element y of the TYPE partition in x

x TYPE.Elements[y] Object denotes the object represented by element y of the TYPE partition in x

x TYPE Elements[y] Object.Salience denotes the visual saliency ascribed to the object represented by element y of the type partition in x

x.Partitions[] denotes the set of basic partitions in x

x.Partitions[y] denotes basic partition y in x

x.Partitions[y].Criterion denotes the differentiation criterion of the basic partition y in x

x.Partitions[y].Elements[] denotes the set of elements in the basic partition y in x

x.Partitions[y] Elements[z] denotes element z of the basic partition y in x

446

x.Partitions[y].Elements[z].Object denotes the object represented by element z of the basic partition y in x

x.Partitions[y] Elements[z].Object.Salience denotes the visual saliency ascribed to the object represented by element z of the basic partition y in x

Section 2: Definitions of functions used in the algorithms.

The function *createReferent()* takes the head noun and any adjectival descriptions supplied by the expression as parameters This function creates a new object in the simulation whose attributes match these parameters Where no value is supplied for a particular object attribute default values are used The *createReferent()* function returns a pointer to the created object's data structure

The function **checkSalience()** takes a set of objects as a parameter (often the set of objects will be a partition in a reference domain) This function returns true if the difference between the saliency of the element with the highest saliency in the set and the other elements of the set exceeds a predefined confidence interval

The function **createPartition()** takes two parameters a reference domain and a string defining a referential expression This function creates a new partition in the domain whose differentiation criterion matches the adjectival descriptions supplied by the referential expression It then fills the partition with the domains elements that fulfil the partition's criterion The function returns the index of the newly created partition

Fulfils(x) is a Boolean function, i e , it returns true or false The parameter to this function *x* is an object in the simulation that is represented by an element in a reference domain This function returns true if the objects attributes matches the linguistic description specified in the NPStr
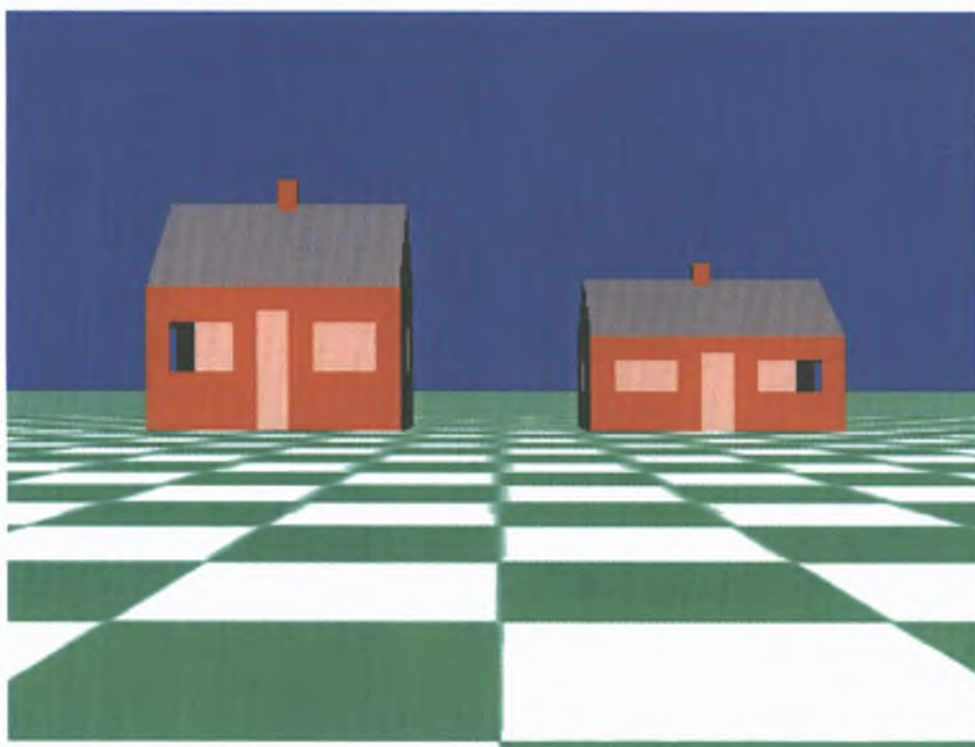
447

MinIndex(x, y) takes two parameters The first parameter is either the LDL or the VDL The second parameter y is a subset of the elements of x This function returns j where $\wedge$ ((x[j] == m), (m $\in$ y), ($\forall$n (x[n] $\in$ (y – m), (j < n)) In effect, the function MinIndex(x, y) returns the index of the most recent domain in the VDL or LDL as specified by x which fulfils the criteria defining the set of domains in y
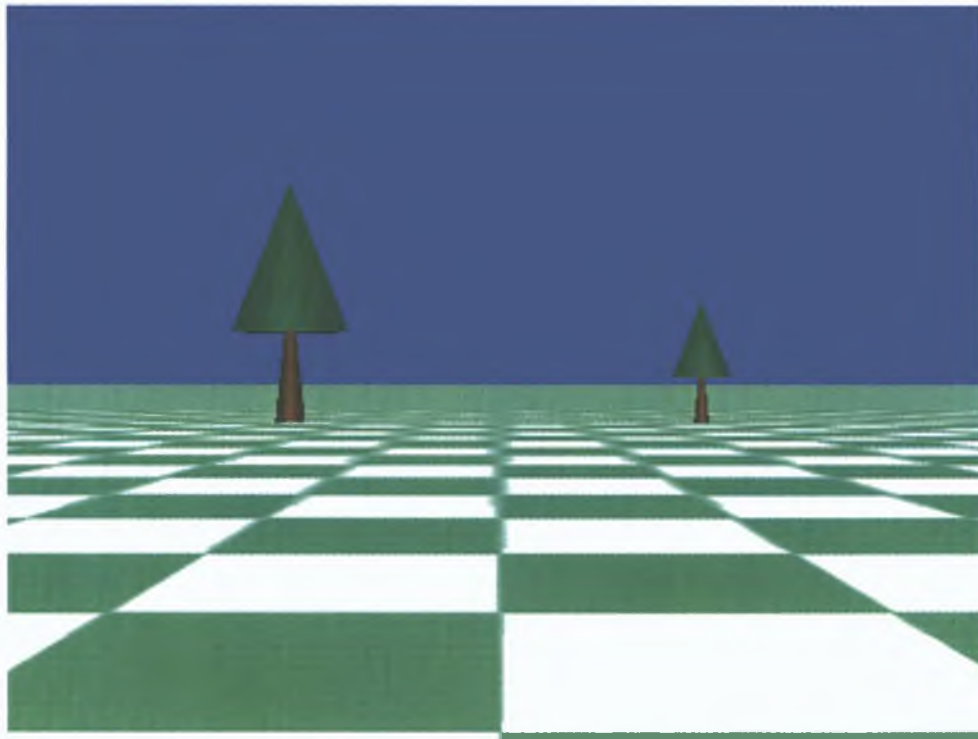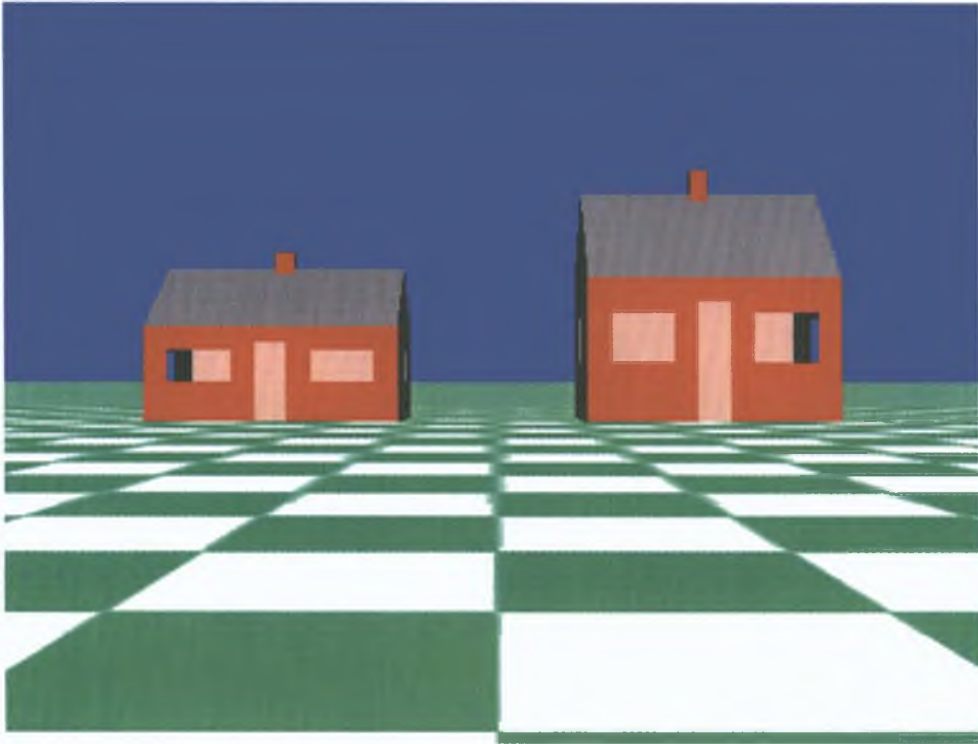
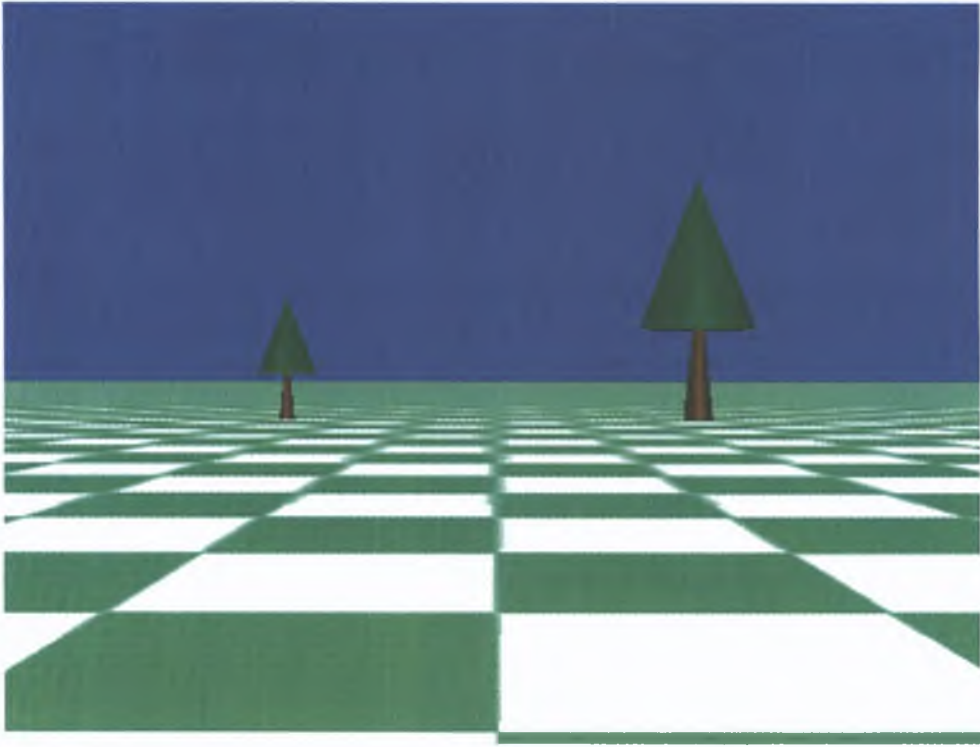The function *random()* takes a partition as a parameter and returns a randomly selected element from the partition

The function *restructure()* takes the most recently created VPD that contains an element representing the newly created object and restructures this reference domain so that the element representing the newly created object is profiled This restructures results in the creation of a new LD

448

Appendix B

This appendix lists the images used in Experiment 1, see Section 10.2.

450

451

Index of definitions of technical terms

Bibliography

Andre, E., G. Bosch, G. Herzog and T. Rist (1987). Coping with the Intrinsic and Deictic Uses of Spatial Prepositions. *In: Artificial Intelligence II: Methodology, Systems, Applications - Proceedings of the Second International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA '86)*, pp. 375--382.

Andre, E., G. Herzog and T. Rist (1986). Natural Language Access to Visual Data: Dealing with Space and Movement. *In: Proceedings of the 1st Workshop on Logical Semantics of Time, Space and Movement in Natural Language, Toulouse, France.*

Andre, E., G. Herzog and T. Rist (1988). On the Simultaneous Interpretation of Real World Image Sequences and their Natural Language Description: The System SOCCER. *In: Proceeding of 8th European Conference on Artificial Intelligence (ECAI-88), Munich, Germany*, pp. 449-454. Pitmann Publishing, London.

Appelt, D. (1985). "Planning English Referring Expressions". *Artificial Intelligence*, Vol. 26(1), pp. 1-33.

Bean, D. L. and E. Riloff (1999). Corpus-Based Identification of Non-Anaphoric Noun Phrases. *In: Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics (ACL-99)*, pp. 373-380.

Bennett, D. (1975). *Spatial and Temporal Uses of English Prepositions An Essay in Stratification Semantics*, 2nd Ed, Longman Group Limited, London.

Bowerman, M. (1996). "Learning how to Structure Space for Language". *In:* M. Garrett ed. *Language and Space*, pp. 385-436. MIT Press, Cambridge.

Byron, D. K. (1998). Understanding Referring Expressions. Available online at: http://citeseer.nj.nec.com/byron98understanding.html. Accessed: 26 March 2003.

Cao, Y., B. Jung and I. Wachsmuth (1995). Situated Verbal Interaction in Virtual Design and Assembly. *In: Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-95)*, pp. 2061-2062. Morgan Kaufmann Publishers, San Francisco.

Carlson-Radvansky, L. (1996). Constructing Spatial Templates: The Influence of Reference Frame Selection. *In: Proceedings of the Seventh Midwest Artificial Intelligence and Cognitive Science Conference (MAICS-96).*

Carlson-Radvansky, L. and D. Irwin (1993). "Frames of Reference in vision and language: Where is above?" *Cognition*, Vol. 46, pp. 223-224.

Carlson-Radvansky, L and D Irwm (1994) "Reference Frame activation during spatial term assignment" *Journal of Memory and Language*, Vol 33, pp 646-671

Carlson-Radvansky, L and G D Logan (1997) "The Influence of Reference Frame Selection on Spatial Template Construction" *Journal of Memory and Language*, Vol 37, pp 411-437

Clark, H (1973) "Space, time, semantics, and the child" *In* T E Moore ed *Cognitive development and the acquisition of language*, pp 65-110 Academic Press, New York

Clark, H H and C R Marshall (1981) "Definite reference and mutual knowledge" *In* I A Sag ed *Elements of discourse understanding*, pp 10-64 Cambridge University Press

Cooper, G S (1968) *A semantic analysis of English locative prepositions* Clearinghouse for Federal Scientific and Technical Information Report 1587

Cooper, R , R Crouch, J van Eijck, C Fox, J van Genabith, J Jaspers, H Kamp, M Pinkal, M Poesio, S Pulman and E Vestre (1994) *The State of the Art in Computational Semantics Evaluating the Descriptive Capabilities of Semantic Theories* The FraCas Consortium University of Edinburgh, Universitat des Saarlandes, Universitat Stuttgart, SRI Cambridge, CWI Amsterdam Report D9

Crystal, D (1985) *A Dictionary of Linguistics and Phonetics*, 2nd Ed, Basil Blackwell Inc

Dale, R (1992) *Generating Referring Expressions Building Descriptions in a Domain of Objects and Processes*, MIT Press

Dale, R and C Mellish (1998) Towards the Evaluation of Natural Language Generation *In the First International Conference on Language Resources and Evaluation*, pp 555--562

Dale, R and E Reiter (1995) "Computational Interpretations of the Gricean Maxims in the Generation of Referring Expressions" *Cognitive Science*, Vol 19(2), pp 233-263

Dennett, D (1991) *Consciousness Explained*, Harmondsworth, Penguin

Dowding, J , E O Bratt and S J Goldwater (1999) Interpreting Language in the Context of CommandTalk *In "Communicative Agents The Use of Natural Language in Embodied Systems", Association of Computing Machinery (ACM) Special Interest Group on Artificial Intelligence (SIGART), Seattle, WA*, pp 63-67

Duwe, I and H Strohner (1997) *Towards a Cognitive Model of Linguistic Reference* Universitat Bielefeld Report 97/1 - Situierte Kunstliche Kommunikatoren

Ericksen, C W (1990) "Attentional search of the visual field" *In* D Brogan ed *Visual Search*, pp 3-19

Fillmore, C J (1997) *Lectures on Deixis*, CSLI Publications, Stanford University

Forgus, R H and L E Melamed (1976) *Perception A Cognitive Stage Approach*, McGraw-Hill

Fuhr, T , G Socher, C Scheering and G Sagerer (1998) "A Three-Dimensional Spatial Model for the Interpretation of Image Data" *In* K -P Gapp ed *Representation and Processing of Spatial Expressions*, pp 103-118 Lawrence Erlbaum Associates

Gapp, K -P (1994a) Basic meanings of spatial relations Computation and evaluation in 3D space *In Proceedings of Twelfth National Conference on Artificial Intelligence (AAAI-94), Seattle, WA*, pp 1393-1398

Gapp, K -P (1994b) From Vision to Language A Cognitive Approach to the Computation of Spatial Relations in 3D space *In Proceedings of First European Conference on Cognitive Science Industry, Luxembourg*, pp 339-357

Gapp, K -P (1995a) Angle, Distance, Shape and their Relationship to Projective Relations *In Proceedings of the 17th Conference of the Cognitive Science Society, Pittsburgh*

Gapp, K -P (1995b) An Empirically Validated Model for Computing Spatial Relations *In Proceedings of 19th German Conference on Artificial Intelligence (KI-95), Berlin, Heidelberg*, pp 245-256 Springer

Gapp, K -P (1995c) Object Localization Selection of Optimal Reference Object *In Proceedings of the 2nd International Conference On Spatial Information Theory (COSIT-95), Semmering, Austria*

Gapp, K -P (1996) *Processing Spatial Relations in Object localization Tasks* Universitat des Saarlandes Report SFB 378 Ressourcenadaptive kognitive Prozesse Bericht NR 135

Garrod, S , G Ferrier and S Campbell (1999) "In and On investigating the functional geometry of spatial prepositions" *Cognition*, Vol 74, pp 167-189

Goldwater, S J , E O Bratt, J M Gawron and J Dowding (2000) Building a Robust Dialogue System with Limited Data *In Proceedings of the Workshop on Conversational Systems at the First Meeting of the North American Chapter of the Association of Computational Linguistics, Seattle, WA*

Greenbaum, S (1996) *The Oxford English Grammar*, Oxford University Press

Grice, H P (1989) *Studies in the Way of Words*, Harvard University Press, London

Grosz, B J , A K Joshi and S Weinstein (1995) "Centering A Framework for Modelling the Local Coherence of Discourse" *Computational Linguistics*, Vol 21(2), pp 203-225

Hajicová, E (1993) Issues of Sentence Structure and Discourse Patterns *In Theoretical and Computational Linguistics, Charles University, Prague*, Vol 2

Hernandez, D and A Mukerjee (1995) Representation of Spatial Knowledge *In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence (IJCAI-95), Montreal, Canada*

Herskovits, A (1986) *Language and Spatial Cognition An Interdisciplinary Study of the Prepositions in English*, Cambridge University Press, Cambridge, London

Herskovits, A (1998) "Schematization" *In* K -P Gapp ed *Representation and Processing of Spatial Expressions*, pp 149-163 Lawrence Erlbaum Associates

Herzog, G (1995) *From Visual Input to Verbal Output in the Visual Translator* Universitat des Saarlandes Report SFB 314 Kunstliche Intelligenz - Wissensbasierte Systeme Bericht Nr 124

Herzog, G (1997) *Connecting Vision and Natural Language Systems* Universitat des Saarlandes Report SFB 314 Project VITRA

Herzog, G (2001) Re CITYTOUR system E-mail, 17 July 2001

Herzog, G and P Wazinski (1994) "VIsual TRAnslator Linking Perceptions and Natural Language Descriptions " *Artificial Intelligence Review*, Vol 8(2-3), pp 175-187

Hewett, M S (2001) Computational Perceptual Attention, Ph D , University of Texas, Austin

Hill, C (1982) "Up/down, front/back, left/right A contrastive study of Hausa and English " *In* W Klein ed *Here and there Crosslinguistic studies on deixis and demonstration*, pp 11-42 Benjamins, Amsterdam

Hirst, G (1994) "Reference and Anaphor Resolution in Natural Language Processing" *In* J M Y Simpson ed *The Encyclopaedia of Language and Linguistics*, Vol 7, pp 3487-3489 Pergamon Press, Oxford

Jackendoff, R and B Landau (1992) "Spatial Language and Spatial Cognition" *In* R Jackendoff ed *Languages of the Mind*, pp 99-125 MIT Press, Cambridge Massachusetts

Jording, T and I Wachsmuth (2002) "An Anthropomorphic Agent for the Use of Spatial Language" *In* P Olivier ed *Spatial Language Cognitive and Computational Aspects*, pp 69-86 Kluwer Academic Publishers, Dordrecht

Kamp, H and U Reyle (1993) *From Discourse to Logic*, Kluwer Academic Publishers, Dordrecht

Kosslyn, S M (1994) *Image and Brain*, The MIT Press

Krahmer, E and M Theune (2002) "Efficient Context-Sensitive Generation of Referring Expressions" *In* R Kibble ed *Information Sharing Reference and Presupposition in Language Generation and Interpretation* CLSI Publications, Standford

Kuffner, J and J C Latombe (1999) Fast synthetic vision, memory, and learning models for virtual humans *In Proceedings of Computer Animation Conference (CA-99) Geneva, Switzerland*, pp 118-127 IEEE Computer Society

Landau, B (1996) "Multiple Geometric Representations of Objects in Language and Language Learners" *In* M Garrett ed *Language and Space*, pp 317-363 MIT Press, Cambridge

Landau, B and R Jackendoff (1993) ""What" and "Where" in spatial language and spatial cognition" *Behavioural and Brain Sciences*, Vol 16, pp 217-256

Landau, B and E Munnich (1998) "The Representation of Space and Spatial Language Challenges for Cognitive Science" *In* K -P Gapp ed *Representation and Processing of Spatial Expressions*, pp 262-272 Lawrence Erlbaum Associates

Landragin, F, N Bellalem and L Romary (2001) Visual Salience and Perceptual Grouping in Multimodal Interactivity *In Proceeding of the International Workshop on Information Presentation and Natural Multimodal Dialogue (IPNMD), Verona, Italy*

Langacker, R L (1987) *Foundation of Cognitive Grammar Theoretical Prerequisites*, Stanford University Press, Stanford

Langacker, R L (1991a) *Foundations of Cognitive Grammar Descriptive Applications*, Stanford university Press, Stanford

Langacker, R W (1991b) *Concept, Image and Symbol the Cognitive Basis of Grammar*, Mouton de Gruyter, The Hague

Langacker, R W (1994) "Cognitive Grammar" *In* R E Asher ed *The Encyclopaedia of Language and Linguistics*, Vol 2, pp 590-593 Pergamon Press, Oxford

Leech, G N (1969) *Towards a Semantic Description of English*, Longmans, London

458

Levelt, W J M (1989) *Speaking From Intention to Articulation*, MIT Press, Cambridge

Levelt, W J M (1996) "Perspective taking and ellipsis in spatial descriptions" *In* M Garrett ed *Language and Space*, pp 77-108 MIT Press, Cambridge, Massachusetts

Levinson, S (1996) "Frames of Reference and Molyneux's Question Crosslinguistic Evidence" *In* M Garrett ed *Language and Space*, pp 109-170 MIT Press, Cambridge, Massachusetts

Logan, G D (1995) "Linguistic and conceptual control of visual spatial attention " *Cognitive Psychology*, Vol 12, pp 523-533

Logan, G D and D D Sadler (1996) "A Computational Analysis of the Apprehension of Spatial Relations" *In* M F Garrett ed *Language and Space* MIT Press, Cambridge, Massachusetts

Lyons, J (1977) *Semantics*, Cambridge University Press

McCawley, J D (1993) *Everything That Linguists Have Always Wanted To Know About Logic* (*but were ashamed to ask)*, 2nd Ed, The University of Chicago Press

Miller, G A and P N Johnson-Laird (1976) *Language and Perception*, Cambridge University Press, Cambridge, London, Melbourne

Mukerjee, A (1998) "Neat vs Scruffy A survey of computational models for spatial expressions" *In* K P Gapp ed *Computational Representation and Processing of Spatial Expressions* Lawrence Erlbaum Associates

Mukerjee, A , K Gupta, S Nauityal, P Mukesh, M Singh and N Mishra (2000) "Conceptual Description of Visual Scenes From Linguistic Models" *Journal of Image and Vision Computing*, Vol 18

Munsell, A H (1905) *A Color Notation*, G P Putnam's Sons London

Noser, H , O Renault, D Thalmann and N Magnenat-Thalmann (1995) "Navigation for Digital Actors Based on Synthetic Vision, Memory, and Learning" *Computer Graphics*, Vol 19(1), pp 7-9

Olivier, P (2001) RE Language Visualizer E-mail, 20 July 2001

Olivier, P , T Maeda and J Tsujii (1994) Automatic Depiction of Spatial Descriptions *In Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI '94), Seattle, WA*, Vol 2, pp 1405-1410

Olivier, P. and J.-I. Tsuji (1994). "Quantitative Perceptual Representation of Prepositional Semantics". *Artificial Intelligence Review*, Vol. 8, pp. 147-158.

Peters, C. and C. O'Sullivan (2002). A Memory Model for Autonomous Virtual Humans. *In: Proceedings of Eurographics Irish Chapter Workshop (EGIreland-02), Dublin*, pp. 21-26.

Pinkal, M. (1986). Definite Noun Phrases and the Semantics of Discourse. *In: Proceedings of the 11th International Conference on Computational Linguistics, Bonn*, pp. 368-373.

Poesio, M. (1994). Discourse Interpretation and the Scope of Operators, Ph.D. Dissertation, University of Rochester, Rochester, NY.

Poesio, M. (2003). Definite NPs Statistics. Email, 10th August.

Poesio, M. and R. Vieira (1998). "A Corpus-Based Investigation of Definite Description Use". *Computational Linguistics*, Vol. 24(2), pp. 183-216.

Poesio, M. and R. Vieira (2000). "An empirically-based system for processing definite descriptions". *Computational Linguistics*, Vol. 26(4), pp. 539-593.

Regier, T. (1996). *The Human Semantic Potential: spatial language and constrained connectionism*, MIT Press.

Reiter, E. and R. Dale (1997). "Building Applied Natural Language Generation Systems". *Natural Language Engineering*, Vol. 3(1), pp. 57-87.

Renault, O., N. Magnenat-Thalmann and D. Thalmann (1990). "A Vision-Based Approach to Behavioural Animation". *Visualization and Computer Animation*, Vol. 1(1), pp. 18-21.

Rets-Schmidt, G. (1988). "Various views on spatial prepositions". *AI Magazine*, Vol. 9(2), pp. 95-105.

Reynolds, J. (2001). Visual Salience, Competition, Neuronal Response Synchrony and Selective Attention. *In: Sloan/Swartz Centers for Theoretical Neurobiology Annual Summer Meeting 2001, Lake, Tahoe, NV.* The Swartz Foundation.

Rieser, H. (1999). Observations on Deixis and Pointing Based on the Bielefeld Corpus of Task-oriented Dialogue. *In: Proceedings of the Workshop on Deixis, Demonstration and Deictic Belief at the Eleventh European Summer School in Logic, Language and Information (ESSLLI XI), Utrecht, The Netherlands*, pp. 10-12.

Russell, B. (1905). "On Denoting". *Mind*, Vol. 14, pp. 479-493. Reprinted Logic and Knowledge (1956), pp. 39-56, R. C. Marsh ed.

Salmon-Alt (2001) Reference calculus and human-machine interaction from linguistics towards a computational model, Ph D (Abstract), Université Henri Poincaré, Nancy

Salmon-Alt, S and L Romary (2001) Reference resolution within the framework of cognitive grammar *In Proceedings of the Seventh International Colloquium on Cognitive Science (ICCS-01), Donostia, Spain,* pp 284-299

Schank, R C (1973) "Identification of Conceptualizations Underlying Natural Language" *In* M C Colby ed *Computer Models of Thought and Language,* pp 187-248 W H Freeman and Company, San Francisco

Schirra, J and E Stopp (1993) ANTLIMA - A Listener Model with Mental Images *In Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI-93), Chambery, France,* pp 175-180

Socher, G and U Naeve (1996) *A Knowledge-based System Integrating Speech and Image Understanding* Universitat Bielefeld Report SFB 360 Situierte Kunstliche Kommunikatoren Report 95/15

Socher, G , G Sagerer, F Kummert and T Fuhr (1996) Talking About 3D Scenes Integration of Image and Speech Understanding in a Hybrid Distributed System *In Proceedings of the International Conference on Image Processing (ICIP-96), Lausanne, Switzerland,* Vol 2, pp 809-812

Spivey-Knowlton, M J , M K Tanenhaus, K M Eberhard and J C Sedivy (1998) "Integration of Visuospatial and Linguistic Information Language Comprehension in Real Time and Real Space" *In* K -P Gapp ed *Representation and Processing of Spatial Expressions,* pp 201-214 Lawrence Erlbaum Associates

Stent, A , J Dowding, J M Gawron, E O Bratt and R Moore (1999) The CommandTalk Spoken Dialogue System *In Proceedings of Thirty-Seventh Annual Meeting of the Association of Computational Linguistics (ACL-99), University of Maryland, College Park, MD,* pp 183-190

Talmy, L (1983) "How Language Structures Space" *In* H L Pick ed *Spatial orientation Theory, research and application,* pp 225-282 Plenum Press, New York

Taylor, H , S Naylor, R Faust and P Holcomb (2000) "Could you hand me those keys on the right? Disentangling Spatial Reference Frames using Different Methodologies" *Spatial Cognition and Computation,* Vol 1(4), pp 381-397

Tomasello, M (1987) "Learning to uses prepositions a case study" *Journal of Child Language,* Vol 14, pp 79-98

461

Tu, X. and D. Terzopoulos (1994a). Artificial Fishes: Physics, Locomotion, Perception, Behaviour. *In: Proceedings of ACM SIGGRAPH, Orlando, FL*, pp. 43-50.

Tu, X. and D. Terzopoulos (1994b). Perceptual Modelling for Behavioural Animation of Fishes. *In: Proceedings of the Second Pacific Conference on Computer Graphics and Applications, Beijing, China*, pp. 185-200.

Tversky, B. (1996). "Spatial perspectives in descriptions". *In:* M. Garrett ed. *Language and Space*, pp. 463-492. MIT Press, Cambridge, Massachusetts.

Ungerer, F. and H.-J. Schmid (1996). *An introduction to Cognitive Linguistics*, Addison Wesley Longman Limited.

van Deemter, K. (2001). Generating Referring Expressions: Beyond the Incremental Algorithm. *In: 4th Int. Conf. on Computational Semantics (IWCS-4), Tilburg*.

van Deemter, K. (2002). "Generating Referring Expressions: Boolean Extensions of the Incremental Algorithm". *Computational Linguistics*, Vol. 28(1), pp. 37-52.

van Deemter, K., R. Power and E. Krahmer (2002). TUNA: Towards a UNified Algorithm for the Generation of Referring expressions. Available online at: http://www.itri.bton.ac.uk/home/Kees.van.Deemter/TUNA.pdf2003.

van Eijck, J. (1994). "Discourse Representation Theory". *In:* J. M. Y. Simpson ed. *The Encyclopaedia of Language and Linguistics*, Vol. 2, pp. 977-982. Pergamon Press, Oxford.

Vandeloise, C. (1991). *Spatial Prepositions: A Case Study From French*, The University of Chicago Press.

Vorwerg, C., G. Socher, T. Fuhr, G. Sagerer and G. Rickheit (1997). Projective relations for 3D space: Computational model, applications, and psychological evaluation. *In: Proceedings of the Fourteenth National Conference on Artificial Intelligence (AAAI-97), Providence Rhode Island*, pp. 159-164.

Wachsmuth, I. and Y. Cao (1995). "Interactive Graphics Design with Situated Agents". *In:* F. Wahl ed. *Graphics and Robotics*, pp. 73- 85. Springer.

Winograd, T. (1973). "Procedural Model of Language Understanding". *In:* K. M. Colby ed. *Computer Models of Thought and Language*, pp. 152-186. W. H. Freeman and Company, San Francisco.

Yamada, A. (1993). Studies in Spatial Descriptions Understanding based on Geometric Constraints Satisfaction, Ph.D., University of Kyoto.

Yee, H , S Pattanaik and D P Greenberg (2001) "Spatiotemporal Sensitivity and Visual Attention for Efficient Rendering of Dynamic Environments" *ACM Transactions on Graphics (TOG)*, Vol 20(1), pp 39-65