

# Reference Object Choice in Spatial Language: Machine and Human Models

Michael Barclay

16th August 2010

Submitted by Michael John Barclay, to the University of Exeter as a thesis for the degree of Doctor of Philosophy by Research in Computer Science, August 2010.

This thesis is available for Library use on the understanding that it is copyright material and that no quotation from the thesis may be published without proper acknowledgement.

I certify that all material in this thesis which is not my own work has been identified and that no material has previously been submitted and approved for the award of a degree by this or any other University.

For my parents, who have always been my heroes

I had no idea when I started this research how important supervisors are and how different they can be in terms of the support and advice they give. All those of us who are supervised by Dr Galton know before long how fortunate we are. There is no one more generous with his time and ideas, or more constructive and supportive with his guidance. To Antony my sincerest thanks for all this, these last years have been the most enjoyable of my working life to date. Thanks also to Jonathan, who, though he only got the post of second supervisor recently, has provided some invaluable pointers and comments. I also need to thank the rest of the research group, Larry, Zena and Max for their help and support and again for making it such an enjoyable time. Last but not least all my love and thanks to Sadie, Tommo and Kitty who have been very forbearing and have put up with a lot of distractedness and not a lot of money for longer than they deserved.

## Abstract

The thesis underpinning this study is as follows; it is possible to build machine models that are indistinguishable from the mental models used by humans to generate language to describe their environment. This is to say that the machine model should perform in such a way that a human listener could not discern whether a description of a scene was generated by a human or by the machine model.

Many linguistic processes are used to generate even simple scene descriptions and developing machine models of all of them is beyond the scope of this study. The goal of this study is, therefore, to model a sufficient part of the scene description process, operating in a sufficiently realistic environment, so that the likelihood of being able to build machine models of the remaining processes, operating in the real world, can be established.

The relatively under-researched process of reference object selection is chosen as the focus of this study. A reference object is, for instance, the ‘table’ in the phrase “The flowers are on the table”. This study demonstrates that the reference selection process is of similar complexity to others involved in generating scene descriptions which include: assigning prepositions, selecting reference frames and disambiguating objects (usually termed ‘generating referring expressions’). The secondary thesis of this study is therefore; it is possible to build a machine model that is indistinguishable from the mental models used by humans in selecting reference objects. Most of the practical work in the study is aimed at establishing this.

An environment sufficiently near to the real-world for the machine models to operate on is developed as part of this study. It consists of a series of 3-dimensional scenes containing multiple objects that are recognisable to humans and ‘readable’ by the machine models. The rationale for this approach is discussed. The performance of human subjects in describing this environment is evaluated, and measures by which the human performance can be compared to the performance of the machine models are discussed.

The machine models used in the study are variants on Bayesian networks. A new approach to learning the structure of a subset of Bayesian networks is presented. Simple existing Bayesian classifiers such as naive or tree augmented naive networks did not perform sufficiently well. A significant result of this study is that useful machine models for reference object choice are of such complexity that a machine learning approach is required. Earlier proposals based on sum-of weighted-factors or similar constructions will not produce satisfactory models.

Two differently derived sets of variables are used and compared in this study. Firstly variables derived from the basic geometry of the scene and the properties of objects are used. Models built from these variables match the choice of reference of a group of humans some 73% of the time, as compared with 90% for the median human subject. Secondly variables derived from ‘ray casting’ the scene are used. Ray cast variables performed much worse than anticipated, suggesting that humans use object knowledge as well as immediate perception in the reference choice task. Models combining geometric and ray-cast variables match the choice of reference of the group of humans some 76% of the time. Although neither of these

machine models are likely to be indistinguishable from a human, the reference choices are rarely, if ever, entirely ridiculous.

A secondary goal of the study is to contribute to the understanding of the process by which humans select reference objects. Several statistically significant results concerning the necessary complexity of the human models and the nature of the variables within them are established.

Problems that remain with both the representation of the near-real-world environment and the Bayesian models and variables used within them are detailed. While these problems cast some doubt on the results it is argued that solving these problems is possible and would, on balance, lead to improved performance of the machine models. This further supports the assertion that machine models producing reference choices indistinguishable from those of humans are possible.

# Contents

<b>1</b>	<b>Introduction</b>	<b>16</b>
1.1	The motivation for, and intention of, this study . . . . .	16
1.2	Locative expressions and the focus of the study . . . . .	17
1.3	Elements of locative expressions . . . . .	19
1.3.1	Forms of locative expression . . . . .	19
1.3.2	Prepositions: functional, locative and both . . . . .	19
1.3.3	Reference objects and reference frames . . . . .	21
1.3.4	Spatial Location and Disambiguation . . . . .	23
1.4	Reference object choice . . . . .	23
1.4.1	Multiple contending influences . . . . .	23
1.4.2	The relationship with preposition choice . . . . .	25
1.4.3	Function and discourse . . . . .	25
1.4.4	Simple and compound references . . . . .	26
1.4.5	Qualified references . . . . .	27
1.5	Machine Spatial Language Generation . . . . .	28
1.5.1	Applications for machine spatial language generation . . . . .	28
1.5.2	Machine learning and machine language generation . . . . .	29
1.5.3	Mimicking human behaviour . . . . .	31
1.6	The scope of this study . . . . .	31
1.7	Contribution of the study . . . . .	32
1.8	Organisation of the study . . . . .	33
<b>2</b>	<b>Spatial language generation systems</b>	<b>36</b>
2.1	Introduction . . . . .	36
2.2	SHRDLU . . . . .	38
2.3	The VITRA System . . . . .	38
2.4	Regier’s constrained connectionist system . . . . .	41
2.5	Situated Artificial Communicators . . . . .	43
2.6	Abella and Kender’s scene describer . . . . .	45
2.7	The ‘Describer’ system . . . . .	46
2.8	Kelleher’s ‘Situated Language Interpreter’ system . . . . .	48
2.9	Automatic landmark detection systems . . . . .	50

2.10	Other systems . . . . .	52
2.10.1	The attentional vector sum model . . . . .	52
2.10.2	The ‘Bishop’ system . . . . .	53
2.10.3	Coventry’s functional/geometric neural net system . . . . .	54
2.10.4	Space Case . . . . .	55
2.10.5	The ‘GLIDES’ system . . . . .	56
2.10.6	The GRAAD system . . . . .	56
2.11	Summary . . . . .	56
<b>3</b>	<b>A hypothesis model for reference object choice</b>	<b>60</b>
3.1	Introduction . . . . .	60
3.2	Fundamental influences on reference object choice . . . . .	63
3.3	Influences on reference locatability . . . . .	65
3.3.1	Specific and categorical knowledge. . . . .	65
3.3.2	Degree of belief in listener’s specific knowledge. . . . .	65
3.3.3	Reference apparency. . . . .	67
3.4	Searching for the target object . . . . .	69
3.4.1	Directed or constrained search. . . . .	69
3.4.2	Reference location. . . . .	70
3.4.3	Search start area. . . . .	73
3.4.4	Reference and target topology. . . . .	74
3.5	Communication cost . . . . .	75
3.5.1	Reference innate cost. . . . .	75
3.5.2	Reference ambiguity. . . . .	80
3.6	Other approaches to determining reference characteristics . . . . .	80
3.7	Practical limits on hypothesis model realisation . . . . .	83
3.8	Model evaluation . . . . .	84
3.8.1	Reference interaction . . . . .	85
3.8.2	Reference pruning . . . . .	85
3.8.3	Transition to compound locative sentences . . . . .	86
3.9	Summary . . . . .	87
<b>4</b>	<b>Design and Validation of a test data set</b>	<b>89</b>
4.1	Introduction . . . . .	89
4.2	Scene corpus design . . . . .	91
4.2.1	Corpus requirements . . . . .	91
4.2.2	Corpus derivation . . . . .	93
4.2.3	Scene corpus construction . . . . .	95
4.2.4	Test case generation . . . . .	99
4.3	Scene construction . . . . .	101
4.3.1	Scene representation . . . . .	101
4.3.2	Object schematisation . . . . .	104

4.4	Data set validation experiments . . . . .	106
4.4.1	Requirements . . . . .	106
4.4.2	Experiment format . . . . .	107
4.4.3	Process . . . . .	108
4.5	Data set validation results . . . . .	109
4.5.1	Selection of performance measure . . . . .	109
4.5.2	Validity of training data . . . . .	114
4.5.3	Human group conformance . . . . .	115
4.5.4	Preposition choice data . . . . .	118
4.6	Listener present scene validation . . . . .	122
4.7	Summary . . . . .	130
<b>5</b>	<b>Machine Learning Methods</b>	<b>132</b>
5.1	Introduction . . . . .	132
5.2	Choice of machine learning method . . . . .	133
5.3	Bayesian network classification techniques . . . . .	135
5.3.1	Principles of Bayesian networks . . . . .	135
5.3.2	Naive Bayes Classifiers. . . . .	146
5.4	Feature combination using interaction information . . . . .	153
5.5	A practical Bayesian classifier construction algorithm . . . . .	163
5.5.1	Forming feature groups . . . . .	163
5.5.2	Combining feature groups . . . . .	169
5.5.3	Terminating the construction process . . . . .	170
5.6	Data handling for the reference choice experiments . . . . .	171
5.6.1	Variable derivations . . . . .	171
5.6.2	Discretisation . . . . .	180
5.6.3	Network training . . . . .	183
5.7	Summary . . . . .	183
<b>6</b>	<b>Results from models of reference object choice</b>	<b>186</b>
6.1	Introduction . . . . .	186
6.2	Experimental conditions and analysis . . . . .	187
6.3	Results from geometric variable models . . . . .	190
6.3.1	Single variables . . . . .	190
6.3.2	Relative size and distance models . . . . .	192
6.3.3	BaseSet plus one (four variable) models . . . . .	193
6.3.4	More complex models . . . . .	195
6.3.5	Further models . . . . .	197
6.4	Extended data set . . . . .	199
6.5	Results from sight-line variable models . . . . .	203
6.6	Effect of scene scale . . . . .	209
6.7	Results from learned structure models . . . . .	212

6.8	Results from listener present models . . . . .	213
6.9	Assessment of over fitting . . . . .	216
6.10	Summary . . . . .	218
<b>7</b>	<b>System performance, limitations and findings</b>	<b>220</b>
7.1	System performance . . . . .	220
7.1.1	Performance of machine models . . . . .	220
7.1.2	Relative performance of Bayesian network variants . . . . .	223
7.1.3	Were the best models found? . . . . .	225
7.2	System limitations . . . . .	226
7.2.1	Limitations of the test data set . . . . .	226
7.2.2	Possible deficiencies of machine models . . . . .	228
7.3	What has been learned about spatial language? . . . . .	235
7.3.1	Influencing factors and variable representations . . . . .	235
7.3.2	Human performance in the reference choice task . . . . .	238
7.3.3	Extension to compound locative expressions . . . . .	239
7.3.4	Better forms of model . . . . .	240
<b>8</b>	<b>Conclusions and directions for further work</b>	<b>242</b>
8.1	Nature and context of the study . . . . .	242
8.2	Achievements of the study . . . . .	243
8.3	Future directions for research . . . . .	244
<b>A</b>	<b>Object types used</b>	<b>246</b>
A.1	Primitive object types . . . . .	246
A.2	List of named objects . . . . .	247
<b>B</b>	<b>Validation results</b>	<b>248</b>
<b>C</b>	<b>Sample file listings</b>	<b>263</b>
C.1	Sample scene files . . . . .	263
C.2	Sample network files . . . . .	263
C.3	Sample validation files . . . . .	263



# List of Figures

1.1	A possible perception to language system . . . . .	17
1.2	Where is the man? . . . . .	24
1.3	A typical scene from the test data set . . . . .	30
1.4	A schematic of the experimental platform used . . . . .	33
2.1	Example trajectories from the CITYTOUR project . . . . .	40
2.2	VITRA system architecture . . . . .	41
2.3	Regier’s preposition learning system architecture . . . . .	42
2.4	Bayesian network used in the Situated Artificial Communicator . . . . .	44
2.5	Abella and Kender’s system - example . . . . .	45
2.6	Kelleher’s Situated Language Interpreter system architecture . . . . .	49
2.7	Derivation of the attentional vector sum model . . . . .	52
3.1	Competing influences on reference choice . . . . .	62
3.2	Three primary influences on reference suitability . . . . .	64
3.3	Influences on reference locatability . . . . .	66
3.4	Influences on search-space optimisation . . . . .	74
3.5	Influences on communication cost . . . . .	78
3.6	Disambiguation or aggregation . . . . .	80
3.7	Realisable hypothesis model . . . . .	84
3.8	Interacting references . . . . .	85
4.1	Typical scenes from the data set . . . . .	90
4.2	Scenes showing manipulated geometries . . . . .	96
4.3	Structure of a scene file . . . . .	102
4.4	Scene corpus interfaces . . . . .	104
4.5	Data-set validation experiment instructions . . . . .	110
4.6	Data-set validation scene display . . . . .	111
4.7	Validation results . . . . .	112
4.8	Consensus of participants on reference choice for each scene . . . . .	113
4.9	Comparison of top one and top three matches . . . . .	113
4.10	Conformity of validation participants to group choice of reference . . . . .	115
4.11	Scenes where the author disagreed with group choice of reference . . . . .	116

4.12	Comparison of human choices with random distribution on reduced object numbers . . . . .	118
4.13	Correlation of human reference choice consensus with scene complexity . . .	119
4.14	Relative frequency of directional preposition selection; comparison between this study and the corpus based study of de Vega et al. [2002]. . . . .	121
4.15	Data-set validation experiment instructions . . . . .	124
4.16	Data-set validation scene display . . . . .	125
4.17	Listener present validation results . . . . .	127
4.18	Listener chosen as reference . . . . .	128
4.19	Scenes in which a listener changes reference choice . . . . .	128
4.20	Conformity to group choice with listener present . . . . .	129
5.1	A joint probability table . . . . .	136
5.2	The general form of a Bayesian network. . . . .	137
5.3	Evaluation of a Bayesian network using message passing. . . . .	139
5.4	An example of entailed dependency . . . . .	142
5.5	The Markov blanket of a variable . . . . .	143
5.6	Variants of naive Bayesian classifiers . . . . .	147
5.7	The effective structure of a tree augmented naive Bayes network (from figure 5.6b) in the fully observed case. . . . .	148
5.8	Less restricted forms of Bayesian classifier . . . . .	150
5.9	Illustration of the effects of redundancy . . . . .	151
5.10	Class discrimination using CMI and II . . . . .	157
5.11	Examples of positive and negative interaction information . . . . .	158
5.12	Class discrimination with high negative interaction information . . . . .	160
5.13	Class discrimination with parents chosen using mRMR . . . . .	162
5.14	Combining variables in to groups . . . . .	166
5.15	Distribution of interaction information variables for a complex network . . .	171
5.16	2-dimensional representation of the different distance measures . . . . .	173
5.17	Schematic of proximal latitude definition . . . . .	174
5.18	Schematic of proximal longitude definitions . . . . .	174
5.19	Derivation of a cast ray . . . . .	177
5.20	Effect of varying the number of bins on test results . . . . .	181
5.21	Effect of using arithmetically spaced bins on test results . . . . .	182
5.22	Effect of varying bin boundaries on test results . . . . .	182
5.23	Effect of increasing numbers of training iterations . . . . .	184
6.1	Network topology for the assessment of single variables . . . . .	191
6.2	Single variable model performance . . . . .	191
6.3	Network topologies for the assessment of size and distance variables . . . .	192
6.4	Size and distance only results . . . . .	193
6.5	Network topologies used for assessing four variable models . . . . .	194

6.6	Four variable models . . . . .	195
6.7	The combined clustered network topology for a more complex model . . . .	196
6.8	Results from more complex geometric models . . . . .	197
6.9	The combined clustered network topology for the final models analysed . .	198
6.10	Results from final geometric models . . . . .	199
6.11	Performance of ‘best’ network on reduced training set sizes for original and extended data sets . . . . .	200
6.12	Performance of key models on series 1 and 2 scenes . . . . .	201
6.13	Distribution of number of objects in scenes . . . . .	202
6.14	Effect of number of objects in a scene on model performance . . . . .	202
6.15	Performance of single sight-line variable models . . . . .	204
6.16	Performance of baseline sight-line variable models . . . . .	205
6.17	Performance of sight-line variables with non-baseline variable models . . . .	206
6.18	Performance of salience plus search space models . . . . .	207
6.19	A more complex network combining sight-line and geometric variables . . . .	208
6.20	Results from the combined sight-line / geometric networks . . . . .	208
6.21	Networks with scene scale variable . . . . .	211
6.22	Results including a scene scale variable . . . . .	211
6.23	Results from learned structure networks . . . . .	212
6.24	Worst performing network produced by the structure learning algorithm . .	213
6.25	Best performing network produced by the structure learning algorithm . . .	213
6.26	Effect of listener related variables in scenes with a listener present . . . . .	214
6.27	A complex network with listener related variables . . . . .	215
6.28	Effect of listener related variables on a complex network . . . . .	216
6.29	Assessment of over-fitting . . . . .	217
7.1	Comparison of human and machine performance in the reference choice task	220
7.2	Cases where machine model and human (group) choices differ 1 . . . . .	222
7.3	Cases where machine model and human (group) choices differ 2 . . . . .	223
7.4	An example tree augmented naive Bayes network . . . . .	224
7.5	Alternative derivations for search distance vectors . . . . .	235
7.6	An extension to the model to generate hierarchical references . . . . .	239
7.7	A possible improvement to model structure . . . . .	241
B.1	Validation results 1 of 7 . . . . .	249
B.2	Validation results 2 of 7 . . . . .	250
B.3	Validation results 3 of 7 . . . . .	251
B.4	Validation results 4 of 7 . . . . .	252
B.5	Validation results 5 of 7 . . . . .	253
B.6	Validation results 6 of 7 . . . . .	254
B.7	Validation results 7 of 7 . . . . .	255
B.8	Listener present validation results 1 of 6 . . . . .	256

B.9 Listener present validation results 2 of 6 . . . . .	257
B.10 Listener present validation results 3 of 6 . . . . .	258
B.11 Listener present validation results 4 of 6 . . . . .	259
B.12 Listener present validation results 5 of 6 . . . . .	260
B.13 Listener present validation results 6 of 6 . . . . .	261
B.14 Scenes from series not used in validation . . . . .	262

# List of Tables

2.1	Spatial Language Systems Summary 1 . . . . .	58
2.2	Spatial Language Systems Summary 2 . . . . .	59
3.1	Talmy’s proposed target and reference object characteristics . . . . .	61
4.1	Subject distribution and derivation of scenes in the corpus . . . . .	100
4.2	Reference choice distribution over position of chosen reference in presented list . . . . .	108
4.3	Age breakdown of participants in the first validation exercise . . . . .	109
4.4	Rating table for group and median human, match most popular choice . . .	116
4.5	Rating table for group and median human, match one of top three choices .	117
4.6	Relative frequency of preposition selection . . . . .	120
4.7	Relative frequency of directional preposition selection . . . . .	120
4.8	Preposition selection in internal and external scenes . . . . .	122
4.9	Age breakdown of participants in the listener present validation exercise . .	126
4.10	Relative frequency of preposition selection with listener present . . . . .	130
5.1	Guide to notation . . . . .	135
5.2	Example of Bayesian network parameter learning . . . . .	144
5.3	Mutual information values . . . . .	155
5.4	Conditional mutual information values . . . . .	156
5.5	Interaction information and mutual information values for a variable subset. The classifier $CL$ is ‘reference suitability’ for all the variable pairs $A_i, A_j$ for the interaction informations $I(A_i; A_j; CL)$ . shown. The mutual information values are $I(A_j; CL)$ . . . . .	163
5.6	Classifier construction algorithm output (1st stage) . . . . .	166
5.7	Classifier construction algorithm output (2nd stage) . . . . .	166
6.1	Illustration of Wilcoxon signed rank test . . . . .	188
6.2	Critical values for the Wilcoxon signed rank test . . . . .	189
6.3	Illustration of model significance and performance . . . . .	189
6.4	Significance of variable to improvement in reference choice prediction over BaseSet variables . . . . .	195

6.5	Correlation of machine model performance with number of objects in a scene for the best model from figure 6.10. (Product moment correlation coefficient).	203
6.6	Performance of the best network from figure 6.20 when trained and tested separately on interior and exterior scenes . . . . .	209
6.7	Bin allocations for reference object volume (materialVolRef) for interior, exterior and combined scene sets using $S^3$ scale factor . . . . .	210
6.8	Bin allocations for reference object volume (materialVolRef) for interior, exterior and combined scene sets using $S^2$ scale factor . . . . .	210
6.9	Performance of the best network from figure 6.20 when trained and tested separately on interior and exterior scenes using $S^2$ scale factor . . . . .	210
6.10	Performance of the best network from figure 6.20 when trained and tested separately on scenes with a listener present and not present. Note, no cross validation has been used . . . . .	214
A.1	List of distinct selectable object types . . . . .	247

## Terminology

Various terms are used in this study which either appear to have no universally accepted definition or are one of a variety of terms used by different researchers. The following list gives the meaning intended by these terms in this study.

**Geometric extension.** This is also sometimes described as a ‘shape factor’. In this study an object with ‘high’ or ‘large’ extension is likely to be long and thin (e.g., a pencil). An object with low extension would be typified by a die.

**Locative phrase, sentence or expression.** In earlier publications by the author the term locative *phrase* has been used. However formally most of the phrases, containing as they do, a subject, object and a verb are actually sentences. In this study the term ‘locative expression’ is adopted in line with other recent researchers [Kelleher and Costello, 2009]

**Target (or target object).** In spatial cognition literature various terms are used for an object that is to be found or located. In this study the term ‘target’ or ‘target object’ is used, rather than ‘trajector’ [Regier, 1996] or ‘Located object’ [Carlson and Hill, 2008] which have the same meaning. So in the expression “the cup is on the table” the cup is the target.

**Reference (or reference object).** Landmark, ground or relatum are used to mean the same thing as Reference in this study. It is the object to which the location of the target is being referenced. In the expression “the cup is on the table” the table is the reference.

**Fixed computational model.** This is taken to mean a mathematical expression, in which any variables and constant parameters used and the relationships between them have been set down (fixed) by human decision.

**Machine learned computational model.** By contrast to a fixed model this is taken to mean a model whose variables, parameters and structure are at least in part derived by a machine employing some algorithm over a starting data set.

In addition the following abbreviations have been used although through personal preference the use of abbreviations is limited as far as possible:

---

CMI	Conditional Mutual Information
II	Interaction Information
MI	Mutual Information
mRMR	minimum Redundancy Maximum Relevance
TAN	Tree Augmented Naive [Bayesian Network]
PMCC	Product Moment Correlation Coefficient

---

The mathematical definitions of these quantities and other notational definitions are given in section 5.3.

# Chapter 1

## Introduction

### 1.1 The motivation for, and intention of, this study

To see that spatial communication is among the most fundamental and important forms of communication in which humans engage, consider only the sentence, "There is a lion under the trees!". Before the emotional, political, financial, cultural, artistic or other realms were invented as subjects for discourse the physical realm existed and required description and discussion. Spatial metaphors influence and structure more areas of human communication than any other (see Lakoff and Johnson [1980]). This is not just a question of linguistic convenience but one which illustrates important aspects of human understanding. We say that "Tom is *in* love" but "Harry is *on* guard". Tom is perhaps immersed, not necessarily in control, as if he was *in* the sea. Harry is positioned possibly with a clear view maybe *on* a watch tower. The importance of spatial information and the urgency with which it is sometimes exchanged also suggests that these expressions will have evolved to achieve a near optimum utility and that studying them will possibly shed light on the development of human communication as a survival tool.

Another aspect of spatial language use, and the cognitive processes behind it, that can reveal much about human thought in general, is the combination of knowledge about the world obtained over years of experience, with the perceptions of the moment in time and space which is being described. A spatial expression will contain immediate judgements about shape, size, distance and, topology along with acquired knowledge of object characteristics and function, and also the experience of 'naive' physics: gravity, friction, inertia and the properties of substances and objects, hardness, deformability mobility, animacy etc. To borrow from Feist and Gentner [2003] a coin is more often described as "in" another object such as a hand or a dish than a firefly. The geometry in each case may be identical but we have prior knowledge of the animacy of the firefly. Further examples of this interplay between knowledge and judgement are discussed in section 1.4.

Given this it is not surprising that spatial language and the cognitive processes that underlie it have received a lot of attention from researchers over the past decades. The elements of spatial language and some (there is a vast amount) of the most important



research into each of them are described later in this chapter. The majority of the work undertaken however, deals with the elucidation of single influencing factors on language, or perhaps the balance between two competing influences, and almost always is situated in highly simplified or completely synthetic experimental environments. This work is of course vital and provides hard evidence about factors influencing human behaviour, however there is a less developed complementary strand of research, which is to assess how these influences interact in human spatial language production in the real world. Although this could be attempted, by interviewing subjects about why they made judgements about spatial language in real world situations, it would be difficult to obtain hard evidence about some of the subtle judgements. To give an example that this study sheds some light on, it seems unlikely that subjects would reliably judge (and communicate the judgement of) whether they had used the distance between object centroids, or the distance between closest points on objects, when making a decision about a spatial description. This would be even more unlikely if the size of the objects, their angular and topological relationships and possibly many other factors were also simultaneously involved in their language production.

This motivated the use of machine learning techniques in this study. By using a machine to model human behaviour and then examining the machine, in terms of its use and organisation of information, we can infer something of the processes being used by humans in producing spatial language, even in complex ‘near real world’ situations. It is important to note however that, even though the machine learning methods have had to be described in some detail, this is a study of spatial language using machine learning as a technique, rather than a study of machine learning using spatial language as an example.

To summarise: Spatial language seems to offer a window on fundamental aspects of human cognition, but the complexities underlying its use in real world situations have hitherto been difficult to research. It is hoped in this study to make a start, if only on a limited aspect of spatial language, in investigating a broader range of interacting influences through the use of machine learning techniques operating in near real world environments.

## 1.2 Locative expressions and the focus of the study

Spatial language is the result of a process of translating perception to language. Figure 1.1 shows a highly simplified schematic of a system that might accomplish this.

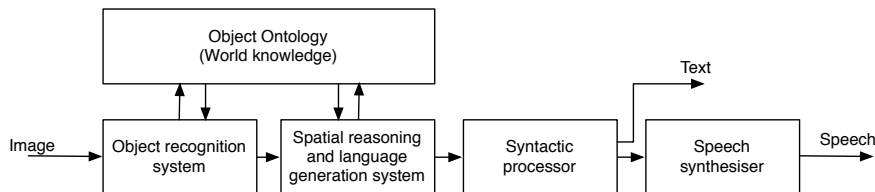


Figure 1.1: A possible perception to language system

A spatially locative expression is one in which a reference object, a spatial preposition

and an object to be located (the “target”) are combined for the purpose of identifying the position of the target to the listener. Hence “The cup is on the table” and “The bird flew over the hill” are both examples of simple spatially locative expressions, with the cup and the bird respectively as targets, and the table and the hill as reference objects, further information and any temporal reference being provided by the verb.

To form a locative expression, given a target object, a speaker has to make three decisions:

1. Decide on an appropriate reference object.
2. Select a reference frame (see section 1.3.3).
3. Assign a suitable preposition.

The focus of this study is the first of these decisions, choosing a reference object. There are two reasons for choosing this, firstly it seems that compared to the other two parts of the process the choice of reference object seems relatively un-researched, and secondly that compared to the other two areas the choice of a reference object most benefits from considering the problem in near real world situations.

The assumption made at the outset of this study is that *if the purpose of the expression is genuinely locative* the first step in the process is, as listed above, the choice of a reference object. This is further discussed in section 1.4; however, if true, it explains why the choice of reference must be considered in more complex environments than the selection of a reference frame or assignment of a preposition.

Herskovits [1998], proposes that the process of generating spatial language is one of progressive schematisation, or perhaps filtering, of irrelevant information from the scene being described. Considering the information bandwidth reduction from the visual scene input to the system in figure 1.1, to the text or speech output, this seems reasonable. The following sequence of steps seems to accord with this process and delivers simplification at each step:

1. The objects in the scene are ‘recognised’, a step which allows us to replace a complex visual representation of an object with its necessary geometric characteristics and a name, which serves to point to its functional and other characteristics.
2. Choosing a suitable reference object effectively removes extraneous objects from the scene leaving only the target, the reference and possibly a few key ‘distractor’ objects that may influence the following decisions. Note that the spatial relationship between the target and the reference will be important in the choice of a suitable reference, but that this is not the same as choosing the appropriate spatial preposition.
3. A reference frame is then selected. This process is often, but not always, trivial (see section 1.3.3). The result however, is that contending reference frames are removed from consideration, leaving the chosen reference frame to fix the axes for the next step.

4. The geometry of the remaining objects together with functional and/or topological relationships between them are combined in the assignment of an appropriate preposition.

If this sequence is accepted it can be seen that it is valid to research preposition assignment and reference frame choice in scenes containing very few objects, whereas research into reference object choice will require more complex scenes containing multiple objects. In other words reference choice will need to be investigated in environments closer to the real world.

### 1.3 Elements of locative expressions

#### 1.3.1 Forms of locative expression

As well as the previously noted ‘canonical case’ where the target is a tangible object (“The cup is on the table”), in the landmark domain a phrase such as “The meeting place is in front of the cathedral” is clearly in the category of a locative expression even though the meeting place is not an object as such. The same is true of the ‘decision point’ in a route description such as “turn right in front of the petrol station”. An expression which combines only one reference and preposition with a target can be termed a simple locative expression. More complex spatial descriptions are typically built up from these simple locative expressions by a few relatively simple processes. (see for instance Porzel et al. [2002] on ‘linearisation’, that is, producing complex descriptions from strings of simple locative expressions, and Plumert et al. [2001] on hierarchical descriptions of object locations). In this study only simple locative expressions with tangible references are used. It is hoped however that a full understanding of the basic building blocks will go some way to enabling insight into complex communication processes.

#### 1.3.2 Prepositions: functional, locative and both

Although not the focus of this study, the use and comprehension of prepositions sheds some light on possible factors influencing the choice of reference object. For this reason, as well as for background on what has probably been the most active area of research in spatial language in recent years, some of the key ideas are discussed here.

As has been noted the group of words that we call spatial prepositions is used far outside the realm of spatial descriptions. These are not the concern of this section, in which the focus is the spatial or apparent spatial usage. Also, as noted by Miller and Johnson-Laird [1976] phrases of the type “the bird is out of the cage” conform to the template of a spatially locative phrase and appear to be describing a spatial description but actually provide no effective communication about the location of the target object. Instead the purpose of the phrase is to convey the information that the cage is not fulfilling its *normal* containment function. The same can be said to be true of the phrase “The flowers are in the vase”, which, if the vase is as mobile as the flowers, conveys only the information that the vase

is performing its containment function. It doesn't help a listener find the 'flowers in the vase', which may be anywhere in the house (say). Although functional relationships are vital in understanding what is meant by spatial prepositions these 'non-locative' senses are not important for this study.

A study by Lakoff [1987] building on work by Brugman [1981] maps the evolution of the word 'over' and its different meanings, arriving at some 14 variants of three basic 'meanings' (over as a path across and above, over as covering and over as above) dependent on the geometric characteristics or topological relationships between the objects involved in the spatial expression. Many of these specifications can also be mapped onto metaphorical uses but the intention is to pin down a complete but finite set of use-cases for 'over', that is, a lexical specification. This approach to prepositions has also been applied in the entirely spatial realm, perhaps most notably by Herskovits [1986] who arrives at a dozen or more different 'uses' of the word 'in'.

These lexical models, complex as they are, are not the full story. Work by various researchers suggests that the geometric form of an object will alter our view as to the space denoted by a preposition. Objects cannot be idealised to their centroids. Gapp [1995b] looks at this with regard to object extensions and projective prepositions, Herskovits [1998] incorporates geometric form and distance into preposition applicability, and Regier and Carlson [2001] (see section 2.10.1) develop the 'attentional vector sum' to explain the appropriateness of prepositions, taking into account the different geometric extensions, as well as positions, of reference and target.

In the above mentioned studies the objects are more or less abstract shapes and as Coventry [1998] points out this is still not sufficient to account for human comprehension of spatial prepositions. The space we understand to be denoted by a preposition does not simply depend upon a canonical geometry related to the cardinal axes (up/down, left/right, front/back) and the form of the objects, but is also heavily influenced by functional relationships between objects and the characteristics of those objects. Coventry and Garrod [2004] provide the comprehensive overview of preposition use and the functional, geometric and other influences involved. Specific experiments by Coventry et al. [2001] demonstrate this with respect to the prepositions 'over', 'under', 'above' and 'below', Garrod et al. [1999] with respect to 'in' and 'on' and Feist and Gentner [2003] investigate 'in' and 'on' with respect to containment and the properties (such as animacy) of the target object.

Although many computational models of preposition use take into account the geometric form of an object (see for instance Gapp [1995b], Regier and Carlson [2001]) seemingly the only attempts to derive computational models incorporating function as well as geometry are due to Coventry et al. [2005], and Lockwood et al. [2006]. These models are described further in chapter 2.

The importance of this is that the functional relationship between the objects involved makes the difficulties of machine generation of spatial language considerable, even if the non-locative and metaphorical senses of spatial language are neglected. Coventry [1998] rules out the possibility of complete lexical specifications of prepositions which, if possible,

would make machine selection of preposition a choice between a relatively small number of prototypes. Instead he points to the evidence from psycho-linguistic studies and in effect proposes, as a requirement for plausible models of spatial language use, a complete ontological account of objects and their functional relationships.

How much of this also applies to the choice of reference object is of considerable interest to this study and is discussed in section 1.4.3 and with relation to the results of the study in chapter 7. In as much as it is relevant the requirement for an object ontology is assumed in this study (see figure 1.1).

### 1.3.3 Reference objects and reference frames

The choice of a reference frame seems at first sight to be a fairly simple matter. Many researchers adopt a system of three reference frames (see for instance Retz-Schmidt [1988], and Carlson-Radvansky and Irwin [1993] for a description and bibliography of the derivations) as follows:

1. Intrinsic, centred on the reference object and adopting the object's cardinal axis arrangement (e.g., "The ball is in front of the car")
2. Absolute (or extrinsic), taking a reference from a global or external object or system (perhaps most usually the globe as in "Exeter is west of London" but also for instance in a local organisation such as "Dave is in front of Eric in the queue")
3. Deictic, centred on the speaker and adopting his cardinal axis arrangement and used when the reference object has no intrinsic reference frame (e.g., "The rock is to the left of the tree")

Levinson [1996] pointed out the important distinction between binary and ternary relationships and classified intrinsic relationships as binary (requiring only the target and reference objects to define the spatial relationship) and deictic relationships as ternary (requiring the orientation of the reference and target relative to the speaker to define the spatial relationship). He extends this scheme to include the special cases of the speaker and listener being the reference object and the listener or a third party being the origin of the ternary relationship to arrive at the following scheme:

1. Intrinsic (speaker centred) for example "The ball is in front of me"
2. Intrinsic (object centred) for example "The ball is in front of the car"
3. Intrinsic (listener centred) for example "The ball is in front of you"
4. Relative (speaker centred) for example "The ball is in front of the tree"
5. Relative (listener centred) for example "The ball is in front of the tree (from your point of view)"

6. Relative (third party centred) for example “The ball is in front of the tree from John’s point of view”

Strangely Levinson then goes on to assert that the absolute frame of reference is a separate binary relationship, rather than a general case of the Relative relationship. It is difficult to accept that the direction ‘north’ in “the tree is north of the rock” is not defined relative to the earth’s magnetic field or that the expression “the hatch is aft of the main mast” is not defined relative to a ships orientation. In both cases a third object is required to fix the relationship between reference and target, as it is with the speaker relative (deictic) case.

In this study the relative-intrinsic distinction will be used with special cases for the speaker and listener leading to the following terminology:

1. Binary relationships (intrinsic)
  - (a) ‘Object-intrinsic’ (or simply ‘intrinsic’), which is equivalent to the standard interpretation of the intrinsic reference frame with the co-ordinates being fixed by the reference object
  - (b) ‘Speaker-intrinsic’ the special case of the speaker being the reference object.
  - (c) ‘Listener-intrinsic’ the special case of the listener being the reference object.
2. Ternary relationships (relative)
  - (a) ‘Object-relative’ which is equivalent to the standard interpretation of the absolute reference frame. In practice the specific object can be used instead of the generic as in ‘world-relative’ to denote a North-South-East-West co-ordinate system.
  - (b) ‘Speaker-relative’ which is equivalent to the standard interpretation of the deictic reference frame.
  - (c) ‘Listener-relative’ the case where the speaker constructs a description from the listeners point of view.

Although there is an argument for including, as Levinson does, a special case for descriptions relative to a third person, as an example of an independently mobile external reference, this is not a case that occurs in this study.

The situation becomes more complex when functional relationships between objects are considered and also when the objects involved are not in their canonical orientations (a chair lying on its side for instance). This has been researched by Carlson-Radvansky and Irwin [1994] who look at non-canonical conditions and conclude that use of the intrinsic reference frame is inhibited if an object is not in its canonical orientation, and by Carlson-Radvansky and Radvansky [1996] who look at the effect of functional relationships on reference frame choice and conclude that the presence of a functional relationship between a reference and target supports the use of an intrinsic reference frame.

It is not clear whether possession of a reference frame has a direct effect on whether an object is a suitable reference or not. This is complicated by the fact that not all intrinsic reference frames are the same ‘strength’, a car or a person will have a strong intrinsic reference frame, a chair or a desk a weaker intrinsic reference frame, the exact strength of which is possibly dependent on its geometric form. This is further discussed in section 3.2. The effect of cardinal axis alignment to intrinsic reference frames, and the alignment of target objects to these cardinal axes, is likely to be important however. For this reason objects in the test data set used in this study that have intrinsic reference frames are denoted as such and their intrinsic cardinal axis orientations are available.

### 1.3.4 Spatial Location and Disambiguation

The issue of reference choice is not the same issue as that of generating referring expressions (see for example Dale and Reiter [1995], Duwe et al. [2002]). In referring expression generation the target is disambiguated from a group of similar objects by adding qualifiers to the target, so for example “The big red dog with the collar” might serve to specify a particular dog in a group of small animals. None of the group may require locating as such and the expression has not helped the listener to *locate* the target. The question being answered in referring expression generation is “Which?” rather than the question “Where?” considered in this study. However there are some areas of overlap between the two fields.

A considerable amount of effort has been expended on generating algorithms which select sufficient adjective sets for disambiguation while being as concise as possible (see for instance van Deemter [2002], Krahmer et al. [2003]). Spatial location can also be used to disambiguate, as in “The small white dog next to the big red dog”. This is a hybrid expression serving to disambiguate and possibly, but not necessarily, locate the target. In as much as the location element is not concerned solely with disambiguation it would fall within the scope of this study, but the question addressed by this study would be why was the ‘big red dog’ chosen as the reference rather than some other object, not how the reference should be disambiguated. Spatial location in disambiguation is addressed in work by Tenbrink [2005] and Vargas [2004] among others, however the factors affecting the reference choice are not investigated.

The target object in this study is uniquely specified: it needs only to be located, not disambiguated. By contrast to the target object, a candidate reference object in this study may or may not be ambiguous, and this ambiguity may have an influence on reference choice. This is further discussed in chapter 3.

## 1.4 Reference object choice

### 1.4.1 Multiple contending influences

The problem addressed in this study, illustrated in figure 1.2, is to identify a suitable reference object from the many present in a scene, and so take the first step in forming a

spatially locative phrase. In figure 1.2, in answer to the question “where is the man?”, the answers “by the skip” or “in front of the pink house” are acceptable answers, but what about “on the sidewalk” or “to the right of the road”? If Talmy’s categorisation (see Talmy [2000] and section 3.1) of reference objects is considered, the road and the sidewalk should be suitable candidates. The sidewalk is actually the closest physical object to the man but its linear extension means that using it as a reference could mean that the man is anywhere along the length of the street.

The situation becomes more complex if the location of the post-box is considered. The same arguments about the suitability of the road and sidewalk apply as they did for the man, but what about the skip? In this case the fact that the skip may not be a permanent feature comes in to play. If the description of the location of the post-box is for a friend going out to post a letter later in the day the skip may not even be present. The skip is a potential reference when it is less mobile than the target but of more dubious worth when it is more mobile than the target.

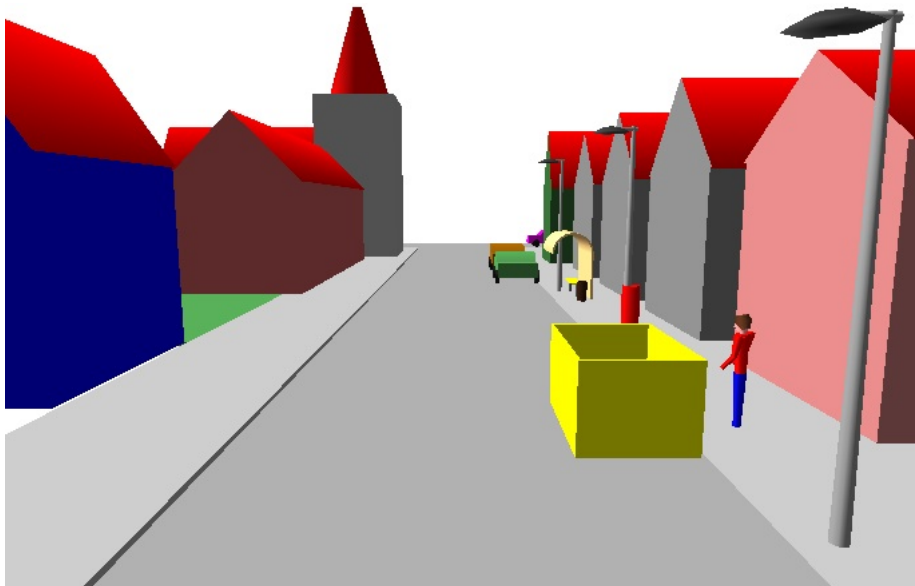


Figure 1.2: Where is the man?

As another example a cup may be *usefully* described as “on the table” even if it is actually in a saucer which is on a tablecloth which is on the table. The cup might not *usefully* be described as “in the saucer”, since the saucer is as mobile as the cup and does not help a listener find the cup. The tablecloth, by its nature not a rigid object, conforms to the shape of the table and in a sense becomes insignificant in defining the area in which the cup can be found. The table may not even be visible if it is covered by the cloth and yet it is still the suitable reference. A human would discard the saucer and cloth as unsuitable references almost without thought but the processes involved in this decision, how they would apply to more borderline cases, and in particular how the different influences on the choice interact, are not properly understood. A human would equally discard the ‘floor’ as



a reference even though the topological chaining of ‘on’ appears to be headed this way (see section 3.4.4 for a discussion of topological relationships and reference suitability).

What can be concluded from these multiple contending influences, and the nature of them in terms of the complex geometries and topologies involved and the characteristics of the objects concerned, is that a near real world environment will be needed in which to base a study of these contending influences. The near real world environments used in this study are described in chapter 4.

### 1.4.2 The relationship with preposition choice

Along with a reference object a speaker must select a spatial preposition and decide on a reference frame in order to form a meaningful spatially locative phrase. A key question for a study such as this is “in what way, if at all, is the choice of reference object independent of the choice of preposition and reference frame?”. To a certain extent this can be answered by the study: if successful models of reference choice can be constructed that are independent of preposition or reference frame choice then there is a strong likelihood that this independence is real. There are also good reasons for thinking that reference choice is independent of preposition choice although there are common factors in the choice of both. On one hand it seems intuitive that when describing the location of a target the preposition is not the starting point: that is, a person will not choose ‘left’ and then look for a candidate reference that conforms to this choice of preposition. On the other hand it is clear that a reference should be close to a target, but this does not mean that a preposition such as ‘near’, ‘next to’ or ‘by’ has first been chosen, only that choice of a good reference object will bias the use of spatial prepositions towards those indicating proximity.

It is true that in interpreting a spatially locative expression the space in which the listener will search for the target is defined by the preposition in combination with the reference and its functional or other associations with the target. This might suggest that what is in fact happening when reference objects are being selected is that reference/preposition pairs are being selected and compared. This is subtly different from the idea that a reference is chosen with regard to various geometric factors which include the definition of a search space and subsequently a preposition is chosen that best fits this reference. In this study the latter is implicitly assumed as a start-point. The reference selection process is independent of preposition selection which is assumed to follow afterwards. Reference choice is not assumed to be independent of the spatial relationship between reference and target, but this is not taken as being equivalent to preposition choice. The soundness of this assumption is further discussed in 3.4.2, and in the light of the results of the study, in chapter 7.

### 1.4.3 Function and discourse

If someone is trying to find the flowers the phrase “the flowers are in the bedroom” is more likely to be helpful than “the flowers are in the vase” even though there is a con-

ceptual/functional link between ‘flowers’ and ‘vase’. If “the bird is in the cage” and the cage is fixed and its location known then the cage is also a good spatial reference for the bird as well as having a conceptual/functional link to the bird. In this study the starting assumption is that the aim of a speaker is to select a good spatial reference, irrespective of any functional relationship, and that the existence of a functional relationship between a target and a reference does not of itself make the reference more suitable (see the discussion of the work by Carlson and Hill [2008] in section 3.4.2).

Another question that bears on this study is that of whether purely spatial descriptions of *static* scenes are ever given. Every scene has a history and most descriptions will be part of a discourse. Both of these processes will alter the salience of some objects in the scene relative to others and the result will be that at best spatial descriptions exist on a spectrum from purely spatial through those partially influenced by discourse or history to those that, while appearing to be spatial have little locative value at all and are entirely concerned with the discourse. The scene in figure 1.2 illustrates this point. Taken as a static scene the descriptions “The man is in front of the pink house” or “the man is behind the dumpster” might be most appropriate. If the man’s activity with a suspect parcel is being monitored “The man has got to the post box” might be better. The phrase has locative value but in a particular discourse context. If the man had previously been in the road with a bus approaching “The man is on the sidewalk” might be the best description. The sidewalk would be used as indicative of a place of safety, and little purely locative value would be conveyed by the phrase.

This study is confined to ‘static’ descriptions, independent of discourse history. While this may seem like a major limitation there is much that is not understood about simple reference object choice in a purely locative sense even before the influence of discourse history is included. That is to say there is still the need to understand how we answer ‘static’ questions such as “where are my car keys?”. If this study can shed any light on this it will be a step towards a full understanding of locative phrases in overall discourse context. It should also be noted that the process of describing object locations in static, context free, pictures did not present any difficulty or apparent strangeness to human participants. The multiple added complexities of describing a scene during, or as part of, a discourse will be the subject of future work.

#### 1.4.4 Simple and compound references

The simple locative expressions discussed so far omit any consideration of the use of ‘between’ which requires two reference objects. Many uses of ‘between’ such as “The train is between London and Exeter” are outside the scope of this study because the references are not derived from a static scene description but are determined by the raised salience of Exeter and London as major destinations on a railway. Although an expression such as “the bench is between the tree and the church” could fall in the scope of the study, an investigation of the circumstances under which the decision is made to incorporate a second reference in this special case will need to be the subject of future work.

Other prepositions such as ‘along’, ‘among’ ‘across’ and particularly ‘opposite’ describe a geometrical arrangement in the scene under consideration which is more complex than that of a projective preposition such as ‘left of’. In this study the starting assumption is that the geometry specified by the preposition is independent of the choice of reference. So for instance in saying that “The pub is opposite the church” the church has been chosen because it is a good reference and the preposition ‘opposite’ has been chosen (rather than ‘near’ perhaps) because the geometry of the scene allows this more specific description. If the pub had been next to the church and opposite a row of houses it is assumed that “the pub is next to the church” would have been the description used. Although this study does not allow for these scene geometries to be taken into account in the reference choice task there are examples in the data set which could suggest whether this omission is justified.

If no suitable single reference can be specified that allows the listener to effectively locate the target, a compound locative phrase is required. This would often contain a hierarchical reference such as “The car keys are on the table next to the telephone directory”. The question of when this transition to a compound expression is required is an important one, incorporating a trade-off between the extra cost and complexity of the communication, against the assistance provided to the listener in the search task. Although the production of compound expressions is not directly addressed in this study some ideas about how the models developed could address this issue are discussed in section 7.3.3.

Also problematic is the case where a reference object requires disambiguation by use of a second reference as in “the keys are on the desk under the window” in the case of a scene containing more than one desk. A second reference must be provided that is suitable for the desired primary reference and unsuitable for any distractors. In practice using the model developed in this study to achieve this may be easier than detecting the problem in the first place and recognising that, though ambiguous, the desk is still a good primary reference because of its conventional use in defining a space (indeed in typifying a scale) where objects are collected.

Cases of reference combinations that are not hierarchical such as “the library is at the intersection of 5th Street and 7th Avenue” will also need to be the subject of future work.

#### 1.4.5 Qualified references

Many spatially locative expressions include what might be described as qualified references:

1. “The tree is on the other side of the river”
2. “The church is at the end of the road”
3. “The post box is by the town hall steps”

It should be noted that these three expressions are subtly different. The first contains the phrase “on the other side of” which denotes a region associated with an object in much the same way as a preposition. (“On the other side” could be replaced with “beyond”

although if “on the other bank” had been used the expression becomes more like the third) This should perhaps be called a qualified preposition rather than a qualified reference. The second expression qualifies the reference by specifying a region of it rather than the whole. The third expression qualifies the reference by specifying a named part of it rather than the whole.

Although there is only one spatial preposition in each expression there is explicitly or implicitly a second non-spatial preposition ‘of’ in each case (in the third expression “the steps *of* the town hall”) and this is where the extra information to aid the search process is introduced. In each of the above cases the process of searching for the target can be further broken down and hence made easier. For example, in the third case the town hall is found, then the steps are found and then the target can be located in a more specific area.

Given this it seems that qualified references are effectively a special case of hierarchical references and hence of compound locative expressions. They are reluctantly consigned to future study.

## 1.5 Machine Spatial Language Generation

### 1.5.1 Applications for machine spatial language generation

Although an investigation of human spatial language in its own right is the motivation for this study there are also important practical applications for machine spatial language generation. These can be broadly divided into two groups. Although considerable work is underway in all the subsystem areas illustrated in Figure 1.1, the problem of object recognition in particular (in anything like cluttered real world environments) cannot be considered as solved. So an important but more futuristic set of applications including scene description from video input for unsighted people, automatic commentary generation, image search given verbal description etc., must wait for this. the second group of applications is, as Kelleher and Costello [2009] point out, virtual environments for which it is not necessary to wait for a solution to the object recognition problem. They list the following applications, to which have been added usage examples:

1. Graphic design and drawing programs. Complex 3-dimensional drawings, particularly of animated construction sequences can be difficult to interpret visually. It can be seen that it would be useful for a system to be able to answer questions such as “how do I access the screws securing the spool-shaft roller bearing?” with perhaps “Under the front cover below the spool arm”.
2. Computer games. As is often the case with new software, games are likely to be the first area for deployment of spatial language generation systems, partly because a game environment is naturally error tolerant. The realism of a computer game would be vastly enhanced if software agents playing alongside humans could, without

reliance on a script, describe their location and surroundings, a facility they currently lack.

3. Navigation aids. It is generally accepted that landmark information improves way-finding instruction giving. Instead of the current sat-nav instruction of “turn right in 200 yards” the more user friendly “turn right in front of the church” could be enabled by machine spatial language generation.
4. Robot systems. This is perhaps the easiest application to visualise. In exactly the same manner as a human, a robot can be directed more easily if it can answer the question “Where are you now?”  
and to these could be added:
5. Training simulators. This is in effect the same requirement as for computer games. The key application seems to be in ‘disaster emergency response management training’. For example in the Kings Cross tube fire co-ordinators above ground quickly lost control of the location and deployment of fire-fighters underground and training simulators for this sort of eventuality require conversational agents that can describe their location.
6. Geographic information system interfaces. The Ordnance Survey have indicated an application which might be termed address generation or address disambiguation. A surprisingly large number of facilities, some of them important, such as electrical substations, have no address or post-code and often need to be referenced to a suitable local landmark. Address disambiguation relates to the reconciliation of commonly used names with official names for places again this can often be achieved by agreeing a reference to a landmark which is known by the same name commonly and officially. This apparently happens sufficiently frequently for an automated process to be worthwhile.

The common factor in these systems is that the objects are known and can be presented to the language generation system as a geometric entity with a name and type (and hence a pointer to the information in an object ontology).

### 1.5.2 Machine learning and machine language generation

Computational models of human spatial language use are not new and various fixed and machine learned models are described in chapter 2. Reflecting the balance of research interest, the majority of systems concentrate on preposition assignment although from the area of geographical information systems there is some work on landmark selection, looking both at fixed and (to a degree) machine learned models.

The advantage of a machine learned model at the most basic level is simply that it is a way of trying out far more models, more quickly, than assigning parameters to a fixed computational model by hand. Using statistical learning techniques, parameters can be

attached to a model that will enable the model to represent the most likely match to the data, given the (almost) inevitable constraints on the degree to which the machine model can represent or be an expression of the ‘true’ model.



Figure 1.3: A typical scene from the test data set and the real world scene from which it is derived

The environments in which the reference selection tasks in this study are performed are ‘schematised’ or ‘pseudo real-world’ scenes such as that shown in Figure 1.3. They contain multiple complex objects in a wide variety of realistic spatial and topological relationships. The objects are represented as collections of 3-dimensional vectors organised into surfaces. It would be very difficult to manually organise and apply weightings to a dozen or more variables derived from these representations to arrive at a fixed computational model of reference selection in this environment. Instead variables expressing the topological and geometric relationships between objects in the scene, as well as variables describing characteristics of the objects themselves, are ‘given’ to a machine learning system. In fact more than 40 different variables are assessed in this study, in a large number of combinations, although some of the variables are variants of each other (three different ways of considering an object’s volume are tested for instance). The machine models used in the study are Bayesian networks and are trained by being given the variables derived from a number of scenes along with a human assessment of what is a good reference object for a target object specified in the scene. The trained models are tested on scenes for which they are not given the good reference to see if they can then match a human choice of reference object.

The results of the study appear to justify the use of machine learned models, in that, even given the number of possible variables and the complexity of the scenes in the test data set, useful information about which variables are important can be gained. The performance of the resultant models and what can be learned from them are discussed in chapter 7.

### 1.5.3 Mimicking human behaviour

There is a complexity in this study which is not present in many machine learning scenarios and it is that there is no absolute right answer to the question “what is a good reference object?” for a particular target.

Clearly there are very bad references in a typical scene and some that are good, but also there are a lot that seem all right to some people but not to others (as can be seen in the results in chapter 4). This creates problems both for training the machine models and also for assessing their performance. The problem of deciding which of the references chosen by humans to give to the machine as ‘good’ references is not too difficult. Deciding whether the machine is modelling human behaviour well, when a group of humans all behave differently, is more difficult. In chapter 4 it can be seen that there is reasonable agreement across a group of people and a group of scenes as to what constitutes a good reference, with some people conforming to the group consensus more than others. This leads to two plausible measures though: that the model should aim for highest conformance to the group although actually this is at one extreme of the spectrum of individual behaviour, or that the model should aim for average conformance to the group which represents the behaviour of the ‘median’ human. Ultimately it might be that a Turing test (on this very limited domain) might be the best way to assess the performance of the machine models; all people behave differently (in this case they are presumed to have slightly different models for reference selection), but all think they can judge what constitutes human-like behaviour. So if a machine can be produced whose reference choices convince a group of humans that they are references chosen by a human this would indicate success.

Given a degree of performance by a machine model there remains the question of what can be learned from it about human behaviour. In an absolutist sense the answer might be nothing. The model might be a ‘Chinese room’ (Searle [1981]) and the behaviour exhibited by the model might be derived from a process utterly unlike anything that a human would use. On the other hand, if a model which uses a certain group of variables achieves a level of performance that is more ‘human’ than a model which does not use these variables, it is difficult to avoid the conclusion that this is telling us something about variables used by humans.

## 1.6 The scope of this study

In section 1.4 various limitations to the study were mentioned. The scope of the study is confined to the choice of a single reference object, appropriate to a given target, in a set of schematised near-real-world scenes. The limitations following from this are summarised here:

1. In this study any discourse context or other history is excluded and the scenes are considered as *static* or *memoryless*; that is, time leading up to the point at which a scene is described is unknown or non-existent. This may seem to rule out a large

portion of the every-day uses of locative expressions, however there is a significant set of questions such as “Where are my car-keys?” which can be answered with a static, discourse independent locative phrase of the type investigated in this study.

2. Only simple locative phrases with single reference objects are being investigated. There is no scope for learning compound phrases with hierarchical references such as “The mug is on the table under the window”. There is no scope for qualifying objects with regions such as “The church is at *the end of* the road”. It is also not possible to choose two references that a target is ‘between’.
3. Only tangible objects in the scene can be chosen as references, “The ball is in the air” would not be allowed as the air is not an ‘object’. Various limitations on the way objects are represented in the scenes which form the test data set and have some bearing on the study are discussed in chapter 4.

As will be seen the problem of reference choice, even in this limited domain, is far from trivial when the complexities of near real world environments are involved.

## 1.7 Contribution of the study

Although little attention to the reference choice problem has been given this is clearly not the first study to look at the issue. However it is the first to look at the problem in anything like near real world scenarios, which given the nature of the reference choice problem outlined in section 1.4, is a significant step. The most similar work, due to Carlson and Hill [2009] uses three or four objects in a study comparing two influences on reference choice and is discussed in section 3.4.2

The development of a structured hypothesis model for reference choice also appears to be new. Many studies, particularly from landmark selection identify factors affecting landmark (and therefore to a certain extent, reference object) choice but no attempt seems to have been made to systematically organise these influences.

This study is the first to report on human performance in the reference choice task in these near real world environments. The studies were not designed to look at comparative performance (in the sense of gender, cultural or linguistic differences) or in themselves to elucidate the factors behind choice of reference object, however the results presented in chapter 4 should be of interest in themselves. Although Carlson and Hill [2009] ask participants to choose reference objects, only a single scenario was used, with the limited range of objects noted above.

The adoption of a machine learning approach to model human behaviour in reference choice and hence to be able to make inferences about factors determining the human behaviour is also new. The results pertaining to the complexities of the models and the extent and nature of the information used by the models to match human behaviour are the single most important contribution of the study.



In addition the nature of the reference choice problem has suggested that some new approaches to Bayesian network classifiers might prove useful. Although this part of the study cannot be considered comprehensive and the new methods developed require further study and testing against other methods on standard machine learning data-sets, they seem promising.

Although a limited amount of research has been carried out in to machine learning of landmark selection (see section 2.9) and the ability of machines to learn the reference choice task in spatial language from ‘grounded’ examples is more meaningfully quantified in this study than in earlier work.

## 1.8 Organisation of the study

Figure 1.4 shows the experimental platform developed for, and used in, this study. All components of the system with the exception of the XML file parser library, and the convex hull generation algorithm used in the topological and geometric analysis elements, were specifically created as part of the project.

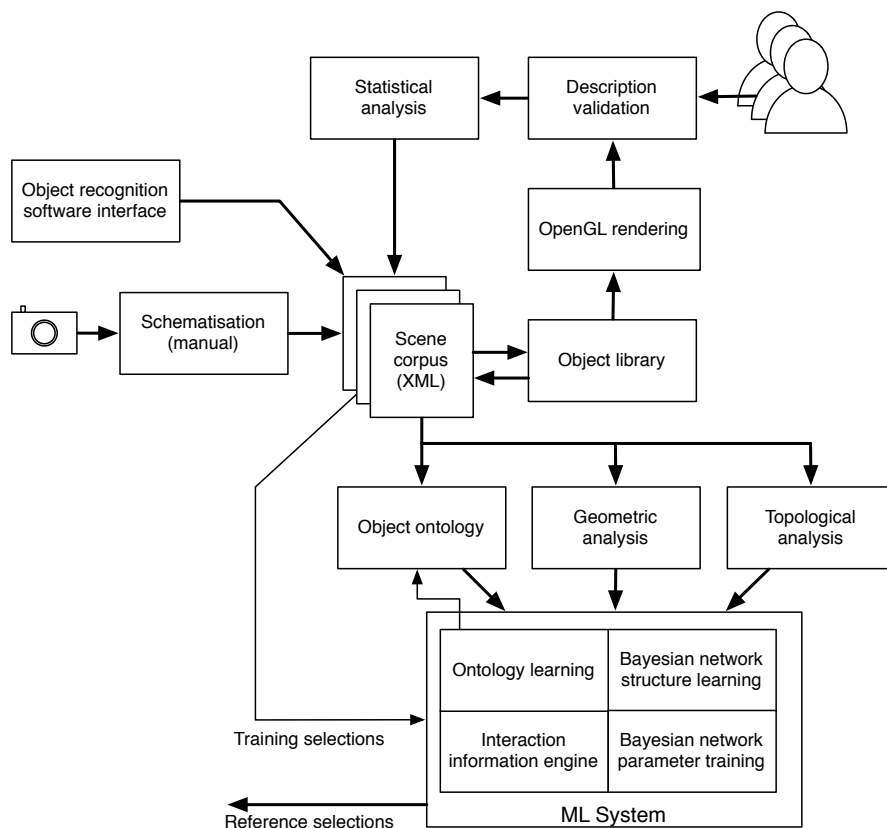


Figure 1.4: A schematic of the experimental platform used

The method of the study largely follows from figure 1.4 and a simple overview can be given as:

1. A test data set of machine and human ‘readable’ scenes is created.
2. Descriptions of object locations are given for each test case in the data set by a group of human validators and/or by the author.
3. Machine models of differing complexity and capability are assembled for testing.
4. The machine models are trained on a portion of the test data set by ‘examining’ variables derived from the scenes and the descriptions given by the humans.
5. The machine models are tested on the remaining portion of the test data set to see if they can reproduce the reference object choices made by the humans.
6. By analysing the results of the machine models inferences are made about the process of reference choice in humans.

The organisation of the remainder of the study is as follows:

**Chapter 2** reviews other systems used for spatial language generation so that the system used in this study can be put in the context of this earlier work. Differences between the approaches and capabilities of these systems and the system used here are highlighted.

**Chapter 3** looks at literature from psycho-linguistics and linguistics that bears on reference choice and work from geographical science and geographical information systems that deals with landmark choice in navigation, a subject closely related to reference object choice. The aim of the chapter is twofold, firstly to identify influences on reference choice (variables) that should be tested in the machine models and secondly to organise these influences into a plausible structure, based on the process of finding a target given a simple locative expression. This structure can then be used, if not as a gold standard, then at least a starting hypothesis for comparison with findings from model tests.

**Chapter 4** deals with the production and validation of the 133 scenes similar to that shown in figure 1.3, and equating to 569 test cases, that make up the test data set. This broadly equates to describing the top half of figure 1.4. The derivation of the scenes and the way in which they are represented to humans and to the machine models is explained. The experiments in which the human validators gave their views as to the description of object locations are described along with the results of these experiments.

**Chapter 5** explains why the Bayesian network based machine learning technique was chosen and why the new additions to the accepted state-of-the-art have been made. This equates to describing the lower half of figure 1.4. The procedures for generating the geometric, topological and other variables (on which the machine learning system operates) from the scene are given. The probable error margins in the system due to discretisation and model training are examined.

**Chapter 6** presents the results from a variety of Bayesian network models of reference choice. Models with designed structure are used along with models whose structure is machine learned. Tests are made on models with one variable to illustrate the individual

significance of influences (such as target/reference distance or object size) and successively more complex models in which multiple variable interactions are introduced.

**Chapter 7** discusses the performance of the machine models and compares this to human performance in the reference choice task. Possible limitations of the system and areas for improvement of the machine models and the system as a whole are examined. What it has been possible to learn about spatial language is assessed with particular emphasis on reference choice.

## Chapter 2

# Spatial language generation systems

### 2.1 Introduction

The intention of this chapter is to review the systems and models that have been produced to generate spatial language from grounded examples, so that their relation to the work in this study can be understood. It is not entirely possible to exclude psycho-linguistic work from this chapter as many of the systems have been developed alongside psycho-linguistic studies, but it is not the intention to review work from psycho-linguistics here. Chapter 3 presents a full review of work relating to the problem of reference choice from the fields of psycho-linguistics and geographic information systems.

Not covered here are systems, such as ‘chat-bots’, that may appear to generate spatial language (e.g., “London is in England”) but do so simply from stored ontological knowledge, without any analysis of a real or virtual scene. This study is also not concerned with generation of syntax or conversion of text to speech and these aspects of the systems discussed (if present) are ignored.

Also neglected are many systems or models whose principal function is referring expression generation (e.g., Dale and Reiter [1995]) even if the models contain a topological (Varges [2004]) or geometric (Tenbrink [2005]) element as part of the disambiguation process; since these ignore the characteristics of the reference and are not primarily concerned with object location. The work of Socher et al. [2000] is discussed, even though its principal focus is referring expression interpretation, as it contains what seems to be the only other Bayesian network approach to scene analysis and highlights some issues of concern in this study.

About half of the systems discussed are principally concerned, not with language generation but language interpretation, however many of the algorithms presented can be used for both purposes. It can be seen that within all the systems, whether for language interpretation or generation, the problem of reference selection has received a tiny fraction of the attention given to modelling spatial prepositions for both language interpretation and

generation. The problem of reference generation is only directly tackled by Gapp [1995a] in a fixed computational model and it seems that no other researchers have tackled the problem using machine learning techniques. Hence although much of the work reviewed may seem only tangentially related to this study, it illustrates the most relevant related work.

The systems described, and the reasons for considering them, are as follows:

**Section 2.2** briefly revisits Winograd’s SHRDLU system (Winograd [1971]) as it serves as a historical anchor and a reminder of how little progress has been made, in some ways, since then.

**Section 2.3** describes the VITRA project which focusses on scene description (language generation as opposed to interpretation) and contains the Gapp model for reference choice. The VITRA system, the Situated Artificial Communicator and Kelleher’s Situated Language Interpreter build on earlier computational models of preposition use including Olivier [1994] and Yamada et al. [1988] which are considered as superseded and are not reviewed here.

**Section 2.4** describes Regier’s [1996] neural net model. This is the first attempt at a machine learned approach to spatial preposition use, it models the use of a small set of static and dynamic prepositions, but looking only at geometry and topology, not functional aspects. The mixture of derived characteristics and machine learning is similar to the approach used in this study.

**Section 2.5** describes the ‘Situated Artificial Communicators’ project which is in many ways a physical realisation and extension of Winograd’s ‘Blocks world’. It is principally a language interpreter but contains Socher’s Bayesian network scene analysis work.

**Section 2.6** describes a system due to Abella and Kender [1999]. This is a language generation system which indirectly tackles reference object selection while trying to uniquely specify the location of target objects.

**Section 2.7** discusses the ‘Describer’ system due to Roy [2002]. While this is principally a referring expression generator it does, in a very limited way, apply machine learning to reference choice.

**Section 2.8** describes Kelleher’s ‘Situated Language Interpreter’. As might be expected this is a language interpreter, however it operates in *potentially* complex virtual 3-dimensional environments similar to those used in this study. Importantly the system includes an algorithm for calculating ‘Visual Saliency’ which, although used for reference resolution, is of direct relevance to the reference selection problem.

**Section 2.9** describes two automatic landmark detection systems which, although operating on a narrow range of scales and objects contain computational models and elements of machine learning.

**Section 2.10** contains brief descriptions of systems that contain points of interest but are less closely related to the current study.

## 2.2 SHRDLU

Winograd's SHRDLU system (Winograd [1971]) is one of the most famous early AI systems and given the limitations of computing equipment and the state of knowledge it is truly impressive in scope and achievement. It combined a state of the art natural language parser with a first order logic theorem prover and a rule database to produce realistic human-computer dialogue within the limited realm of a blocks world (Minsky [1986]). The system concentrated on syntactic and semantic (first-order logic) processing and did not make judgements or reason about space beyond a very simple set of logical rules. All spatial descriptions (left of, in front etc.) are based on the centroid positions of objects in the half spaces defined by fixed cardinal axes and fixed definitions of 'in' and 'on'. To answer a "where is..?" type question the system would always resort to returning 'on' with the name of whatever was supporting the object in question. The system could not, for instance, learn the concept 'near' from a set of examples, although it could combine 'left' and 'behind' to identify an object from a compound positional description. The rule database and dictionary are extensible so for instance a 'steeple' could be defined as 'a stack of two blocks and a pyramid' and this new compound object can then be used in subsequent instructions, as in, for instance, "move the steeple into the box". Winograd anticipates many issues that have concerned AI research over subsequent decades such as ambiguity in language, uncertainty and the link between meaning and grounding of symbols in some external reality. SHRDLU however, does not learn from grounded example but only through symbolic assertions extending a rule database. It does not make judgements in conditions of uncertainty or deal with language ambiguity beyond anaphora resolution.

## 2.3 The VITRA System

The VITRA (VISual TRANslator) project had the intention of producing "integrated knowledge-based systems capable of translating visual information into natural language descriptions" (Herzog and Wazinski [1994]). This and other overview papers including Wahlster [1989] and Herzog [1995] describe the different application areas covered by the project including:

1. Automated commentary generation of football matches
2. Analysis of traffic movements at fixed locations
3. Communicating with autonomous mobile robots
4. Route descriptions in 3-dimensional model environments

The VITRA project extends and integrates various earlier projects such as CITYTOUR (Andre et al. [1986]) and SOCCER (Andre et al. [1988]). Some of the applications deal with photographic input but the object recognition elements seem limited, only trajectories of moving objects being automatically determined. The system concentrates on language

generation from processed or generated scenes in which objects are already known and labelled, as is the case in this study.

Spatial relationships between objects are calculated using fixed computational models such as those described in Gapp [1995b] which create ‘volumes of applicability’ for given spatial prepositions. They are extended using logical or fuzzy-logical models into concepts relevant for path and movement description such as ‘past’ and ‘along’ (as in “he ran past the church and along the river”).

So for a trajectory defined by a number of discrete times  $t_i$  and associated positions  $P_i$ , between  $t_{begin}$  and  $t_{end}$  for the target, Andre et al. [1986] would define the relation “passed in front of” as being satisfied if the following (equations 2.1, 2.2 and 2.3) hold:

$$(P_{begin} \in \mathcal{L} \wedge P_{end} \in \mathcal{R}) \vee (P_{begin} \in \mathcal{R} \wedge P_{end} \in \mathcal{L}) \quad (2.1)$$

where  $\mathcal{L}, \mathcal{R}$  are regions defined as being ‘left of’ and ‘right of’ the reference object, so to pass in front the target must start on the left and finish on the right or vice versa.

$$\forall t_i \in ]t_{begin}, t_{end}[ : (P_i \notin \mathcal{L} \cup \mathcal{R}) \wedge (P_i \in \mathcal{F}) \quad (2.2)$$

$\mathcal{F}$  is a region defined as ‘in front of’ the reference so that at all points between the start and end of the trajectory (and excepting when the target is still to the regions to the left or right of the reference) the target’s position must be in the region defined as in front of the reference.

$$\forall i : t_{begin} < t_i < t_{end} : distance(P_i, Ref) < k(size(Ref)) \quad (2.3)$$

$size(Ref)$  is defined as the arithmetical mean of the sides of the bounding box of  $Ref$  and it is suggested that  $k = 2$  produces reasonable results. This condition in effect requires the trajectory to be at all times within a certain distance of the reference, the distance being relative to the size of the reference. Clearly many scenes such as that in figure 2.1 can be envisaged where a trajectory satisfying these constraints could (as illustrated) include travelling through a building on the opposite side of the street. This suggests that first choosing the correct reference is important, for trajectory T2 in 2.1, for instance, the church is not a good reference. Having done this some constraints on the trajectory to satisfy a given description (or the constraints on the choice of a motion preposition) will be unnecessary. In the case in figure 2.1 the constraint in equation 2.3 is likely to be unimportant if a suitable reference is adopted.

The choice of reference objects is discussed in Gapp [1995a], where one of the very few computational models for reference choice is proposed. The feature variables used in the Gapp model and how they relate to those used in this study are discussed in sections 5.6.1 and 6.3.1. The model computes the Euclidian distance between a vector of scaled attributes derived from the feature variables, and the attribute vector of an optimum reference, for all candidate reference objects, as in equation 2.4:

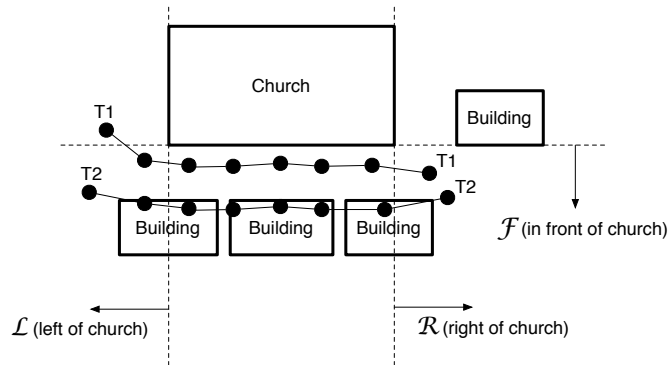


Figure 2.1: Example trajectories from the CITYTOUR project. Both trajectories, T1 and T2, satisfy the constraints for the description ‘passed in front of the church’

$$Q(Ref) = \sqrt{\sum_{i=1}^n (Sc_i(f(i)) - f_i^{opt})^2} \quad (2.4)$$

where  $f(i)$  are the features determining reference quality (size, proximity etc.) and  $f_i^{opt}$  is a co-ordinate of the point in the  $n$ -dimensional space occupied by the ‘optimum reference’. The best reference is then that with the minimum distance,  $Q(Ref)$ .  $Sc_i$  is a vector of “context dependent” scaling factors, where the context could depend on whether the listener was present for instance. No clue is given as to how these factors might be arrived at. The likely performance of this model is discussed in 7.1.2. Gapp does not give any application examples or results for the model, but reference selection in the football commentary application is discussed in Blocher and Stopp [1998]. This contains a simplified version of the Gapp model in the specific context of labelled regions of a football field. The CITYTOUR example, which appears to contain multiple possible references, does not make specific mention of reference choice.

Discourse planning and listener modelling modules are also included in the VITRA system to allow it to support the production of human acceptable descriptions of image sequences and routes, intentional inference is also included as illustrated in figure 2.2.

The VITRA system’s intended comprehensive coverage of spatial language generation is limited by the fixed nature of the models used. In contrast to the work presented here there is no use of machine learning and there is little evidence of system performance in cluttered, near real world environments. No formal tests of the acceptability of system output to human participants appears to have been undertaken which makes judgement of model performance difficult. The use of more flexible models, particularly for the spatial cognition aspects of the system, may have made the system easier to deploy outside of its initial application areas.



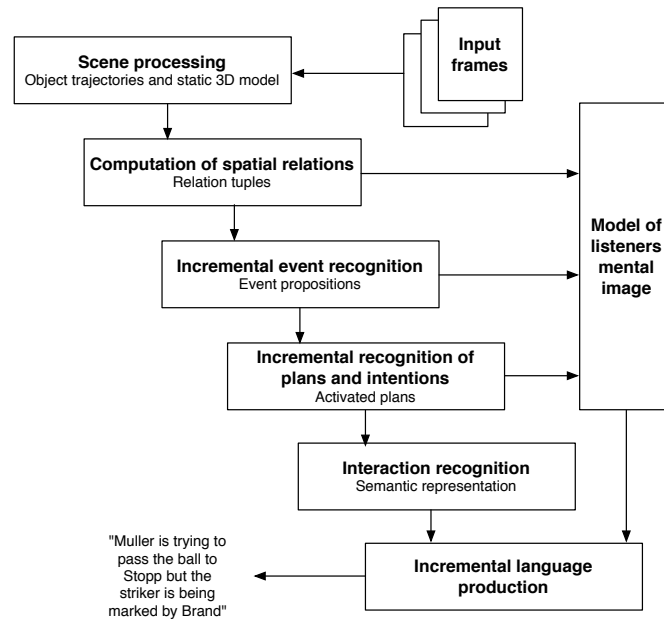


Figure 2.2: Cascaded processing architecture of the VITRA system (SOCCER example) adapted from Herzog and Wazinski [1994]

## 2.4 Regier's constrained connectionist system

A machine learning approach to spatial language generation is taken by Regier [1996]. Although he concentrates on spatial preposition assignment the work is closely related to this work in that the machine learned elements act on features (variables) that have been extracted from the scene rather than directly on the drawn elements. In Regier's case the drawn elements are represented in a 2-dimensional pixelated image as opposed to the 3-dimensional vector based representation in this thesis. Regier uses an image sequence as input enabling the learning of motion prepositions such as 'through' and 'into' as well as static prepositions. Note that although distance between reference and target is available in the training sequences it is not used and the preposition set contains only topological and projective prepositions. Prepositions such as 'near' in which a determination based on distance would be required are not used. A combination of geometric calculation and dedicated neural net structures are used to extract features as opposed to the all-geometric calculation used in this thesis.

The architecture of Regier's system is shown in figure 2.3. The angular information is calculated from the scene geometry and the topological relationship between the target and reference is derived from the 'feature maps' which are effectively a parallel array of local image filters. The features are combined in the general purpose neural net labelled 'current' with output nodes corresponding to the different spatial prepositions. Motion prepositions are handled with another special purpose structure which stores the maximum, minimum and average activations of the 'current' output nodes. Hence 'through' in the final output layer should have a high activation if the current (last image) and source (first image)

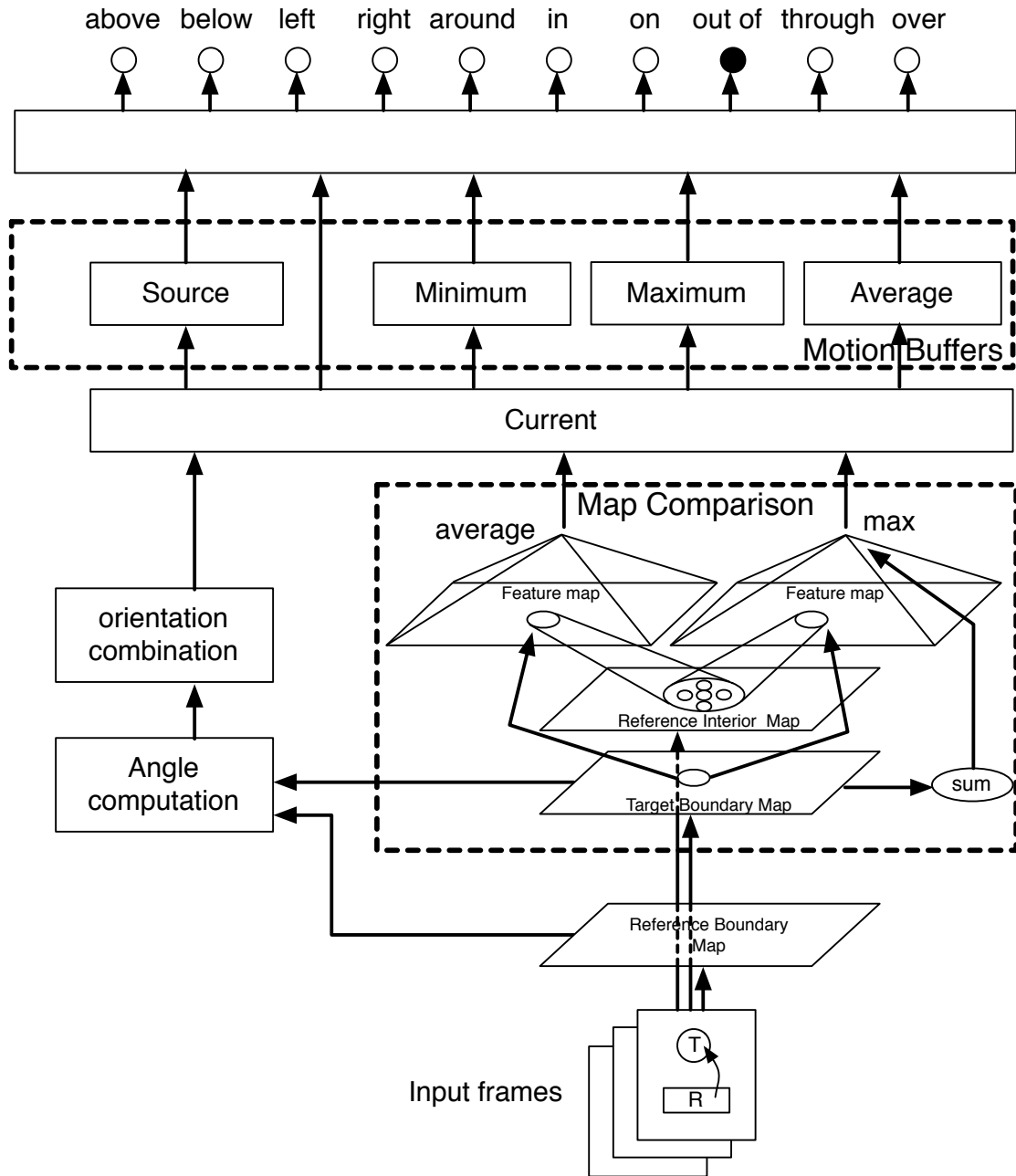


Figure 2.3: Architecture of Regier's constrained connectionist preposition learning system, after Regier [1996]

values of 'in' are low but the maximum value of 'in' at some point in the image sequence is high.

Regier's system produces good results from the test sets given although these are not compared to human performance. The fact that the scenes are not cluttered is probably not a significant limitation in Regier's work, if it is accepted that the choice of a reference effectively filters out other objects from a scene prior to assignment of a preposition, Regier's reference object being pre-assigned. The single reference scenario would be more

problematic if prepositions such as ‘near’ were being considered. The use of very simple convex shapes is a limitation although Regier demonstrates, and proposes some solutions to, the problem of preposition assignment when highly extended reference objects are used.

## 2.5 Situated Artificial Communicators

Like the VITRA project this large research effort covers many aspects of spatial language processing and modelling with results described in Rickheit and Wachsmuth [2006]. Much of the work concerns interaction with a robot operating on a table-top world populated with objects from a child’s construction set; nuts bolts, struts and wheels. Although the environment could be considered cluttered the objects tend to be arranged so that a minimal amount of overlapping or occlusion of objects takes place. Thus although many of the models are developed as 3-dimensional they are operating in a mainly 2-dimensional environment. This is probably due to limitations in the vision sub-system since this is a ‘real world’ environment, viewed through a camera, not a ‘virtual world’ as used in this study.

Within the overall project Fuhr et al. [1998] present a computational model of preposition applicability. This is a fixed (not machine learned) model like that of the VITRA project or Kelleher’s Situated Language Interpreter system although it differs from both in detail. The volumes of applicability for prepositions are not defined as fields with graduated acceptability, instead transitions between areas relating to given projective prepositions are handled by considering the bounding boxes (rather than centroids), of both reference and target. As the target moves across a region boundary the volume fraction of the target in each region is calculated to arrive at the appropriate preposition. Regions are defined for each individual preposition and for all combinations of prepositions for the cardinal axes in 3-dimensions, (above, above-left, above-left-behind, above-left-infront etc.), 78 regions in all plus one for the reference object itself. The regions are defined relative to the reference bounding box and then weighted according to the orientation of the bounding box with respect to the selected reference frame. Tests on human participants using a 2-dimensional situation containing two objects over a range of orientations suggest 90% acceptability of the generated prepositions. However Fuhr et al. [1998] suggest that objects with high extension in certain orientations produce some sub-optimal preposition assignments.

Of most relevance to the current study is an object identification task described by Socher et al. [2000]. It compares output from the neural net based object recognition system (not described here) with verbal descriptions in a Bayesian network. This is in effect a referring expression interpreter, but the scene representation in the model is interesting, and very different from that used in this study. The structure of the Bayesian network is shown in figure 2.4, it contains a node cluster for each object present in the scene and various nodes whose value range equals the number of object types in the scene. This will become unworkable in both storage and computational loading terms as the scene becomes realistically complex. A classifier type approach in which each object is sequentially applied

to a network which models only the scene characteristics (not the objects themselves) is used in this study and largely overcomes this problem. The network of figure 2.4 would probably perform better if actual perception data from the cameras (r, g, b) were modelled as having a direct statistical dependency with the stated colours from the verbal descriptions and likewise the recognised types from the images were modelled as having a direct dependency with the geometric parameters (size, shape) from the verbal descriptions. Currently these parameters appear to be independent. Note also that some post-processing is performed on the output of the network before the most likely target is identified.

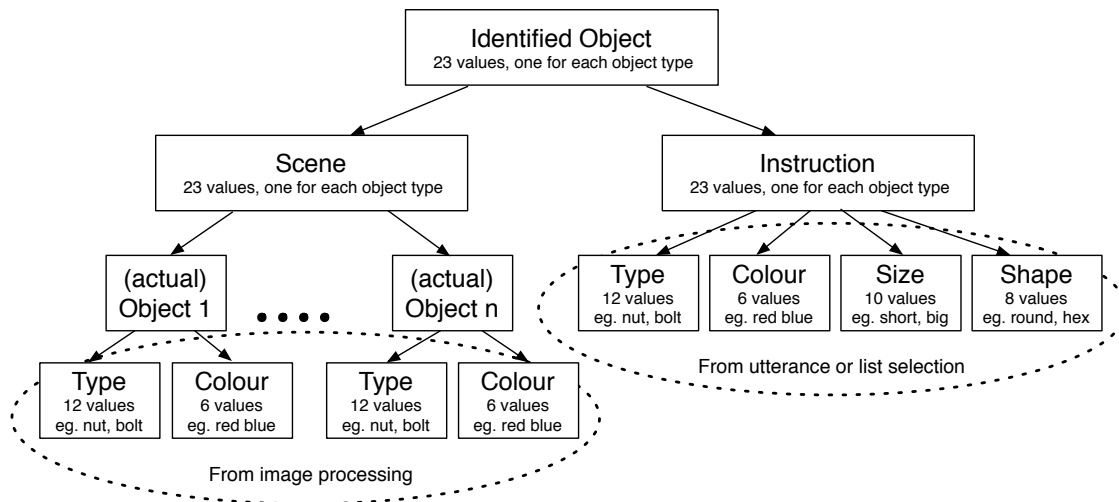


Figure 2.4: Bayesian network architecture used in the Situated Artificial Communicator for object identification, after Socher et al. [2000].

Additionally spatial information in the form of a locative sentence can also be interpreted by the system (e.g., "the blue nut left of the strut"), however the spatial information is not included in the Bayesian network. In this case of spatial disambiguation the output of the Bayesian network is combined with the output of the fixed spatial model described above in a simple algebraic model to identify the intended target. Note the system does not choose the reference (the 'strut' in the example given), this is always supplied by the human instructor.

In summary the 'Situated Artificial Communicators' system does not address the reference selection problem and contains only a fixed computational model for generating prepositions from objects in images. It is implemented in a limited range of staged scenarios although it should be noted that it is using real, not virtual, images. The application of the Bayesian network to referring expression interpretation is rare in systems of this nature and contains various drawbacks, some of which have been addressed in the current study.

Much of the work on the situated artificial communicators project has been incorporated into 'MAX' the virtual conversational agent [Wachsmuth, 2008] but it does not appear that the spatial language element has been significantly advanced.

## 2.6 Abella and Kender's scene describer

Reference selection is inherent in the scene description system described by Abella and Kender [1999] although it is not tackled as a problem in its own right but as part of a process which is aimed at arriving at the most unique (or possibly least vague) description of a target object. The system operates on strictly 2-dimensional scenes containing discrete labelled objects. The objects are defined by bounding boxes although these are derived from the object's moments of inertia, not their maximum extensions along orthogonal axes as is the case with most other systems that use bounding boxes (including this study). It is not clear that any advantage is gained from the added complexity introduced. Spatial relationships are defined as fuzzy regions around ideal locations. For prepositions such as 'near', a fixed parameter related to the object size is used to define the 'ideal' region.

To select the best description for a target object all possible prepositions are calculated for all object pairs in the scene, yielding effectively a 3-dimensional matrix. Each possible reference is considered and the preposition is chosen for it that most uniquely describes the target, that is, could be used to describe the fewest objects in addition to the target. This process may not result in a unique descriptor and various additions to the algorithm are used to improve matters. Firstly the fuzziness of the situation is exploited by applying adverbs to appropriate prepositions (e.g., 'very near', 'somewhat near'). Secondly compound phrases can be used. So in figure 2.5 the target would be "above Reference 3 and below Reference 2" as there is no single reference that uniquely specifies the target with a projective preposition. The use of "near Reference 2" will often not be possible because of the thresholding of the fuzzy regions and the size of the objects on which the regions depend through a pre-determined formula. In figure 2.5 the target may be 'very near' all four references.

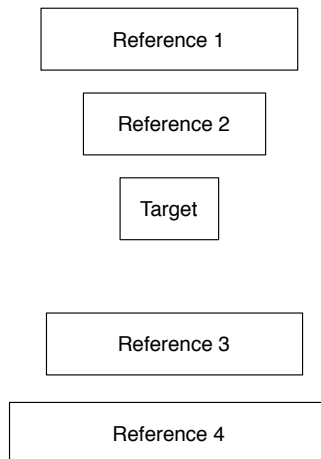


Figure 2.5: An example of locative sentence generation in Abella and Kender's scene description system

Abella and Kender [1999] also suggest a 'user model' which would contain the degree of knowledge of a scenario expected from a user. An example of this is given as a consultant or

a trainee doctor examining an X-ray. This user model would be used to prune some of the longer hierarchical or path descriptions that the system is prone to producing. Hand coding the levels of expertise for each scenario seems to be the only suggested way of achieving the goal however.

The system is tested in the following scenarios;

1. Describing the locations of kidney stones in (digitally processed) x-rays
2. Producing path descriptions in a schematic map

The emphasis on uniquely specifying the target means that the descriptions generated tend to be long-winded (particularly for the path descriptions) and not very human like. It is difficult from the results presented to say anything meaningful about the system's effectiveness. Some 70% of people were able to follow route descriptions on the map but the route descriptions are only given for two cases so the difficulty of the task is not clear.

Although the system may be useful in some areas the insistence on uniquely specifying the target and only categorising potential references by their bounding boxes will prevent the system providing useful 'human like' descriptions in cluttered 3-dimensional environments.

## 2.7 The 'Describer' system

The task attempted by the Describer system [Roy, 2002] is to arrive at a phrase which uniquely identifies a target object and as such it is principally a referring expression generator. Reference selection is incorporated however, in complex referring expressions which involve both attributes of the target and location of the target relative to a reference. The reference selection is somewhat crude, even so the system warrants some examination as it may be the only other example of an implemented system of machine learned reference selection.

The scene containing the object to be located is a 2-dimensional computer generated arrangement of 10 rectangles of arbitrary colour, size and extension. Eight features are extracted from the scene: colour(r,g,b) position(x, y), height/width ratio, area, and max/min ratio. A corpus of 500 scenes was used. These were paired with spoken descriptions of a target rectangle in the scene, from a subject unfamiliar with the experiment. After analysis another speaker was used to provide a further 157 descriptions to try and improve the coverage of colour and geometric terms used (preference in term usage from a single subject is unsurprising). Of the total utterances 185 were 'complex', that is, they contained a spatially locative component.

Words from the descriptions in the corpus were classified by two learning algorithms - first by distributional clustering based on the idea that words from a class were unlikely to be used together in a simple description (i.e., one colour and one shape describing word will each be used rather than two colour words) -second by correlating word usage with the feature vector for the object in a scene. Combining these two techniques yields very

reasonable word classes in a training task that seems in many ways harder than a human child might encounter. (Human children are often given specific colour and shape oriented training sessions, as evidenced by the widespread availability of infant books and toys used for this purpose.)

Feature association with words in word classes is then performed using a multivariate form of the correlation used in word classification. (so high r, g and b, should all correlate with the word ‘white’ for instance). The features for a word class are a conjunction of the features for the words in the class.

A probabilistic model for word order is derived from analysis of the corpus as the final requirement for generating syntactically as well as semantically correct descriptions. This is modelled as a transition probability network between the word classes.

Thus from the training data supplied the machine learns word associations with features, word classes and a simple syntax simultaneously.

The machine then attempts to generate object descriptions, first by building a simple expression. An ambiguity measure defined by the difference between the fit of a description to the target object and the fit to the next best candidate is used to score the simple descriptions.

Use of a complex expression is determined by the ambiguity measure exceeding a manually assigned threshold. At this point a reference object is selected based on its ambiguity and three factors related to the spatial relationship between reference and target. The three spatial factors are centroid distance, proximal distance and the angular relationship between the target and reference. The angular relationship used is based upon the ‘attentional vector sum’ [Regier and Carlson, 2001]. The process for using these factors is fixed and the elimination of ambiguity is given precedence (i.e., a reference is only used if it can be uniquely described). For each unambiguous (uniquely describable) reference the most appropriate spatial relation is generated from the learned correlation between the spatial factors and words. Then the reference with the spatial relation which has the best fit to the utterance corpus is used, the best fit being a probabilistic function incorporating word usage and word sequence likelihood.

The system was evaluated by comparing 200 human generated descriptions and 200 descriptions generated by the trained system. Three human judges tried to identify the correct object in the scene from the descriptions. 89.8% accuracy was achieved for the human descriptions and 81.3% for the machine generated descriptions. Although the use of ambiguity as the primary discriminator for reference selection produced some distinctly odd references (including many harder to identify than the target, because of their small size and distance from candidate targets) it seems that misinterpretations of colour terms generated most errors.

These results are impressive because the system is modelling a learning task that would typically be simplified into multiple stages in humans. Words are learnt and classified but the classes themselves are not learned independently and assembled in to an ‘ontology’. The system is ‘end-to-end’ and audio processing, syntax parsing and semantic problems

are combined when they could justifiably be separated.

In comparison to this study the system operates in a highly simplified environment; the fact that it is 2-dimensional and contains only a limited range of objects not being as important as the limited number of factors used to select the appropriate references. The training of the system does not create a map between good references, the characteristics of those references and their relationship to the target. Instead the system uses the most likely descriptions of the spatial relationship between target and reference as the discriminating factor and relying on the fact that the human choice of reference descriptions will contain enough information to learn the reference characteristics. Characteristics of the reference, aside from ambiguity (its similarity to other candidate reference objects) are ignored and vague distance terms such as ‘by’ and ‘near’ are excluded from consideration. Note however that ambiguity is assessed on a scale of similarity, not the polar identical / not identical distinction used in this study.

## 2.8 Kelleher’s ‘Situated Language Interpreter’ system

The system was initially described by Kelleher [2003] with aspects expanded on in further papers as described below. As the name suggests most of the implementation of this system is aimed at interpreting rather than generating spatial language, however there are various reasons for considering the system:

1. The system architecture is clearly bi-directional, encompassing language generation as well as interpretation.
2. Some of the individual algorithms, and in particular the ‘visual salience’ algorithm, are applicable to language generation.
3. The system attempts to integrate discourse history with visual information during language interpretation.

A schematic of the system architecture adapted from Kelleher [2003] is shown in figure 2.6. The system operates on a 3-dimensional virtual reality scene which could contain any number of objects although in practice the number of object types appears to be limited to two (houses and trees) and the total number of objects used appears to be no more than ten and more usually three or four. The user can issue verbal commands to the system to navigate around the scene and to add, remove or change objects in the scene. The correct identification (by the system) of the user’s intended referents is the main focus of the work. To this end a variety of spatial reasoning sub-systems are integrated with a natural language parser as shown in figure 2.6. The ‘interpretive module’ relates mainly to generating volumes of applicability for spatial prepositions. Psycho-linguistic experiments validating these are described in Kelleher and Costello [2005] and Costello and Kelleher [2006]. As with the VITRA project this is a ‘field’ model with acceptability falling to 0 at the boundary of the volume. The models are fixed not learned and comprehend distractor



objects (i.e., objects other than the target and reference) enabling applicability areas for prepositions such as ‘near’ to be dependent not just on the reference and target but also other local objects. Although this represents an advance on earlier systems, the details of the preposition models are not important for this study.

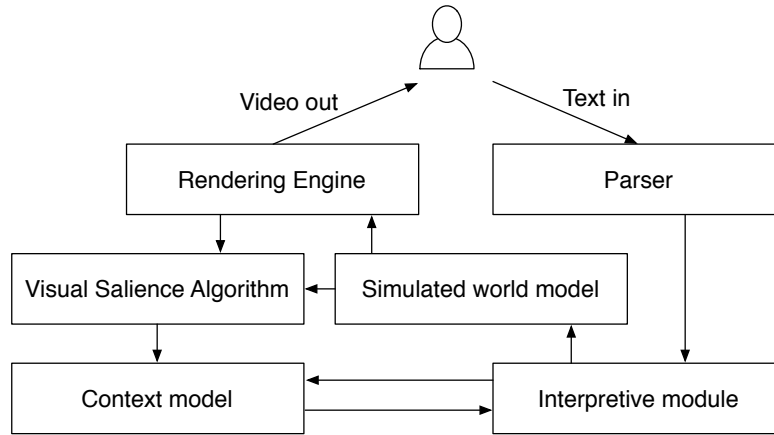


Figure 2.6: The architecture of the Situated Language Interpreter system, after Kelleher [2003]

The visual saliency algorithm is described in Kelleher and van Genabith [2004] and Kelleher and Costello [2009] and is of direct relevance to this study. Although it is principally used for reference resolution (i.e., to decide which object in a scene is intended in a linguistically ambiguous sentence) it can be used for reference selection or as part of a reference selection process. As implemented by Kelleher the visual saliency of an object in a scene is defined as:

$$Saliency(Obj) = \sum_{pixel \in Scene} F(pixel) \left( 1 - \left( \frac{d}{1-D} \right) \right) \quad (2.5)$$

where  $d$  is the distance of the pixel from the scene centre point and  $D$  is half of the diagonal length of the scene.  $F(pixel)$  is defined as:

$$F(pixel) = \begin{cases} 1 & \text{if } Colour(pixel) = Colour(Obj) \\ 0 & \text{otherwise} \end{cases} \quad (2.6)$$

Each object in the scene is assigned a different uniform colour (a ‘false’ colour in Kelleher’s terminology) prior to rendering. So saliency is in effect the sum of all the pixels ‘belonging’ to the object weighted by their distance from the user’s current focal point, which in Kelleher’s implementation is the centre point of the rendered scene. If the ‘user’s focal point’ is replaced by the ‘target’ object it can be seen that the algorithm combines representations of the distance between target and reference, reference size and reference visibility (or degree of obscurity). Because the 3-dimensional scene has effectively been projected onto a 2-dimensional plane in this algorithm the distance and size representations

are combined for objects at different depths in the scene. That is to say an object further in to the scene will have a smaller projected area and therefore a lower visual salience than an object of the same size at the front of the scene. Section 5.6.1 contains discussion of possible drawbacks of this algorithm and descriptions of variants of it used in this study.

The context model sub-system contains linked visual and linguistic context models. The models contain histories of objects mentioned with described discriminating characteristics as well as objects seen with their visible discriminating characteristics. The model allows resolution of anaphoric references in the normal way but using the visual information also allows the anaphoric reference to be negated. So for instance if a red house in a group of houses had been mentioned, but the point of view had been shifted so that the red house was no longer visible but a green house was, the sentence “move towards the house” would be interpreted with the visible house as the referent. Further, in combination with the visual salience algorithm references to “the other house” can be correctly interpreted even in a scene containing multiple houses if one of the houses is more ‘salient’ than the others. These ideas are incorporated into the CoSy project<sup>1</sup> and further described in Kruijff et al. [2006]. The context model illustrates the limits of what can be achieved in a discourse without the use of a related ontology in which for instance tables and chairs could be referenced as ‘furniture’. Indeed the system would not interpret a group of similar objects referred to using a plural noun correctly (as in ‘houses’ for instance). However the level of reality encountered in the system, in particular in the limited number of object types, means that this is less of a problem in the Situated Language Interpreter system than it seems to be in this study.

The system was tested in a very limited way on human participants. Fourteen participants were shown five instances of the system responding to selected instructions and asked to answer ‘yes’ or ‘no’ to the question “did the system respond as you expected?”. Only one negative answer was received.

The system does not address the reference selection problem directly but does present a potentially useful algorithm and usefully sophisticated visual/linguistic framework. As with other systems described the usefulness of the fixed computational models of spatial relationships in realistically cluttered or differently scaled environments is open to question.

## 2.9 Automatic landmark detection systems

Elias and Brenner [2004], describe a system for automatic landmark selection that uses machine learning to a certain degree. The system uses 3-dimensional mapping data and information from a geographic information system on building type, building use, building (conventional) orientation and occupancy of ‘land parcels’. The information from the geographical information system is fed into a decision tree which is used to decide which of the buildings in the vicinity of a route decision point can be most uniquely described.

---

<sup>1</sup>The CoSy project deals with many aspects of cognitive systems architecture but does not seem to extend the work on spatial language generation significantly and so is not described here. Documentation on the CoSy project can be found at [www.CognitiveSystems.org](http://www.CognitiveSystems.org)

‘Uniqueness’ is used as a substitute for ‘is a good landmark candidate’ as, apparently, no training data was available which would have directly indicated which buildings were considered good landmarks and which were not. All ‘unique’ candidates from this process have their visibility assessed using calculations based on the 3-dimensional mapping data and the direction of approach of the ‘listener’. (Although perhaps in this context ‘navigator’ is more appropriate than ‘listener’.) A candidate with high visibility is then chosen as the appropriate landmark.

The system appears to be more comprehensive than many others and possibly the closest in many respects to that used in this study. However it is described in the conclusion to the paper as a ‘concept’ and it is not clear how much of it has been implemented, or to what degree the different components have been integrated. No results are given.

Also as it stands the system has several drawbacks, some of which are noted by the authors. In particular good candidate references may be excluded because they are not unique. Two churches, for example, one of which was behind the other, leaving the most prominent as the best overall reference, would be discarded before the visibility analysis. Incorporating the visibility aspects into the machine learning system along with the ‘uniqueness’ characteristics (as attempted in this study) could overcome this problem. Also the measures of uniqueness do not take into account some visual aspects which might be considered most important, for instance, colour, but this is also a drawback in this study.

Nothegger et al. [2004] investigate landmark selection in urban surroundings in a study on pedestrian navigation in Venice. They use a fixed computational model based on visual characteristics and two aspects of ‘semantic attraction’ which relate in effect to how likely the navigator is to know the building, or class of building, in question. A vector of differences to the median value of attributes is used with the differences weighted by the standard deviation from the median, for each attribute, of all the candidate references. This gives another ‘uniqueness’ measure. Nothegger et al. [2004] point out that using deviation from a local mean or median value to represent salience does not hold for ‘asymmetric’ quantities such as size, where bigger is usually better than smaller for instance. Hence they skew data for asymmetric values before incorporating them into the uniqueness measure.

The model was tested against human participants who were asked to choose ‘the most prominent facade’ from panoramic displays of intersections in Vienna. The machine model matched the consensus choice of the humans in 7 out of the 9 test cases. The number of potential facades is not given but appears to be no more than 7 or 8 in each case and sometimes less.

The system does perform reference choice, although in a highly restricted domain; it is useful because it provides a data-point of machine model to human matching of 78%, although there is little else that is comparable in the approach taken by Nothegger et al. [2004], with the approach of this study. The system is very unlikely to be portable to other environments.

## 2.10 Other systems

### 2.10.1 The attentional vector sum model

This model for preposition acceptability is described in Regier and Carlson [2001]. The intention is to find a computational model that accounts for the experimental data gained from human participants on the acceptability of directional prepositions such as ‘above’, ‘below’, ‘left’ ‘right’. In particular the model should be able to account for reference objects that have high geometric extensions or irregular forms. The model effectively derives a new angle for comparison with the canonical direction for the preposition in question by a process of summing vectors between the reference and the target which are weighted by the attention given to the point on the reference from which they ‘originate’ by the speaker (initially) and then by listeners who judge the acceptability of the preposition. This is illustrated in figure 2.7.

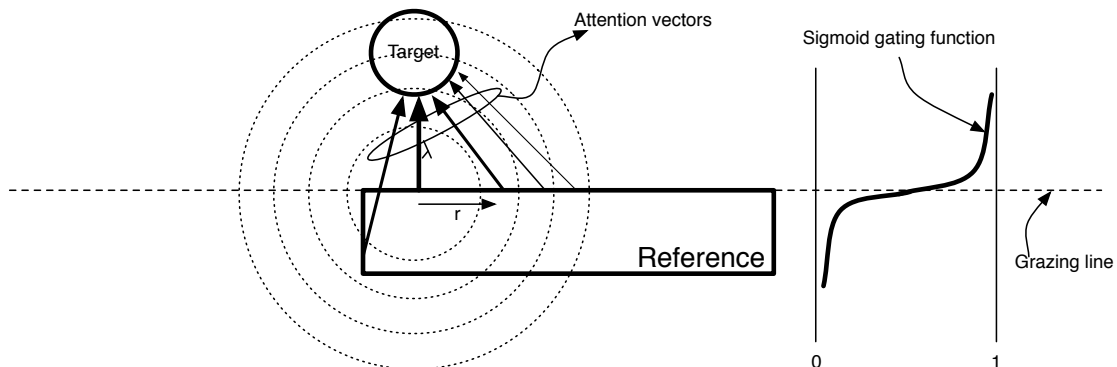


Figure 2.7: Derivation of the attentional vector sum model, the attention vectors are schematically represented with their weight falling off as the distance,  $r$ , from the focus increases. The sigmoid function related to the grazing line for the preposition ‘above’ is also shown

The peak of the attention function is at the nearest point on the reference to the target and the weight applied to the attention vectors,  $W$ , decays exponentially in the manner given by equation 2.7.

$$W = e^{\frac{-r}{\lambda\alpha}} \quad (2.7)$$

where  $r$  is the distance from the proximal point on the target to the point in question,  $\lambda$  is the magnitude of the proximal vector and  $\alpha$  is a free parameter. The result is to produce a vector with an angle between that of the centroid vector and the proximal vector. The angle varies depending on the rate of ‘fall-off’ of the attention function,  $\alpha$ . As  $\alpha$  approaches

0 the attentional vector sum returns the proximal vector, at high values of  $\alpha$  it approaches the centroid vector. To account for the rapid fall off in the acceptability of prepositions as the target falls the wrong side of the relevant ‘grazing line’ an additional factor is included in the model, unrelated to the attention vectors. This takes the form of a sigmoid function as shown in figure 2.7 which is multiplied with the attentional vector sum. The grazing line shown in figure 2.7 relates to the preposition ‘above’. Grazing lines for other prepositions can be constructed as perpendicular to the canonical direction of the preposition passing through the extreme point on the reference in that canonical direction

The necessity for these two factors could be explained by the compounding of two slightly different prepositional uses by the human participants, which might be termed the ‘locative sense’ and the ‘comparative sense’. The comparative sense of above would then equate to ‘above the level of’ and might be used, for instance, in comparing two mountain summits. This use would be largely independent of horizontal separation but critically dependent on the grazing line.

In the purely locative sense what is important is not the acceptability of a single preposition but the transition between prepositions, for example, when does ‘above’ become less effective than ‘left’ when a listener is searching for the target, and is there a transition region in which a compound preposition (above and left) would be used? In this case low values of acceptability of a single preposition, and associated grazing line effects will presumably be unimportant. The situation is more complicated if competing references are being considered, as is the case in this study, and further complicated if proximity prepositions (‘near’ or ‘by’ for instance) are allowed to compete with directional prepositions. It is also true that, when a listener is searching for a target the focus of attention is not established, as this is dependent on knowing the target location, and so the attentional vector sum cannot be directly used by the listener. This does not mean that the speaker will not have taken account of the angular relationship between the reference and target in choosing a reference of course. It is true however, that the model for reference choice may involve different considerations to the model for assigning a preposition. The way that the angular relationship between reference and target is modelled in this study is discussed in section 3.4.2.

The case of competing prepositions is not addressed in Regier and Carlson [2001] however the attentional vector sum model, with the correct choice of  $\alpha$ , fits experimental data well in the case of acceptability of single prepositions. The model is extended to take account of functional relationships between target and reference in Carlson et al. [2006], however functional relationships between target and reference are not modelled in this study (see section 1.4.3)

### 2.10.2 The ‘Bishop’ system

Gorniak and Roy [2004] describe the ‘Bishop’ system, in work leading on from the ‘Describer’ system (section 2.7). The experiments are conducted on scenes composed of a randomly synthesised layout of up-to 30 identically sized cones in one of two colours. The

scenes are effectively 2-dimensional though rendered in 3-dimensions.

The system is effectively a referring expression interpreter however, because of the lack of discriminating qualities between the objects themselves, it is forced to use predominantly spatial location for object identification. As with the ‘Describer’ project the ‘Bishop’ system takes in information as audio from human participants and so incorporates possible errors from speech processing and language parsing as well as interpreting the intended semantics.

The parsed speech input is passed to a series of pre-programmed ‘semantic composers’ - effectively simple mathematical functions that weight objects according to descriptions such as ‘leftmost’ or ‘middle’. These composers can be combined with the aim of leaving a single referent object as the result of the description. Note that no learning takes place in the system and some of the errors could be addressed by this - in particular the parameters of the composers could be easily adapted to remove some errors.

The overall performance of the system is stated as 72.5% correct identification on a clean test set (with speech and parsing errors removed) and 58% on the set including the errored cases. Errors in parsing typically occurred in more complex phrases as would be expected and errors in the semantic composers occurred typically during the linear process of combination (the ‘leftmost one at the front’ would fail if the object was a good example of ‘front’ but a poor example of ‘leftmost’). The performance of the system, given that 78% of the descriptions contained a reference to a positional extreme, may not seem so impressive. It does not require a complex model to interpret “The leftmost green cone”.

The failure of the system when contextual or historical information is required is noted by the authors. Human listeners can infer from incomplete information reasonably easily. For instance the word ‘middle’ can refer to the middle of a scene or the middle of a group. In the case where there is a single group of objects not centred in the scene ‘the cone in the middle’ would normally relate to the middle of the group. The ‘Bishop’ system cannot make this inference that a human makes automatically.

### 2.10.3 Coventry’s functional/geometric neural net system

The use of a neural network based system to correctly assign prepositions, taking into account both functional and geometric aspects of a situation, is described in Coventry et al. [2005]. In contrast to the GLIDES system (see section 2.10.5) the system designed by Coventry et al. uses multiple neural net structures to tackle different aspects of image processing in a manner based on structures in the human vision processing system. The image processing system is described in Joyce et al. [2002] and Joyce et al. [2003].

The neural system is trained and tested on data from experiments on human participants described in Coventry et al. [2009]. These concern the applicability of the prepositions over/above and below/under when used in situations where there is a functional link between the target and reference objects. For example a teapot pouring tea into a cup would be said to be ‘over’ the cup even if it was not vertically above the cup, if the tea was actually ending up in the cup. If the teapot was exactly positioned on the vertical axis above the cup but the tea, as a consequence, was missing the cup the preposition ‘over’

is less applicable. Note that what is being learned here is the correlation between a successful pouring operation and the applicable preposition, not the ‘ontological’ relationship between the specific objects, the system once trained on a teapot and cup should work equally well on a watering can and a flower. This distinction has a bearing on this study discussed in section 7.3.4. The system was able to detect and replicate human judgement of the acceptability of the prepositions in these types of situation with high accuracy.

In contrast to Regier’s system (section 2.4) or the system developed for this study, that both ‘mechanically’ derive geometric and topological variables from an image which are then fed into a machine learning system, Coventry’s system uses machine learning all the way from the image to the linguistic output, in this case, preposition assignment. The system in this study works on more complicated 3-dimensional images which would be challenging for any object recognition software however.

#### 2.10.4 Space Case

A novel approach to preposition assignment is presented in Lockwood et al. [2005] and Lockwood et al. [2006]. A narrow range of prepositions (‘in’ and ‘on’) is used but the functional as well as the geometric aspects of preposition use are investigated. The input to the system is from a sketch analysis package, although it is not clear that this is integral to the experiments. The objects in the sketch (typically just the target and reference) are tagged by the user and not subject to a recognition system.

The system derives the geometric characteristics from the sketch and the functional relationships and other characteristics of the objects from looking up the ‘Cyc’ knowledge base<sup>2</sup>. These factors are combined in a simple Bayesian network which classifies the situation as being appropriate for description using ‘in’, ‘on’ or neither. Characteristics of objects used included curvature of the reference object, the topological relationship between target and reference, animacy<sup>3</sup> of the target and reference and whether the reference is characterised as a container. The system was given the same input stimuli as those used on human participants by Feist and Gentner [2001] and Feist and Gentner [2003]. The correlation between the human use and machine use of ‘in’ and ‘on’ was generally very good. Some instances where the machine system failed include describing a firefly as being ‘in’ a hand where the human participants used on, attributed to Cyc returning similar animacy values for the hand and the firefly and the case of a block being put ‘on’ a building where the machine used ‘in’ again as Cyc suggests that buildings ‘contain’ things more than they ‘support’ things.

This is a different approach to that taken by Coventry et al. [2005] who operate entirely on perceptual information to derive the functional and geometric applicability of prepositions. It is probable that a combination of both approaches would be needed to truly match human behaviour.

---

<sup>2</sup>Information on the Cyc knowledge base can be found at [www.cyc.com](http://www.cyc.com)

<sup>3</sup>Not the same as the characteristic ‘mobility’ used in this study; in Space Case a car would not be animate but a dog would be, in this study both would be more or less ‘mobile’

Although Space Case does not address the reference selection problem its use of, and illustration of some of the pitfalls of using, large knowledge bases or ontologies is of interest to this study. (See section 7.2.2)

### 2.10.5 The ‘GLIDES’ system

The GLIDES system (Grounding Language in DEscriptions of Scenes) is described by Williams and Miikkulainen [2006]. The system operates on two input vectors, one derived from a 20 by 20 pixel image displaying one or two simple geometric objects and the second derived from a 31 word/phrase vocabulary containing object names (square, triangle, cross etc.), object sizes (small, medium, large) object positions (top left, bottom middle etc.) and for scenes with two objects the spatial relation between them (above, to the left of etc.). Two self organising maps are generated from the input vectors and these maps are fully interconnected, the strengths of these interconnections being learned to associate the images and the vocabulary.

Training and test data sets are randomly generated by computer. Results suggest that the self organising map derived from the image is unable to separate object type and object positional information and that this is because it does not have sufficient degrees of freedom to do so. It is not possible to say whether this is a decisive argument in favour of a system such as that due to Coventry et al. [Coventry et al., 2005] but having separate sub-systems recognising object types and positions (what and where) as Coventry et al. do, would seem to be a route forward.

Williams tries to combine all of this into a single connectionist construct and demonstrates the difficulty of this as a learning task. As noted in the case of the Describer system (section 2.7) it would seem that in humans these learning tasks (object type, positional description, spatial relation) can be undertaken separately, and probably are since it is clearly a lot easier.

### 2.10.6 The GRAAD system

The GRAAD system [Moulin and Kettani, 1999] is a fixed computational model for generating route descriptions or directions. It chooses references (landmarks) on a map by means of influence areas of the potential reference objects and taking into account calculated view angles of the listener. Influence areas seem to be based solely on proximity and the system only uses landmarks for added ‘user comfort’ relying on street names and turn descriptions at intersections for actual navigation. In common with most other landmark selection systems this is of limited relevance to the current study as it is not readily portable to application areas outside of urban navigation.

## 2.11 Summary

This chapter has described the systems that have attempted spatial language generation or interpretation that are the most relevant to the development of the system used in this



study. It can be seen that there is nothing very similar to the system used here. Although the different systems described undoubtedly have their own strengths and weaknesses their difference of purpose makes detailed discussion of these less important than a discussion of the ways in which they differ from the system used in this study. Tables 2.1 and 2.2 summarise the differences between the surveyed systems and this work. Some strengths and weaknesses can be inferred from the table. The lack of performance reporting in some systems (in particular the VITRA system) is clearly a weakness. The relatively small size of some data sets (for instance in Nothegger et al. [2004]) would be a weakness if the purpose was general reference object selection but could be judged satisfactory for the purpose at hand.

In the tables the word ‘schematised’ is taken to mean created, as opposed to photographic, images which are intended to be representative of the real world. Abstract images are random collections of simple geometric shapes. Tagged images are those in which the key objects are identified and named so the system in question does not have to recognise or identify them.

The only system that directly addresses reference object choice for location across a range of scales and environments is that due to Gapp. No results or performance measures are apparent for the proposed Euclidian distance model employed. Of the various landmark selection systems (operating over limited scales and environments) only that due to Elias and Brenner [2004] uses machine learning and again no results are apparent. Perhaps the nearest approach to this system for which some performance measure is available is that due to Nothegger et al. [2004] but this uses such a small data set in such a limited environment that it is difficult to attach significance to it as a general measure of machine reference selection. The Describer system (Roy [2002]) selects references for the purposes of disambiguation (identification of a target, rather than principally location) and it is not possible to derive a measure of its effectiveness in location. Anecdotally however it seems to perform quite badly in this respect. So it seems that to date there is no system whose purpose is to select reference objects for locating a target, of which it can be said, “it used a model of this type and performed to this level”, where ‘performed to this level’ is any measure, whether of human acceptability, human conformance, or a direct measure of target search effectiveness.

Table 2.1: Summary of the various spatial language systems described in this chapter which directly address reference choice

System	Language generation or interpretation?	Addresses reference selection?	Uses machine learning?	Environment or test scenarios used	Performance reported?
VITRA	Generation, including dialogue models	Uses Gapp's reference selection model in some scenarios	No, Gapp's model is Euclidian distance between feature vectors	Various 2-D and 3-D moving/static image and schematic inputs are used	No formal performance assessments are visible
Situated Artificial Communicators	Principally interpretation	Not directly, but includes spatial referring expression interpretation	Bayesian network for referring expression interpretation	3-D images of blocks world with 10-20 objects	Measures of object identification accuracy are given
Abella and Kender's scene describer	Generation only	Yes, within the context of complex referring expressions	No, fixed models based on fuzzy regions	2-D schematised, tagged images, 10 - 20 objects	Limited assessment of human interpretation accuracy
Roy's describer system	Generation only	Yes, within the context of spatial referring expressions	Yes, but limited in the case of reference selection	2-D abstract images with 20-30 simple shapes	Yes, full assessment of human interpretation accuracy
Gorniak and Roy's Bishop system	Interpretation only	Yes, within the context of complex referring expressions	No	2-D abstract arrangement of 30 objects rendered in 3-D	Full assessment of interpretation accuracy
Kelleher's situated language interpreter	Interpretation, including dialogue models	No, but includes a relevant reference resolution method	No	3-D schematised, tagged images 5 - 10 objects	Very limited assessment of interpretation acceptability
Elias and Brenner's landmark selection system	Generation only	Yes, in urban street scenes only	A simple decision tree	3-D map data, tagged and schematised	No performance figures given
Nothegger et al's landmark selection system	Generation only	Yes, in Viennese street scenes only	No, a Euclidian distance between feature vectors is used	Only 9 panoramic photographs, each with 7 or 8 buildings tagged as potential references	Directly tested against human reference choices
The system from this study	Generation only	Yes, reference selection for target location only	Yes, Bayesian networks and fixed computational units	3-D schematised, tagged images with 10 to 40+ objects	Directly tested for conformance to human reference choices

Table 2.2: Areas addressed by the indirectly relevant spatial language systems described in this chapter

System	Language generation or interpretation?	Addresses reference selection?	Uses machine learning?	Environment or test scenarios used	Performance reported?
Winograd's SHRDLU	Interpretation only	No	Has an extensible rule set but no statistical learning	3-D schematised blocks world	Not really relevant, no scope for errors except in language parsing
Regier's constrained connectionist system	Generates preposition assignments	No	Yes, neural nets working with fixed computational units	2-D abstract images with target and reference objects only	Yes although no assessment against human preposition assignment
The attentional vector sum model	Generation only	No, preposition assignment only	No	2-D abstract arrangement of 2 objects	Full assessment of assigned preposition acceptability
Coventry's functional/geometric system	Judges preposition acceptability, could be used for generation or interpretation	No, prepositions (above, over, under, below) only	Yes, neural networks from image processing to preposition judgement	2-D moving images but with only reference, target and functional indicator objects	Compared to human preposition acceptability measures
Lockwood et al's Space Case	Generation only	No, prepositions (in, on) only	Yes, simple Bayesian network and object characteristic look-up in Cyc ontology	2-D sketch inputs but with only target and reference objects	Compared to human preposition usage data from Feist and Gentner
Williams and Miikkulainen's GLIDES system	Generation only	No, object type and relative positions are produced	Yes, self organising maps	2-D abstract image with one or two objects	Description accuracy assessed against authors judgement

## Chapter 3

# A hypothesis model for reference object choice

### 3.1 Introduction

The intention of this chapter is to derive, from a review of relevant literature, the factors that affect human choice of reference objects, and to understand as far as possible how these factors are related so they can be organised into a reasonably comprehensive model of reference choice.

Two distinct bodies of literature on reference objects exist;

1. From linguistics or psycho-linguistics. As has been noted the work in this field relating to reference selection is far less than that relating to preposition assignment or even reference frame selection. Since the comprehensive account of preposition use in Coventry and Garrod [2004] however, more attention is now being switched to the problem of reference selection. In particular work by Carlson and Hill [2008] and Carlson and Hill [2009] is starting to experimentally verify the importance of some factors used in reference selection which had been identified but not tested by linguists. Hitherto this has been almost totally lacking.
2. From the area of geographic, or spatial information, theory as far as it relates to landmarks. A landmark is a reference object although some caution should be observed in treating the literature on landmarks on the same footing as the literature from psycho-linguistics. Firstly the literature on landmarks deals with a subset of references that occupy a single environment and environmental scale, typically urban street level, and does not address table top or room scale environments. Secondly landmarks as reference objects tend to locate ‘places’ not objects, which alters the nature of the search task once the reference (landmark) has been found. While this is a valid use of a spatially locative sentence, over-reliance on characteristics of landmarks as opposed to reference objects in general might lead to bias in resultant models. The relative scarcity of work on reference objects in general means that work on landmarks cannot

be ignored however. The study of landmarks in general goes well beyond way-finding into areas such as cognitive mapping and structuring of space. The literature relating to these topics is not discussed here.

Largely missing from the literature on reference objects is any *organisation* of the factors involved in reference choice. Gapp [1995b] makes some attempt at ranking the factors and Carlson and Hill [2009] investigate the relative importance of reference object salience and the spatial relationship between the reference and the target. Other researchers in both linguistics and spatial information have so far been content simply to produce lists of factors, or characteristics, of reference objects. One of the earliest and most influential is due to Talmy [1983] (and slightly amended in Talmy [2000]) who proposed that target and reference objects would have the characteristics listed in table 3.1.

Table 3.1: Talmy’s proposed target and reference object characteristics after Talmy [1983]

Primary object (target)	Secondary object (reference)
Has unknown spatial (or temporal) properties to be determined	Acts as a reference entity, having known properties that can characterise the primary object’s unknowns
More moveable	More permanently located
Smaller	Larger
Geometrically simpler (often point-like) in its treatment	geometrically more complex in its treatment
More recently on the scene/in awareness	Earlier on the scene/in memory
Of greater concern/relevance	Of lesser concern/relevance
Less immediately perceivable	More immediately perceivable
More salient, once perceived	More backgrounded, when primary object is perceived
More dependent	More independent

If the process of the communication between the speaker, who is aiding the listener to find the target, is analysed, the simple list of factors appears inadequate as a cognitive model of reference choice. If the scene in figure 3.1 is considered it can be seen that there are many potential references that could be used to construct an answer to the question “Where is the post-box?”. The lorry is larger than the post box, but mobile. The gate and the street lamp are permanently located but about the same size as the post-box and so may need locating in their own right. The blue house is larger than the post box and permanently located so might be the best reference. However if these are the only criteria, the church, which is even larger than the blue house would presumably be an even better reference, although it is further away and does not seem to locate the post-box very well. The wall is bigger than the post box and at some points very close but because of its geometric extension also does not locate the post box well. The trees are also large and

fixed but largely hidden by the blue house which makes them a less obvious choice from this angle at least. A model which can properly represent the reference choice process needs to take into account the relationships between the different items on Talmy’s (or any other) list and how they affect different aspects of the reference choice process.

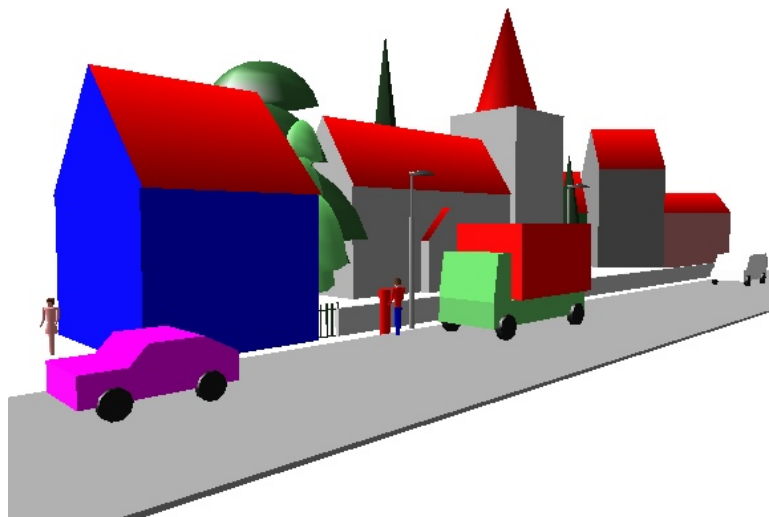


Figure 3.1: Where is the postbox? An illustration of the competing influences of proximity, mobility, reference size and perceivability on reference choice.

A paper arising from this study (Barclay and Galton [2008]) tackled this issue and covers some of the same material as in this chapter, although the model has been slightly improved since then and the material is arranged differently here.

The organisation of the factors affecting reference choice is based on an examination of the steps a listener must take on hearing a spatially locative phrase. It is presented here as an influence diagram, although some of the influences could be considered computational factors as much as statistical relationships. It is termed a ‘hypothesis model’ for the purposes of this study as it forms a basis for comparison with machine learned models. Although it seems plausible there is little other external support for it at present.

How comprehensive the model can be and whether in being comprehensive it better reflects human thought processes is discussed in section 7.3.4. It is certainly intended to be comprehensive in its coverage of scale and to be usable from tabletop scenes up to at least street scale, so it should be remembered that when object size measures are referred to they are always in a sense scaled to the size of the scene they are part of. In this sense the model developed in this chapter assumes that the reference choice process is scene-scale invariant, although this issue turns out to be problematic (see section 6.4). A discussion of the ways in which scenes at different scales might be perceived and in particular the work of Montello [1993] is given in chapter 4 as part of the rationale for the different scene scales in the test data set.

The rest of this chapter, dealing with the construction of the model, is arranged as follows:

**Section 3.2.** Looks at the steps a listener must take on hearing a locative sentence (assuming he is also going to act on it) and derives three fundamental influences on reference choice.

**Section 3.3** examines the factors which make the reference object itself locatable by the listener.

**Section 3.4** examines the factors which affect the listeners search for the target object once the reference object has been found.

**Section 3.5** examines the factors which affect the communication cost of using a particular reference object.

**Section 3.6** evaluates some approaches to the reference choice issue which do not readily fit within the framework presented in the previous sections.

**Section 3.7** discusses how much of the hypothesis model can be realised in the current study and why this is so.

**Section 3.8** looks at potential issues with using the hypothesis model as developed as the basis for a computational model.

## 3.2 Fundamental influences on reference object choice

Presented with a locative phrase and the task of finding the target object the listener must do two things:

1. Locate the reference object.
2. Search for the target object in the region constrained by combining the reference object location with the spatial preposition in the appropriate reference frame.

Making the assumption that the speaker intends his communication to be effective, or at least is trying to co-operate with the listener, it will follow that the speaker will have chosen the reference object to be easily *locatable*: and also that, in conjunction with the preposition and reference frame, the reference will *optimise* the region in which the listener must search for the located object. Work by Schober [1995] supports the assumption that the speaker is trying to make the task of the listener easier in the context of the adoption of reference frames in spatial descriptions. He found that speakers translate the reference frame to the listener's perspective, even though this increased their own cognitive load. Schober does not rule out the possibility that the speaker's strategy is designed to reduce the total effort required in the communication. Further evidence for this co-operation with (or consideration for) the listener in spatial communication is found in a cross cultural (American and Japanese) study of reference frame adoption due to Mainwaring et al. [2003] and by Tenbrink and Winter [2009] in adjusting route description 'granularity' to the difficulty of a navigation task.

Note that Carlson-Radvansky [1996] proposes a similar model of listener response to a locative phrase, but adds selection of a reference frame as a separate step between locating

the reference and searching for the target, whereas here selecting the appropriate reference frame is considered as part of the search for the target object. This is because if the locative sentence does not contain a projective preposition, the reference frame selection step becomes irrelevant. It seems unlikely that an influence that is only sometimes operative would be considered by a speaker as of equal importance to the always-relevant issues of reference locatability and definition of a search space. Also in some cases, where the object is weakly associated with an intrinsic frame of reference, (Carlson-Radvansky and Radvansky [1996] use the example of a post box) it is possible that the regions associated with more than one reference frame will need to be searched. This might suggest that an object with a weak intrinsic reference frame may be a less good choice of reference as it will have a less well defined search area. On the other hand there is no evidence in the literature to suggest that possession of a strong intrinsic reference frame makes an object a better or worse reference, independent of consideration of possible search areas for the target, again it does not seem appropriate to place this as one of the fundamental influences on reference choice. Possession of a reference frame is left as a possible influence on the optimisation of the search space (see section 3.4).

Continuing the argument from co-operation and effectiveness in communication leads to the addition of a third criterion for reference object choice, namely the *communication cost* of using a given reference. Grice [1975] outlines general principles on brevity and giving the optimum amount of information during communication; in this study there is a specific requirement that the communication should ‘match’ the difficulty of the search task. This is perhaps easier to illustrate in cases where complex locative sentences containing hierarchical or qualified references are used. To say “The mug is to the back and left of the desk” when the desk is relatively uncluttered may well be an over-specification that delays the overall search. Equally to use a single reference which under-specifies the object location (as in Miller and Johnson-Laird’s “The ashtray is near the Town-hall”), will result in a difficult search that delays the finding of the object. In the case of a single, unqualified, reference there are fewer possible trade-offs between communication cost and search task difficulty but the need to include this in the model remains and this is further discussed in 3.5.

These three primary influences on reference object choice are shown in figure 3.2. The factors which in turn influence these primary influences are described in the following sections.

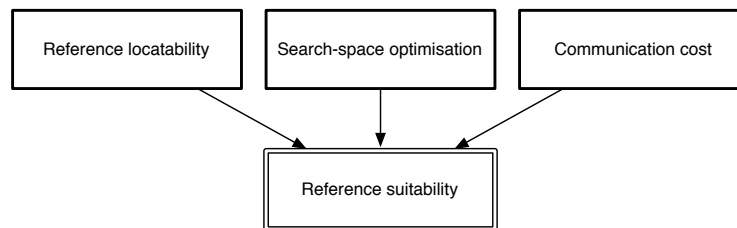


Figure 3.2: Three primary influences on reference suitability



### 3.3 Influences on reference locatability

#### 3.3.1 Specific and categorical knowledge.

For a reference object to be locatable the listener must either have specific knowledge of the object in question, or have categorical knowledge of the type of object in question so that it is apparent when visually encountered. Specific knowledge may substitute for or enhance categoric knowledge: for instance, in the case of “The National Gallery” specific knowledge of the building in question would be required by the listener if we accept that there is no visual category for art galleries. “St Paul’s Cathedral” may be specifically known but it is also clearly a member of its category and hence could be referred to as “the Cathedral” in some circumstances. Since the influence model is for reference choice it is more appropriate to term these two primary influences on reference locatability as

1. “Degree of belief in listener’s specific knowledge” for the case where specific knowledge is relied on.
2. “Reference apparency” for the case where categoric knowledge only is assumed.

These are shown in figure 3.3 and the various influences on these two factors are discussed in the next two sub-sections.

#### 3.3.2 Degree of belief in listener’s specific knowledge.

Various studies in landmark selection identify “historical or cultural significance” (Sorrows and Hirtle [1999]) or “semantic attraction” (Raubal and Winter [2002], Nothegger et al. [2004]) as contributing to the usefulness of a landmark. Examples of significant or semantically attractive landmarks would appear to include iconic individual buildings such as the Eiffel Tower or St Paul’s Cathedral and also buildings with universal identifying marks such as a MacDonalld’s restaurant. These would be taken to be references of such note that there is a good possibility that any listener would have prior knowledge of them and therefore be able to identify and locate them. This is termed “reference general significance” in figure 3.3. Note that Sadalla et al. [1980] identify ‘landmark familiarity’ as an important determinant of landmark choice in a speaker’s own cognitive process. This however, is a different consideration from direction giving, where the listener’s likely familiarity with a landmark is important.

The second influence on the speaker’s degree of belief in listener specific knowledge comes from the speaker’s knowledge of the listener rather than simply a judgement about the significance of the landmark. For instance in Shaftesbury, “Gold hill”, a well known landmark, would be useful in giving directions to visitors or locals; “Shooters hill”, less well known, would only be of use if the speaker knew that the listener was local to Shaftesbury. Sorrows and Hirtle [1999] give a similar example relating reference choice to frequency of visits to a building. This influence is included in Fig. 3.3 as the speaker’s “knowledge of listener’s past locales”. It can be seen that there is a scale of ‘universality’ that at one end

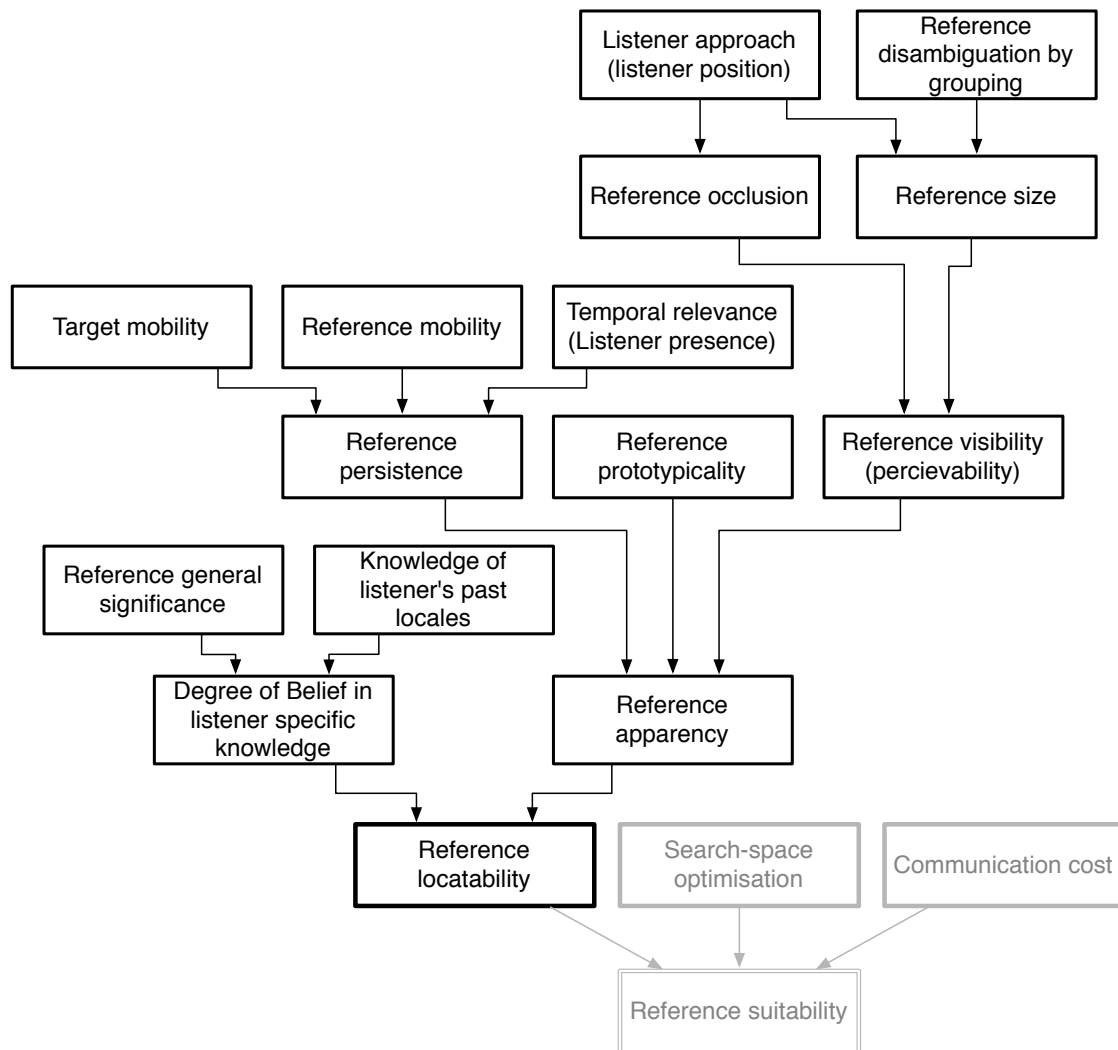


Figure 3.3: Influences on reference locatability

would have the Eiffel Tower and at the other, perhaps, “Mike’s desk”. The latter would be a useful reference only to people working in the same office as Mike, and only useful as an object of which the listener has specific knowledge, rather than as a member of its category, (there may be many desks in the office that fit with the ‘desk’ category).

Note that this factor may influence more than just reference object choice for a simple locative phrase and may influence whether a simple or complex locative phrase can be used (see Sect. 3.8.3). For example “In front of St Martin in the Fields” is likely to be replaced by “In front of the church at the north east corner of Trafalgar square” for a listener unacquainted with the church in question.

In an extension to this Tezuka and Tanaka [2005] propose a reinforcement mechanism whereby landmark usage effectively improves the *goodness* of the landmark. The initial choice of a landmark which subsequently becomes much used would presumably have been made because it displayed other characteristics of a good landmark. Otherwise although it

is possible that a landmark is famous for being a landmark, as far as a speaker is concerned it will still be assessed on the likelihood of the listener having prior knowledge of it. A related case is noted in Sorrows and Hirtle [1999], “turn left where the red barn used to be”, where the use of the landmark outlives the landmark itself. In this case also prior knowledge in the listener is being invoked, although it is knowledge of a past not a present scene.

### 3.3.3 Reference apparency.

If no specific knowledge of the object can be assumed by the speaker then the reference must be ‘apparent’. For a reference to be apparent to a listener who has categoric knowledge of the object in question it must;

1. Be a good representative of the category, that is, it should be prototypical (Sorrows and Hirtle [1999]),
2. Be visible.
3. Remain in place until it is no longer required as a reference, that is, it should be persistent, or permanent (Burnett et al. [2001]).

Note that although ambiguity may be thought to have a direct influence on apparency, the way it is proposed to deal with ambiguous references results in them influencing *either* communication cost or search space optimisation. This is discussed in section 3.5. The items from the list above are discussed further in the following sub-sections and illustrated in figure 3.3.

#### **Prototypicality.**

This is a complex area and the computer implementations of the hypothesis model do not include this parameter. Size, geometry and presence or absence of features will all influence prototypicality. It should be noted that this is also likely to be a cultural or regional issue, a prototypical English church is not a prototypical Texan church. Further study of relevant literature and consideration of methods of representation will be required before this can be brought within the scope of the model. In this study reference objects are assumed to be *recognisable* members of their category and prototypicality is not further considered.

#### **Visibility.**

In studies of landmark use the terms ‘Visual Salience’ (Denis et al. [1999], Tezuka and Tanaka [2005]) ‘Visual Characteristics’ (Sorrows and Hirtle [1999]) or ‘Visual Attractiveness’ (Nothegger et al. [2004]) are typically used instead of the simple term ‘visibility’ to capture a range of factors influencing the use of a landmark. The following attributes: size, degree of occlusion, brightness, colour contrast, shape factor (geometric extension) and

possession of distinctive features are all variously included in the assessment of this characteristic in the studies mentioned. The speed of travel of the listener (Tezuka and Tanaka [2005]) and the direction of approach of the listener (Winter [2003]) are also considered important.

Looking at this list of attributes it appears that they can be divided into two groups which can be considered differently. Firstly we have size, degree of occlusion and possibly aspects of geometric extension (principally ‘height’, although possibly only in larger scale, outdoor contexts), which can simply be considered as contributing to visibility. Secondly colour contrast, brightness, geometric extension, possession of distinctive features which all contribute to the ease or succinctness with which a reference can be described. This second set seems inextricably bound up with the idea of ambiguity and the way it relates to communication cost and so is further discussed in section 3.5.2.

This approach is reflected in figure 3.3 which shows the factors “reference size” and “reference occlusion” only affecting visibility. The factor “listener approach” (after Winter [2003]) is taken as influencing the degree of occlusion and, by some measures, the perceived size of the object. For listener approach in a static scene, such as those used in this study, we can substitute listener position, that is, how large and how occluded is the reference from the listener’s point of view.

Visibility is discussed by Gapp [1995a] who makes the important point that it is dangerous to use a simple measure of degree of occlusion. If the parts of an object which make it easy to identify, (the spire on a church for instance) are hidden then, even if this is a small fraction of the visible area, the degree of occlusion is high. This effect is linked to considerations of prototypicality and so modelling it is outside of the scope of the present study, however an assessment of whether it makes a lot of practical difference, when taken along with other considerations, can be made.

Gapp [1995a] also suggests that objects are differently categorised for size, giving the example of a man being characterised by his height and a road by its width. Whether this descriptive preference (which is sometimes seen, particularly in roads) translates into the cognitive process for reference selection is unclear as Gapp gives no evidence. Although this study tests many different size measures there is currently no facility for using different size measures preferentially with different object types (this is further discussed in section 7.2.2).

The apparent size, the area projected toward the speaker, may well be more important than the ‘actual’ size. Raubal and Winter [2002] and Elias and Brenner [2004] use this measure in the case of selecting building façades for use as landmarks. Degree of occlusion, size and position are all combined in the ‘visual salience’ calculation of Kelleher and van Genabith [2004]. Although it is tempting to say that an occluded object would always be a poor reference, as the researchers above implicitly do, the situation is not that straightforward. If the case of the cloth covering the table is considered, the occluded table may still be the better reference than the cloth. Talmy’s term ‘perceivable’ is probably better than the term ‘visible’, used by the landmark community, and so is also shown in figure 3.3.

The range of possible size, visibility and combined size/visibility measures used in this study allows evaluation of many of the assertions made by the researchers cited above. These are described in detail in section 5.6.1 but are as follows; bounding box volume, convex hull volume, actual (material) volume, maximum dimension, minimum dimension, simple visibility measures and variants on Kelleher’s visual salience algorithm to give different weights to visibility and position. These possible size measures are omitted from figure 3.3 for simplicity.

In figure 3.3 it can be seen that “reference disambiguation by grouping” is included as an influence on reference size. This stems from the way ambiguity is treated in this study. One of the possible ways of dealing with potential reference objects that are ambiguous is to group them together, a row of (more or less) identical houses can be used as a reference by referring to them as “the houses”. In this case of course the reference is larger than it would be if a single house had been used. A full discussion of the treatment of reference ambiguity is given in section 3.5.2.

### **Persistence.**

Following Talmy [2000] and the work by de Vega et al. [2002] it is clear that the mobility of both the target object and candidate reference object are expected to influence reference choice. Intuitively the reference object is expected to be more ‘stable’ (see Vandeloise [1991]) than the target. Also important, as pointed out by Burnett et al. [2001], is when the listener will need to use the reference to find the target. It was noted in section 1.4 that if, in figure 1.2, the target object is the post box and the listener will not be at the scene for some time, then the pink house, rather than the skip (which may be removed) will be a better reference even though the skip is nearer and plainly visible. These factors are summarised as “Temporal relevance (listener presence)” in figure 3.3. No experimental evidence appears to exist to support these assertions at present, this study provides some indications as to the practical relevance of mobility using measures described in section 5.6.1.

It is an open question as to whether persistence should be a direct influence on apparentcy or considered an influence on visibility as “visible at the time required”, it is left as a direct influence at present.

## **3.4 Searching for the target object**

### **3.4.1 Directed or constrained search.**

Searching for the target object is a different process to that of the initial search for the reference in that it is a directed search. It has a start point (the reference object) and a direction (in the case of a projective preposition such as ‘above’ or ‘left’) or a constrained region (in the case of a topological or proximity preposition, such as ‘on’ or ‘near’). Not all of the possible searches will be of equal difficulty and the choice of reference, as well as

taking into account the locatability of the reference, must take into account the difficulty of the search for the target once the reference is found. This is termed ‘search-space optimisation’ and the factors influencing it are shown in figure 3.4.

### 3.4.2 Reference location.

Reference location is likely to affect search-space optimisation in two ways.

1. The simple proximity of the reference to the target reduces the search space.
2. The presence of the target on a cardinal axis (where the reference is the origin) appears to make the search easier, (as well as the communication cost lower). This is apparent in experimental work by Carlson-Radvansky and Logan [2001], Carlson and Hill [2008] and Carlson and Hill [2009].

#### Proximity.

If the listener, having found the reference object, starts his search for the target at some point on the reference then the closer the target is to the reference the less space will have to be examined before it is found. It is not the purpose of this thesis to examine visual search processes or assess the vast literature on the subject. Much of the work in visual search does not seem analogous to the problem encountered in this study, which is finding a more or less unique object in a more or less defined space. A typical problem in the visual search field would be to locate the letter ‘T’ in a field of letter ‘L’s in various orientations. Wolfe [1998] gives a review of the theories advanced for human behaviour in this area and associated experimental work. Recently researchers in robotic vision (see for instance Söo et al. [2009]) have looked at the problem of efficient search for target objects in real world environments. This is more closely related to this study and currently the first steps of using locative expressions to assist in the robot search process are being taken. In Aydemir et al. [2010] the robot visual search processes are augmented with the knowledge that the target is ‘on’ a second object, whose location the robot already knows. Although improvements in search time and success rate are reported it seems difficult to conclude much from this work as yet. For the purpose of this study it is taken as self evident that in the absence of other constraints, reducing the search space will reduce the time taken to search for the target.

In Herskovits [1985] it is postulated that ‘nearness’ is assumed in a locative expression unless some evidence is given to the contrary (as in “the fountain is 100 meters to the left of the city hall”). However it is not entirely clear whether this nearness (assumed to be relative to the sizes of the objects involved) is only really relative to the distance to other potential references. Gapp [1995a] suggests that the nearness assumption may be related to the fact that the nearer a target is to the reference the less possibility there is that other distracting objects (potential targets) will come between the reference and the intended target. There appears to be no experimental evidence for this proximity requirement, in

reference selection, from the psycho-linguistic community, however proximity sensitivity is shown in judgement of preposition applicability by, for example, Costello and Kelleher [2006]. There is considerable supporting evidence from the landmark selection community, where among others Denis et al. [1999], Burnett et al. [2001] and Nothegger et al. [2004] all confirm a preference in humans for landmarks that are close to the places at which route choices need to be made.

As with reference size there are a variety of different measures for the distance between two objects that can be used, these differences become more important if the target and reference are modelled as complex 3-dimensional entities as is the case in this study. The different distance measures considered in this study are;

1. The distance between the target and reference object centroids,
2. The distance between the closest ('proximal') points on the target and the reference
3. As proposed by Gapp [1995a], the distance between the target centroid and the closest point to the target on the reference. The rationale given for this is that the geometry of the target is not considered until it is located; this may be true for the listener but may not be for the speaker.

### **Cardinal axis placement**

Recent experimental work by Carlson and Hill [2008] and Carlson and Hill [2009] indicates that the geometric placement of a reference relative to a target is a more important influence than a conceptual link between target and reference in the choice of reference. Proximity and location of the target on a cardinal axis defined by the reference (for example, target directly above or directly to the left of reference) are preferred in reference selection (see Sect. 3.4). The work of Carlson and Hill [2008] used very simple geometrical arrangements containing a target and two potential references, one of which was on a cardinal axis, and the other at 45 degrees. The references were visually similar although one had a 'functional' relationship with the target. (A burger, a mustard jar and a tub of pesticide would be a typical object set, the burger and the mustard jar being conceptually linked). Carlson finds that the reference for which the target is located on one of its cardinal axes is preferred irrespective of any functional relationships. The experiments were carried out using 2-dimensional object representations of similar size, on a 2-dimensional grid which gives a rather coarse granularity (objects are directly on the cardinal axis or at 45 degrees to it) however the results were emphatic even given the distraction of functionally linked objects. The experiments also do not entirely control for proximity (the reference objects on the cardinal axis were also closer to the target, if not dramatically so), but it is difficult to see this affecting the results.

Carlson and Hill [2009] also includes descriptions of experiments where participants were asked to describe the location of a target from a photograph of a desk with objects arranged on it and, similarly, in a real life situation of a room containing a desk with objects

arranged on it. The experiments were intended to test whether a ‘good’ spatial relationship between reference and target was more important than a ‘salient’ reference in deciding on the choice of reference. In this case salient means larger and more easily distinguishable because of a colour contrast; a red binder and a somewhat smaller greyish stapler are the ‘critical’ reference candidates. The experiments certainly demonstrate that the process of reference selection is complex and indicate that both object salience and spatial relationship are involved in the process, although no statistical analysis of the results is presented. It does not seem clear from the experiments that the spatial relationship is more important than salience as suggested by Carlson and Hill, as other factors, not fully incorporated in to the design of the experiments, and in particular the overall difficulty of the search task, make the data difficult to interpret. This is indicated by the fact that the desk, on which the target and the critical reference objects (whose salience and spatial relationship to the target are being compared) are placed, is itself chosen as a reference more frequently than the critical objects. The desk is itself a highly salient object (the largest in the scene) but very poor in terms of locating the target for this reason. However it is a good enough locator for many of the participants, because the target (a calculator) is large enough to be almost as evident as the critical reference objects. This lack of difficulty in the search task is likely to bias participants against the communication cost of more lengthy locative expressions, which may take longer to utter than the time taken for the listener to perform the unguided search, and among these would be included complex spatial expressions such as “behind and to the left”. It should also be pointed out that the difference in salience between the two critical references may not be that significant. They (and indeed the target as well) do not vary in linear size or volume by anything like an order of magnitude, although the desk is greater in linear size by roughly an order of magnitude than all three. To be able to come to firm conclusions about influences in reference choice will require further experiments which take into account, at least, a wider range of target sizes relative to the reference candidates and a greater spread of sizes in potential reference objects (or in this case controlling for the salience and spatial relationship of the desk which is a much larger reference candidate). Again the proximity of the reference and target objects was not controlled, the critical object in a ‘good spatial relationship’, being also closer to the target. These are clearly preliminary studies, and are the first psycho-linguistic studies to look at reference object choice in locative expressions, so it would be expected that more comprehensive experiments will follow.

Intuitively, given a preposition “above” and a reference the listener will locate the reference and move his eyes up from there until the target is encountered. Given a reference and the direction “above and to the left” the process is much more involved and the search space potentially larger, being in some senses 2-dimensional rather than 1-dimensional. The communication cost is also greater for the speaker and possibly the listener in the sentence “The bird is above and to the left of the barn” as opposed to “The bird is above the tree”. However it is dangerous to conclude that it is the nature of the spatial relationship, as expressible by the normal three axis projective preposition set, that is the dominant factor



in determining reference suitability. Saying “The speed camera is above the road” is of marginal usefulness even if the ‘above’ relationship is very good. It would be better to say “The speed camera is in front of the town hall” even if the ‘in front’ relationship is merely acceptable because the town hall limits the search space much more effectively.

However, even if indirectly, through the association of cardinal axes with projective prepositions, the placement of a potential reference on a cardinal axes will affect reference choice. Proximity and cardinal axis placement are shown in Fig. 3.4 as influencing reference location which in turn influences search space. The angular measures used in this study for assessing cardinal axis placement are described in section 5.6.1

### 3.4.3 Search start area.

As already noted, Miller and Johnson-Laird [1976] point out that the scale of the reference and located objects are important in determining whether a reference is appropriate. It is proposed here, following Plumert et al. [1995], that this is due to the influence on the search space. Choosing a large reference may make the reference more apparent but may leave the listener a difficult task finding the target object as, along with any preposition, it defines too large a region of interest (for example “the table is near Oxford”).

Reference size must be treated carefully as, dependent on the geometry of the reference and target objects, the search space may vary considerably. For instance, to say a target object is “to the left of the train” defines a large area from which to start a search but to say that it is “in front of the train” defines a much smaller area. In the case of a target object “near the train” the search area is determined by all the relevant surfaces on the reference from which the search can start. So the search space in the case of a projective preposition is the product of “projected area in the direction of the target” and “expectation of distance to target”. For the case of a topological or proximity preposition it is the product of “surface area for search” and “expectation of distance to target” where, to some degree, the expectation is that the distance to the target will be smaller than that for a projective preposition. Computational models illustrating this can be seen in Gapp [1995b] and Kelleher [2003]. Although these models are created as volumes of applicability for prepositions the results are similar. It can be seen that these possible search spaces are not well characterised by simply considering the volume of the reference object; the geometric extension of the reference and its orientation with respect to the target are also important. The variables used in this study to represent the search areas are discussed in section 5.6.1. In figure 3.4 the influence of this search area on the search space is shown as ‘reference search area’.

Clearly the search space so defined will be ‘measured’ relative to the size of the target object. To say “the suitcase is by the train” results in a more difficult search task than saying “the lorry is by the train”. The same considerations relating to the characterisation of the size of the reference (section 3.3.3) also apply to the size of the target, which is included in figure 3.4 as ‘target size’ for simplicity.

As Plumert et al. [1995] point out, if the target object is a safety pin and the listener

is more than a few yards away, there may be no single suitable reference. In the model developed in Barclay and Galton [2008] the location of the listener relative to the target was included as a separate influence on search space, leading to some confusion as to whether the model was to be interpreted as being from the speaker’s or listener’s point of view. Here the entire model for reference suitability is defined as ‘an influence model for reference suitability *to the listener*’ (and for this reason a reference is chosen (or not) by the speaker). In effect the whole scene is now assumed to be scaled by the speaker to account for the listener’s point of view. So if there is no reference that is sufficiently apparent (to the listener) and that defines a realistic search-space, this will force the decision to use a compound locative phrase containing more than one reference. This is discussed in section 3.8.3.

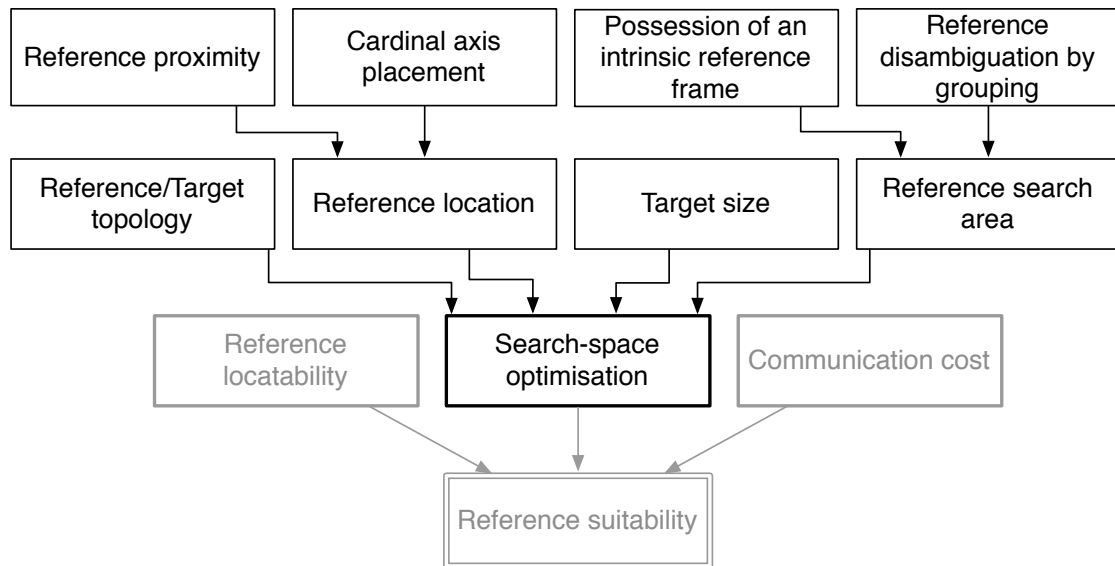


Figure 3.4: Influences on search-space optimisation

### 3.4.4 Reference and target topology.

From the study by Plumert et al. [1995] it is clear that as well as the geometry of the reference and target, the topological relationship between them is also important in forming a locative expression. If a target object was “on the book on the table” the book was more likely to be included as a reference than if the target was “near the book on the table” (in which case the the target was simply “on the table”). The reduction in search space would appear to be comparable for using either “on the book” or “near the book” but in the case where the target was “on the book” the extra communication cost of using the two references was more often considered worthwhile by the speaker. It is possible that there is a perceived chance of confusion in that an object “on A which is on B” is not necessarily seen as “on B” (that is, ‘on’ is not always accepted as transitive). This might then lead to a listener looking only for objects directly ‘on the table’ instead of ‘on objects on the table’,

however this does not always seem to hold considering the case of ‘the cup on the saucer on the table’ for instance. Note that this possible lack of transitivity is not necessarily the same as Miller and Johnson-Laird’s limited transitivity (Miller and Johnson-Laird [1976]) which is used to explain why the book on the table cannot be described as “on the floor” if the table is, as usual, standing on the floor.

While some sort of (possibly functional) topological relationship is implied by prepositions such as ‘on’ or ‘in’, it isn’t clear that the ability to use these prepositions improves the suitability of the corresponding reference, beyond the corresponding likelihood that the distance between the reference and target is low. Further, considering the alternative descriptions, “The boat is in the sea” and “The boat is by the headland”, neither the topological relationship between the boat and the sea, or the proximity of the boat and the sea means that the sea is the best reference object, as the search space it defines may be vast.

However cases where the target is occluded by a container will often lead to the container being specified as a reference. To say ‘the bowl is on the dishwasher’ suggests that the listener can look for the dishwasher and then will see the bowl. To say “the bowl is in the dishwasher” carries the implication that the listener should look for the dishwasher and then won’t see the bowl - any other reference would be confusing as the listener might expect to see the bowl. Note however that targets not visible *behind* references are dealt with in the same way and in this case the target and reference may have a variety of topological relationships, certainly including touching and disjoint.

As there appears to be no conclusive evidence for including or excluding the influence of topological relationships as an influence on search space the reference/target topology influence is included in the model at present pending further testing of its relevance.

## 3.5 Communication cost

### 3.5.1 Reference innate cost.

Communication cost is a complex issue and goes well beyond the quantity of syllables that must be processed. There is general agreement (see Grice [1975] and Burnett et al. [2001] for instance) that brevity is important, but increased cognitive load for either the speaker or the listener can come from a variety of other sources. Some of these have been investigated by the psycho-linguistic community, others have been noted by landmark researchers but typically with less experimental support.

The issue of the cognitive load associated with establishing a reference frame is investigated by Schober [1995] and Mainwaring et al. [2003]. Schober suggests that listener-relative or listener-intrinsic reference frames are more difficult to take (for a speaker) than either object-intrinsic or speaker-relative / speaker-intrinsic frames. Both Schober [1995] and Mainwaring et al. [2003] find that the speaker will usually take on the more difficult task out of consideration for the listener (or try to circumvent the reference frame issue altogether). Schober also notes that once a reference frame is decided both parties in a

dialogue will tend to maintain its use, if possible, and suggests from this that communication effectiveness is what is being sought by both parties. The differing cost of establishing different reference frames is illustrated in figure 3.5 as “reference frame orientation”, an influence on the innate cost of using a given reference, although it is to an extent an influence on a wider communication strategy and somewhat independent of the choice of reference. In this study the effect of reference frame selection is limited to an assessment, by the speaker, of the strength of an object’s intrinsic reference frame and, in the case where a listener is present in the scene, to the decision as to whether to use the listener-intrinsic reference frame (for example “it is in front of you”).

The relative difficulty of processing spatial relationships on the three cardinal axes has been investigated by Franklin and Tversky [1990] and Logan [1995]. They find that people are quickest in interpreting above/below relationships, slower with front/back and slowest of all with left/right relationships. This is in line with the strength of the environmental cues from gravity in the above/below axis and bodily asymmetry in the front/back axis. This may result in a preference for references placed above or below a target and a bias against those placed to the left or right. This is included in figure 3.5 as ‘axial location’. It can be seen in section 4.5.4 that the projective prepositions ‘left’ and ‘right’ are certainly used less often than the other projective prepositions, but whether this means that references are chosen less often along this axis, or whether proximity prepositions are substituted for projective prepositions in this case, is less clear.

Other areas, more specific to this study, where the communication cost of using a reference can be increased but with the result of reducing the difficulty of the search task are as follows:

1. The reference can be qualified by specifying a part or region of it (or associated with it) such as “the end of the road” or “the other side of the pond”.
2. The reference can be specified by description, typically by adding adjectives or appending descriptive phrases to arrive at “the tall green house with the gable” for instance.
3. The reference can be specified by a count, such as “the second grey house” or the “third set of traffic lights”.
4. The preposition associated with the reference requires quantifying distance or direction, as in “about 30m away from the tree at 11 o’clock”.
5. The reference frame associated with the reference may need to be established or oriented in a manner not covered by the dialogue situations mentioned above. This would include cases where the listener is not necessarily in place in the scene but in which local object relative reference frames are being established such as “to the left of the fountain looking from the town hall”

These items are expanded in the following paragraphs however in all of the cases the aim in trading communication cost against search difficulty must be to arrive at an effective overall communication of location, minimising the time for the listener to locate the target.. The second and third items in the list above are all concerned to some degree with disambiguating the reference. Disambiguation in the reference choice task is slightly complicated by the options it gives to the speaker and this is discussed in section 3.5.2. However all the items in the list affect the search task difficulty and the key question for this section is how the search is affected and how it should be included in the model.

### **Reference qualification**

As noted in section 1.4.5 references can be parts of objects (see de Vega et al. [2002]) such as “the town hall steps” or regions such as “the end of the road”. Qualified references of this nature are more like compound references in terms of their impact on the search task. The listener is able to perform the search in three steps as though he had two hierarchical references for example: first, find the road; second, look towards the end of the road; third, find the post box in the reduced search space ‘at the end of the road’. How the model might be extended to address compound references, and the costs associated with them, is discussed in section 7.3.3. The cost of reference qualification however is not included in the hypothesis model for suitability of a *single* reference. Note that in some cases where prior knowledge is being relied upon the search might be a single step and the reference should be treated as a single not a compound reference. For instance “next to the Tower of London” will probably not require the listener to first locate London, then the tower.

### **Reference specification by description**

Although reference specification by adding adjectives or descriptive phrases is only likely to be necessary if the reference needs to be disambiguated it is not clear whether this is a case similar to that of reference qualification or not. Faced with the expression “The post box is in front of the shop with the green awning” it might be thought that the search can be broken down into separate stages as in: first, find the shops; second, find the shop with the green awning; third, find the post box. However a street scene may contain different groups of shops along with pubs, churches and other visual elements, not organised in any particular way. Here it would seem that no assistance to the search is given by the description except to disambiguate the reference. Although it is possible to conceive of scene arrangements where some assistance to the search might be given in addition to disambiguation, the cost of reference specification is included in the model (figure 3.5) as ‘disambiguation by specification’ Whether the cost is due simply to the increased length of the utterance or to some extra cognitive load imposed on the listener is left as an open question.

The measure of brevity does seem to be similar to the consideration of reference specification however. Brevity is not associated with disambiguation but is represented by the length of the reference name prior to any specification. This may turn out to be a negligible

factor but is left in the model as an influence on ‘reference innate cost’ pending testing. It is possible that, all other factors being balanced a preference for saying “to the left of the bandstand” (2 syllables) as opposed to “to the right of the Winston Churchill memorial” (7 syllables) might be detectable.

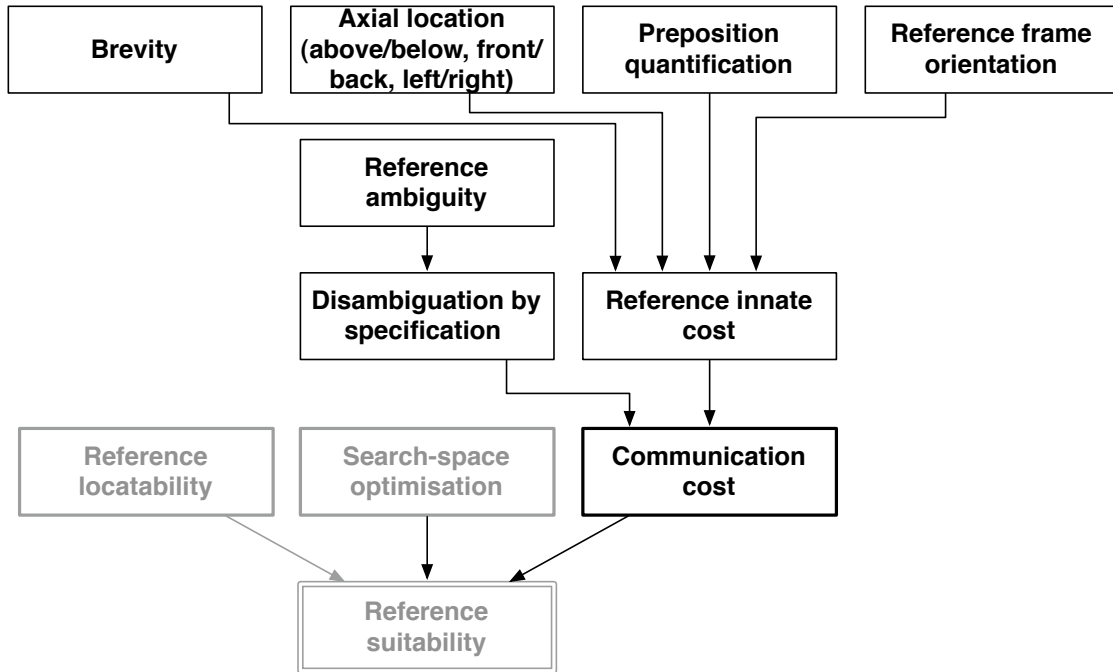


Figure 3.5: Influences on communication cost

### Reference specification by count

If the ‘third house’ is also the ‘tall green house’ it would seem that both descriptions are equivalent and therefore the treatment of count specification of a reference should be the same as that for specification by description. This does not seem to capture the whole picture however. For a count specifier to be appropriate there has to be some geometric order imposed on the objects, typically they would be in a line. This order is likely to make the whole (in this case ‘the row of houses’) easier to identify as well as *necessary* to locate, prior to locating the specified member of the group. So the reference appears to be hierarchical in a way that the search for the ‘shop with the green awning’ is not. The situation is complicated further by cases such as noted by Burnett et al. [2001], “turn right at the third set of traffic lights”. In this case there may be a requirement to search for each set of lights, which may not be in a row that can be discerned from a single point. In this case the description can be seen to be serial rather than hierarchical and it could be expressed as “go past the first lights, go past the second lights and turn right at the third lights” While there may be no clear distinction between the case of the houses and that of the traffic lights, except in as much as they can all be viewed (or not) from a single point,

it seems that the cost of count specifying a reference should not be included in the model for choosing a *single* reference for either case.

### **Reference frame orientation**

As an extension to the work by Schober [1995] some locative expressions have to take into account the fact that the position of the listener is not known at the time when he or she needs to make use of the information in the locative expression. In some cases, for instance “to the left of the fountain when looking from the town hall”, a reference without an intrinsic reference frame and in a place with no obvious object-relative reference frame may still be the best reference although it can only be used with the added cost of providing an orientation so that the resultant object-relative or speaker-relative reference frame is not ambiguous.

Another case is noted by Winter [2003] in selecting landmarks, that it is sometimes possible, but not preferable, to select a reference that is likely to be behind the listener when they come to use it. This not only makes the search task for the reference more difficult, but when it is found, can lead to confusion as to the orientation of any reference frame related to it. That is “did [the speaker] mean to the left of it from where I am now, or from some other point?” Winter’s example is of a navigator coming to a square, in which case a good landmark for a left or right turn would be on the side of the square facing them as they approach, not on the opposite side.

The cost of this is clearly associated with the reference used and is included, along with the reference frame considerations already noted, as ‘Reference frame orientation’, in figure 3.5.

### **Preposition quantification**

Use of angle quantification of a preposition (or replacement of a preposition) is more often seen in specific task oriented situations than in everyday life, and often where access to supporting instrumentation is available as in the case of navigation by compass, for example, “the yacht was North North West of the headland”. Prepositions may also need to be combined to achieve a similar effect as noted in the discussion on cardinal axis placement “The bird is above and to the left of the barn”. Prepositions can also have a distance measure attached to them as in “the treasure is buried 20 paces to the left of the tree” . The angular relationships between target and reference are modelled in this study which may show a general preference for cardinal axis placement when single prepositions are used. There is no facility however, in the computational model in this study, to reflect the reduced search area that would accompany a fully quantified preposition such as “the post box is 100 yards left of and 20 yards in front of the museum”, and balance its considerable extra cost of communication. This contributor to communication cost is included in the hypothesis model as ‘preposition quantification’ (see figure 3.5).

### 3.5.2 Reference ambiguity.

Two possibilities exist for a speaker confronted with an ambiguous reference in the case of a spatially locative phrase, as opposed to the case in which the object is the intended referent in a referring expression, when disambiguation is mandatory. Consider a scene such as that in Fig 3.6. The speaker can choose to disambiguate with a count as in “The bus-shelter is in front of the second grey house” or potentially by a specification as noted in section 3.5.1. However the speaker also has the option to aggregate the ambiguous references into a single unambiguous reference as in “The bus-shelter is in front of the grey houses”. This creates a reference with different size and geometry and hence will affect both the apparency of the reference and the search space associated with it. These influences are included in figures 3.3 and 3.4.

Methods for disambiguation and algorithms for arriving at suitable phrases are addressed in the literature on referring expressions, see for instance Dale and Reiter [1995] and for an empirical study of disambiguation using spatial location see Tenbrink [2005].

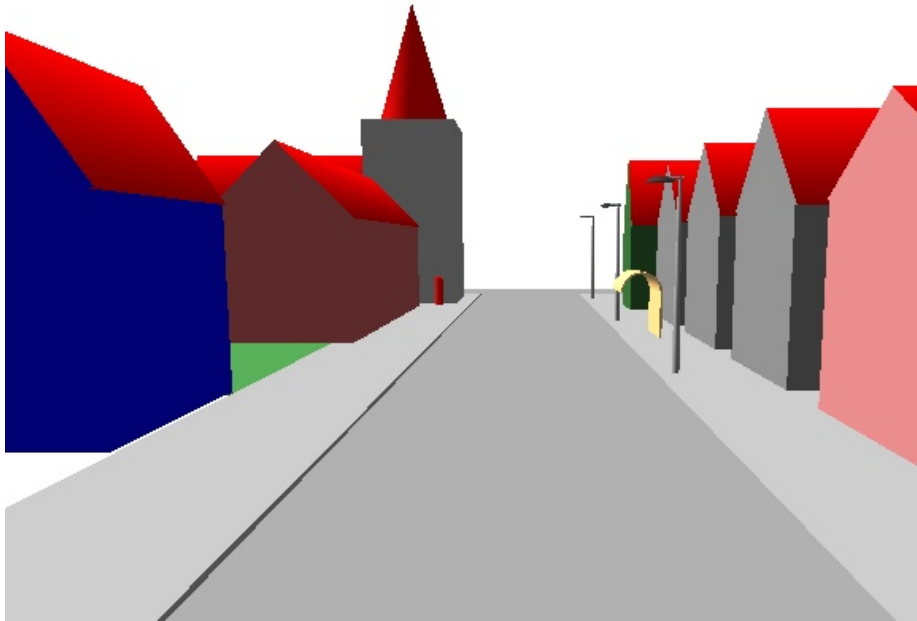


Figure 3.6: Disambiguation or aggregation: “The bus shelter is in front of ?”

## 3.6 Other approaches to determining reference characteristics

A few other researchers have approached the definition of reference characteristics from angles other than the process of interpreting locative expressions. The most relevant of these is due to de Vega et al. [2002] who analyse Spanish and German text corpora and make links between characteristics of reference and target objects and the prepositions that



are associated with them. The text corpora are taken from novels and are not necessarily descriptions of real scenes or, if they are, the scenes in question are not known. For this reason the study cannot be considered to be ‘grounded’, that is, it is not looking directly at the process of describing a scene and drawing conclusions about the reasons for a given description. However it is possible to derive some useful information about the characteristics of reference and target objects from a study of this sort, for instance the relative size of reference and target objects chosen and the tendency for humans to use mobile or animate objects as references might be expected to be reflected in such corpora. Learning what characteristics of spatial relationships between target and reference are important from this source is more problematic. There is for instance, no quantifiable information about distance between reference and target, and no contextual information about the spatial organisation of other reference candidates to allow conclusions to be drawn about why a particular reference has been chosen, rather than any other. In the case of directional prepositions, although preferences for object organisations along one or more of the cardinal axes might be gained, there is no information about whether a particular relationship was good or simply acceptable (to use the terminology of Carlson and Hill [2009]). So it is not possible to say that one reference was chosen rather than another because it was closer to the cardinal axis for instance.

The study selects phrases with a directional preposition connecting a target and reference (including at least some senses of the English ‘on’) discarding those that had a metaphorical or temporal rather than spatial meaning. The prepositions ‘right’ and ‘left’ were excluded from the study because too few examples of the use of these prepositions were found. Just over 2000 sentences in German and 2000 in Spanish remained. Characteristics of the objects such as mobility, animacy, solidity, countability, whole or part (that is, a hand is a part of a body), and relative size of target and reference were coded with binary variables. Various differences between the languages are noticeable and it is not clear why these should be so, or how they would translate to English. For instance Germans are (apparently) twice as likely to use a mobile reference than the Spanish in the vertical axis, although the figures for the use of mobile references in the horizontal (front-back) axis are similar. For brevity the aggregate of the German and Spanish figures are used in the examples here. The study finds that, in general, reference objects are less likely to be mobile or animate than targets, in line with Talmy [2000]. Targets and Grounds are both more likely to be whole objects rather than parts of objects. This being particularly true of targets for which partite examples are only 16% of the total. This confirms the expectation that, if a part of an object is attached to the whole, the whole would be the subject of an enquiry as to its location and then the whole being found, locating its parts is often a trivial matter. Perhaps surprisingly only a minority of targets (24%) are smaller than references although it is not clear how the coding was performed and objects of broadly similar sizes may have been grouped with those in which the target was larger than the reference, in which case this might be in line with Miller and Johnson-Laird [1976].

Correlations are then made between (and within) the cardinal axis directions and char-

acteristics of reference objects. Two characteristics have high correlations and these are:

1. Contact between reference and target (vertical 61%, horizontal 8%).
2. Adoption of a projective view by the speaker (vertical 0.5%, horizontal 36%).

There is an element of circularity apparent in both of these. The projective view is effectively defined to be in the horizontal direction as it refers to objects which can only be described using a speaker centred reference frame, and this is taken as excluding above and below in which the reference frame is defined by gravity. Contact between target and reference is presumably defined by the preposition equating to the English ‘on’ which is assigned, in this study, to the vertical axis. Other typical characteristic differences between the axes are, for instance, reference countability which occurs in vertical descriptions 81% of the time and horizontal descriptions 93% of the time. Taken as a group the differences in characteristics between the dimensions, evident in the corpus, allow a very good prediction of the spatial relationship (vertical or horizontal), given the characteristics. Note though that quite a lot of this predictive power is explained by the two cases mentioned. It is not clear however that the reverse can be said to be true, that given a target, a reference should be found that has certain characteristics if it is in a vertical spatial relationship with the target, or other characteristics given a horizontal spatial relationship with the target. This is particularly true since, with the two exceptions listed above, the individual characteristic differences between the axes are relatively small.

A second experiment is reported in which participants, are given target and reference objects and asked to complete a description by choosing either a vertical (above, below) or horizontal (in front, behind) preposition. The target and reference objects are chosen to have characteristics that should make either the vertical or horizontal preposition ‘more sensible or appropriate’. So for example the ‘vertical’ references are ‘inanimate’ and the majority (12 out of 16) ‘uncountable’ (firewood, sawdust, straw, grain, mud, snow, grass, sand, ice, ashes, debris, sea water, balcony, blanket, awning, brick). The ‘vertical’ targets are all ‘inanimate’ and ‘partite’ (newspaper page, bicycle pedal, door lock, guitar string, pot lid, watch strap, computer keyboard, motorcycle wheel, hatband, glasses sidepiece, picture ground, coat button, jar handle, pencil lead, tree branch, chain link). These objects generate designedly unfamiliar descriptions, however the participants select the vertical preposition 93% of the time. It is clear that, ‘sea water’, ‘snow’, ‘sand’ and ‘ice’ for instance all tend to have extended horizontal surfaces and that therefore descriptions including horizontal prepositions are less likely simply from the geometric arrangement. This is also true of the selected references ‘balcony’, ‘blanket’ and ‘awning’. It seems probable that the participants in the experiment are selecting the most likely visualisation and matching the preposition to this. This does not mean that the correlation between the object characteristics and prepositions is not genuine, but that it is explained by something other than the object characteristics listed in the paper. Most obviously this explanation would be the likely geometric extensions involved. In the case quoted of “The bicycle pedal is in front of the snow”, it may be an unlikely description because of the usual geometries,

but it is perfectly possible, in a particular situation, given a pile of snow with a detached bicycle pedal in front of it, that it is the best description. Instances like this cannot be explained by a statistical model of ‘what is most likely to be true’, which is all that can be obtained from a non-grounded study, but they must have an explanation. This explanation will be based on geometric aspects of the specific scenes not available in a text based corpus.

The key problem though, in relating this non-grounded study to a model such as the one being developed, is that it cannot include consideration of the candidate reference objects that were not selected as part of the descriptive process. That is to say there is no way of telling why the given reference was chosen as opposed to other candidates in the scene being described. As an example “The seagull is above the sea” and “The seagull is in front of the fishing boat” might be equally good examples of target, reference and preposition fit according to de Vega et al. [2002] but in different real world scenes one or other may have more value, or appropriateness. In reality, in a strictly locative sense, the fishing boat is far more likely to be an appropriate reference.

### 3.7 Practical limits on hypothesis model realisation

The model, derived from literature and from the proposed ‘process’ of locative expression interpretation, which was described in the previous sections, is intended to move towards a complete map of the influences on reference object choice. The extent to which the model developed in the preceding sections can be realised in the computational model for this study is limited by the nature of the training data set and the time available for the study. What might be termed the ‘testable model’ is shown in figure 3.7. Even considering figure 3.7, some of the variables available do not address the full extent of the influences as described in the previous sections. The following simplifications and omissions to the full hypothesis model have been made:

1. There is no way at present of including learnable measures of “reference general significance” or “speaker knowledge of listener’s past locales”. A simple mechanism for tagging some of the objects in the scene as specifically known to the listener could be used but this would not be the same as the speaker genuinely knowing the listener.
2. As noted, there is no attempt to measure or learn “prototypicality”. Prototypicality by itself would, in all probability, require a model of similar complexity to the one developed here. All of the objects in the test data set are assumed to be readily recognisable members of their categories.
3. Listener approach and listener position is limited to placing a static listener figure in the scenes. Further limitations on the position and orientation of the listener were found necessary for the scenes to be suitable for validation by human participants and this is further discussed in section 4.6.
4. Also, although not strictly limited by the training data, only a simple measure of brevity in communication cost, related to utterance length will be used. There is

no facility for assessing the cognitive load of preposition quantification. Reference frame orientation is not modelled as a cognitive load but the measures of angular orientation between the reference and target should allow preferences for front/back or above/below orientations as opposed to left/right orientations to be expressed.

5. The realised model allows for disambiguation by grouping, but only for objects with identical names. Three “red houses” can form a single grouped reference “the red houses”, however there is no facility for free aggregation of objects into groups so a “red house and a “blue house” cannot be grouped as “the houses”. The model allows ambiguous objects to be chosen as references but does not allow them to be disambiguated by counts (as in “the second red house”). This clearly does not express many aspects of the cognitive load involved in disambiguation, which will need to be the subject of future work.

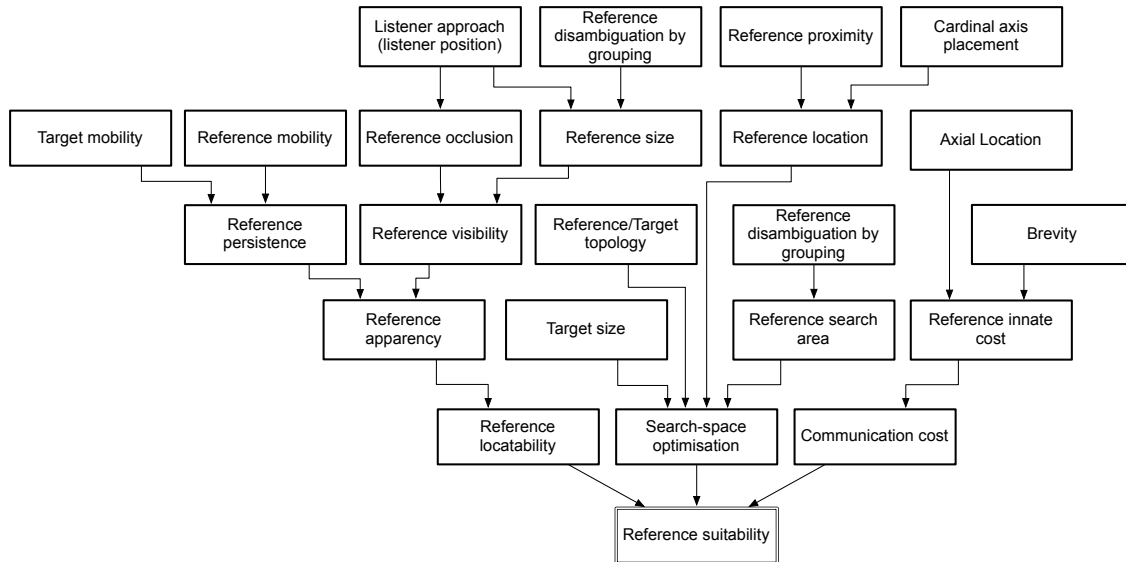


Figure 3.7: The extent of the hypothesis model realisable in the study

### 3.8 Model evaluation

The model as described is not a model of an entire scene such as that used by Socher et al. [2000], rather, for a given candidate reference object it is a model of that object’s suitability as a reference. For a given target object the model as described is ‘evaluated’ for each potential reference in the scene. A figure for the suitability of each candidate reference can be obtained and then the reference with the highest suitability figure will be chosen. Several potential problems arise with this approach.

1. The interaction between potential references is not modelled, each reference is considered in isolation.

2. There is no inherent way of limiting the search to a ‘realistic’ subset of potential reference objects, ‘reference pruning’.
3. There is no satisfactory way of rejecting all references and making a transition to a model which constructs a compound locative sentence.

### 3.8.1 Reference interaction

Both Wazinski [1992] and Gapp [1995a] suggest that the interposition of objects between a candidate reference and the target will reduce the acceptability of the candidate reference. Neither provide any evidence for the assertion although from figure 3.8 it can be seen that the presence of reference 3 might tip the balance between the otherwise identical candidates (references 1 and 2) in favour of reference 1. This is of course making the difficult assumption that reference 3 is not itself the best reference for some reason. Equally it will not always be true that interposed objects are problematic even if they are poor references. If the case of a group of people waiting to get on a bus is considered, the bus may well be the best reference for an individual in the group even though there are several other people between him and the bus.

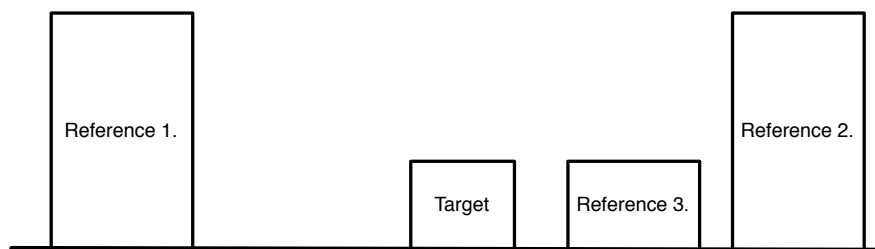


Figure 3.8: The effect of interposing objects between candidate references and the target object

The model described, and used in this study, treats individual candidate references on their own merits without considering neighbouring or interposed objects. There appears to be no clear evidence in the literature that the best reference when considered on its own is not actually the best reference and some evidence on this is provided by this study. This evidence is discussed in the context of results from running the model in section 7.2.2.

### 3.8.2 Reference pruning

If the model is not simply to evaluate all objects in the scene then it must evaluate the candidate reference objects in some order and various possibilities present themselves for this. Perhaps the most obvious would be to start with the closest object to the target and work ‘outward’. The evaluation order may prove to be important if the first suitable reference is to be returned or if pruning of the evaluation is required (that is, ignoring

references that are clearly unsuitable) as may be the case in a real world scenario with hundreds rather than tens of identifiable objects.

Various possible methods for implementing pruning are possible;

1. If a threshold for reference suitability had been established then the search could terminate when a given number of references had exceeded this threshold. If the number of references exceeding the threshold was set at more than one, then the most suitable of the set would be returned. This method could be applied to a search that proceeded outward from the target.
2. The previous method effectively performs an initial ranking by distance but a ranking by size could also be considered. References larger than the target by a given amount could be tried first, again with the search terminating if a suitable number of references exceeded a threshold of acceptability.
3. The search for a reference could be performed with a ‘simple model’ being used to dispose of definitely unsuitable references without full evaluation. Using some product of size and distance would seem intuitive. If a candidate reference’s size-distance product was less than some fraction of an already considered reference it could be discarded before the full model was evaluated. This does not require an acceptability threshold as such and the ‘best’ reference would be returned, however all objects would potentially be assessed, even if only in a computationally trivial manner.
4. A combination of the above methods of thresholding and filtering might be the best solution as it could return a good reference quickly but could also terminate without considering all objects in the scene, even trivially.

All of the methods above require assumptions to be made about reference suitability which at the outset may have little support. However there is no reason why pruning rules cannot be learned and applied after ‘experience’ has suggested, for instance, that objects smaller than the target and more than halfway across the visual field are never good references.

Currently no pruning is performed in the model as it is not necessary for the practical purposes of the study.

### **3.8.3 Transition to compound locative sentences**

Although not part of the computational model used in this study, which is only choosing single references for simple locative sentences, consideration of how the hypothesis model could be used to form compound sentences is instructive. It is clear that some threshold must be used for the acceptability of single references and that if no reference meets this threshold then a compound locative phrase must be considered. Work by Plumert et al. [2001] suggests that when giving directions compound sentences are organised with the

references in descending order of apparency. So a typical sentence, given to someone who was going to look for a handbag (say) would be;

“The handbag is in Harrison building, in lecture room 3, on the desk under the window”

Plumert et al also find a preference for giving references in ascending order of apparency if the purpose is simply for location description not direction giving. The descending order of reference apparency is also supported in the way-finding literature in the work of Tomko and Winter [2009] and Tenbrink and Winter [2009] who consider ‘destination descriptions’. These use a hierarchy of references, in the manner of a compound locative sentence, rather than the linear sequence of single references typically used in a route description.

Intuitively the easiest way to generate a compound locative expression using the model would be to choose a reference which defined a reasonable search space for the target whether or not it was suitably apparent and then use this object as the target for a further iteration of the model. The process would terminate when a reference that was suitably apparent was chosen. This would naturally produce an ascending order references as in “The handbag is on the desk under the window in lecture room 3 in Harrison building”. It is not clear whether operating the model in this manner to generate references in ascending order is in any fundamental sense ‘wrong’ in the case of direction giving, or whether humans also tackle the problem in this manner, with the sentence being re-arranged as required after the references are decided.

What is clear from the studies mentioned is that at some point in the construction of a compound locative sentence the focus switches from using reference objects to using regions or at least objects that define what might be termed a descending order of ‘scale spaces’. There is no scope in the model as currently presented, for raising the salience of these types of objects or regions which define the spaces which appear to have a psychological significance for humans.

The extension of the model to compound locative expressions is further discussed in section 7.3.3.

### 3.9 Summary

There is a distinct scarcity of literature pertaining to reference object selection for target object location. This is certainly true when compared to other elements of spatial language such as preposition usage. Considering that the problem is far from straightforward and presumably, in the context of understanding locative expressions, at least as important as other elements of spatial language, this seems strange. However to enumerate, two recent papers by Carlson and Hill [2008], Carlson and Hill [2009], a proposed model by Gapp [1995a] and mentions in passing by Talmy [2000] and Miller and Johnson-Laird [1976] form nearly all of the relevant literature outside the specific field of landmark selection. The study by de Vega et al. [2002] is interesting but its relevance is diminished by the lack of physical context in its source material. The work of Plumert et al. [1995] does not look directly at reference selection, but at the order of use of hierarchical references, however

it contains some interesting observations. So, although limited in its field of application, considerable dependence has to be placed on the literature concerning landmark selection in wayfinding for detailed experimental investigation in to reference selection. In particular Burnett et al. [2001] and Nothegger et al. [2004] compare lists of reference (landmark) characteristics with human behaviour in navigation and Tezuka and Tanaka [2005] compare landmark characteristics with information stored on the web.

The central point of this chapter is to demonstrate that the lists of landmark or reference characteristics produced by the researchers mentioned, although useful, are not sufficient to explain human behaviour in reference choice. The example given in the introduction to this chapter shows the requirement for understanding the interactions between the different characteristics along with the process of using the reference to locate the target. The most important factor in organising the model is to note that, in using a reference object, the search for the target has been broken into two steps. The different characteristics of the reference object, the target object, the geometry of the scene, as described in this chapter, then contribute to the difficulty of first finding the reference and secondly searching for the target given the reference and a spatial preposition. It is also postulated that the communication cost of using a reference will be an important consideration and will be balanced against the overall difficulty of the search task.

Although there is support for certain elements of the model in the literature surveyed, the overall structure is clearly untested, although not without some basis in reason. The model can be used as a start point for other investigations as well as this study and can be amended as evidence suggests.



## Chapter 4

# Design and Validation of a test data set

### 4.1 Introduction

In this chapter the design, construction and validation of the test data set is described. The test data set is a collection of scenes such as the two shown in figure 4.1. (The term ‘scene corpus’ is used interchangeably with ‘test data set’). An object to be located, the ‘target object’, is identified in the scene and the task for the machine model and for human participants is to commence the process of forming a locative expression by choosing a suitable reference object to locate the target. The scene corpus consists of 133 such scenes. Each scene has up to four target objects defined giving a total of 529 training and test cases. An average of 27 objects are present in each scene, any of which could be chosen as a reference for the identified target. The actual objects chosen by a group of human volunteers and the corresponding choices of prepositions for two example scenes can be seen in figure 4.1. Preposition choice is not yet part of the machine models and is not the focus of this study. However it seemed sensible to collect preposition choices as part of the study to allow for future extensions to the work.

The key point about the test data set is that it should enable reference choice in a manner close to that of the real world. That is to say that the influences on reference choice identified in chapter 3 as being present in the real world should also be present, as far as possible, in the test data set. Although some influences from chapter 3 have been ruled out of scope for this study, the thesis that machine learned models can be used to model human spatial language generation, in a situation where many influences are at work, is still testable provided the data set is sufficiently realistic and representative.

For complete coverage of the reference object selection problem, human participants would describe ‘very many’ real world locations and the computer model would be trained and tested through analysing stereoscopic images of the same scenes. This is not yet possible for a variety of reasons. Firstly it expands the problem of generating spatial language to include the error prone task of object recognition in potentially highly cluttered envi-

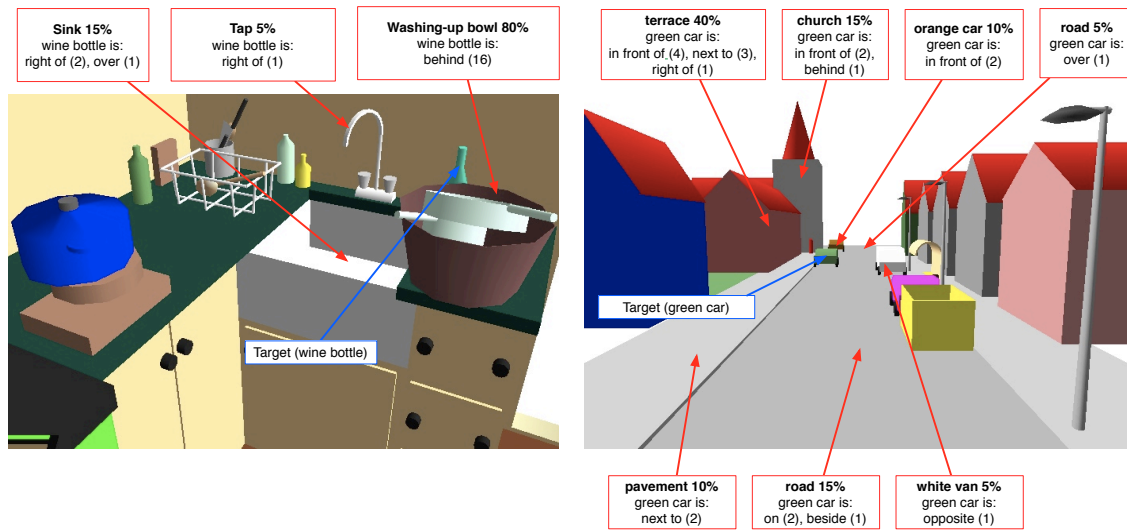


Figure 4.1: Two scenes from the data set at table-top and street scale. An example target object for each scene and the opinions of human participants as to suitable references and prepositions are also shown. The percentage against each reference object is the proportion of people who chose the reference, the number in brackets against the prepositions is the actual number of people who chose that preposition to go with the reference

ronments. Secondly it increases the computational load unnecessarily for the investigation at hand and makes the collection of training and test data from humans far more costly.

However questions still need to be asked about the ‘realism’ and ‘representative nature’ of the test data set. Clearly the scenes are not ‘real’, they are ‘cartoons’, but this does not mean they are not real enough for the purpose of the study. Also the test data set does not ‘represent’ the entirety of situations in which humans make judgements about object location and generate suitable spatial language; there are (for instance), no maritime scenes or railway stations. However a wide range of scene scales and locations are provided that might span the daily experience of many humans. It will be shown that the data set does present a more realistic and diverse set of situations for grounded language production than any others described in the literature but whether this is sufficient is one of the main questions addressed in this chapter.

The second major issue addressed in this chapter is the need to validate the data set. The machine learning systems used in the study are ‘supervised’, that is, they learn directly from examples provided by humans rather than by calculating a ‘goodness factor’ from some set of heuristics or a scoring system. In this study the examples provided are of ‘good’ references for a given target object. The problem is that the data set is too large to be entirely supplied with example references by human participants other than the author. Relying solely on the author’s judgement of what constitutes a ‘good’ reference, however, would devalue the exercise as the author might be biased by his own ideas of what influences on reference choice should be present or dominant in a particular case. To overcome this as far as possible a subset of the test data set has had reference choices provided by groups of

volunteers who are largely unfamiliar with the study (as shown in figure 4.1). The author’s opinions, which are necessary for the bulk of the example reference choices in the data set, are compared with those of the volunteers to ensure that no marked bias is present. All of the scenes that have had reference choices provided by volunteers are shown in appendix B. This also gives a good overview of the diversity and complexity of the corpus as a whole.

The rest of this chapter is arranged as follows:

**Section 4.2** describes the design of the test data set (scene corpus) including the requirements that were considered and the sources from which the scenes were derived.

**Section 4.3** describes how the scenes are constructed and represented to the machine learning system and to the human validators.

**Section 4.4** describes the first validation exercise by which reference choice examples are gathered from the groups of volunteers.

**Section 4.5** presents and discusses the results of the first validation exercise.

**Section 4.6** presents the results from a second validation exercise in which groups of volunteers provide reference choice examples in scenes with and without a figure representing the listener present.

## 4.2 Scene corpus design

### 4.2.1 Corpus requirements

The corpus was designed with the potential to be used for more experiments in spatial language than solely the current study. Although the presentation of near real world scenes in some ways makes this wider usage inescapable there are some limitations to what can be achieved with the corpus in its current state. The requirements considered in the design of the corpus are as follows:

1. Representation of the real world. This study is not designed to examine the influence of one or two factors on reference choice in a balanced set of experiments where the influence of other factors can be cancelled. It is designed to encompass as wide a range of factors as possible to model human behaviour in an environment as close to reality as possible. The scene corpus therefore needs to represent reality as closely as possible. How closely the corpus and the individual scenes approximate reality is not a measurable quantity and is therefore open to debate. For the reference choice task what is necessary is to represent as many of the different influences identified in chapter 3 as possible in as many different combinations with other influences as possible. A complete enumeration of these combinations is also impractical but an overview of object relationships and characteristics present in the corpus is attempted in section 4.2.3.
2. Human and machine ‘readability’. For the study to achieve its goals it is necessary that the scenes in the corpus are understandable by humans to the extent that a judgement about an appropriate spatial description can be given and also that the

scene is interpretable by a machine learning system in such a way that the machine can make the same judgement about spatial language.

3. Coverage of spatial language. Although reference choice is the focus of this study the scene corpus should allow similar multi-factorial experiments on other aspects of spatial language. Elements of spatial language that could be covered include:
  - (a) Selection of appropriate reference object(s). This is essential to the current study.
  - (b) Adoption of appropriate reference frame. This is covered to an extent in the corpus. Objects with intrinsic reference frames have an assigned vector denoting their ‘frontal’ direction. ‘Strength’ of intrinsic frame is not modelled and currently no object-relative reference frame indications (such as compass directions) are given.
  - (c) Use of correct spatial prepositions. This is implicit in the corpus for geometric preposition usage but no direct assistance is provided for interpretation or assignment of prepositions taking into account functional relationships between objects. To test preposition use taking account of object relationships would require additional data in the form of an extended ontology, or illustrative animations as in Coventry et al. [2005], and possibly a discourse model.
  - (d) Incorporation of gesture, emphasis or other non-verbal communication. No consideration was given to this factor during the corpus design.
  - (e) Integration of listener models. Listener position and orientation are modelled in the corpus but no consideration was given to modelling the listener’s knowledge of the scene or objects in it.
  - (f) Strategies for construction of multi-phrase descriptions. The scenes in the corpus are complex enough to allow modelling of hierarchical locative sentences with multiple references and ‘linear’ descriptions of scenes. Animation of the scenes in the corpus is allowed for in the design of the underlying software structures, which allow for multi-frame sequences, but this has not yet been realised, so experiments on path descriptions are not yet feasible.
4. Size of corpus. Although it is trite, the only answer to the question “how large should the corpus be?”, if the corpus is supposed to represent reality is, “as large as possible”. It would also be dishonest to say that any reasoned consideration set the corpus size to its actual level. Available time limited the size of the corpus. Whether the corpus is large enough to address the needs of the study is discussed in section 7.2.1. The necessary size of a test data set for a machine learning exercise depends on a number of factors including, with particular reference to this study:
  - (a) The number of variables and number of values each variable can take.
  - (b) The statistical dependencies between the variables.

- (c) How ‘noisy’ the data set is, and how well it represents the statistical distribution of variable values in the real world (the entire population in statistical terms).
- (d) The degree to which the actual variable dependencies are represented in the data

If a good model for reference suitability turns out to require 10 variables which can each take 5 values, if none of the variables are statistically independent and if 10 occurrences of each combination of variable values was thought to provide a sufficient measure of the likelihood of that combination representing a good reference (or not) then a well distributed and noise free data set would need to contain  $10 \cdot 5^{10}$  or just short of 100 million examples. In fact the full test data set consists of 529 test cases, for each test case there are about 3 examples of good references and 24 examples of bad references. So there are just over 14 thousand, poorly distributed, examples in the test data set. Fortunately there is a considerable degree of statistical independence of the variables and many combinations of variable values are irrelevant (or even impossible).

5. Knowledge requirements. From the literature reviewed in section 3 there is an expectation that characteristics of the objects in the scene as well as their specific geometric relationships within a given scene will affect their suitability as references (as well as judgements about other aspects of spatial language). The obvious example of this is whether and to what degree objects are mobile, (or indeed animate, able to move under their own volition) which cannot be determined from a static scene. Other object characteristics which may affect spatial language use include function and functional relationships, ubiquity (how common is an object) and plurality (do many instances of an object occur together).

#### 4.2.2 Corpus derivation

The corpus used in the study is constructed by hand as in the end there seemed no other way of addressing the needs of the study; other possible sources of material were considered for the study as noted in the following paragraphs.

##### **Photographs.**

Use of photographs as a direct input to the study would have enabled a corpus of any required size and diversity, and the best possible representation of reality. Unfortunately photographic source material is not satisfactorily machine readable with current state of the art object recognition software. Whilst staged, uncluttered scenes containing more or less discrete objects could be ‘read’ by a machine these would not have fulfilled the requirements of representing reality. Direct use of photographs (video clips) is used in some spatial language systems including the VITRA project (see section 2.3) and Tellex and Roy [2009]; these deal with trajectory descriptions related to simple pre-assigned references. No

attempt is made at recognition of static objects in cluttered environments, a more difficult problem and a requirement for this study.

### **Existing scene corpora.**

It does not appear that anything similar to the corpus used in this study exists at present. A corpus for the study of referring expression generators was developed by van Deemter et al. [2006] but this is 2-dimensional and not representative of reality in a spatial sense (objects are arranged on a simple geometric grid). A corpus of annotated maps for the study of automatic route description generators is used by Schuldes et al. [2009], which is rich in terms of its geometric complexity but again is only 2-dimensional. Use of 3-dimensional mapping data would come close to fulfilling the requirements of the study (see Elias and Brenner [2004]), however the range of object types (basically buildings) and scene scales is limited and permits only landmark, not general reference object, selection. A recent development in testing natural language generation systems, the ‘GIVE’ (Generating Instructions in Virtual Environments) challenge, is described in Byron et al. [2009]. This is a fully annotated 3-dimensional virtual environment which could potentially fulfil the requirements for this study. It emerged after the decision had been taken to generate the scene corpus internally and currently does not yet cover the required range or complexity of environments or diversity of object types. There seems no reason why it could not do so however, and any further studies should consider collaboration with the ‘GIVE’ research team.

### **Virtual reality game systems**

There are many different on-line multi-player games that offer a huge number of scenes of the required complexity and diversity for a language generation study such as this. The most obvious of these is probably ‘second life’<sup>1</sup> particularly since this has its own 3-dimensional graphic content generator available to users. There are several problems with using scenes culled from games such as these however;

1. Accessing the data in a form that is usable by an external system is not straightforward.
2. Annotation is often not complete, meaning that a manual process of object identification and tagging has to be undertaken before the scenes can be used.
3. The number of vertices used in a scene is difficult to control leading to problems with computational load. This is a recognised problem in second life even for the basic process of rendering scenes on a display and would present a more serious issue for geometric and topological calculations.

---

<sup>1</sup>see [www.secondlife.com](http://www.secondlife.com)

None of these problems is insuperable and in future the placement of conversational agents in these environments is likely to be an important application area. However the greater control over data structures and certainty of outcome led to the decision to continue with an ‘in house’ scene corpus.

### 4.2.3 Scene corpus construction

Some of the requirements outlined for the corpus in section 4.2.1 are addressed by the scenes, in their complexity and the diversity and number of objects within them. Some of the requirements are addressed by the construction of the corpus. In particular the factors described in the following paragraphs must be taken into account;

#### **Approach to realism through diversity of scene provision.**

As noted the question of how diverse the subjects need to be to ‘represent reality’ does not have a simple answer. Also as noted a minimum requirement for this study is a representation of the influences on reference choice in sufficiently rich combination to enable training and testing of models to test different combinations of influences. The approach to this in the scene corpus has been twofold. Some scenes have been created to specifically contain a range of geometric or topological relationships and ensure a minimum degree of coverage of these. Other scenes have been taken from photographs of different locations to try to provide a diversity of objects and arrangements of objects that genuinely reflect those found in the real world. The scenes which have been staged do not appear unreal and still contain a wide range of objects among which only two or three will have been manipulated to give some coverage of particular spatial relationships. Some of these are illustrated in figure 4.2 where the box can be seen to be adopting a variety of angular relationships with respect to the books and the pen and the post-it notes are at varying distances between and around the mug and the bowl.

In particular the scenes in the desktop series (see table 4.1) contain arrangements of a target object and key candidate reference objects which:

1. Contain eight deliberately varied angular relationships
2. Contain a graduated range of separations
3. Are contained in vertical stacks and horizontal rows of objects in different orders and at different degrees of separation
4. Are in differing degrees of (convex hull) topological overlap

Although not directly relevant to this study the scenes in the ‘park’ series contain various arrangements of objects to illustrate the prepositional arrangements of ‘beyond’, ‘along’, ‘around’ and ‘among’.

The following list gives some idea of the range of values (of the variables that constitute the influences on reference choice) which are covered in the scene corpus. Note that these

value ranges were not an a-priori design criterion, they arise from the attempt to include a wide range of object types and arrangements within the scenes in the corpus.

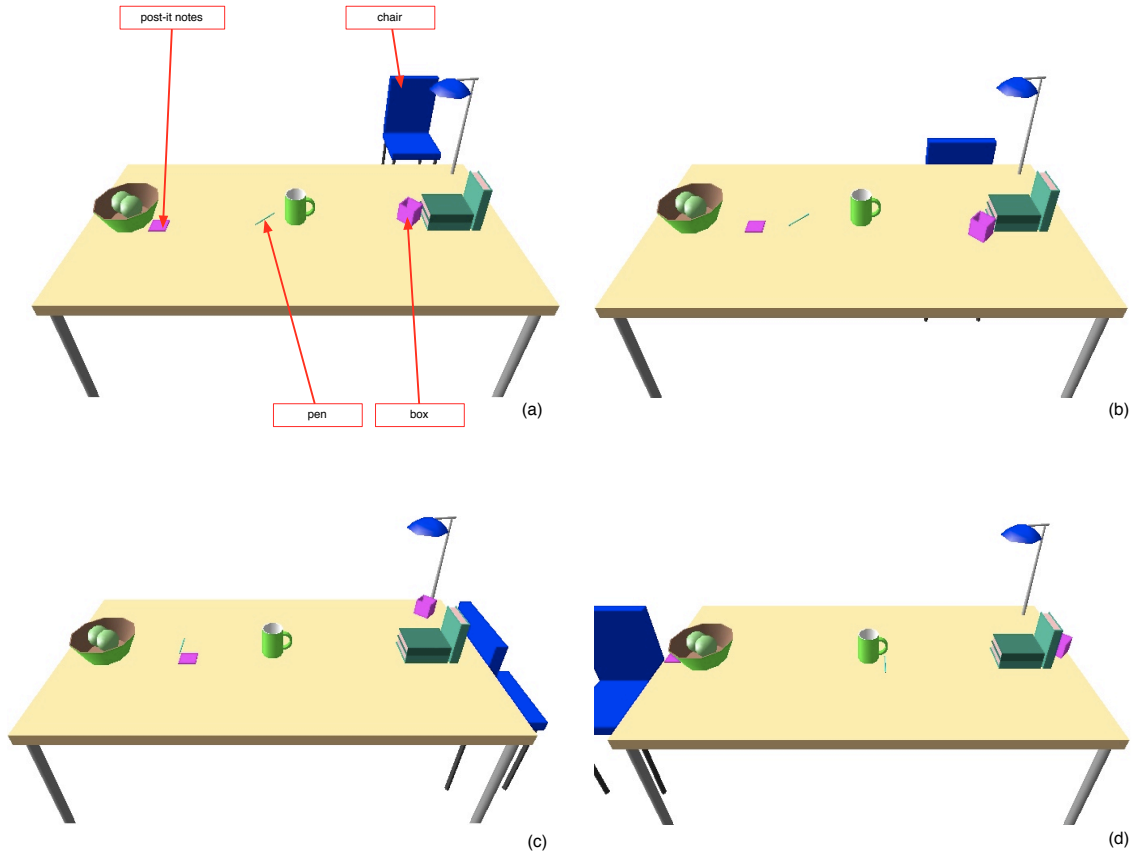


Figure 4.2: Scenes in which objects are placed in graduated geometric arrangements. The pen and the post-it notes vary in separations from the mug and the bowl. The box varies in angular placement around the books and the chair varies in angular placement and separation around the desk

1. Volume of objects. The largest object is a church at  $\approx 850m^3$ , in the same scene the smallest object is a rucsac at  $\approx 0.04m^3$ . The smallest object in any scene is a pen at  $\approx 0.000003m^3$
2. Extension of objects. The object with the lowest geometric extension is a ball with maximum to minimum dimension ratio of 1 the object with the greatest extension is a road with the maximum to minimum dimension ratio of 5000. Of objects that are not normally perceived as surfaces the greatest extension is a park wall with a dimension ratio of 600. Objects which are near 1-dimensional (pens, street-lamps) and near 2-dimensional (mats, paintings) are included.
3. ‘Density’ of objects. Objects are included which have a high ratio of convex hull volume to material volume (such as chairs, draining baskets) as well as objects which have a similar convex hull volume to material volume ratio but which are perceived as



more solid (thin walled containers such as boxes and bowls). Note that some objects, in particular trees, are poorly represented in this way being portrayed as much more dense than they actually are.

4. . Topological relationships. The relationships of proper part of, overlapping, touching and disjunct are covered for convex hull volumes of objects and overlapping, touching and disjunct for material volumes. This is not sufficient in itself however. Sequences of topological relationships are required to illustrate cases such as the “cup is on the table” when in fact “the cup is on the saucer on the mat on the table”. A wide variety of these are available in the corpus: fruit is on other fruit in bowls; saucepans are on frying-pans in washing-up bowls in sinks on cupboards; men are in (or under) bus-shelters that are on pavements, and so on.
5. Object occlusion, A few objects are totally occluded in the corpus. For practical reasons these examples have not been used by human validators (see section 4.4.2), except possibly in error. Other degrees of occlusion have not been manipulated but are variously present including, due to the perspectives available, large objects which are nearly occluded by much smaller objects.
6. Geometric relationships. As well as the manipulated relationships it should be clear from the various scenes illustrated that objects of all shapes and sizes at all angles and distances from each other are present in the corpus.
7. Ambiguity and plurality. Some objects usually appear in reasonably large numbers in scenes (books and street-lamps). Others tend to occur singly (sinks in kitchens, churches).
8. Mobility. Objects range from totally immobile (buildings) to effectively always mobile (people, dogs). Animacy, as distinct from mobility, is not apparent to the machine models. Objects in scenes are not tagged as animate and there is no provision of examples for learning it.

#### **Provision of a variety of scene scales.**

Montello [1993] provides an overview of ideas relating to perception of scenes at different scales. He defines four different ‘psychological spaces’ a term he prefers, with some justification, to scales:

1. Figural space. Spaces smaller than the body that can be comprehended from a single point without moving. Note that Montello uses projected size so that both a distant landmark and a picture would qualify as figural spaces.
2. Vista space. Spaces larger than the human body but which can also be comprehended from a single point without appreciable movement. Rooms and town squares are given as examples.

3. Environmental space. Spaces larger than the human body and surrounding them so that the space can only be comprehended by moving around in them. Buildings and Towns are given as examples.
4. Geographical space. Much larger than the human body such they can only be comprehended through symbolic representations, such as maps, which are in figural space.

Montello cites Ittelson [1973] who also distinguishes ‘environmental’ space as being the class of spaces that require movement to comprehend and for which knowledge must be built up over a period of time, as distinct from ‘object’ spaces that can be comprehended, effectively immediately, from a single point. In this study although a range of scales are provided within the test data set, which represent Montello’s figural and vista spaces, they all occupy Montello’s figural space or Ittelson’s object space when portrayed on a computer monitor and there is no scope for learning them by moving around within them. Montello would be sceptical about them being perceived (in a psychological sense) as different types of space. The term scale is used in this study as, in reality, the difference is only one of size and does not include ideas about whether movement, time or symbolic representations are required to comprehend them.

It seems clear that some aspects of spatial language are affected by the scale of the scene. For example it would seem that the preposition ‘beyond’ is more often used in large scale scenes than when referring to objects on a table top. A study by Lautenschütz et al. [2007] also suggests that preposition use may vary across scene scale. This finding needs to be treated cautiously, pending further research, as the study was unable to dissociate object scale from object characteristics. Whether variations in reference choice are similarly affected by scale is not clear. The structure of the hypothesis model in chapter 3 is effectively scale independent but when translated to a machine learning system the parameters may be scale dependent. Provision of scenes at different scales in the corpus will allow this to be tested to a certain extent. The classification of scales in the corpus shown in table 4.1 does not correspond directly to any of those reviewed by Montello but relates to the organisation of the principal objects in the scene and how they in turn might relate to the reference choice problem as follows.

1. Table top scale spaces have a single dominant object (the ‘table’ around, or on, which all the other scene objects are placed, but which in itself is often not a good reference as it does not serve to focus the search for the target
2. Room scale spaces are confined by a single dominant object (the ‘room’ in which all the other scene objects are placed. In a similar manner to the table top scale the room is usually a poor reference.
3. Street scale spaces are confined although ‘external’, however unlike the room scale spaces the confinement is provided by many, often regularly spaced, objects (buildings) which may be good references.

4. Vista scale spaces are (more or less) unconfined external spaces. No single objects dominate to the extent that they are almost by definition poor references. Provision of potentially good references is not so certain as it is with the buildings in a street scene.

Ultimately the rationale for providing scenes at different scales in the corpus is that this accords with human experience which in most cases encompasses the scales provided in the corpus (and maybe more) on a daily basis.

### **Ability to learn object characteristics**

Although the descriptions of the scenes given by humans are intended to be ‘history-less’ it is thought that pre-acquired knowledge of object characteristics, and particularly their mobility, will affect judgements as to their suitability as references, even in a static context. To enable the computer to learn these characteristics the corpus is comprised of groups (sets) of scenes which have the same viewpoint. Typically there are ten scenes in a set, for the photographically derived scenes the time intervals between photographs were varied from between a few seconds up to a few hours, although the time intervals are not used in the model. Within these scene sets some objects will change positions and some will remain stationary. In this way the machine model can judge the degree of mobility of different objects (learning animacy is, as noted, another matter). Other characteristics of objects could be ‘learned’ including ubiquity, plurality and associations between objects, but these have not been used in the study so far.

### **Scene corpus extension.**

The full data set as described was arrived at in two stages. Initially 93 scenes containing 369 test cases were created. Initial experiments did not yield as many statistically significant results as had been hoped with this size of data set, the sample size being effectively too small to allow discrimination between different machine models of similar performance. This is further discussed in section 6.4. For this reason a further 40 scenes containing 160 test cases were produced. This second series contained 30 exterior and 10 interior scenes giving totals of 63 exterior and 70 interior scenes, thus allowing for comparative experiments on interior and exterior scenes. The second series of scenes are also more complex than the first in terms of number of objects and object representation to facilitate future work using parts and regions of objects.

This extension to the scene corpus had unforeseen consequences which are described in section 6.4. Which sets of scenes belong to which series, their derivation method, and the distribution of scene scales, is shown in table 4.1.

#### **4.2.4 Test case generation**

Several test cases (typically four) are generated for each scene. A test case is generated by specifying a target object whose location is to be described. So for instance in figure

Table 4.1: Subject distribution and derivation of scenes in the corpus

Scene subject	Scale	Derivation	Number of scenes	Series
Desk	table top	extrapolated	30	1
Kitchen	table top	photograph	10	1
Living room	room	extrapolated	10	1
Kitchen	room	photograph	10	1
Pub interior	room	photograph	10	2
Residential street	street	imagined	10	1
Shaftesbury high street	street	photograph	10	2
Shops	street	photograph	10	2
Park walk	vista	photograph	10	2
Park	vista	imagined	16	1
Parish church	vista	imagined	7	1
TOTAL			133	

4.2(a) the four target objects corresponding to the four test cases are shown. The target objects are selected (rather than being chosen at random) as being ‘reasonable’ subjects for an enquiry as to location. The following rules are applied to the selection:

1. A dominant object, such as the ‘table’ in a table top scene will not be chosen as a target object as an enquiry as to its location (in the context of the scene) is nonsensical.
2. Ambiguous objects (objects of which there is more than one instance, such as one of a row of street lamps) are not chosen as targets as this would lead to confusion between the need for a referring expression and a locative expression.
3. An object is not chosen twice as a target (within a given scene set) unless it has ‘moved’.

Figure 4.7(b) in section 4.4 shows that a wide range of cases have been covered from those in which a single reference was clearly superior to those in which seven or eight different references were selected by a group of 20 human participants. In practice most objects in a scene that are not dominant and not ambiguous are chosen as targets at some point. An alternative strategy, of using all ‘valid’ objects in each scene as targets, would have led to unequal duplication of cases between scenes with more or less objects in them and a bias in training towards describing the locations of more frequently occurring objects (such as houses, cars or books, which tend to appear multiple times, often in regular geometric arrangements).

## 4.3 Scene construction

### 4.3.1 Scene representation

Given the requirement to represent reality as nearly as possible it is clear that only a 3-dimensional representation of the scenes will be adequate and that therefore a ‘vector’ representation will be the only practical solution. No significant thought was given to other representations such as 3-dimensional bit maps. The ubiquity of OpenGL led to its adoption and although to an extent this affects the way the vertex data is generated for the corpus scenes there would not be a high cost in transferring to other graphics interfaces.

The full scene file listings are given in appendix C.1 (on attached disc) but the structure of a file is shown in figure 4.3. As with all data files used in the study it is an xml file and its main constituent parts are as follows:

1. A list of objects which make up the scene (each object between tags `<object>`, `</object>`), the way these are defined and handled is described in section 4.3.2.
2. Animation information (between tags `<animationVectors>`, `</animationVectors>`) is not currently implemented.
3. The description section (between tags `<descriptionStrings>`, `</descriptionStrings>`) which contains descriptions of target objects in terms of reference objects and prepositions. These define the test cases, at present they are not in the form of complete sentences but simply defined by the name of the target object, the name of the reference object and the associated preposition, hence a ‘pre-verbal’ message. In future natural language descriptions or descriptions in languages other than English could be added to the scenes. The two possible sources of description strings are contained in this section as follows:
  - (a) All scenes contain test cases defined by the author (between tags `<preVerbal>`, `</preVerbal>`). Up to three ‘good’ references are defined for each target (test case), with each being given an associated preposition and a rating, although this rating data is not used in training the machine learning system at present. The machine models are given all of the author’s chosen references (or the top three references chosen by the validation participants) as ‘equally good’ references. The top rated reference given by the author is the only one used to compare with the validation group choice to decide whether the author agrees with the group choice (see section 4.5). This puts the author in the same position as the validators who are only able to choose one reference.
  - (b) If the scene has been validated by a group of human participants the ‘validation section’ (between tags `<validationSection>`, `</validationSection>`) contains the references and prepositions chosen by the validators, in the same pre-verbal format as for the author’s descriptions. In this case, for each reference chosen, the ‘rating’ is the fraction of validators who chose the reference. As noted

```

<?xml version="1.0" encoding="UTF-8"?>
<scene version="1.1" numFrames="1">
  <!--list of objects in the scene-->
  <object name="church" class="COMPOSITE" type="SOLID_CHURCH">
    <dimensions length="12" width="6" height="10" spireHeight="15"/>
    <colours wall="GREY3" roof="RED"/>    <position x="-3" y="0" z="-60"/>
    <orientation pitch="90" roll="0" yaw="0"/> <conventionalFront x="1" y="0" z="0"/>
  </object>
  <!-- .....further arbitrary number of objects.....-->

  <!-- For each moving object, for each frame the object's position and normal vector-->
  <animationVectors></animationVectors>

  <descriptionStrings> <!-- annotation strings for target objects (test cases) -->
    <preVerbal located="flower bed"> <!-- target object-->
      <locator reference="bench (s)" preposition="in front" rating="0.6"/>
      <locator reference="tree" preposition="left" rating="0.4"/>
    </preVerbal>
    <!-- .....further target (located) objects with up to 3 `good' references each-->

    <!-- references given by human subjects in a validation exercise-->
    <validationSection runDate="4, 2010 1 13 11:17:46 GMT+00:00">
      <preVerbal caseNum=" 0" locatedObjectNum=" 10" located="flower bed">
        <locator refObjectNum=" 26" reference="bench (s)" preposition="left" rating=" 0.1"/>
        <!-- .....further arbitrary number of references if selected by validators.....-->
      </preVerbal>
    </validationSection>
  </descriptionStrings>

  <!--View point, focal point and view angle for the `camera'-->
  <displayParameters></displayParameters>

  <AnalysisSection runDate="6, 2010 1 29 11:40:22 GMT+00:00">
    <ConvexHullData></ConvexHullData> <!-- Convex hull volume for all objects-->
    <!--For each frame the topological relationships between all objects-->
    <!--bounding box, convex hull and tight space topologies-->
    <Topology></Topology>
    <!-- For each frame the proximal vector between this object and all other scene objects -->
    <Geometry LocatedObjectNumber="10" LocatedObjectName="flower bed"> </Geometry>
    <!--.....Geometry for all other target objects.....-->

    <PerceptionData>
      <!-- For each frame, for each object, its viewability -->
      <ObjectViewability></ObjectViewability>
      <!-- Saliency of all other scene objects with respect to this target object (5 measures)-->
      <SaliencyMeasures LocatedObjectName="flower bed"> </SaliencyMeasures>
      <!-- .....Saliency measures for all other target objects.....-->
    </PerceptionData>
  </AnalysisSection>
</scene>

```

Figure 4.3: Structure of a scene file

this rating information is not used in the current experiments. The method by which the validation data is collected from the human participants is described in section 4.4.3. The author selected references from the same scene display as the validation participants but entered the reference choices directly into the scene xml file rather than selecting from a list of objects and having the choice machine transcribed. Although the process is not identical it should have no influence on the data generated. As an example figure 4.11 shows two scenes with the authors choices of reference (with ratings) and the validation participants reference choices. All of the validation participants choices of reference for all of the experimental scenes, and the corresponding choices of the author, can be seen in appendix B.

4. The analysis section (between tags `<analysisSection>`, `</analysisSection>`) contains variables computed from the object vertex lists. This is simply to remove repetitive calculation for high computational load routines such as convex hull derivation and ray casting (for the salience measures). Each time the scene is examined by a machine model this data is available without requiring calculation. If the scene's object list is edited the data in the analysis section must be recalculated.

An important point to note is that the scene is entirely defined by its objects. The scene bounding box is the bounding box of the aggregation of its objects. Only objects or parts of objects can be named entities. This leads immediately to various divergences between human language use and the interpretations available to the machine learning system in this study. For instance an aperture in a wall representing a window or a door cannot be named as such. A window or door must be explicitly provided and named if required. The case of what a human might mean by a 'street' for example, is even more complex. In this study the 'street' would strictly refer to the object defined to represent the tarmac surface, in human usage a 'street' may mean the volume enclosed by the buildings on either side, and to a certain extent the buildings themselves, as well as the carriage-way.

Theoretically an arbitrary number of objects can make up a scene, however there are various reasons in practice for limiting the number:

1. The effort required to construct the scenes is more or less proportional to the number of objects in the scene.
2. The computational load increases with the number of objects in the scene. This increase is linear in the number of objects for training the machine learning system with pre-calculated variables, but the time to pre-calculate the variables is  $O(N^2V^2)$  where  $N$  is the number of objects and  $V$  is the number of vertices in an object.
3. Scenes with too many objects are difficult for human validators to 'read' given limited resolution computer screens and the simple nature of the object rendering and lighting. The time taken to validate a scene increases and the number of scenes each validator can reliably annotate reduces, limiting the amount of test data.

In practice the average number of objects in a scene is 27.5 with the maximum number being 42 (a busy street) and the minimum 10 (a sparse desk-top) This is the number of ‘top level’ objects which, in the current study, are considered as candidate reference objects. The objects themselves can be constructed from multiple named parts (see section 4.3.2 but parts of objects (as in, for instance, “the mouse was by the table leg”) are not used in the current study. Since the objects omitted from the representations tend to be small or largely redundant for choosing references (an easy simplification is to omit 4 books from a row of 12 for example), it is felt that the resultant scenes still contain all the major candidates for reference object choice and present a realistic task for the human validators and the machine learning system.

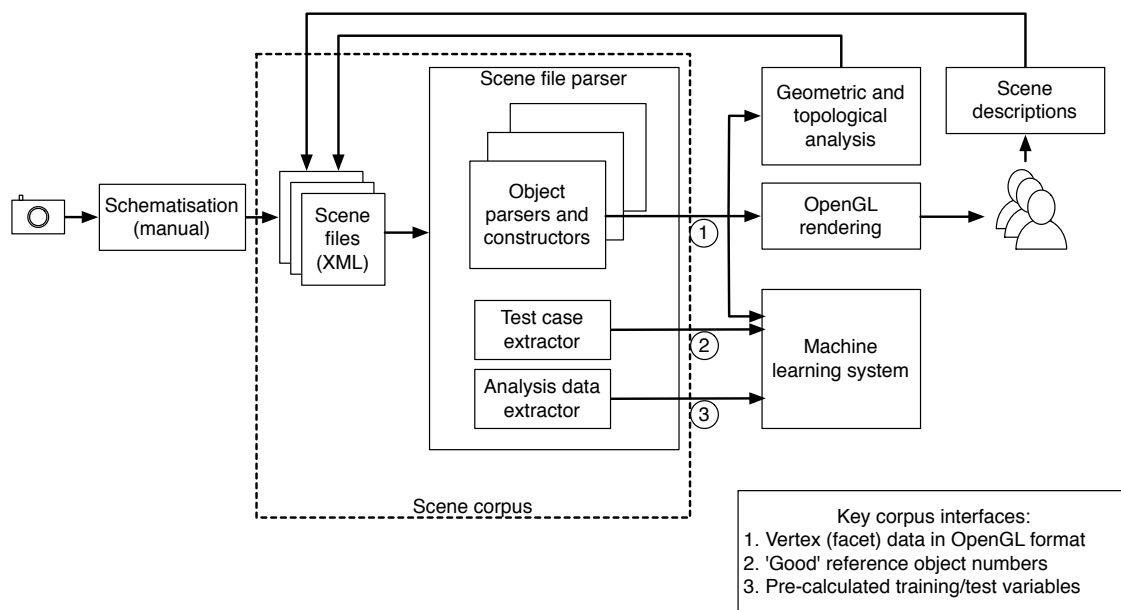


Figure 4.4: Structure and interfaces of the scene corpus

### 4.3.2 Object schematisation

In order to be realistic scenarios for the generation of spatial language the scenes must be composed of objects which are more than the abstract shapes of Regier [1996] or Roy [2002]. To be suitable for reference object selection the objects must be recognisable and have their major geometric features present; volume, extension and concavity (in particular for containers) are all likely to be important considerations. They certainly cannot be tested for their importance if they are absent. Hence map based or 2-dimensional schematisations such as used by Abella and Kender [1999] will not be adequate. As noted the data derived from 3-dimensional geographic information systems will not be sufficient due to its limited number of object types and scales. However the computational cost of using over-complex object representations needs to be taken in to account.

The definition of an object in the scene corpus can be seen in figure 4.3. Objects are



constructed recursively from other objects starting with a base set of primitive objects for which the vertex positions and facet organisations are explicitly calculated in openGL format. This gives rise to three classes of object (the ‘class’ attribute of the ‘object’ element in figure 4.3 as follows:

1. ‘PRIMITIVE’. Simple solids for which the vertex positions and facets can be easily derived from geometric calculations, such as prisms, cones and spheres. There are 10 types of primitive object and they are listed in appendix A.1. The list of primitive objects was not pre-assembled but developed as required and so may not contain all the expected primitives for a general purpose drawing application.
2. ‘COMPOSITE’ Objects built from primitive objects and/or other COMPOSITE objects. This includes most of the frequently occurring objects such as tables, houses, cars.
3. ‘USER DEFINED’ These objects are assembled from groups of objects in the scene file which may themselves be composite or primitive objects. This helps to simplify the construction of single objects which are rarely seen or are specific to a scene. A wall with a dog-leg in for example is easily constructed from three rectangular prisms. It would be tiresome to create the construction code for a specific composite object for this purpose, but also linguistically odd to have three separate objects all described as a ‘wall’. A user-defined object allows these to be grouped as a single object called ‘wall’.

It is important to note that, resulting from the object construction method, all objects in a scene are solid; that is they have volume. How this affects the machine model representation of objects that have undefined volume (such as a ‘road’) or objects that have a volume but are perceived as largely 2-dimensional (such as a piece of paper or a picture) is discussed in section 7.2.2. In the scene corpus thin objects such as sheets of paper are simply given thin dimensions, objects such as roads are given a ‘depth’ related to a typical surface texture depth. So a tarmac road might have a ‘roughness of perhaps 2 centimetres and this would be used as its third dimension after its width and length.

Given this object construction all object manipulation code (object rotation, translation etc.) and related geometric calculations (proximity, collision etc.) can be defined as recursive routines. An object defined in the scene file is structured as a tree with primitive objects at the ends of all branches. The limited depth of object nesting prevents any of the stacking overheads of recursive routines becoming problematic.

The ‘type’ attribute of an object in figure 4.3 calls a specific routine to construct the object using the parameters in the body of the object definition. There are 61 different object types in the corpus at this point. The parameters are largely self explanatory defining the dimensions, colour, position and orientation of the objects. The ‘conventional front’ parameter is used for a limited number of objects that have an intrinsic reference frame denoted not by the object itself but its position relative to other objects in a scene.

Hence a house in any orientation would tend to have as its front the side which is facing the street. Each object also has a vector defining its canonical orientation with respect to gravity but this has not been used in the study so far.

The number of ‘named’ object types (the ‘name’ attribute of the ‘object’ element in figure 4.3) is not limited by the number of primitive and composite object definitions. For example a ‘pen’ and a “post box’ are both simple cylinders, a sofa is simply a ‘soft chair’ extended in width. There are 141 different named objects (excluding those that are differentiated by addition of an adjective) in the scene corpus at this point. These are listed in appendix A.

## 4.4 Data set validation experiments

### 4.4.1 Requirements

The requirement for the data set validation experiments is to obtain as much useful information as possible within the limited resources of the study. Useful information in this context can be defined as: “reliable examples of human spatial language generation from the test data set, with particular emphasis on reference choice”

The validation experiment needs to present scenes from the corpus to a human participant and record answers to a “where is the ⟨target object⟩?” question as given by the participant. Broadly speaking two alternative routes to obtaining the data were available; an open web based experiment, or a more observable but closed experiment on a ‘local’ computer. The advantage of the web based approach would have been that it could potentially yield more data, the disadvantages were that it would have taken more time to create the experimental platform (given the author’s skill set) and more time to process the data, particularly if this had to be filtered in some way to remove suspect data. The decision was taken to use participants in a local experiment where they could be observed, at least to the degree that they could be judged to have understood the purpose of the experiment, and to be taking it seriously. Whether this was the correct decision is still an open question. The majority of the test cases in the corpus have not yet been annotated with reference object and preposition choices supplied by independent participants and rely on the author’s opinion. (The term ‘annotate’ is used from here on as a shorthand for ‘supply a reference object and preposition choice’.) However it can at least be said that a future web based experiment will have a more formal set of results for comparison and the amount of data collected, some 1600 opinions on reference choice, is comparable with, and in most cases considerably more than, similar studies (see for instance Abella and Kender [1999], Roy [2002], Kelleher [2003]). What has been achieved however is a body of independent opinion which allows a good comparison with the author’s choices to check for any bias.

Two separate validation experiments were carried out on the scene corpus. The first, described in this section simply collected 800 opinions on reference and preposition choice from 40 participants. The second used a selection of paired scenes to try to establish whether the presence of a ‘listener’ in the scene made any discernible difference to reference

choice and is described in section 4.6. Note that the first validation experiment was carried out on the series 1 scenes only (see table 4.1) but the second validation experiment used some series 1 scenes and some series 2 scenes.

#### 4.4.2 Experiment format

It can be seen from figure 4.7b that, on average, annotation of a test case (providing an opinion as to a suitable reference object and preposition) takes 33 seconds. To annotate the entire test data set would take 4.9 hours non-stop work. In reality the task could not have been undertaken in a single sitting at that throughput rate. Within the resources available to the project, annotation of the entire test set by a sufficient sample of volunteers (ideally at least 20), was not possible. Informal experiments suggested that annotation of 30 test cases by a volunteer in a single session was leading to complaints about the length of the task and possibly to errors from Corrected lack of concentration. Each volunteer was therefore asked to annotate 20 test cases.

Given a group of 40 volunteers it would therefore be possible to annotate each test case twice (almost) or 40 of the test cases (just under 10%) 20 times, with the other 90% of cases only having the author's annotation. The advantage in the first approach, of having one or two independent reference choices for all test cases, was outweighed by the fact that no moderated assessment could be made of whether these individual choices were valid. Since some of the responses are, inevitably, spurious (highly idiosyncratic), or errors (not what the participant intended) and only the author's judgement as to the validity of these could be used in practice, this seems little better than having the author's sole opinion in the first instance. If no moderation by the author is used some spurious or erroneous reference choices are likely to be embedded in the training data. The approach of having 10% of the test cases annotated 20 times was chosen, as useful information is obtained on the distribution of reference choices by a group of participants and their agreement (or lack of it) on what constituted a good reference. This approach also allows spurious or erroneous reference choices to be rejected by the 'majority' judgement (without recourse to the author's judgement of validity) and allows the author's opinions to be tested against the majority judgement in a significant subset of the test cases.

The scenes given to the participants were chosen from the corpus by the following method. About 20% of scenes (typically two from a set of ten) were chosen at random from each subject set (without replacement) to give good coverage of different scene scales and subjects. From each scene chosen one of the four test cases (target objects) was chosen at random. If the test case was not suitable for presentation to a human participant the next test case in sequence was used. (The only reason for unsuitability is complete or near complete occlusion of the target object, making it difficult for a human to identify.). This process was performed twice to obtain two groups of twenty scenes and test cases. The test cases in each set were different except for a deliberate overlap of two test cases, which could be used to provide comparative data between the groups of test participants.

### 4.4.3 Process

The instructions sheet presented to the volunteers is shown in figure 4.5. It gives an outline of the research and the reasons for the volunteer’s participation. For the 20 cases the participant was shown a scene such as that in Figure 4.6 which asked the question “Where is the ⟨target⟩?” (the ‘knife’ in the figure shown). They then had to choose a reference object and preposition from drop down lists to complete a simple locative phrase of the form “The ⟨target⟩ is ⟨preposition⟩ the ⟨reference⟩”. Note that the reference objects are not tied to any particular preposition or set of prepositions. Any of the 24 prepositions could be chosen with any object in the scene giving, on average, about 540 possible locative expressions. The reference object drop down list was always above that for the preposition but there was no forced order of selection for the reference or preposition, the majority of participants, though not all, appeared to choose the reference object first.

The reference objects in the list were not in a random order which would make the selection of reference objects from the list a much harder task for the participants. Instead the objects in the list appeared as far as possible, clustered as they were in the scenes, with for instance, all the vehicles in one part of a street adjacent on the list. The entry point to the list in the drop down box is chosen randomly, however there is a possibility that this list ordering might lead to bias in object selection. Looking at the distribution of reference choices against their position in the list however suggests that this is not a problem. In table 4.2 the decile position of reference choices (with respect to the total number of objects in the particular scene) is shown. The correlation between the list position of the reference, and the number of times a reference in that position is chosen, is insignificant ( $PMCC = 0.02$ ,  $p = 0.47$ ).

Table 4.2: Reference choice distribution over position of chosen reference in presented list

	Reference choice distribution									
List position decile	1	2	3	4	5	6	7	8	9	10
Reference choices (%)	10.2	8.7	9.5	11.7	10.5	15.5	2.9	7.6	11.2	12.2

The target object and selected reference were highlighted in the scene being shown. The target object blinked red and the object in the scene corresponding to the object selected in the drop down list of potential references blinked blue. All objects in the scene are potential references and can be selected from the list. Identical objects in the scene (for instance three identical houses) appear in the list grouped together to form a separate reference as in “the bus stop is in front of the house (s)”. This is in addition to the individual houses being in the list as candidate references in their own right. No other object aggregations were available for selection (see section 4.5.2 and section 7.2.2 for comment on this), and parts of objects (for example, table legs) were also not available.

The scenes were presented in random order on a laptop computer, the environment was not controlled in any other way. No practice examples were provided but it was possible

to review and amend answers at any point. There was no time limit imposed on the test overall or on the time taken for an individual scene. The overall time taken for the test was recorded for each participant but the times taken for individual selections were not recorded. Some qualitative guidance was provided in the instruction “There is no ‘right’ answer and we are looking for your initial thought rather than careful analysis - after all in reality you would probably take no more than a few seconds to answer the question”.

The participants were also asked to indicate their gender and age (within bands) and to state whether English was their native language. This was mainly included for diagnostic purposes should any of the results appear suspect. It should be noted that this experiment was not designed to be a psycho-linguistic experiment in its own right, although some useful information can be derived from it. The purpose of the exercise was to provide validation data for the machine learning test set and to ensure that the author’s annotations were not obviously biased or idiosyncratic.

A group of 20 volunteers from among Exeter university research staff and students and a further 20 from the authors’ acquaintance each provided opinions on one of two groups of 20 test cases. Of the volunteers 3 were acquainted with the research in more than outline form, however the non-targeted nature of the study makes it impossible for these participants to unconsciously anticipate any results. That is to say there is no specific relationship or result, (such as the dependence of reference suitability on spatial relationship, as opposed to salience, in Carlson and Hill [2009]) that is intended to be proven or disproven prior to the experiment being performed. The exact nature of the (over 20) variables used and combined in the study were not known, even to the three more familiar participants.

Of the participants 22 were male and 18 female, the age breakdown of participants is shown in table 4.3.

Table 4.3: Age breakdown of participants in the first validation exercise

Age bands						
< 14	15 – 24	25 – 34	35 – 44	45 – 54	55 – 64	> 65
0	10	10	7	5	7	1

Of the participants three were not native English speakers. Their contribution is discussed in section 4.5.2. The participants were not paid but coffee and biscuits were provided. The experimental format was presented to the college ethics committee who decided that there were no ethical issues involved.

## 4.5 Data set validation results

### 4.5.1 Selection of performance measure

The graph in Figure 4.7(a) shows the number of different reference objects chosen by the validation participants. Clearly in some scenes there are as few as two, obviously superior,

Scene Description Introduction

Start Quit

Instructions – please read!

This exercise is intended to help us understand why people describe scenes the way they do. In particular we are interested in simple responses to questions such as 'Where are the flowers?'. These questions are often answered with a phrase such as "The flowers are on the table" or "The flowers are by the stairs". In these cases the words 'on' and 'in' are prepositions and 'table' and 'stairs' are reference objects.

In the exercise a series of scenes will be presented along with a "Where is the \*\*\*?" type question. Lists of possible reference objects and prepositions are provided and you should select a reference object and preposition that you think you would be likely to use in answering the question. There is no 'right' answer and we are looking for your initial thought rather than careful analysis – after all in reality you would probably take no more than a few seconds to answer the question.

The subject of the question will blink red in the scene to help you identify it. Once you have identified the reference object you want in the scene and clicked on it in the list it will blink blue. If it isn't the object you intended (we all have different naming preferences and some objects are duplicated in the scene) you can have as many tries at identifying it as you want.

A possible 'answer' is presented once you have selected a reference object and preposition. If you are happy with it click 'OK/next' to go to the next scene. You can click back to review earlier answers. There are 20 scenes in all.

Some points to note:

1. Some objects in the list may not be visible in the scene because they are hidden behind other objects – as some objects may be in a real scene, they can still be selected if you wish.
2. If you want to use a duplicate object (for instance in a street scene there may be several streetlamps) make sure the one you want to use is highlighted, don't choose a random instance from the list.

Please indicate your age

5 – 14

15 – 24

25 – 34

35 – 44

45 – 54

55 – 64

65 up

Please indicate your gender

Male

Female

Are you a native English speaker?

Yes

No

Figure 4.5: The instruction page for the data-set validation experiment

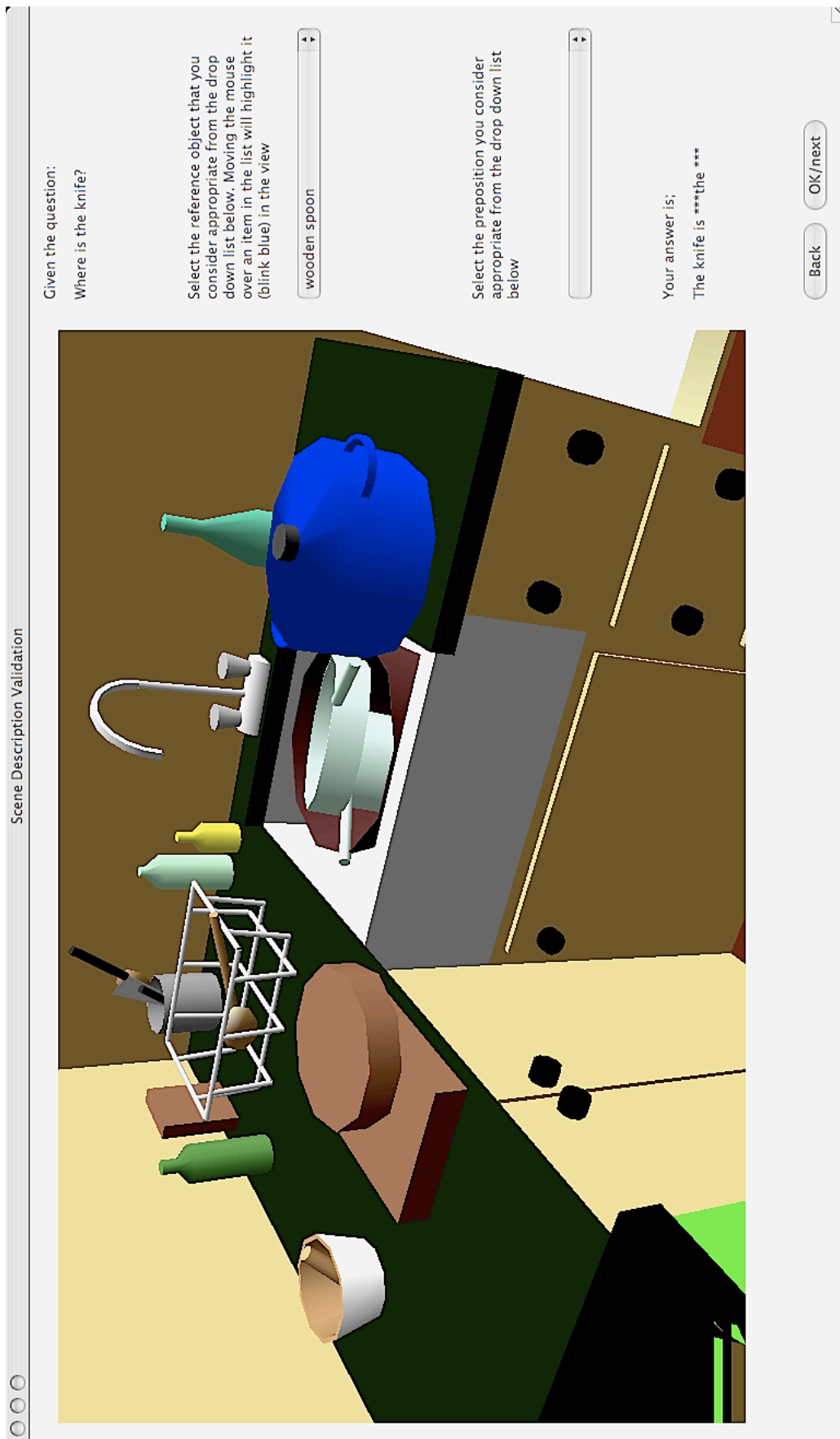


Figure 4.6: A scene display for the data-set validation experiment

references and in some scenes there are more candidates none of which is obviously superior to the others. This raises the question of whether there is a good single performance measure for reference selection across a range of scenes. Should a machine model be asked to match the most popular choice of the human group? Or one of the two or three most popular? Or a reference that is chosen by a certain minimum fraction of participants?

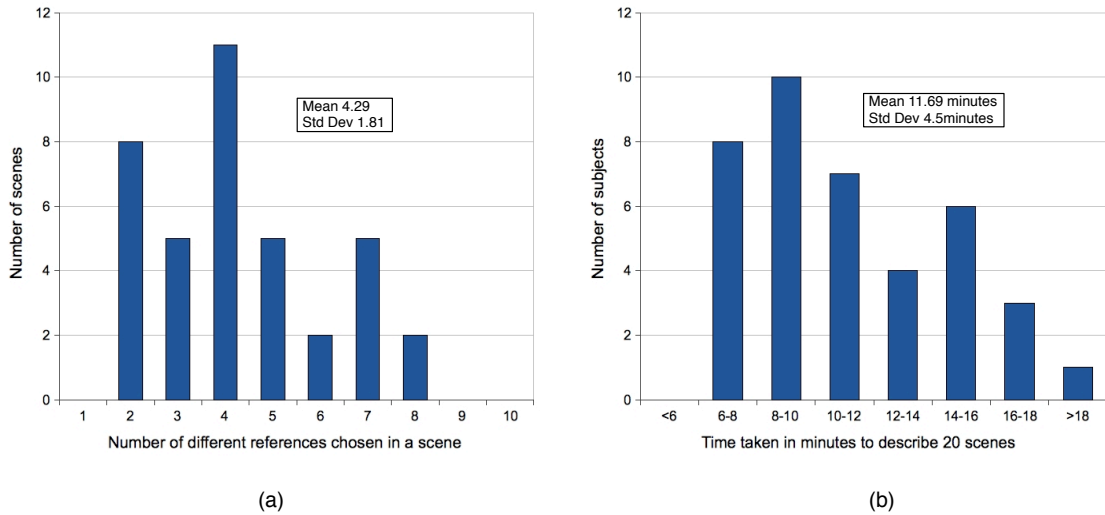


Figure 4.7: (a) Distribution of number of different references chosen in a scene, (b) Distribution of time taken to describe scenes

Figure 4.8 shows the number of times the most popular, second most popular and third most popular reference choices of the group were selected. Again it does not seem from this that there is an obvious point at which to place a single measure for whether a machine model has matched a human choice of reference. Figure 4.9 shows the correlation between the two measures of matching the top one, and matching one of the top three, choices of reference, both for the human validators and for a random selection of machine models from section 6.3. In both cases there is a significant correlation between the two measures ( $p = 0.01$ , or better, with  $df = 38$ ). This shows that for the machine models in particular the two measures are broadly equivalent and either could be used satisfactorily.

An argument for using one of the top three reference choices rather than the top choice is possible though. Given that the average number of references chosen in a scene by the human validators is  $\approx 4.5$  it would certainly not be sensible to use a number of references higher than this as a measure for the machine model performance. The average figure will contain some erroneous and idiosyncratic choices and so allowing the machine to match even one of the top four human choices might still be suspect and make discrimination between models difficult (even poor models might perform well on this measure). On the other hand asking the machine to match only the top one (or even two) human choices could mean that a model was sometimes penalised while actually producing valid reference choices. For this reason the criterion of matching one of the three most popular reference choices is principally used in this study. It should be remembered that if only one, or two,



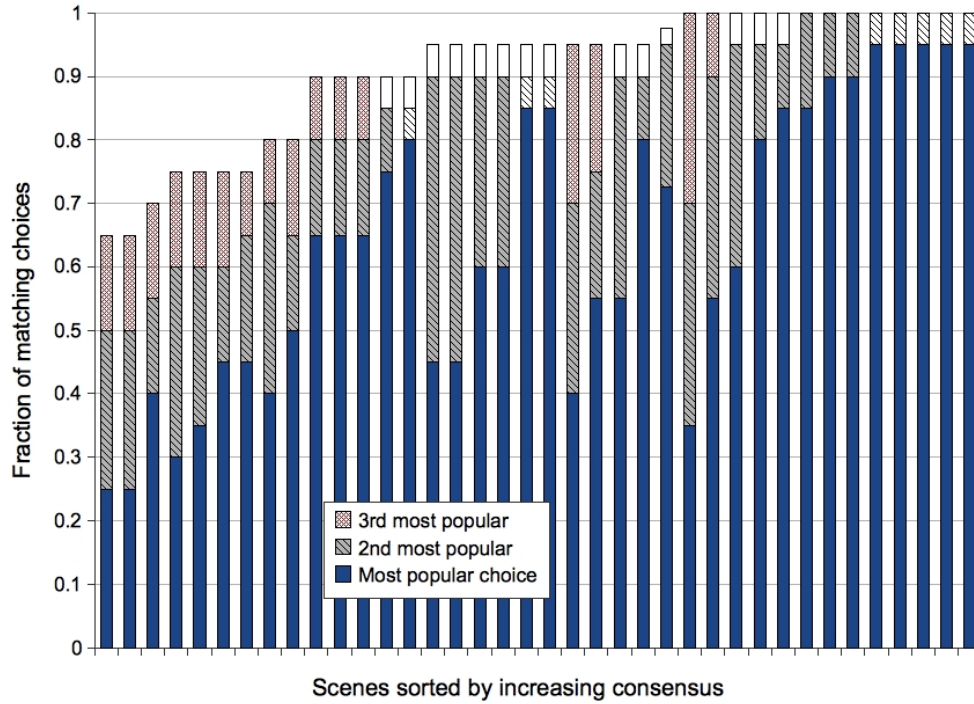


Figure 4.8: Number of participants choosing the most popular, second most popular and third most popular group choice of reference for each scene. The unshaded selections are the choice of a single participant and are excluded from the training and test data even though they were in theory one of the three most popular choices.

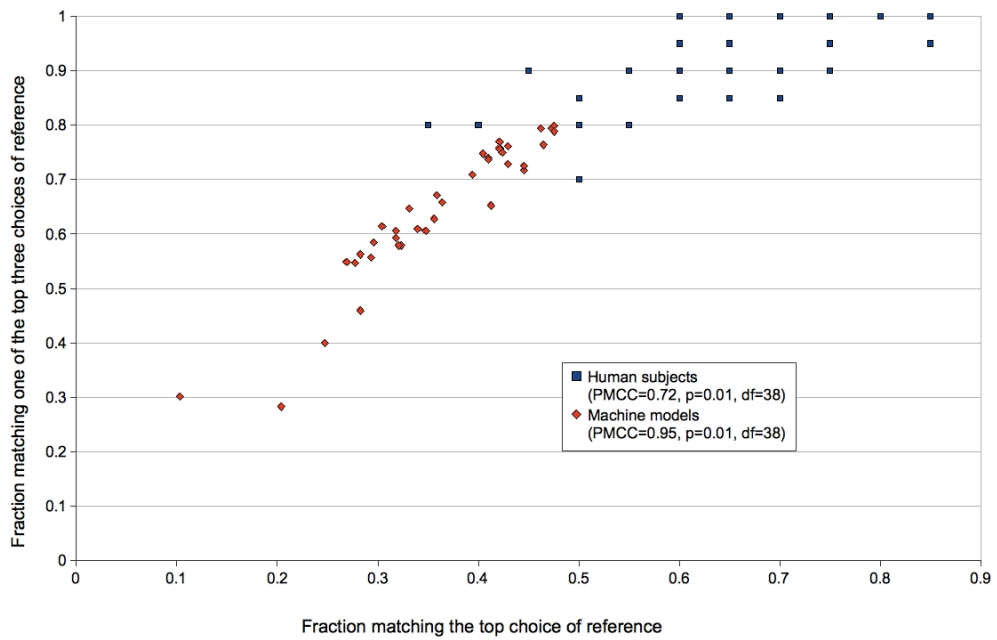


Figure 4.9: Correlation of top one and top three match measures for human validators and machine models

references are chosen by the human group, or the author, the machine must choose one of these to be deemed successful. Also, in the case where test or training data from the human group is used, more than one individual must have chosen the reference for it to be considered valid. So in this sense there is a 5% minimum on the number of validators that are required to have chosen a reference before a machine model matching it is deemed to have made an acceptable choice.

Both matching the top choice, or matching one of the top three choices, are measures of ‘conformance’ to human behaviour, and whether a conformance measure rather than some measure of ‘performance’ for the machine models is appropriate, is discussed in section 7.3.2.

#### 4.5.2 Validity of training data

The graph in figure 4.10 is the important result of the validation exercise. It shows a ranking of the validation participants by their conformity to the group as a whole, and also the place of the author in this ranking. The ‘match 1 of top 3 choices’ graph is generated by recording the three most popular reference choices for each test case and then, for each individual participant, recording the number of times they selected one of these choices. If fewer than three references are chosen by the group excluding the individual the individual’s choice is not allowed to constitute the next most popular choice. This is to say that, to be acceptable a reference must be chosen by more than one participant. This is illustrated in figure 4.11(a) where the author’s choice of the green chair as reference is not considered to be (and hence match) the third most popular choice. An individual is allowed to ‘tie break’ what would otherwise have been equally popular third choices. The match top choice trace is similar but only the most popular object for each test case is used.

The rankings are separately sorted for the cases of matching the most popular reference (i.e., the one most frequently chosen by the whole group of participants) and matching one of the top three most popular references, although as noted above there is a strong correlation between the two measures.

The match of the non-native English speakers to the group consensus is shown in figure 4.10. Although taken as a whole they are marginally below the median level of consensus there seems no reason to discard their opinions. All of them appeared to be fluent in English and several native English speakers behaved in a more idiosyncratic fashion. It should also be remembered that the exercise is principally intended to collect opinions on reference suitability, specific linguistic capabilities may be less important than visio-spatial understanding. In the extreme case, no knowledge of English would be required to select the reference object from the list in the validation exercise, the position of the object in the list alone (rather than the object name) can be linked to the highlighted candidate reference object in the scene. Clearly though this does not apply to the preposition selection.

The author’s opinions on reference suitability appear to be reasonably in line with the group as a whole matching one of the three most frequently chosen references  $\approx 95\%$  of the time. The two cases in which the author did not choose one of the three most frequently

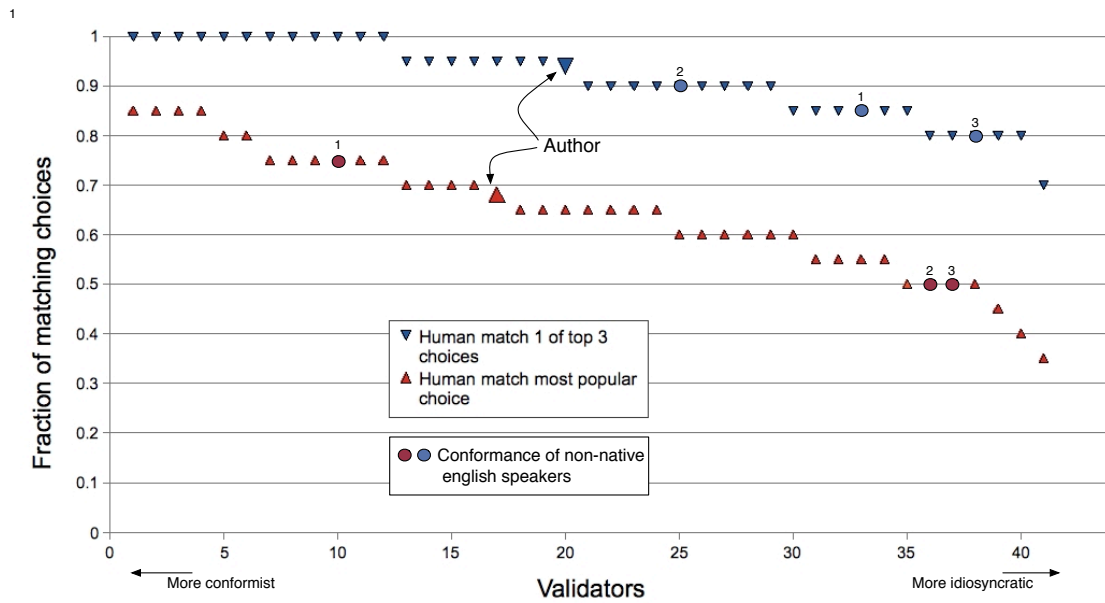


Figure 4.10: Conformity of validation participants to group choice of reference, the position of the non-native English speakers is shown

chosen references are shown in figure 4.11. In figure 4.11(a) the phrase “The orange is in front of the green chair” chosen by the author appears to be perfectly functional although the proximity of the mug appears to have swayed the majority of the other participants. In figure 4.11(b) an issue with the test data ‘representation of reality’ is highlighted. It seems probable that the most popular reference would be ‘fruit bowl’ (or ‘bowl of fruit’), however this choice of reference is not available as currently only objects with the same name (‘books’ or ‘apples’ in this scene) can be aggregated into group objects. Some participants have therefore chosen to represent ‘fruit bowl’ with the bowl and some with the fruit (‘apples’). If this is the case the bowl, as chosen by the author, is a reasonable reference even if, for the purposes of this study, it is considered unacceptable (not one of the three most popular). This is further discussed in section 7.2.2. There appear to be no significant reasons why training the machine learning system on the author’s opinions should not produce meaningful results. Nonetheless, training against one individual’s opinions must limit the likely conformity of a machine model to the group. Further validation of the test data will take place as resources allow.

### 4.5.3 Human group conformance

Quantifying the performance of human reference choice models is difficult and statistical measures of agreement do not appear to have uniformly agreed interpretation. Cohen’s Kappa is a widely used model of agreement between two judges independently assigning cases to categories. Using this measure it is possible to assess the likelihood that the median human participant’s agreement with the group could be explained as a chance happening,

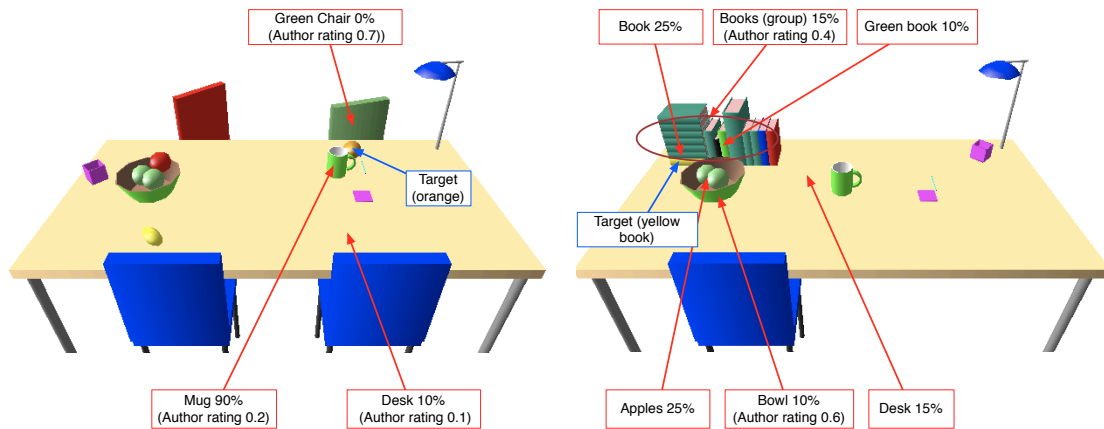


Figure 4.11: Scenes where the author disagreed with group choice of reference

at least for the match of the top choice. In this case the median human matches 65% of 20 choices (see figure 4.10) of good reference from an average of 23 references in a scene, or 460 in total. This leads to a rating matrix as shown in table 4.4. This gives  $K = 0.64$  with a standard error  $SE = 0.096$ . This allows the null hypothesis of  $K = 0$  to be rejected ( $z = 6.7$ ), but it does not tell us much about the quality of the agreement except that it is vanishingly unlikely to be due to chance (as calculated).

Table 4.4: Rating table for group and median human, match most popular choice

Group reference classification	Median human reference classification	
	Good	Poor
Good	13	7
Poor	7	433

The case for matching one of the top 3 references is more complicated as the median human has still only rated one reference in each scene as good. A lower bound on the agreement can be calculated though, using the matrix in table 4.5. Here the group has rated 60 references as good (3 in each scene) but the median human has only rated one reference in each scene as good. This gives  $K = 0.51$  with a standard error  $SE = 0.086$ . Even this lower bound allows the null hypothesis of  $K = 0$  to be rejected ( $z = 5.9$ ). In reality if the median human had been able to choose up to three references there may have been considerably more agreement.

There are various reasons for thinking this may not be a useful picture of the human consensus. Firstly many of the objects in a scene are likely to be hopelessly poor references and should not be given the same weight in the calculation as more likely candidates. Secondly the number of objects in a scene varies and the higher probability of chance agreement in scenes with low numbers of objects may be skewing the results. Another indication of consensus is given in figure 4.12. This shows the performance of the human

Table 4.5: Rating table for group and median human, match one of top three most popular choices

Group reference classification	Median human reference classification	
	Good	Poor
Good	18	42
Poor	2	398

participants compared with the same number of software ‘agents’ choosing references at random from a small selection of candidate reference objects. The number of candidate reference objects is one more than the corresponding performance measure for the humans, That is to say that software agents have four objects to choose from when the human performance level is to match one of the top three choices of the group as a whole and two objects to choose from when the humans are measured by how often they match the top choice of the group as a whole. It can be seen that, taken as a group, the human models for reference choice lead to greater consensus than a model which effectively removes unacceptable references and chooses at random from the remainder, even if there is only one ‘extra’ reference in the random process. The difference between the software agents and the humans for both cases is significant at the 0.0001 level (Students ‘t’ test, unequal variance,  $t > 4.5$   $df > 77$ ). This offers some evidence that humans are using a ‘sophisticated’ model that attempts to find the best reference rather than a simple model that merely prunes unacceptable references before making an arbitrary decision from those remaining. It also suggests that the issue of varying object numbers across scenes is not likely to be significant. Further evidence for this from the variation in the performance of human participants as the number of objects in a scene varies is shown in figure 4.13.

As can be seen there is no significant correlation between the number of objects in a scene and the propensity of the human participants to agree on the best choice of reference object. This suggests that, however cluttered a scene, there is only the possibility for a limited number of objects to be suitably positioned, and have the right characteristics, so as to be a candidate reference object for a given target. It is clear that this limited number can in some circumstances be higher than three but that the distribution will tail off above three or four (see also figure 4.7). Note that the same lack of correlation is observable in the machine models (see figure 6.14). A couple of points should be taken into account, firstly the range of object numbers is  $10 < N < 40$  and other effects may become apparent outside this range, secondly it may be true that scenes, not captured in the data-set, may contain object arrangements that allow for more good candidate references for some target objects.

It will also be the case that if the number of human participants is increased the number of references chosen will also increase, probably to include every object in a scene eventually, through error and idiosyncrasy. However there is no reason to suppose that the

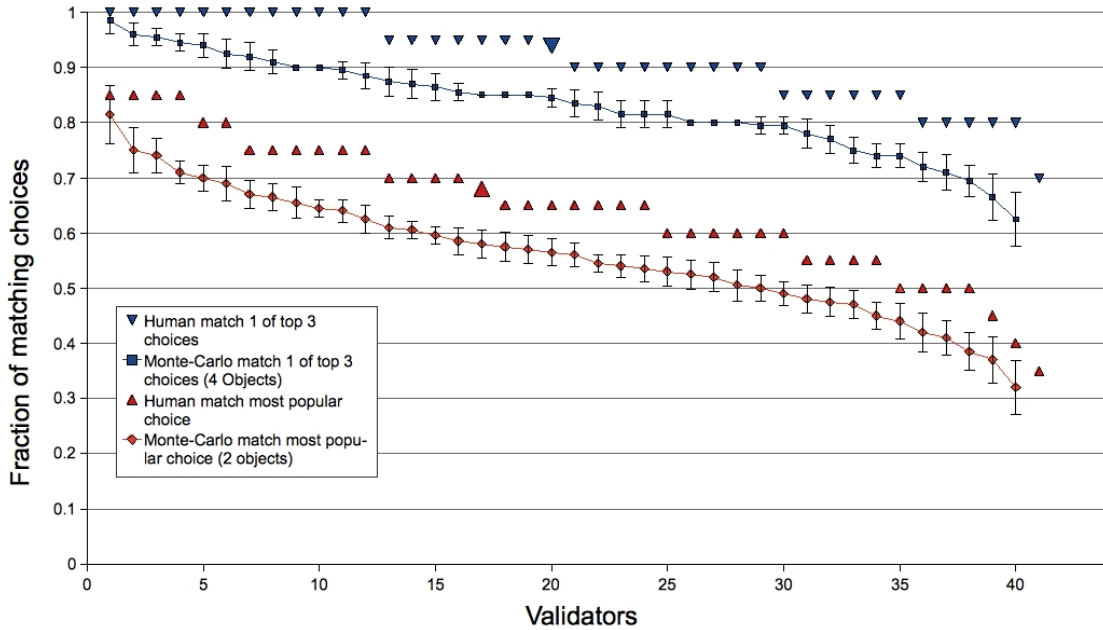


Figure 4.12: Comparison of human choices with random distribution on reduced object numbers

distribution of reference choices among the most popular candidates will vary significantly as the sample number of participants is increased from the current 20. Although the lower threshold of what constitutes an acceptable reference as chosen by a group of humans (that it be chosen by more than one participant) may have to be raised as the number of participants increases, there is no reason to expect that the (lack of) correlation between the number of objects in a scene and the group consensus will change.

There is certainly no indication from this level of correlation that the performance of human participants or of machine models needs to be adjusted to take into account a varying level of difficulty due to the varying number of objects in a scene.

It should be noted that there appear to be a few cases where the human participants have given answers that they may not have intended. No attempt has been made to remove these from the data set. The statistical learning processes used for the machine models will tend to disregard these cases, however they will be reducing the conformity among the human participants in a manner that may not be present outside of the experimental setting.

#### 4.5.4 Preposition choice data

Although the validators' choice of preposition is not used in training or for assessing the performance of the machine models it is a useful extra measure of the reliability of the validation process. If the choices of preposition were, in a significant number of cases, obviously unreasonable, or even if the overall distribution of preposition choices was different

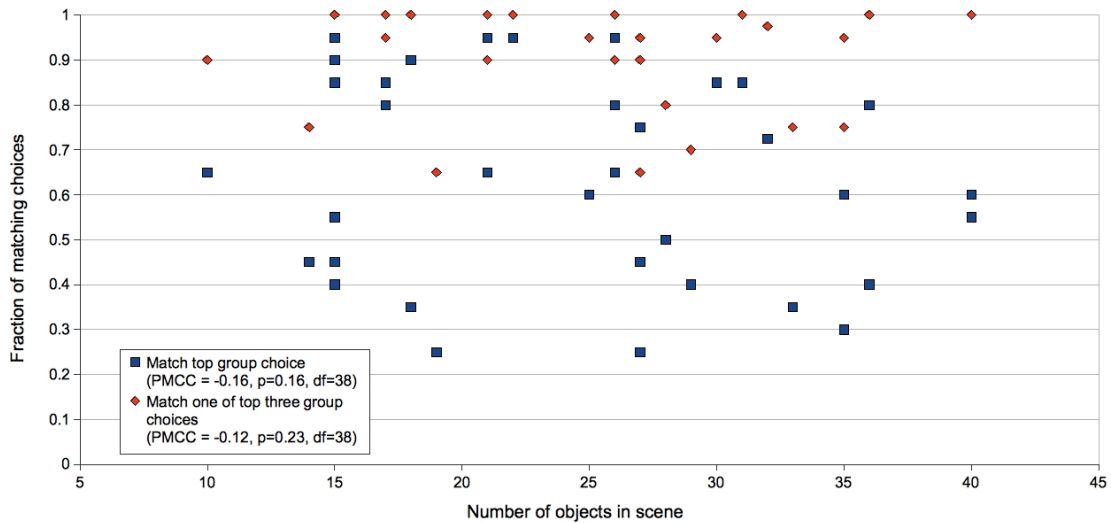


Figure 4.13: The correlation (lack of) between the number of objects in a scene and the number of participants choosing the most popular, or one of the three most popular, references

from other studies, it might suggest that something about the experimental process was leading to errors or bias.

From the validation results shown in Appendix B it can be seen that the vast majority of prepositions chosen will be reasonably effective. Some prepositions, attached to possibly erroneous reference choices, are clearly inappropriate and there are a few cases where the validator has mentally swapped the target and reference before assigning a preposition, resulting in the choice of ‘in front of’, say, when ‘behind’ would have been expected.

The distribution of preposition choices across all those available to the validators is shown in table 4.6. Also shown is the frequency of preposition use taken from the ‘Brown’ corpus, a standard database of written English from a variety of sources<sup>2</sup>. As can be seen there is little similarity between the usage distributions because there is no filtering of the metaphorical or otherwise non-spatial uses of some of the prepositions, particularly ‘to’, ‘in’, ‘by’, and ‘at’, from the usage frequencies in the Brown corpus.

The most appropriate spatial preposition usage data for comparison with the current study would appear to be that from the study by de Vega et al. [2002]. This is limited as only prepositions relating to the six cardinal directions are used (grouped together where necessary) and non-directional prepositions are neglected, except possibly for ‘on’ which is denoted ‘on top’ and assigned to the ‘above’ and ‘over’ directional group.

The preposition usage distributions in table 4.7 and plotted in figure 4.14, appear to show a much more even spread in the current study than that evidenced in de Vega et al. [2002]. In particular the usage of left and right in this study, although lower than the prepositions denoting the other axes (as expected from cognitive load, see section 3.5.1), is

<sup>2</sup>The corpus and associated documentation is available at [www ldc.upenn.edu](http://www ldc.upenn.edu) The word frequency counts used here were obtained from [www.edict.com.hk/lexiconindex](http://www.edict.com.hk/lexiconindex)

Table 4.6: Relative frequency of preposition selection; comparison between this study and the Brown English language corpus

Preposition	This study		Brown corpus	
	Usages	% of total	Usages	% of total
on	115	0.144	6742	0.151
in front of	110	0.138	221	0.005
behind	97	0.121	258	0.006
next to	90	0.113	394	0.009
under	82	0.103	707	0.016
in	67	0.084	21345	0.479
right of	60	0.075	613	0.014
by	38	0.048	5307	0.119
near	37	0.046	198	0.004
left of	33	0.043	480	0.011
above	16	0.020	296	0.007
beyond	14	0.018	175	0.004
over	12	0.015	1237	0.028
below	12	0.015	145	0.003
at	8	0.010	5377	0.121
along	3	0.004	355	0.008
opposite	2	0.003	81	0.002
around	2	0.003	561	0.013
beside	1	0.001	78	0.002
across	0	0.000	282	0.004
facing	0	0.000	282	0.004
from	0	0.000	4370	0.057
to	0	0.000	26154	0.344
with	0	0.000	370	0.005
TOTAL	800	100	76028	100

Table 4.7: Relative frequency of directional preposition selection; comparison between this study and the study of de Vega et al. [2002].

Preposition	This study		deVega (Spanish)		deVega (German)	
	Usages	% of total	Usages	% of total	Usages	% of total
above, over, on top	143	26.58	1027	48.4	769	37.17
below, beneath, under	94	17.47	375	17.67	379	18.32
in front of	110	20.45	483	22.76	632	30.55
behind	97	18.03	204	9.61	245	11.84
left of	34	6.32	19	0.9	20	0.97
right of	60	11.15	14	0.66	24	1.16
TOTALS	538	100	2122	100	2069	100

much greater than that observed by deVega et al. It is not possible to say how much of this



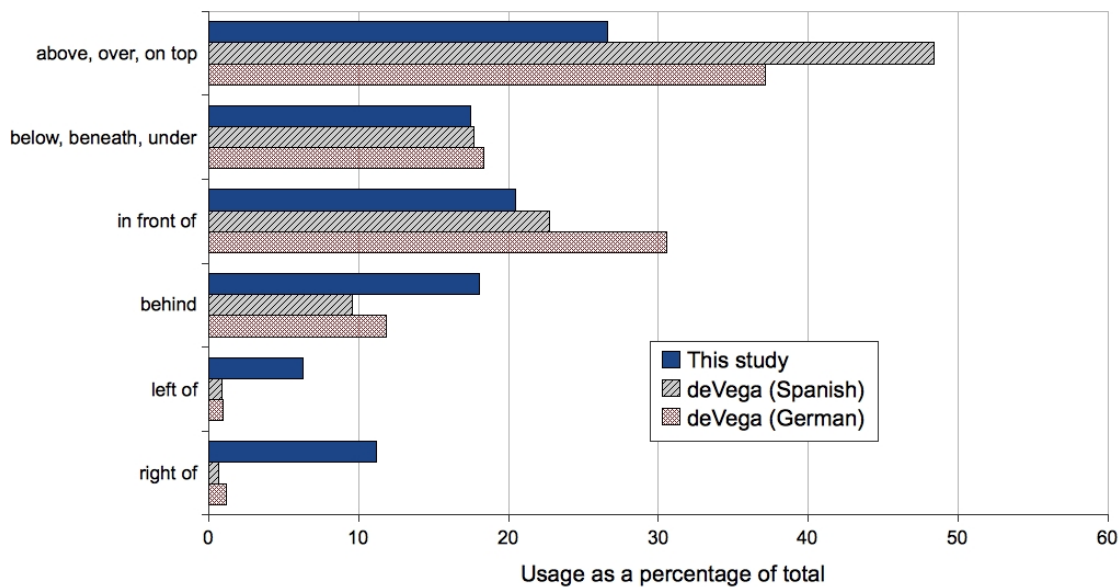


Figure 4.14: Relative frequency of directional preposition selection; comparison between this study and the corpus based study of de Vega et al. [2002].

difference might be explained by the difference in language, however the difference might be explained by the difference between what might be termed the narrative imperative, as opposed to the grounded descriptive imperative. Remembering that the preposition usage in de Vega et al. [2002] is taken from novels it might be expected that the motivation for the spatial descriptions would be different from that of the participants in this study. In particular the above/below and front/back axes contain an ordering that may have more narrative value than the left/right axis. Possibly in the same way that it is often difficult to tell whether a photograph has been reversed left to right, it is difficult to establish significance in a verbal description along this axis which suppresses the tendency to use it. In the case of a grounded, locative description the use of left and right is more pertinent, although the other axes may still be preferred because of the reduced cognitive load.

The distribution of preposition use between external and internal scenes is shown in table 4.8. Some interesting differences between the nature of the internal and external scenes are evident in the table. The most obvious difference is highlighted by the more frequent use of vertical axis prepositions in internal scenes and the compensating use of other axes, in particular the front/back axis, in external scenes. From inspection of the scenes in the figures in this chapter and in Appendix B it is clear that this distribution of preposition use reflects the likelihood of objects being organised (having spatial relationships) in the vertical axis, in internal and external scenes. This is to say there are simply more objects on top of each other in the internal scenes, books on desks, chairs under tables, pictures above fires etc. How much this is an artefact of the corpus is not possible to say, as has already been noted the external scenes do not include hills for instance, but intuitively it seems possible that this is a genuine reflection of object organisation in different environ-

Table 4.8: Preposition selection in internal and external scenes

Preposition	External		Internal	
	Usages	% of total	Usages	% of total
on	24	7.5	91	19.0
in front of	69	21.6	41	8.5
behind	48	15.0	49	10.2
next to	42	13.1	48	10.0
under	19	5.9	63	13.1
in	19	5.9	48	10.0
right of	25	7.8	35	7.3
by	16	5.0	22	4.6
near	13	4.1	24	5.0
left of	17	5.3	16	3.3
above	1	0.3	15	3.1
beyond	12	3.8	2	0.4
over	3	0.9	9	1.9
below	1	0.3	11	2.3
at	6	1.9	2	0.4
along	1	0.3	2	0.4
opposite	2	0.6	1	0.2
around	1	0.3	1	0.2
beside	1	0.3	0	0.0
TOTALS	320	100	480	100

ments. It is also possible that this is a scale issue rather than strictly an internal/external environment issue, however this is difficult to establish from the corpus in its current state (see discussion in section 7.2.1). However there is clearly a sense in which the external scenes in this corpus are more 2-dimensional than the internal scenes and the effect of this on the machine models is investigated in section 6.4.

The difference in object organisation between indoor and outdoor scenes is also reflected in the distribution of the topological preposition ‘in’ as might be expected. Also, although the number of usages is low, the prepositions ‘beyond’ and ‘at’ appear to occur more frequently in external scenes.

## 4.6 Listener present scene validation

Although the validation of a second set of scenes contributes further to the overall amount of multiply annotated training data, the purpose of this second experiment was principally to try to find out if the presence of a ‘listener’ in the scene makes any discernible difference to the choice of reference object. Each of the 133 scenes in the corpus was duplicated and a ‘listener’ figure placed in the duplicate. Effectively this creates 133 scene pairs each having one scene with a listener in and one scene without. The scene pairs used are shown in Appendix B. In each of the ‘listener present’ scenes the figure of a woman in a

black dress has been placed prominently and the participants are alerted to the fact that this figure represents a person who has addressed the “Where is the ⟨target⟩?” question to the participant. In the scenes without a listener, as used in the first experiment, the participants are left to decide for themselves the location of the person to whom they are addressing their locative expression.

The production of scenes with a listener figure in was not intended to facilitate experiments which examine the effect of single parameters or pairs of parameters on reference choice. The ‘listener’ figure was not deliberately placed at a range of distances from, or angles with respect to, a target object, but instead was placed at random in the scenes, subject to some restrictions which, from informal experiment, seemed necessary to make the scenes ‘visually sensible’. These restrictions were:

1. For table top scenes the listener is always facing the table, but can be placed at any point around the table. This is because it seemed unnatural, given that all of the objects in the scene are clustered on the table, to have the listener ‘looking into space’. It would be likely that before the locative question was asked the listener would have turned to face the table, which is a clear focus.
2. For room scale scenes there is no restriction on the orientation or position of the listener.
3. For street and vista scale scenes the listener is always facing the speaker (that is, the camera position) and is always within 5 and 15 metres from the speaker. Conversation with a listener placed outside of these parameters did not seem natural, probably because in practice it would be quite difficult. Again it would seem likely that before the locative question was asked the speaker or listener would adjust their positions. Identification of a distant listener in the scenes is also difficult for the human participants.

Given these restrictions the listener was placed at random. If this resulted in a conflict with an existing scene object the listener’s position was adjusted to the nearest point which avoided conflict.

As before the scenes for the validation experiment were chosen to ensure a coverage of all scene scales and situations. 20 pairs of scenes were chosen, each scene being paired with its corresponding scene in which the listener is present. Except for a deliberate overlap of 2 scenes (each of which has now been validated 30 times) the scenes chosen were different from those in the first validation exercise. The overlap of two scenes allows some comparison of the behaviour of the different groups of participants, no significant differences were apparent. The scenes were organised into two groups of 20 such that each group had 10 scenes with a listener present and 10 without, the other scene from the pair being in the other group of scenes.

As with the first experiment 20 volunteers from among Exeter university research staff and students and a further 20 from the authors’ acquaintance each provided opinions on one

Scene Description Introduction

Instructions – please read!

This exercise is intended to help us understand why people describe scenes the way they do. In particular we are interested in simple responses to questions such as "Where are the flowers?". these questions are often answered with a phrase such as "The flowers are 'on' the 'table'" or "The flowers are 'by' the 'stairs'". In these cases the words 'on' and 'in' are prepositions and 'table' and 'stairs' are reference objects.

In the exercise a series of scenes will be presented along with a "Where is the \*\*\*?" type question. Lists of possible reference objects and prepositions are provided and you should select a reference object and preposition that you think you would be likely to use in answering the question. There is no 'right' answer and we are looking for your initial thought rather than careful analysis – after all in reality you would probably take no more than a few seconds to answer the question.

The subject of the question will blink red in the scene to help you identify it. Once you have identified the reference object you want in the scene and clicked on it in the list it will blink blue. If it isn't the object you intended (we all have different naming preferences and some objects are duplicated in the scene) you can have as many tries at identifying it as you want.

A possible 'answer' is presented once you have selected a reference object and preposition. If you are happy with it click 'OK/next' to go to the next scene. You can click back to review earlier answers. There are 20 scenes in all.

Some points to note:

1. In some scenes a woman dressed in black with reddish hair is present. This is "Annie" the person who has asked you the "Where is the \*\*\*?" question and to whom you are addressing your answer. Annie also appears in the list of reference objects as `Annie' so you can answer (for instance) "The dog is behind you" by selecting `Annie' as the reference object.
2. Some objects in the list may not be visible in the scene because they are hidden behind other objects – as some objects may be in a real scene, they can still be selected if you wish.
3. If you want to use a duplicate object (for instance in a street scene there may be several streetlamps) make sure the one you want to use is highlighted, don't choose a random instance from the list.

Please indicate your age

5 – 14  
 15 – 24  
 25 – 34  
 35 – 44  
 45 – 54  
 55 – 64  
 65 up

Please indicate your gender

Male  
 Female

Are you a native English speaker?

Yes  
 No

Figure 4.15: The instruction page for the data-set validation experiment

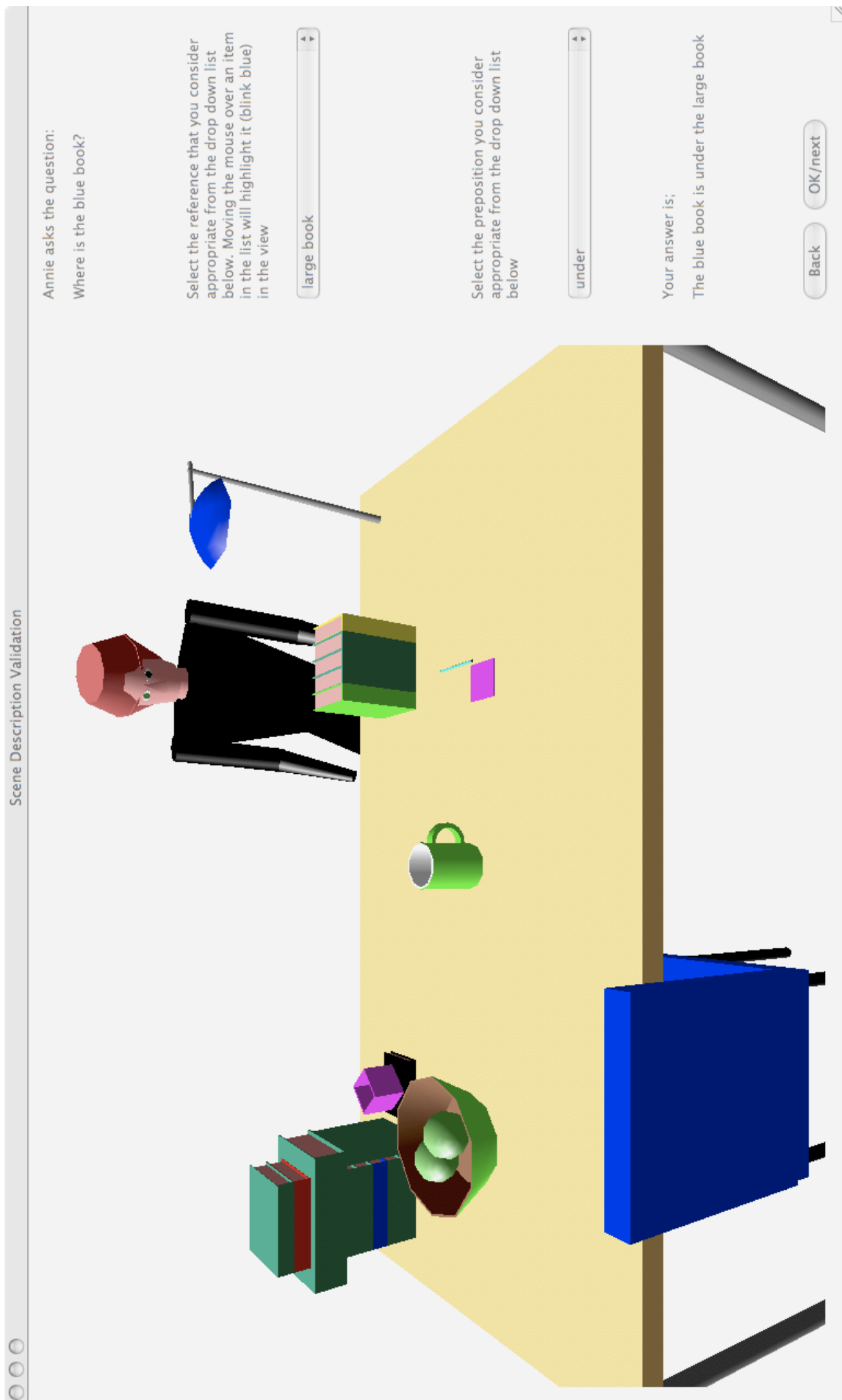


Figure 4.16: A scene display for the data-set validation experiment

of two groups of 20 test cases. The volunteers had not participated in the first experiment except for 3 and these 3 were acquainted with the research in more than outline form. As described in section 4.4.3 the nature of the task led to the conclusion that it was not necessary to exclude these three participants from the final data set.

Of the participants 21 were male and 19 female and two were not native English speakers. The age breakdown of participants is shown in table 4.9.

Table 4.9: Age breakdown of participants in the listener present validation exercise

Age bands						
< 14	15 - 24	25 - 34	35 - 44	45 - 54	55 - 64	> 65
0	11	8	5	9	4	3

The conditions were the same as for the first experiment (see section 4.4.3), except for two changes to the instructions. In figure 4.15 it can be seen that an extra instruction relating to the listener figure has been added. In figure 4.16 it can be seen that when a listener figure is present the “Where is...” question is introduced as “Annie asks the question:” If no listener figure is present the introduction reverts to “Given the question:” as used in the first experiment.

Analysis of the results showed that due to a typographical error 1 pair of scenes was not in fact a pair and so this pair has been eliminated from the results leaving 38 scenes in total. The time taken for the task is still valid and still includes all 40 scenes and the results for each scene are still valid in terms of providing training and test data for the machine learning system. The graphs in figure 4.17 show that, compared with the case when no listeners were present in the scene (see figure 4.7), the validators took longer on average to complete the task of choosing references and prepositions and chose (as a group) more references for each scene.

This indicates that the validators found the task more difficult with the listener present in the scene than the similar task without the listener. However it would be wrong to read too much in to this as it might be that the extra time was taken registering whether there was a listener present (each validator had 10 scenes with a listener present and 10 without, presented in random order) rather than in responding to the fact of the listener’s presence. Also it is possible that the choice of preposition was a higher contributor to the increased time required than the choice of reference object, which is the principal concern here.

As expected the reduction in the group consensus regarding reference suitability shown in figure 4.17 is also reflected in figure 4.20. The addition of a significant potential reference object (as discussed the listener figure is always fairly prominent in the scene to make the idea of a conversation seem plausible) might account for this without the object in itself being a cause for confusion. The number of times the listener figure was chosen as the reference is illustrated in figure 4.18. This is the number of times a locative sentence of the form “The ⟨target⟩ is ⟨preposition⟩ you” was given. The total number of times the listener was chosen (neglecting the author) is 22 from a total of 360 possible choices (each of 40

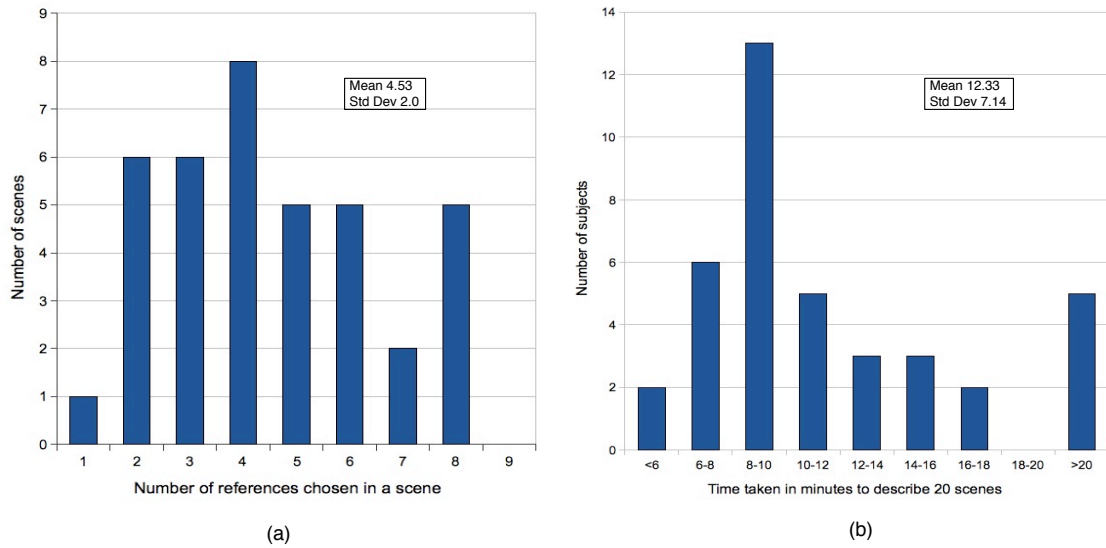


Figure 4.17: Validation results with listener present: (a) Distribution of number of different references chosen, (b) Distribution of time taken to describe scenes

validators saw 9 scenes with the listener present). This compares with the 12.8 times that would be expected for an object chosen at random, given that a scene contains about 28 objects. The non-homogeneity of the data is significant however, it is illustrated by figure 4.19(a). It can be seen that 11 of the 22 uses of the listener as a reference came from this single test case, where the listener was by far the most obvious reference. Without this case the listener is chosen slightly less often than would be expected from a random selection. It seems difficult to say anything either way about whether the listener, simply by being a prominent object, is likely to change the choice of reference.

The reduction in conformance of individual participants in this experiment with respect to the first experiment is significant, but only at the 0.05 level, in both cases. For the case of choosing the top reference  $p = 0.03$ ,  $t = 1.87$ ,  $df = 79$ , for the case of choosing one of the top three references  $p = 0.04$ ,  $t = 1.77$ ,  $df = 78$ . Student's t-test with unequal variances was used. Examination of the scenes in appendix B shows that 6 of the 21 times the listener figure was selected the listener was not one of the top three references. If the assumption is made that in these cases one of the top three references would have been chosen (instead of the listener figure that was actually chosen, or another unpopular choice) the difference in conformity between this experiment and the first would not have been significant at the 5% level ( $p = 0.09$ ,  $t = 1.53$ ,  $df = 79$ ). On its own though, this still doesn't seem to amount to clear evidence that the presence of a listener is making a significant difference to reference choice behaviour.

It could be that the presence of a listener is not making a difference in terms of the listener being chosen as a reference in a listener-relative locative expression, but is causing changes in the suitability of other objects in the scene. It is difficult to discern any pattern from the cases in figure 4.19, except as already noted in (a), where the listener herself is

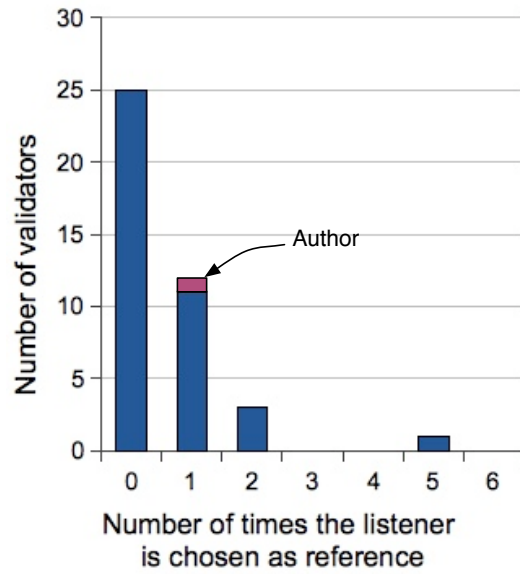


Figure 4.18: Propensity of validation participants to choose the listener as the reference object

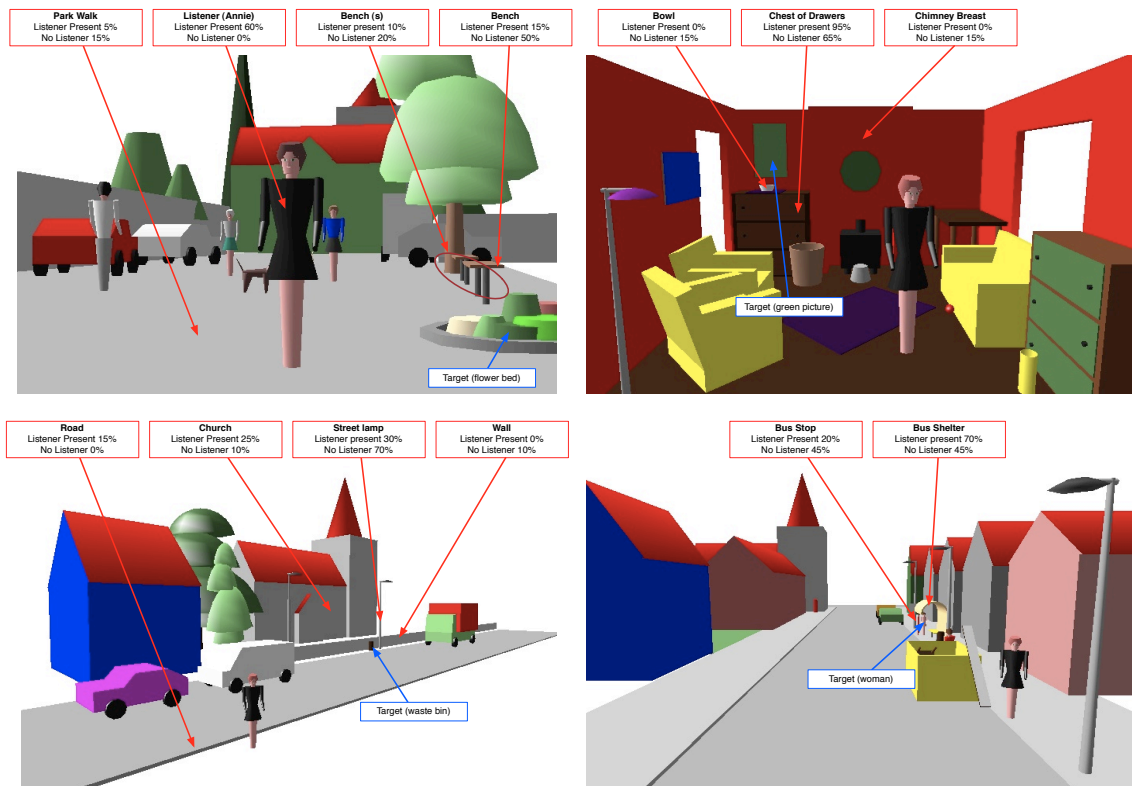


Figure 4.19: Scenes in which the presence of a listener most changes the group choice of reference



clearly the most suitable reference. To quantify the difference between listener present and no listener cases for comparison with changes within the no listener (or listener present) scenes would lead to small data sets. This has not been considered worthwhile as the likelihood of finding significant differences in the noisy data is small. What can be learned from fitting machine models to listener present and no listener cases is described in section 6.8.

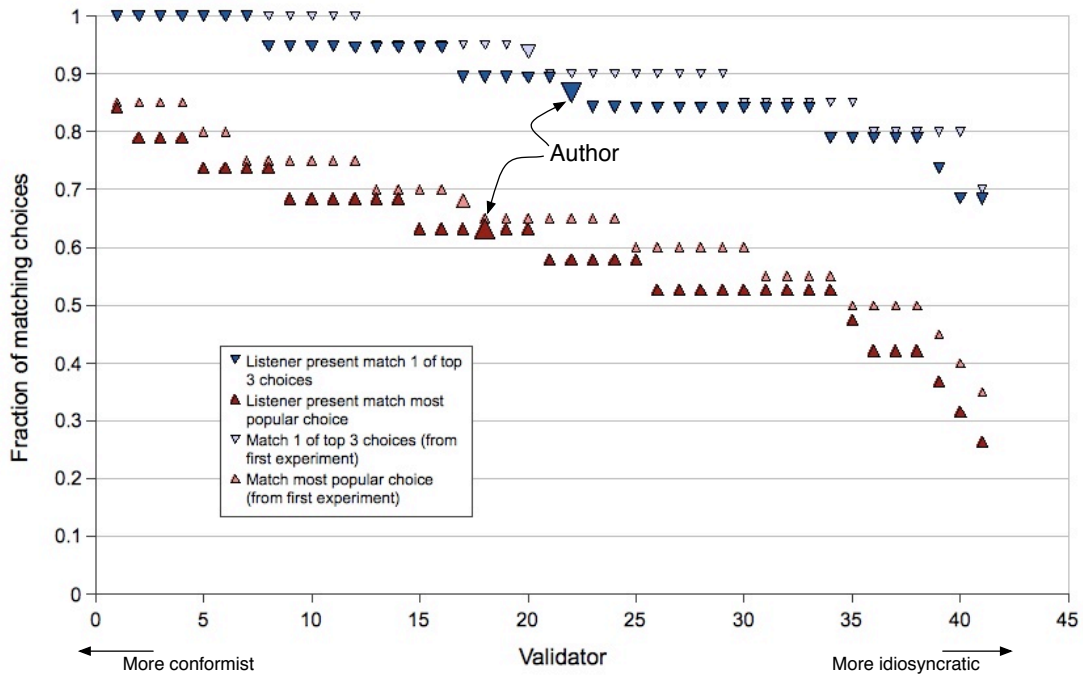


Figure 4.20: Conformity of validation participants to the group choice of reference with a listener included in the scene

Table 4.10 compares the frequency of preposition selection in the first experiment, the listener present cases in the second experiment and the no-listener cases in the second experiment. Even if there had been clearly detectable changes in reference choice with a listener present in the scene it is not clear that this would necessarily have been reflected in preposition choice. However one interesting, if not dramatic, difference in preposition use can be detected. The difference in the use of ‘behind’ between listener present and no listener cases is explained in part by the preponderance of the use of the term in listener relative expressions. The expression form, “The  $\langle$  target  $\rangle$  is behind you”, accounts for 9 of the 22 expressions using the listener as a reference, and 9 out of 11 of those not from the scene in figure 4.19. This is possibly indicative of participants using the listener as the first in what might have been (had the experiment allowed) a sequence of references. The phrase “behind you” is in a sense free as it the next part of the locative expression using a second reference can be given as the listener is turning to face the target. Note that 6 of these cases are not cases where the listener is one of the top three reference choices and

so are not given to the machine learning system. Other differences in preposition use seem insignificant.

Table 4.10: Relative frequency of preposition selection; comparison between the listener present, listener absent and test cases from the first validation exercise

Preposition	No Listener		Listener present		Experiment 1 results	
	Usages	% of total	Usages	% of total	Usages	% of total
on	40	0.11	42	0.11	115	0.14
in front of	33	0.09	30	0.08	110	0.14
behind	35	0.09	61	0.16	97	0.12
next to	53	0.14	45	0.12	90	0.11
under	39	0.1	47	0.12	82	0.1
in	32	0.08	30	0.08	67	0.08
right of	24	0.06	15	0.04	60	0.08
by	35	0.09	25	0.07	38	0.05
near	17	0.04	8	0.02	37	0.05
left of	20	0.05	16	0.04	34	0.04
above	14	0.04	24	0.06	16	0.02
beyond	3	0.01	6	0.02	14	0.02
over	5	0.01	2	0.01	12	0.02
below	3	0.01	6	0.02	12	0.02
at	6	0.02	4	0.01	8	0.01
along	1	0	1	0	3	0
opposite	0	0	1	0	2	0
around	9	0.02	0	0	2	0
beside	9	0.02	8	0.02	1	0
across	0	0	2	0.01	0	0
facing	0	0	0	0	0	0
from	0	0	0	0	0	0
to	0	0	0	0	0	0
with	2	0.01	7	0.02	0	0
TOTAL	380	100	380	100	800	100

## 4.7 Summary

The design and construction of the test data set necessarily involves compromises. In this case (and having decided that no pre-existing source of data was suitable) the time needed to produce a sufficient set of data is the dominant constraint. This determines both the amount and the realism of the test data, although the secondary consideration of computational load in processing the scenes, for both humans and machine models, has some bearing on the realism.

It has been argued that the realism is sufficient both in terms of the scene complexity and the coverage of a wide range of real world objects, scene scales and locations. There is some support for there being sufficient complexity from the lack of correlation between

the number of objects in the scene and the apparent difficulty of choosing a reference. The degree of conformity of either individual humans (or machine models), to the human group consensus does not decrease as the number of objects in a scene increases from about 10 to over 40. Sufficiency of coverage is less easy to defend but it is possible to say that this study represents a significant step forward in this respect over anything previously attempted.

The sufficiency of the amount of test data is not established in this chapter. It appears that it is sufficient for some purposes and less so for others and is variously discussed in chapters 6 and 7.

The second major issue addressed in this chapter is the validity of the training examples, given that the author, for practical reasons, needs to supply the majority of the ‘good’ reference choices that accompany the data set. Over 10% of the test cases have been provided with ‘good’ reference choices by 20 volunteers and the choices of these volunteers have been compared with the author’s choices. From this comparison it appears that the author is not detectably biased or idiosyncratic in his choices of reference, agreeing with the group choice of reference slightly more often than the median member of the group of volunteers. Although more independent opinions on reference choice would be preferable, the number (1600) of opinions obtained here is more than many comparable studies.

The collection of opinions on reference choice was not designed as a psycho-linguistic study in its own right and no direct information is derived from it relating to factors influencing reference choice. A second study looking at the possible influence of the presence of a ‘listener’ figure on reference choice, in the event did not produce any clear evidence, in its own right, that the presence of the listener makes a consistent difference.

The choice of the measure of success for humans or machine models in choosing a reference object has been covered in this chapter. This is an issue because there is no definable ‘right’ answer, although there are clearly good and bad examples of reference objects. For the purposes of this study the machine models need to produce a ‘human-like’ reference choice and so a measure of ‘conformance’ to human reference choice is used. This enables highly conforming machine models to be examined and the inference made that they contain variables and variable organisations that humans are using in reference choice. In particular the measure of ‘matching one of the top three group choices’ of reference is decided upon, although it is shown that this is closely correlated with other possible measures.

## Chapter 5

# Machine Learning Methods

### 5.1 Introduction

The intent of this chapter is to introduce and describe the machine learning system used in this study. Note that the term machine model used in the title of this thesis could be taken as encompassing both machine learned and fixed computational models. An assessment of the problem and the hypothesis model developed in chapter 3 suggested that trying to devise a fixed computational model might be time consuming and ultimately unsuccessful. On the other hand examination of a machine learned model might enable the definition of a fixed computational model if the situation proved less complex than initially thought.

This is not, as already noted, primarily a machine learning study. The intent is not to produce a system that gives the best possible results for this particular problem compared to other state of the art systems. There are always additions that could be made to a system that might improve absolute performance on a given task, not all of them can be tried and so ultimate or optimum performance is unlikely to be demonstrable.

As noted in section 1.5.3 optimum performance in the reference choice problem is not easy to define as there are no absolute right and wrong answers. The task is to match human performance, which has been defined as agreeing with the consensus of a group of human subjects. However the majority of individual humans with their own models for reference choice only agree with the group consensus some of the time (see figure 4.10). A machine model that agreed with the group consensus in 100% of cases would, at least in this characteristic of conformity, not match the models used by the majority of humans. The test for the machine model must be that it appears to be ‘a member of the group’ and some form of imitation game or indistinguishability test will be required to assess this. This is further discussed in chapter 7.

In the absence of a simple performance criterion the requirements for the machine learning system for this study are as follows:

1. To provide results sufficient to demonstrate that a machine model can simulate human behaviour in the reference choice task (i.e., it could be argued that it would pass a ‘Turing type’ test). Note that other factors as well as the machine learning system

may prevent this goal being reached.

2. To be able to capture and demonstrate the complexity of the problem and enable an assessment of whether simple or fixed computational models could provide satisfactory solutions.
3. To allow inferences to be made about the information being used by humans in the reference choice task and how this information is organised.

The initial supposition was that existing systems would be satisfactory and that adoption of a combination of ‘off the shelf’ techniques would allow the study to proceed. Examination of the reference choice problem suggested that this would not be the case and some new variations of Bayesian network structure learning methods have been devised and employed alongside standard Bayesian network learning methods.

The rest of this chapter, dealing with the development of the machine learning system used in this study, is arranged as follows;

**Section 5.2.** Briefly surveys existing machine learning techniques and gives the reasons for choosing Bayesian networks as the basis for the system

**Section 5.3.** Discusses the various types of Bayesian network currently used for classification tasks and their limitations with regard to the reference choice task.

**Section 5.4.** Outlines the use of interaction information as a feature clustering technique and discusses why this should lead to improved performance in some Bayesian classifiers.

**Section 5.5.** Describes how interaction information is applied to a practical Bayesian network construction algorithm

**Section 5.6.** Describes how the variables used in the Bayesian networks are derived from the test data set and discusses the practical limitations of these processes.

## 5.2 Choice of machine learning method

From the nature of the data in the scene corpus it is clear that the learning task is one of ‘statistical learning’. Given the different variables derived from the scene corpus the task is to decide, by looking at the distributions of variable values, which combinations of values most probably represent suitable and unsuitable references.

Although learning the reference choice task in humans might be supervised (a human listening to others has access to examples of good references and, by inference at least, bad references), there is also the possibility of unsupervised learning in that a human can assess the goodness of his own reference choice by reconstructing the search process. In the machine models in this study the learning is supervised, examples of good and bad references are available to the machine models. This is further discussed in section 7.3.2.

The task of choosing a suitable reference is clearly equivalent to a classification task for both humans and machines. A simple example of classification might be to classify wine into ‘Red’, ‘Rose’ or ‘White’ given values for the attributes of colour, opacity, tannin content

etc. In this study a scene is ‘surveyed’ and out of all the objects in the scene one is chosen as a reference to form a simple locative expression. It is clear (see section 4.5) that for the majority of scenes there is a subset of possible references which are ‘human acceptable’, rather than there being a single universally preferred reference; members of this subset can be placed in the class of ‘suitable references’ while the others are ‘unsuitable references’. In fact, for the machine learning task at hand, a ranking of potential references is produced and the highest ranked is compared to the subset of human selected alternatives that are considered to be acceptable. The task is deemed to have been successfully completed if the machine chosen reference matches one of the human acceptable subset. This is more analogous to the task of answering a “where is the ⟨target⟩?” question than the alternative of trying, for each object in the scene, to place it into the ‘suitable’ or ‘unsuitable’ classes. It also does not require a threshold of acceptability to be defined or learned.

Although boundaries between types of classifiers are sometimes blurred and in reality most practical classifiers are assemblages of techniques, each of which can be used across a range of classifiers, for the purpose of description the following list seems a reasonable division of types;

1. Neural networks
2. Decision trees
3. Bayesian networks
4. Kernel machines (‘support’ or ‘relevance’ vector machines)

A thorough investigation of all four types of classifier would not be possible here. The rationale for the choice of classifier comes from the nature of the problem and the motivation for the investigation. The issue is not one of image processing or object recognition, operating on pixel level or vector level information, for which a neural net or kernel machine would have been more appropriate. The geometry of the scenes is converted from vector level to characteristic variables (object size, minimum separation etc.) by deterministic routines which are then used as input to the machine learning system. This is similar to the approach of Regier [1996], and although he uses a neural net approach, it is clear from the limited number of parameters used, and the directed acyclic nature of the connections between them, that his ‘constrained’ neural nets could easily be substituted by Bayesian networks.

An intuitive explanation for the choice of a Bayesian network framework for the machine learning system might run as follows: the characteristic variables used are very much at the symbol level, although they are represented by numbers, the reasoning involved in selecting a reference can easily be interpreted at the symbolic level. For example “I chose the church as reference even though it is a bit large and further away because although the car is by the post box you can’t really see the post box from the street corner”. Although it can be argued that a face recognition procedure could have symbolic components (e.g., “it looked

like Mike but the ears were too small and the nose was too straight”), most systems of this sort don’t start with the characteristic features having been recognised and characterised, but with pixels or vectors.

A decision tree approach could also dig out rules of the sort expressed in the example above but the the ability to easily model latent (or hidden) variables in a Bayesian network with the possibility that these can represent composite or derived variables (or concepts) is an important factor in choosing Bayesian networks.

## 5.3 Bayesian network classification techniques

### 5.3.1 Principles of Bayesian networks

In this and following sections the notation used is outlined in table 5.1.

Table 5.1: Guide to notation

$X$	An independent variable
$CL$	In the case of a classifier the <i>Class</i> variable. As an example in this study the class variable is ‘reference suitability’ and has two possible values, ‘suitable’ and ‘not suitable’
$A_1, A_2 \dots A_n$	In the case of a classifier the attribute variables used to determine the class
$x_1, x_2, x_i$	Possible values taken by variable $X$
$x_{obs}, a^i_{obs}$	The observed value of a variable, or a set of attribute variables, in a particular case
$P(X = x_1)$	The probability that $X$ takes on one of its values
$P(X)$	The probability that $X$ takes on any of its values, that is the probability mass function, or probability density function for $X$
$P(X, Y)$	The joint probability function for $X$ and $Y$
$P(X Y)$	The probability function for $X$ conditional on the probability function of $Y$
$\alpha$	A normalising constant, typically used to ensure a set of conditional probability values sums to 1
$H(X)$	The entropy of variable $X$ . Joint and conditional entropies are expressed in the same way as probability functions
$I(X; Y)$	The mutual information between $X$ and $Y$
$I(X; Y Z)$	The conditional mutual information between $X$ and $Y$ , conditioned on $Z$
$I(X; Y; Z)$	The interaction information between $X$ , $Y$ and $Z$

The full joint probability distribution for a set of variables will allow the calculation of any probability value for the variables involved. This is easiest to illustrate in the context of discrete variables (which can take on a finite number of different values) where the joint probability distribution is represented by a probability mass function which in turn may be represented by a table such as that shown in figure 5.1. For any combination of values for the variables there is an entry in the table for the probability of this combination occurring.

The sum of all the values in the table is 1. For convenience the table can be organised so that all the variables whose values are known in a given case are on one axis and the unknown variables are on another. Given the known values, and using the product rule for conditional probabilities:

$$P(X, Y) = P(X|Y)P(Y) \quad (5.1)$$

or for the specific case in figure 5.1,

$$P(X, Y, E, F) = P(X, Y|E, F)P(E, F) \quad (5.2)$$

where for example,

$$P(E = e1, F = f2) = p9yz + p10yz + p11yz + p12yz = \frac{1}{\alpha} \quad (5.3)$$

the probabilities of the unknown variables adopting any of their values can be simply ‘read off’ the row in question, for example;

$$P(Y = y1, Z = z2|E = e1, F = f2) = \alpha \cdot p10yz \quad (5.4)$$

F	E	Y	y1	y1	y2	y2
		Z	z1	z2	z1	z2
f1	e1		p1yz	p2yz	p3yz	p4yz
f1	e2		p5yz	p6yz	p7yz	p8yz
f2	e1		p9yz	p10yz	p11yz	p12yz
f2	e2		p13yz	p14yz	p15yz	p16yz
f3	e1		p17yz	p18yz	p19yz	p20yz
f3	e2		p21yz	p22yz	p23yz	p24yz



As an example,  $P(E=e1, F=f2, Y=y1, Z=z2) = p10yz$

$$P(E=e1, F=f2) = p9yz + p10yz + p11yz + p12yz = 1/\alpha$$

$$P(Y=y1, Z=z2 | E=e1, F=f2) = \alpha \cdot p10yz$$

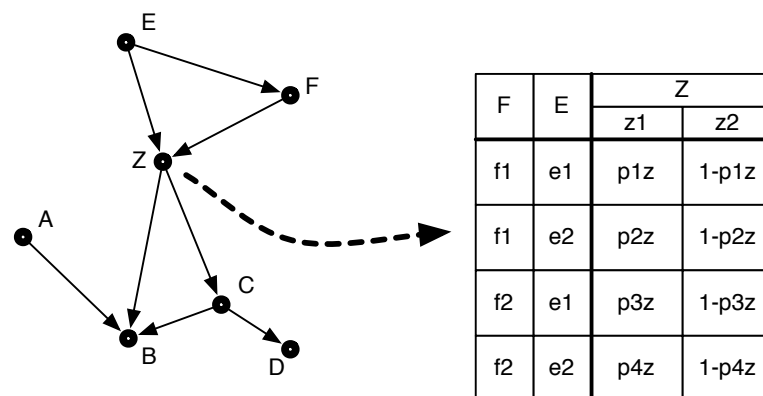
Figure 5.1: An example of a full joint probability table. The table could be extended to include any number of variables each with any number of values

Clearly the issue with this representation is that it becomes unworkably large even for a relatively small number of variables. The representation is also inefficient if some of the variables are statistically independent of each other. A Bayesian network is simply an efficient representation of the joint probability distribution between the variables in question



which takes advantage of the independence of variables where possible. The basic principle is illustrated in figure 5.2. A Bayesian network is, in general, a directed acyclic graph with nodes representing variables. A direct statistical dependence between two variables is represented by a directed arc between them. A cycle is present in the graph if it is possible to start at any of the variables in the graph and return to that variable following a path in the direction of the arrows. The presence of such a cycle would require dynamic relationships between the variables to be represented and the graph would not be considered a valid Bayesian network. As shown in figure 5.2 the network contains conditional rather than joint probability tables. The probabilities associated with a variable are those conditional on the values of its ‘parents’, that is, the variables from which it has an incoming arrow. In figure 5.2  $F$  and  $E$  are parents of  $Z$  and  $B$  and  $C$  are children of  $Z$ . Given probabilities for the values of  $E$  and  $F$ , however, the information corresponding to the joint probability table for  $E$ ,  $F$  and  $Z$  is present in the network. Similarly in figure 5.2 the joint probability table for  $A$ ,  $B$ ,  $C$  and  $Z$  is present but the joint probability table for  $A$ ,  $B$ ,  $C$ ,  $E$ ,  $F$  is not. Note that there is a practical limitation on the number of parents a variable may have. If a single variable in a network had all the other variables as parents the Bayesian network would have no computational advantage over the full joint probability table.

The network shown in figure 5.2 could be extended to contain an arbitrary number of variables and an arbitrary number of arcs, limited only by the computational feasibility of the result. In general a variable can be defined using continuous instead of discrete probability distributions, but in this study only discrete distributions are used.



As an example,  $P(Z=z1 \mid E=e1, F=f1) = p1z$

Figure 5.2: The general form of a Bayesian network.

### Bayesian Network Evaluation

Typically for a given ‘case’ some of the variables in a network will have known (observed) values and are termed ‘evidence’ variables. As with the full joint probability table the requirement is to calculate the probabilities of some or all of the remaining variables taking

on any of their values, given the evidence variables. The calculation of these probabilities from the individual conditional probability tables defined for each variable is, for most networks, NP hard (Cooper [1990]). If, however, the network structures can be restricted to ‘trees’ or ‘poly-trees’, in which there is only one path between any two variables in the network, exact calculations can be performed in polynomial time. The network shown in figure 5.3 is a poly-tree, the network shown in figure 5.6(b), for example, is not, with multiple paths connecting most of the variables.

An example of network evaluation using an algorithm known as ‘message passing’ (Pearl [1982]) can be illustrated with reference to figure 5.3. In this method variables pass their probability values to adjacent variables in the network in the direction of the arrows (down) and against the arrows (up). A variable can only pass its values if one of the following conditions holds:

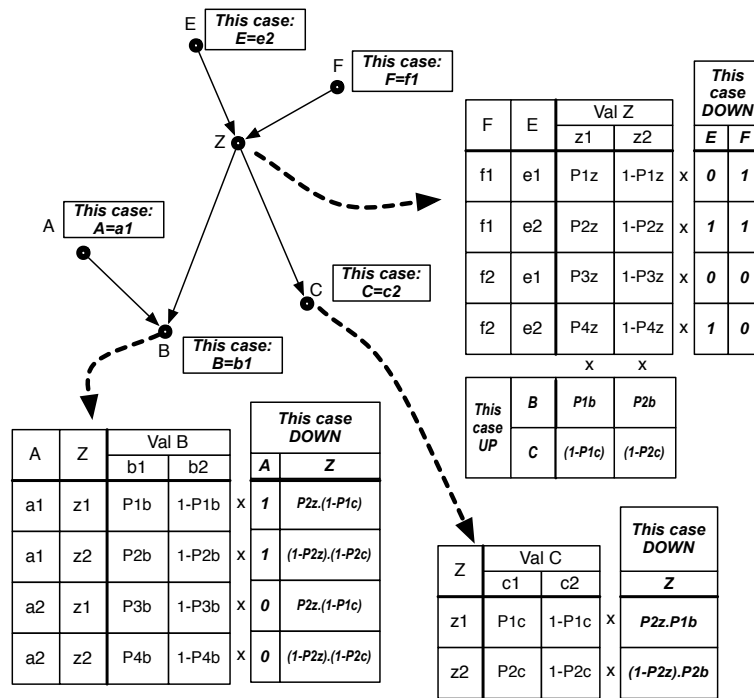
1. A variable with no parents and only one child may pass its probabilities down to its child.
2. A variable with no children and only one parent may pass its probabilities up to its parent (although unless the variable is an evidence variable these will have no effect)
3. A variable with an arbitrary number of parents and children may pass its values to all its parents and children if it has received values from all of them.
4. A variable with an arbitrary number of parents and children which has received values from all except one parent or child may pass its values to that parent or child. This is true as the values passed by a variable are not required to calculate the values to be passed to it.

The effect of this message (value) passing can be seen in figure 5.3, which extends the conditional probability tables shown in figure 5.2 to include the values passed up and down to adjacent variables in the particular case of the observed variables having the values shown. The calculation of the probability values for the unknown variable  $Z$  proceeds as follows:

From point 1 in the list above it can be seen that it is immediately possible for variables  $E$ ,  $F$  and  $A$  to pass their values down to their child variables. Note that this allows us to perform the calculation in the simple case of conditional probability  $P(Z) = P(Z|E, F)P(E, F)$  for the case where  $P(E = e_2, F = f_2) = 1$ . From point 2 it can be seen that variable  $C$  can also pass its value up to its parent (variable  $Z$ ). The explicit use of Bayes’ rule in a Bayesian network can be seen at this point. Bayes’ rule is derived from the observation that:

$$P(A, B) = P(A|B)P(B) = P(B|A)P(A) \quad (5.5)$$

from which it can be seen that:



$$P(z=z1 | A=a1, B=b1, C=c2, E=e2, F=f1) = a.P2z.P1b.(1-P1c)$$

$$P(z=z2 | A=a1, B=b1, C=c2, E=e2, F=f1) = a.(1-P2z).P2b.(1-P2c)$$

Figure 5.3: Evaluation of a Bayesian network using message (value) passing. The probability of unknown variable Z taking on either of its values is determined from the known values of the other network variables

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \quad (5.6)$$

The network of figure 5.3 does not directly contain values for  $P(Z|C)$ , however the table associated with variable  $C$  contains figures for  $P(C|Z)$ . The value that is passed up to variable  $Z$  for each  $z_i$  is actually:

$$\frac{\alpha P(Z = z_i|C = c_j)}{P(Z = z_i)} = P(C = c_j|Z = z_i) \quad (5.7)$$

The value of  $P(Z = z_i)$  required to calculate  $P(Z = z_i|C)$ , given the values of its parent variables, is contained in the table for  $Z$ . The value of  $\alpha$  which corresponds to  $P(C = c_j)$  is applied as part of an overall normalising constant once all values have been passed. Note that if there is no observation of the value of  $C$  (or in the more general case one or more of  $C$ 's descendent variables if  $C$  is not observed) then the values passed up will all be 1. This simply confirms that in the absence of information about its descendent variables there is no modification of the probabilities of the parent variable ( $Z$  in this case).

Finally from point 4 in the list above, with the values passed down from variable  $A$ , variable  $B$  can pass its probability values up to variable  $Z$  for each  $z_i$  as follows:

$$\frac{\alpha P(Z = z_i|B = b_j, A)}{P(Z = z_i)} = \sum_k P(B = b_j|A = a_k, Z = z_i)P(A = a_k) \quad (5.8)$$

Note that the value of variable  $A$  changes which of variable  $B$ 's conditional probabilities are passed to  $Z$ , even though the value of  $B$  is fixed. Following point 3 in the list above, values for variable  $Z$  could be passed down to its child variables as shown in figure 5.3 although in this case, with the values for the child variables being known, they are redundant. At this point the probability values for  $Z$  taking on any of its possible values can be calculated as shown in figure 5.3.

Although only two-valued variables are shown it can be seen that the procedure generalises to any number of values for any variable simply by extending the conditional probability tables. The algorithm extends to any complexity of poly-tree and this can be demonstrated by induction as follows:

1. For an existing poly-tree of any complexity a further variable can be added to any existing variable by a single arc and the result is a poly-tree network.
2. The added variable may immediately pass its values to the variable to which it is attached.
3. The variable to which the new variable has been added can now proceed with value passing in exactly the same way as it did before the new variable was added, that is, it does not need to wait for any further values in addition to those it previously required. The existing variable can pass its values to the new variable at any convenient point.
4. The process clearly works for a poly-tree of two variables.

Different proofs of this are given by other researchers, see for instance Neapolitan [2004]. Other algorithms are also used for exact evaluation of poly-tree networks such as variable elimination Shachter [1986] or the related symbolic probabilistic inference algorithm due to Li and D'Ambrosio [1994]. Pearl's message passing algorithm is not the most efficient, but once completed allows all unknown variables to be evaluated in a given case, rather than just evaluating a specific variable. An implementation of the message passing algorithm is used in this study as only poly-tree networks, or networks that can be converted to poly-trees, are used.

For Bayesian network structures that are not poly-trees approximation techniques can be used to evaluate unknown variables. These include Markov chain Monte Carlo algorithms (Pearl [1987]) or Gibbs sampling (Geman and Geman [1990]). Loopy (or iterative) belief propagation (Pearl [1988]) is an extension of the message passing algorithm described above in which variables pass their values without necessarily having complete information, in practice this often converges to a stable result. Alternatively the reality being modelled (presumably necessarily and accurately) by the non-poly-tree Bayesian network can be approximated by a poly-tree and the parameters can subsequently be calculated exactly. Techniques are available for exactly converting some simple network forms to poly-trees including variable combination (or clustering, see Pearl [1990]).

### Independence and conditional independence of network variables

In the example illustrated in figure 5.3 it can be seen that the known values of variables  $A$ ,  $B$ ,  $C$ ,  $E$ ,  $F$  all affect the value of the unknown variable  $Z$ . An important consideration, for a general network, is which variables will affect the value of an unknown variable given some other known variables. This is important for two reasons. Firstly if a variable can be shown to have no effect on an unknown variable whose probabilities are being calculated, that is if the two are independent of each other conditional on other variable's values being known, then the computation can be made more efficient by ignoring it. Secondly if it is expected or required that a variable should have an effect on an unknown variable then it is important that it should be placed in the network so that the unknown variable is dependent on it.

Looking again at figure 5.3 the following points can be ascertained:

1. If the value of an unknown variable's parent is known ( $P(X = x_i) = 1$  for a given  $i$ ) then no parents or children of this parent node will be able to effect the value of the unknown variable. Values passed up or down to this parent variable will not effect the probability of it having this value, which is what is passed down to the unknown variable.
2. If the value of a child variable of the unknown variable is known then no children of the child node will be able to affect the value of the unknown variable. Values passed up to this node do not effect the probability of its value *conditional on its parent's values* which is the value passed up to the unknown variable.

3. If the value of a child variable of the unknown variable is known then parents of this child variable will still be able to affect the value of the unknown variable. This is again because the probability of the child variables value *conditional on its parent's values* is what is passed up to the unknown variable. A classic example of this is illustrated in figure 5.4. If the car is observed not to start then initially there might be considered an equal possibility that the battery or the fuel pump was at fault. However if the battery is observed to be flat then the probability that the fuel pump was faulty would be assumed to be much lower. If the battery is observed to be good then the probability of the fuel pump being faulty is raised. Note that if the value of the child is not known then the parent of the child variable has no effect on the unknown variable.
4. If the value of a parent of a child of the unknown variable is known then no parents or other children of this variable can affect the value of the unknown variable. This follows from point 1.

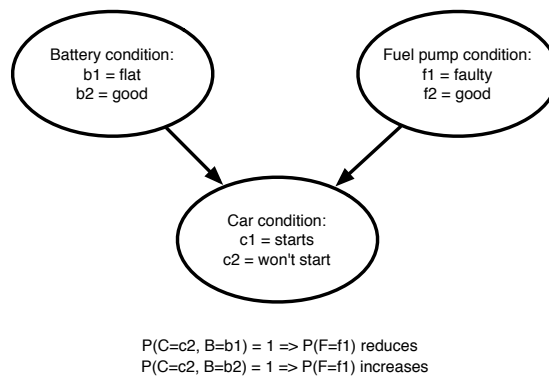


Figure 5.4: A case where the parent of a child of an unknown variable affects the probabilities of the values of the unknown variable (the condition of the fuel pump in this case)

The end result of this is shown in figure 5.5. If the parents, children and parents of children of an unknown variable are known then no other variables in a network can affect the value of the unknown variable. The parents, children and parents of children of a variable are termed the variable's Markov blanket. The unknown variable is conditionally independent of all variables outside the Markov Blanket given known values for variables in the Markov blanket.

In this study and in many Bayesian networks used for classification, the concept of the Markov blanket is essential. The classifier variable is typically the single unknown variable and the attribute variables which are used for classification are always, or nearly always, known (in many cases). It follows from this that the attribute variables should always be in the Markov blanket of the classifier variable. Given the limitation on the number of parent variables, due to the computational infeasibility of large conditional probability

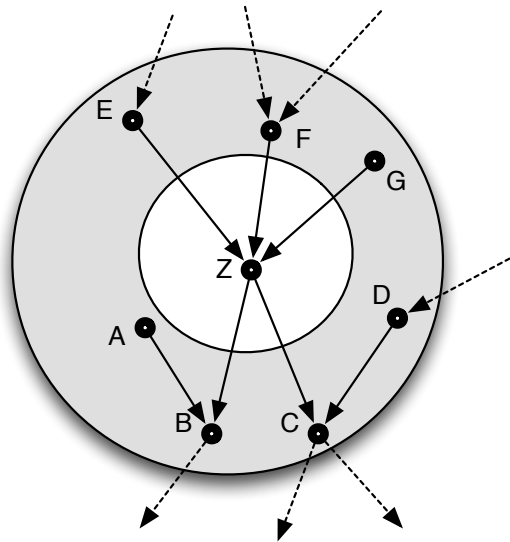


Figure 5.5: The Markov blanket of a variable in a Bayesian network. Given values for all variables in the shaded region the unknown variable  $Z$  is independent of all other variables in the network

tables noted above, it is not surprising that many Bayesian network classifiers put most or all attribute variables as children of the classifier variable. This is further discussed in section 5.3.2.

An extension of this concept, to the case of the independence of an arbitrary sub-set of variables, from a second sub-set of variables, given a third sub-set of variables in the network, is called d-separation. This is not relevant in the case of this study however.

### Learning values in a Bayesian network

The evaluation example given above assumed the values in the conditional probability tables were already known. Historically values have been assigned to Bayes net conditional probability tables from expert knowledge or opinion but, where possible, using data from actual observations is more appropriate. Learning the values in a Bayesian network given complete data is simply a matter of recording the frequencies of the different combinations of evidence values. This can also be performed locally in part of a network; whenever the value of a variable and all its parents are observed, the relevant probability value of the variable can be updated. With reference to figure 5.3, if ten observations of the values of variables  $E$ ,  $F$  and  $Z$  are made with the outcomes shown in table 5.2, then with only these observations the values of  $P_{1z}$  and  $P_{3z}$  can be set to 0.4 and 0.2 respectively. Nothing is (or needs to be) known about variables  $A$ ,  $B$  and  $C$ .

In the case of these observations nothing has been learned about  $P_{2z}$  or  $P_{4z}$  and the question arises as to what value these, or any other unobserved probabilities, should be set. This is of concern if there is a possibility of a network needing to be evaluated for some

Table 5.2: Example of Bayesian network parameter learning, observed counts of values of  $Z$  given values of  $E$  and  $F$ 

$E$	$F$	$Z = z_1$	$Z = z_2$
$e_1$	$f_1$	2	3
$e_1$	$f_2$	1	4

combination of evidence variables that have not yet been observed. Two approaches are generally taken in practice:

1. A default probability distribution is assigned that is replaced by observed values as soon as these become available. If no observed values are encountered the default values are used in network evaluation.
2. The equivalent of a set of observed values are ‘invented’ to which the actual observed values are added when they become available, the probabilities being calculated from the observed and ‘invented’ values.

In either of these approaches, although perhaps more usually in the second, the ‘counts’ of the observed values (of a two valued variable) can be used to define a ‘beta’ probability density function, which allows confidence in the resulting probability values to be assessed:

$$beta(\theta|a, b) = \theta^{a-1}(1 - \theta)^{b-1} \quad (5.9)$$

where  $\theta$  is the domain of the function ( $0 \leq \theta \leq 1$ ) and  $a$  and  $b$  are the counts of the two values of the variable. The simple example of this is the judgement of the degree to which a coin is unbiased when flipped. After only three or four flips a bias might not be ruled out, after 100 flips (hopefully) some confidence in the equal probability of heads and tails should be established. For variables with more than two distinct states (for instance a traffic light which may have different probabilities of being red, green or amber) the Dirichlet probability density function, a generalisation of the beta probability density function, is used.

A more usual reason for adopting the second approach is that of its limiting of extreme probability distributions in the case of limited data from which to learn the network parameters. If the prior assumption on the probability mass function for a variable is that all values are equally likely, then as in the case of the coin an invented set of (say) three heads and three tails can be used as a start point for the probability values. In the case of a single observed flip of the coin, the probability mass function is still not far from  $P(heads) = P(tails)$ , in contrast if no invented values were used the single flip would give a probability mass function  $P(heads) = 1$  or  $P(tails) = 1$ . This effect is sometimes referred to as ‘smoothing’ and can improve network performance in some cases (see Friedman et al. [1997] for example), however there are drawbacks to the approach. In particular, if the prior assumption is wrong, then more observations may be required to arrive at the



correct probability mass function, overcoming the initial bias of the ‘invented’ data. It further seems that if the probability mass function is skewed the number of ‘invented’ values needed to represent it will be large, further increasing any initial bias of the ‘invented’ data. For instance in this study with possibly 3 suitable references out of a total of 27 objects in a scene, the prior probability of a combination of attribute values denoting a suitable reference should be represented by a minimum of 9 invented counts, 8 for the case of unsuitable and 1 for suitable reference.

Overall the introduction of prior counts would introduce further uncertainty as well as complexity in presentation of the results and has not been included in this study. Default probability values are used in the conditional probability tables which are discarded as soon as an observed case (combination of relevant attribute values) is observed. The default probability values for the classifier variable in this study are set so that an unobserved case is less likely to indicate a good reference than a randomly chosen reference. If one of the top three references is being matched with an average of 27 objects in a scene then on average an object would be randomly chosen as a reference one in nine times. Unobserved attribute combinations have the corresponding reference suitable probability set to 0.1 (and reference unsuitable to 0.9).

Any possible improvement in the absolute performance of the system through the use of smoothing is not likely to be significant in terms of affecting the findings from the study. An analysis of the results to assess how many are dependent on cases where limited training data is present could produce useful diagnostic information and may be included in future. An indication of the effect of unobserved and rarely observed combinations of attributes is discussed in section 6.9. Section 6.4 contains an assessment of whether the test and training data set for this study is of sufficient size overall.

### **Bayesian network structure learning**

The ‘structure’ of a Bayesian network refers to the connections between the variables that make up the network, including whether variables are connected at all and are hence included in or excluded from a network. The structure of a Bayesian network as opposed to the values in the probability tables can be derived from ‘expert’ knowledge of the situation (see for instance Pearl [1988]) but machine learning techniques can also be used. Learning the structure of a general Bayesian network is sometimes attempted using search techniques although these need to be constrained as an exhaustive search in a network containing more than about 5 variables is not feasible (see Cooper and Herskovits [1992], there are for instance about 29,000 possible network structures containing 5 variables). Buntine [1991] combines expert knowledge of a domain encapsulated in an initial network structure with search techniques to arrive at revised parent variable sets for each network variable. Cooper and Herskovits [1992] use a ‘greedy’ search for the most likely parents of each variable starting with an ordered set of variables such that a variable cannot be a parent of a variable ‘higher’ in the ordering. Starting with an unconnected network, the variable which maximises the likelihood of the network ‘matching’ the data set is added as a parent

to the lowest variable in the ordering. This procedure continues until no improvement is seen or a maximum number of parents is reached. The procedure is repeated for all variables.

Cheng et al. [2002] use mutual information between variables along with conditional independence checks to construct networks. The algorithm due to Chow and Liu [1968] is used to create an initial tree structure which maximises the sum of the mutual information associated with the adjacent (directly connected) variables in the tree. Further connections between variables are then added or removed dependent on whether all of the mutual information between non-directly connected variables in the tree can be accounted for by the existing connections. Given sufficient representative training data this algorithm will produce a Bayesian network (not necessarily a poly-tree) that approaches the optimum representation of the joint probability table although it is computationally intensive. A constraint needs to be incorporated to ensure the resulting network is feasible in terms of probability table size and thresholds have to be set for accepting or rejecting conditional independence checks as significant, which may be a problem with limited or noisy training data. Reasons for doubting the practicality of the algorithm in real situations and this is further discussed in section 5.4.

Elements of these techniques are further discussed in the following sections, although the specific nature of the networks, used in this study (for classification) in which the attribute variables need to be within the Markov blanket of the classifier variable, mean that detailed discussion of accurately learning general Bayesian network structures is not entirely relevant.

### 5.3.2 Naive Bayes Classifiers.

As noted, for reasons of computational feasibility many widely used Bayesian network classifiers place most or all of the attribute variables as children of the classifier and can be considered as variants of the ‘naive’ Bayes classifier (see figure 5.6a). In the case of the basic naive Bayes classifier no attempt is made to model the correct dependencies and independencies between the network variables, making the assumption that the classifier variable is dependent on all the attribute variables and that the attribute variables are all independent of each other. Duda and Hart [1973] introduced the concept of what is now called the naive Bayesian classifier as a statistical construct long before the concepts behind Bayesian networks were formalised by Pearl [1988].

Using Bayes rule the probability of the classifier variable  $CL$  given the feature variable set  $A_1, A_2 \dots A_n$  can be expressed as;

$$p(CL|A_1, A_2 \dots A_n) = \frac{p(CL)p(A_1, A_2 \dots A_n|CL)}{p(A_1, A_2 \dots A_n)} \quad (5.10)$$

The naive Bayes assumption leads to the following classifier with the denominator in the above equation (being the same for all values of the classifier), regarded as a normalising constant ( $\alpha$ );

$$\operatorname{argmax}_i P(CL = cl_i | A_1, A_2 \cdots A_n) \approx \alpha P(CL = cl_i) \prod_j p(A_j = a_{j_{obs}} | CL = cl_i) \quad (5.11)$$

where  $a_{j_{obs}}$  is the value taken by the  $j$ th attribute variable in this case. The case is assigned to the class (value of the classifier variable) for which the expression is maximised. Note that the assumption made is that:

$$\prod_j p(A_j) = P(A_1)P(A_2)\dots P(A_n) = P(A_1, A_2 \cdots A_n) \quad (5.12)$$

This is the definition of statistical independence and the assumption is only correct if the attribute variables  $A_1, A_2 \cdots A_n$  are independent. Naive Bayesian network classifiers however are seen to perform well in practice and why this should be so has been investigated by Langley and Sage [1999] who consider that real world data often doesn't contain significant dependencies and Zhang [2004] who demonstrates that where dependencies do exist the naive Bayes network will still perform well if the dependencies are distributed across the classifier values. Both of these findings are questionable in the context of this study. The unexpectedly good performance of naive Bayes classifiers has led to a lot of work being undertaken to improve them further by removing some of the assumptions.

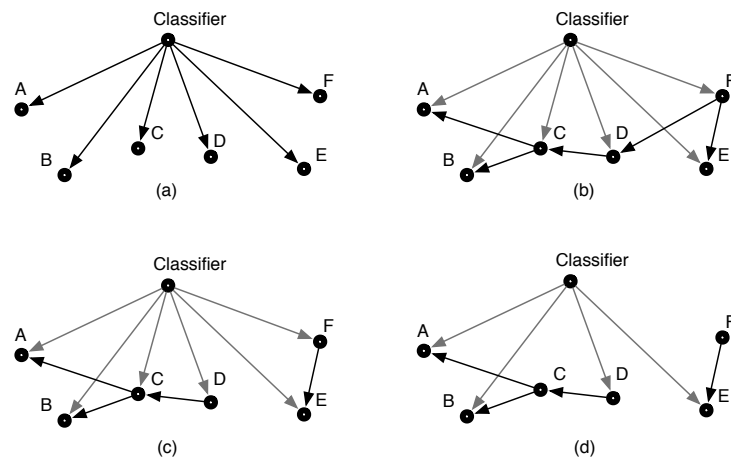


Figure 5.6: Variants of naive Bayesian classifiers: (a) Naive Bayes, (b) Tree augmented naive Bayes, (c) Forest augmented naive Bayes, (d) Selective forest augmented naive Bayes.

### Augmented naive Bayes classifiers.

The first step in removing the independence assumptions is to allow each of the attribute variables a single dependency on one of the other attribute variables, resulting in a 'tree augmented' naive (TAN) Bayesian network (figure 5.6b). This classifier was introduced by Friedman et al. [1997] and uses the algorithm due to Chow and Liu [1968] for maximising

mutual information in a tree network to connect the attribute variables. Friedman et al. [1997] show that the tree augmented naive Bayesian network is an optimum representation of the joint probability given the constraints of the tree structure (each variable in the tree having only one parent in addition to the classifier variable, all attribute variables present in the tree and a child of the classifier variable). The actual effect of the structure, given that all feature variables are observed (as is the case in Friedman et al. [1997]), is shown in figure 5.7. This can be seen to be the case if the Markov blanket of the classifier is considered, that is to say, a node is independent of all other nodes given its parents, its children and its children's parents. Each of the attribute variables has two parents (one being the classifier variable) and is not dependent on the parents of the parent variables, or its child variables. This in effect means that the classifier is combining 'clusters' of two variables, rather than single variables in an otherwise naive Bayesian structure. Note that in this model some of the attribute variables will be instantiated more than once.

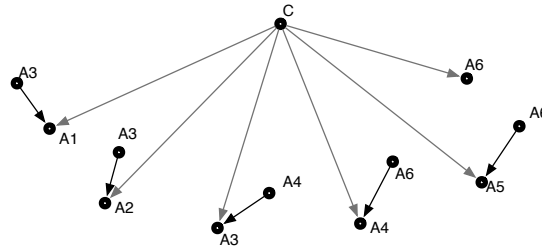


Figure 5.7: The effective structure of a tree augmented naive Bayes network (from figure 5.6b) in the fully observed case.

The tree augmented naive Bayes network is not a poly-tree and cannot necessarily be exactly evaluated except in the case where all the attribute variables are observed, when the decomposition shown in figure 5.7 to a poly-tree is possible. Given this the tree augmented naive Bayes classifier can be expressed as a simple extension to the naive Bayes classifier;

$$\operatorname{argmax}_i p(CL = cl_i | A_1, A_2 \dots A_n) \approx p(CL = cl_i) \prod_j p(A_{j_{obs}} | CL = cl_i, A_{k_{obs}}) \quad (5.13)$$

where  $k$  may not be different for each  $j$  but  $j \neq k$

The tree augmented naive classifier outperforms (if not strikingly) naive Bayes classifiers and slightly outperforms the benchmark C4.5 decision tree algorithm due to Quinlan [1993].

Extensions to the tree augmented naive Bayes concept include Keogh and Pazzani [1999] who use a hill climbing algorithm to add augmenting edges to a naive Bayes classifier, terminating when classification performance ceases to improve. This does not necessarily connect all the feature variables in a single tree and results in what Ziebart et al. [2007] call a forest augmented naive Bayes network (see figure 5.6d). The result, in the fully observed case, is still the same as for the tree augmented network except that not all of the attribute

variables will be ‘clustered’ into pairs.

### Variable grouping or clustering.

The comprehensive way to avoid the independence assumption is to have (as far as feasible) the joint probability tables for dependent variables represented within a naive Bayes framework, resulting in a network such as shown in figure 5.8a. This cannot necessarily be achieved by the tree augmentation approach. Clustering (or combining) variables in naive Bayes classifiers was directly investigated by Kononenko [1991] who uses a direct assessment of statistical independence to decide whether to cluster (join in his terminology) variable values:

$$\epsilon = \sum_j P(C_j) \cdot \left| P(C_j|A1 = a1_i, A2 = a2_k) - \frac{P(C_j|A1 = a1_i)P(C_j|A2 = a2_k)}{P(C_j)} \right| \quad (5.14)$$

Where  $j$  denotes different possible states of the classifier and  $i$  and  $k$  different states of two attribute variables on which the state of the classifier is dependent. Note that Kononenko is not joining (using a joint probability table for) the attributes as a whole. He is choosing either to use the joint probability or the product of probabilities for individual attribute value combinations, depending on whether the attributes are independent when averaged over the classifier values. The factor  $\epsilon$  increases as statistical dependence increases. Kononenko also uses a measure of reliability for joining attribute values based on the number of times the conjunction of values occurs in the data-set and uses the joint probability if the following inequality is satisfied:

$$1 - \frac{1}{4\epsilon^2 N(A1 = a1_i, A2 = a2_k)} \geq 0.5 \quad (5.15)$$

Thus a lower level of statistical dependence that is seen more frequently may still be considered a reliable indicator for joining the attribute values. Kononenko notes that his method can be applied iteratively, resulting in the clustering of three or more attribute values. There is some similarity between Kononenko’s method and that used in this study and this is discussed in section 5.4. In Kononenko’s tests however there was little improvement over the standard naive Bayes classifier.

Pazzani [1997] compare a hill climbing forward (adding features to the classifier independently or to a group) to a backward (removing features from the classifier or combining them into groups) algorithm. The algorithms are computationally expensive but the backward variant in particular delivers significant improvement over naive Bayes networks on a range of data sets. The difference between the results of Pazzani [1997] and Kononenko [1991] could be due to the fact that Pazzani is also performing feature selection (thereby eliminating redundancy, see below) or that Kononenko’s joining criterion is non-optimum (see section 5.4), or both.

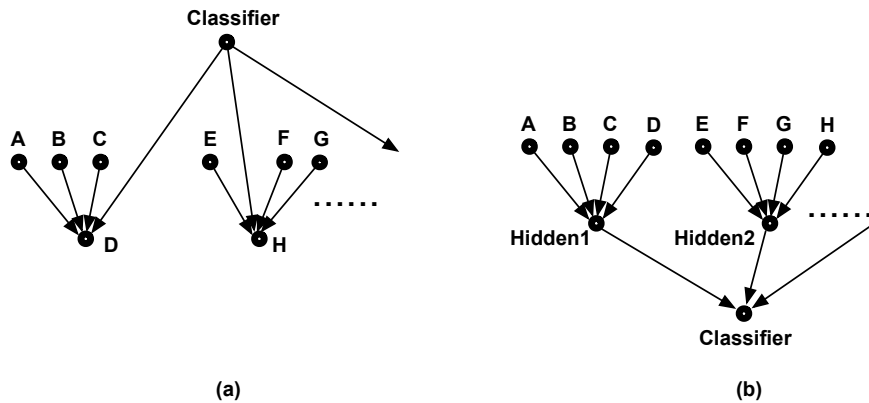


Figure 5.8: Less restricted forms of Bayesian network classifier: (a) a Pazhani type clustered naive classifier, (b) a classifier using variable clusters attached to hidden nodes.

### Feature variable selection.

As well as potentially excluding some important statistical dependencies between attributes the assumption, in a naive Bayesian classifier, that the classifier is dependent on all attribute variables, also potentially includes dependencies that incorporate redundancy into the classifier. The damaging effect of redundancy is illustrated by the simple example shown in figure 5.9 where a classifier is simply learning whether a majority of three attributes is present. Making the assumption that the training data in the table is representative of the population of attribute value cases, the conditional probability tables shown will correctly reflect the class probabilities in the population. The classifier will correctly identify all subsequent cases. However if a fourth variable is added, which is fully redundant with one of the first three, the probability tables will remain unchanged under training but the classifier performance will degrade as shown. In general, even if not in so clear cut a way as in the example, redundancy in real world classifiers will reduce the tolerance of the classifier to noise and lack of correct population representation in the training data set.

Note that if the attribute variables in figure 5.9 are parents of the classifier rather than children then the problem illustrated disappears. However the feasibility constraints on the probability table size in this case still apply and the redundancy leads directly to inefficiency in the representation of probabilities.

Ziebart et al. [2007] use feature selection to remove another class of unwanted attribute variable, those that have low mutual information with the classifier, so are in that sense irrelevant (a random variable for instance). These variables are removed from direct connection to the classifier variable (though not from the network), and the result is termed the selective forest augmented naive Bayes network. (figure 5.6d). Their technique would not necessarily remove redundant attributes either from the network or direct connection to the classifier however.

Feature selection was investigated by Langley and Sage [1994] who used a greedy search to add attribute variables to a naive Bayes network, terminating when classification accu-

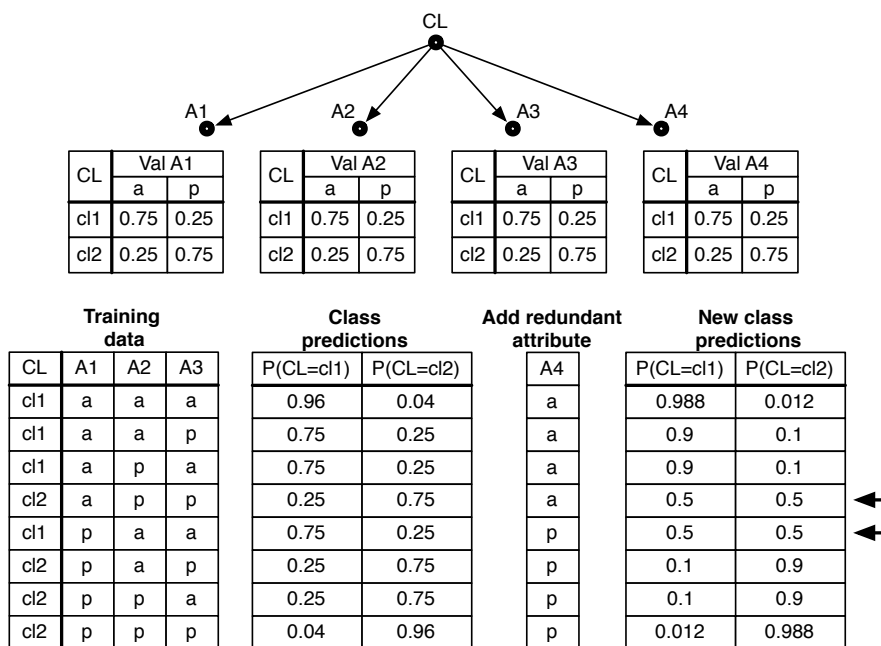


Figure 5.9: An illustration of the effect of a redundant variable on naive Bayesian classifier performance. The network learns the ‘majority vote’ of attributes A1, A2, A3 which are either absent ‘a’ or present ‘p’. If the variable A4, which is a duplicate of A1 is added, the training of the classifier is unaffected but the classification performance is degraded with two cases out of 8 undecidable as indicated.

racy ceases to improve. This results in some improvement in classifier performance. As noted Pazzani [1997] use backward attribute variable elimination starting with a full set of variables. These ‘wrapper’ methods (so called because the search techniques can be wrapped round many different classifier types) are computationally expensive however, involving the training and testing of multiple classifiers for each step of the search. However both redundant and irrelevant variables can be removed, an overview is given by Kohavi and John [1997].

Peng et al. [2005] describe a feature selection technique termed minimum Redundancy Maximum Relevance (mRMR), which selects variables for a classifier on the basis that their mutual information with respect to the classifier is high and their mutual information with already selected variables is low, thus avoiding the addition of redundant variables and irrelevant variables. This technique is used as part of the classifier construction algorithm for this study and is further described in section 5.4.

### **Markov blanket classifiers**

Friedman and Goldszmidt [1996] compare naive Bayesian networks with general Bayesian networks constructed using a greedy search algorithm and the minimum description length (MDL) score as a measure of network quality. Minimum description length balances the fit of a network to the training data against the complexity of the network. They find that there is no real advantage over a range of data sets in using an unrestricted network. However they note that the unrestricted Bayes nets perform worse with data-sets with large numbers of features and attribute this to the restricted number of features in the classifier variables’ Markov blanket. As noted in a fully observed case only the features in the Markov blanket of the classifier will determine the distribution of the classifier variable.

Madden [2002] creates a ‘near-general’ Bayesian network classifier in which all attribute variables are in the classifier’s Markov blanket and a limited number can be parents of the classifier. They use the local search algorithm of Cooper and Herskovits [1992] to add a number of parents to a variable (up to some practical limit). In spite of this “increased expressiveness” over simple tree augmented naive Bayes classifiers the performance is not significantly better over a range of data sets.

In the field of bio-informatics, where Gene expression databases can have many thousands of features, the practical problems of learning full and accurate Bayesian networks have led to a series of algorithms for learning the Markov blanket of a given variable. These tend to be truncated versions of the algorithm due to Cheng et al. [2002] and have the same problems trading network accuracy with feasibility and practicality with limited training data. A comparative overview is given by Peàa et al. [2007].

### **Hidden variables**

Looking at figure 5.8b it can be seen that attribute variables outside of the classifiers Markov blanket can still influence the value of the classifier if a deliberately unobserved or ‘hidden’ node is all that separates them in the network. This allows more attributes



to be (indirectly) parents of the classifier whilst retaining feasible probability tables, but it introduces the problem of defining values for the hidden variables. The advantage of this network topology over that of figure 5.8a is that it will be more robust to redundancy between the hidden variables.

Various methods have been proposed for learning hidden node probability tables. Connolly [1993] uses a mutual information technique to determine the topology (structure) of a network containing hidden variables and a ‘conceptual clustering’ method, similar to the ‘COBWEB’ algorithm (see Fisher [1987]), to assign values to the hidden variable probability tables. The COBWEB algorithm operates on the training data and creates a decision, or classification, tree based on ‘category utility’, a measure of how many attributes of an instance in a category can be correctly ‘guessed’ given a class (value of a classifier variable, or a hidden node in this case). If for all existing values of the hidden node a new combination of attribute values is poorly predicted, then a new class, or hidden node value, may be created. Connolly does not provide extensive results for his method although it formed the basis for the commercially available ‘TANTRA’ system. Russell et al. [1995] uses a hill climbing algorithm to learn probability values for hidden variables. Although good results are reported on a number of networks it is hard to see how this technique would not in general suffer from the local minima problems associated with expectation maximisation.

Expectation maximisation has been used for constructing hidden variables by Lauritzen [1995] although the form of the probability distributions is constrained and even then they note the problem of local minima leading to “unsuitably extreme probabilities”. Expectation maximisation is used to learn network topology by Friedman [1998], although again in the reported results the topologies were highly constrained initially (variables were effectively ‘ordered’ as parents or children of a hidden node layer).

Expectation maximisation was the intended method for learning hidden variable values in this study. But even with a fixed and fairly simple network topology it proved impractical due to the number of local minima in the representations. Experiments, even on simple networks such as a 4-XOR problem (using a network with two hidden nodes, each with three values), showed that a large number of restarts was required to obtain a satisfactory result. Although an evolutionary algorithm might have overcome these issues, a much simpler approach was used in which all the hidden variables are effectively versions of the classifier variable. This is described in more detail in section 5.5. As well as seeming to work reasonably well in practice this technique might prove to be a good starting point to solve the problem of what is the ‘meaning’ of the hidden variable, in a manner similar to that of Fisher [1987].

## 5.4 Feature combination using interaction information

A quick examination of the reference choice problem suggested that the methods for learning classifiers outlined above would probably not produce good results. It is immediately apparent that a variable related to the target object, its volume for example, would not

have any *direct* effect on the classifier variable, which is ranking reference objects from more suitable to less suitable. It would be expected to have a very low value of mutual information with respect to the classifier and hence might be discarded by many of the feature selection or local search based classifier construction algorithms described in section 5.3. In combination with variables related to the reference object and the geometry of the scene however, we would expect this to be a vital parameter in determining reference suitability.

The mutual information  $I(X; Y)$  of two discrete variables  $X$  and  $Y$  which take values  $(x_1, x_2 \dots x_n)$  and  $(y_1, y_2 \dots y_n)$  is defined as:

$$I(X; Y) = \sum_{i,j} P(X = x_i, Y = y_j) \log_2 \left( \frac{P(X = x_i, Y = y_j)}{P(X = x_i)P(Y = y_j)} \right) \quad (5.16)$$

Conditional mutual information ( $I(X; Y|Z)$ ) is defined as:

$$I(X; Y|Z) = \sum_{i,j,k} P(X = x_i, Y = y_j, Z = z_k) \log_2 \left( \frac{P(X = x_i, Y = y_j|Z = z_k)}{P(X = x_i|Z = z_k)P(Y = y_j|Z = z_k)} \right) \quad (5.17)$$

Mutual information and conditional mutual information can be defined in terms of the fundamental quantity entropy  $H$  which is defined for a single variable  $X$  as:

$$H(X) = - \sum_i P(X = x_i) \log_2(P(X = x_i)) \quad (5.18)$$

and for a joint variable  $(X, Y)$  as:

$$H(X, Y) = - \sum_{i,j} P(X = x_i, Y = y_j) \log_2(P(X = x_i, Y = y_j)) \quad (5.19)$$

mutual information is:

$$I(X; Y) = H(X) + H(Y) - H(X, Y) \quad (5.20)$$

and conditional mutual information is:

$$I(X; Y|Z) = H(X, Z) + H(Y, Z) - H(X, Y, Z) - H(Z) \quad (5.21)$$

Tables 5.3 and 5.4 show values of mutual information and conditional mutual information for some of the variables used in the study, as extracted from the annotated test data set. The variables are defined in section 5.6.1 but for convenience here, the names including ‘Target’ refer to the target object and those including ‘Ref’ refer to candidate reference objects. MaxDim is an object’s maximum dimension and MaterialVol is the volume of the material in an object. The classifier variable is ‘RefSuitability’ and ‘ProxDist’ is the smallest distance between a target and candidate reference objects. The values in the tables are scaled so that the maximum value is 10, values of less than 0.5 are insignificant

and probably due to ‘noise’ in the data (see section 5.5.3). Scaling is required for these values as, since the variables do not all take the same number of values, the maximum mutual information is different for different variable combinations. For instance the maximum value for mutual information between two, 2-valued variables, is 1 bit, the maximum for two 4-valued variables is 2 bits. The maximum information is defined by the variable with the lowest number of values so the scale factor applied to the mutual information values  $S$ , to arrive at the values in the table, is:

$$S = \frac{10}{\log_2(N_{min})} \quad (5.22)$$

where  $N_{min}$  is the number of values for the variable with the lowest number of values. The constant 10 is simply used for convenience.

It can be seen that, as feared, mutual information between target size variables and the classifier are negligible. The only significant values of mutual, or conditional mutual, information between the feature variables shown, relate the target size measures (variable numbers 3 and 5) to each other and likewise for the reference size variables (variable numbers 4 and 6). This gives a strong indication that using these functions will tend to cluster together redundant or near redundant variables and it is hard to see how this can help in the classification process.

Table 5.3: Mutual information values for some key variables

		Mutual information values				
Variable No. (Ai)	Variable Name	Variable number (Aj)				
		1	2	3	4	5
1	RefSuitability					
2	ProxDist	3.95				
3	MaxDimTarget	0.20	0.19			
4	MaxDimRef	1.07	0.12	0.21		
5	MaterialVolTarget	0.25	0.12	2.78	0.18	
6	MaterialVolRef	1.49	0.1	0.29	2.14	0.28

High mutual information between two variables (neither of which are the classifier) does not say anything about their relevance to the classifier. Figure 5.10 shows a simple case of 2 binary variables  $A1$  and  $A2$  and their joint probability with a classifier variable  $CL$ . The diagram is in effect a three dimensional probability table. In figure 5.10a it can be seen that high conditional mutual information can identify attribute variables which provide no information about the classifier variable. Changes in the classifier variable are not ‘signalled’ by changes in the attribute variables  $A1$  and  $A2$ . The result is a classifier probability table with very poor prediction performance. What is required is to be able to detect variables which interact to provide information about  $CL$  (where the pattern of  $A1$  and  $A2$  changes as the classifier variable  $CL$  changes) as in 5.10(b). This leads to

Table 5.4: Conditional mutual information values for some key variables

Conditional mutual information values						
Variable No. (Ai)	Variable Name	Variable number (Aj)				
		1	2	3	4	5
1	RefSuitability					
2	ProxDist	0.00				
3	MaxDimTarget	0.00	0.42			
4	MaxDimRef	0.00	0.39	0.50		
5	MaterialVolTarget	0.00	0.38	2.88	0.48	
6	MaterialVolRef	0.00	0.33	0.60	2.32	0.55

a classifier probability table with good prediction performance. As can be seen this also corresponds to a high conditional mutual information,  $I(A1; A2|CL)$  but high conditional mutual information does not select for this case. A function which selects combinations of parent values which correlate with *changes* in the classifier variable value is required. This is provided by the interaction information (II) function introduced by McGill [1954] and expanded on by Bell [2003] (although termed ‘co-information’);

$$I(X; Y; Z) = I(X; Y|Z) - I(X; Y) \quad (5.23)$$

or in terms of the probability mass functions;

$$I(X; Y; Z) = \sum_{i,j,k} P(X = x_i, Y = y_j, Z = z_k) \cdot \log_2 \left( \frac{P(X = x_i, Y = y_j|Z = z_k)}{P(X = x_i|Z = z_k)P(Y = y_j|Z = z_k)} \cdot \frac{P(X = x_i)P(Y = y_j)}{P(X = x_i, Y = y_j)} \right) \quad (5.24)$$

making substitutions of the form  $P(a|b)P(b) = P(a, b)$  the interaction information function can be seen to be symmetric in  $X$ ,  $Y$  and  $Z$ .

$$I(X; Y; Z) = \sum_{i,j,k} P(X = x_i, Y = y_j, Z = z_k) \cdot \log_2 \left( \frac{P(X = x_i, Y = y_j, Z = z_k)P(X = x_i)P(Y = y_j)P(Z = z_k)}{P(X = x_i, Z = z_k)P(Y = y_j, Z = z_k)P(X = x_i, Y = y_j)} \right) \quad (5.25)$$

Higher orders of interaction can also be defined for a variable set  $V = (X_1, X_2 \dots X_n)$  in terms of entropy  $H$  as follows:

$$I(V) = - \sum_{T \subseteq V} -1^{|V|-|T|} H(T) \quad (5.26)$$

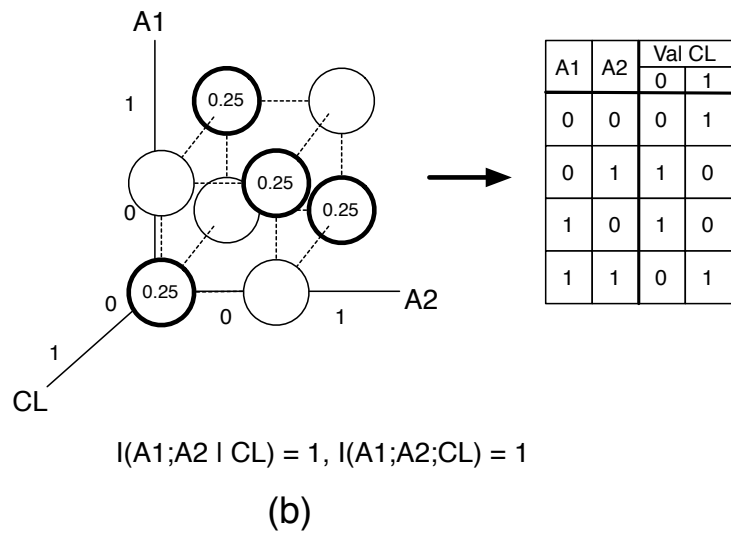
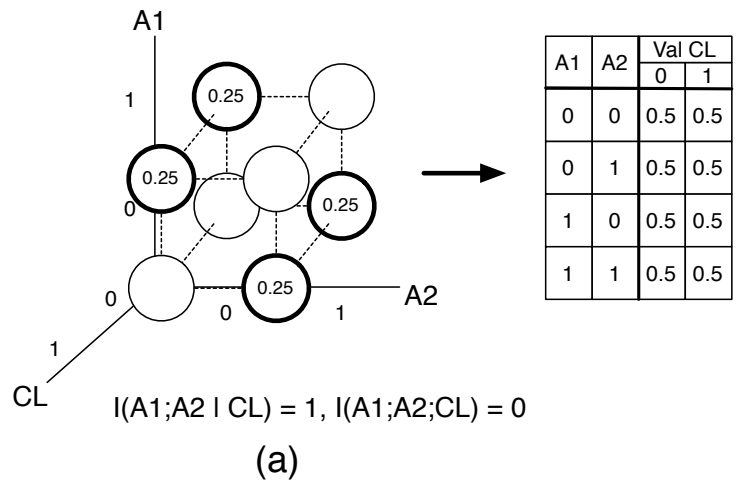


Figure 5.10: Class discrimination with cluster variables chosen using conditional mutual information and interaction information, the numbers in the balls represent  $P(a1, a2, cl)$  values, the diagrams are 3-dimensional probability tables

Note that the interaction information can have negative values, unlike mutual information or conditional mutual information. Bell [2003] notes that this has possibly inhibited its use in applications such as Bayesian networks. The positive and negative interaction information values arise from cases such as those illustrated in figure 5.11. Cases (a) and (c) in figure 5.11 are often used in illustrating interaction information. Cases (b) and (d) come from examples related to or directly from this study. Intuitively it would be expected that the presence of a building (in a digitally photographed image) would cause the features of parallel lines to be present and that if the parallel vertical lines were present so would be the horizontal ones, they would have high mutual information. To detect the presence of a building a Bayesian network cluster following this structure might be used. Also intuitively, in this study a candidate reference “bigger but not too much bigger” (Miller and Johnson-Laird [1976]) than the target might be a suitable reference, but the sizes would not be *caused* by the reference suitability. The mutual information between the reference and target object sizes would be low, which it is.

Values for interaction information between key variables in the study are shown in table 5.5. It can be seen from this table that the desired effect has been achieved and that higher interaction information values exist between non-redundant variables (such as target and volume size measures), than between possibly redundant measures.

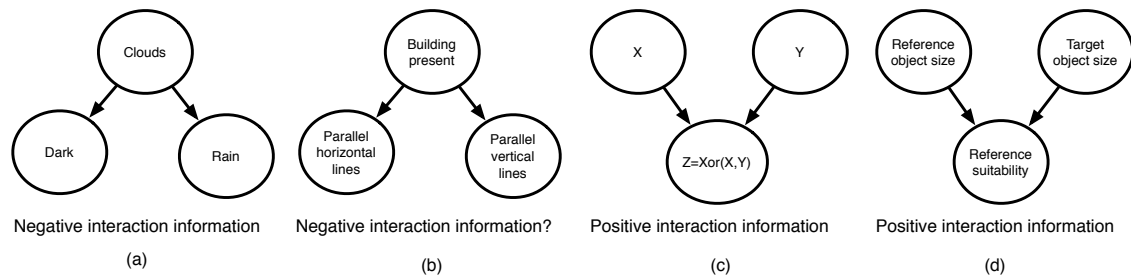


Figure 5.11: Examples of positive and negative interaction information: (a) clouds cause rain and darkness, which will exhibit high mutual information, (b) From object recognition the presence of a building may cause the presence of parallel horizontal and vertical lines, (c) The XOR function between independent variables (low mutual information) has high positive interaction information (d) From this study a combination of reference and target object size may contribute to a reference being ‘suitable’.

A requirement for a variable cluster to produce high ‘discrimination’ between classes can be seen from figure 5.10 to be the requirement for low  $H(CL|A1, A2)$ . This is the same as saying that averaging over all rows in the probability table defined by a given  $A1$  and  $A2$ , the entropy of the classifier  $CL$ , in each row, is low. The ability of interaction information  $I(A1; A2; CL)$  to satisfy this requirement, when used to cluster variables, can be shown as follows;

$$I(A1; A2; CL) = I(CL; A2|A1) - I(CL; A2) \quad (5.27)$$

from the definition of conditional mutual information in terms of entropy;

$$I(A1; A2; CL) = H(CL|A1) + H(A2|A1) - H(CL; A2|A1) - I(CL; A2) \quad (5.28)$$

using the chain rule for entropy,  $H(CL; A2) = H(CL|A2) + H(A2)$ ;

$$I(A1; A2; CL) = H(CL|A1) + H(A2|A1) - H(CL|A1, A2) - H(A2|A1) - I(CL; A2) \quad (5.29)$$

from the definition of mutual information in terms of entropy;

$$I(A1; A2; CL) = H(CL) - I(CL; A1) - H(CL|A1, A2) - I(CL; A2) \quad (5.30)$$

Although initially it may not seem that equation 5.30 proves that  $I(A1; A2; CL)$  does satisfy the requirement it is clear when the following are taken into account:

1. We are interested principally in the case when  $I(CL; A1)$  and  $I(CL; A2)$  are close to zero. If either or both of these are significant we will be dealing with this direct interaction between the classifier and an evidence variable on its own merit.
2. If  $H(CL)$  is not a large value then the classification task becomes trivial. In the limit if  $H(CL) = 0$  all instances belong to a single class.
3. Note that low  $H(CL|A1, A2)$  is a necessary, but not on its own sufficient, condition for clustering variables  $A1$  and  $A2$ . This is because  $H(CL|A1, A2) = 0$  can occur with  $H(CL) = 1$ ,  $I(CL; A1) = 0.5$  and  $I(CL; A2) = 0.5$  for instance.

Given equation 5.23 it is clear that the equivalent expression for conditional mutual information contains the extra term  $I(A1; A2)$ ;

$$I(A1; A2|CL) = H(CL) - I(CL; A1) - H(CL|A1, A2) - I(CL; A2) + I(A1; A2) \quad (5.31)$$

hence high conditional mutual information does not guarantee low  $H(CL|A1; A2)$  because it is also dependent on the value of  $I(A1; A2)$ . (see figure 5.10(a), the entropy of the classifier in all rows is maximised)

Implicit in the assumptions of low  $I(CL; A1)$ ,  $I(CL; A2)$  and high  $H(CL)$  is that the interaction information will be positive. However negative interaction information as illustrated in figure 5.12 can be significant.

High negative interaction information between two attributes ( $A1, A2$ ), requires high mutual information between the classifier and the attribute variables, if the classification task is not trivial (high  $H(CL)$ ) and, as required, the entropy of the classifier given the attributes is low (low  $H(CL|A1, A2)$ ). These values of mutual information with the classifier should be identified without needing to look at interaction information and it appears that it is the positive values of interaction information that are the most important to identify.

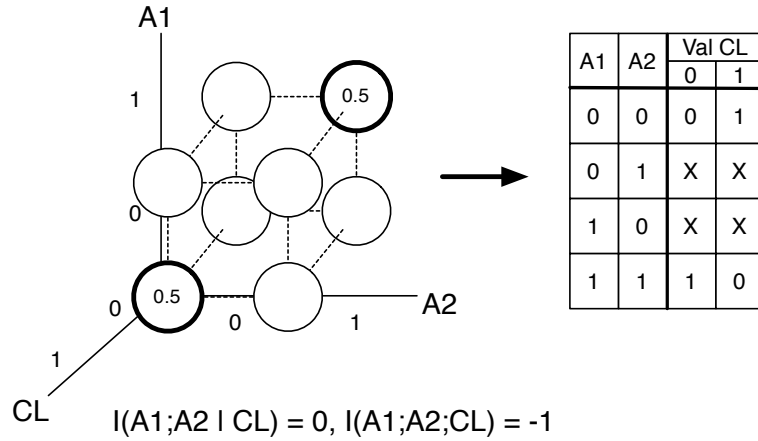


Figure 5.12: Class discrimination with cluster variables chosen using high negative interaction information, the numbers in the balls represent  $P(a1, a2, cl)$  values, the diagrams are 3-dimensional probability tables.

Although not acknowledged, the use of interaction information is indirectly included in the methods of Kononenko [1991] and Cheng et al. [2002]. In the algorithm due to Cheng et al. [2002], the initial structuring of the variables into a tree using mutual information, followed by the adding of connections between variables by testing for conditional mutual information, allows variable interactions to be discovered. The conditional mutual information tests are carried out between a pair of variables in the network conditional on a set of variables which separate them, that is, block all the dependence paths between them. So referring back to figure 5.2, to understand whether there could be a connection between variables  $A$  and  $Z$  the quantity  $I(A; Z|B; C)$  would be calculated. This highlights a practical problem with the algorithm in that, depending on the network structure (the number of paths *in to the Markov blanket* of either of the two variables in question) matrices of large dimension may need to be produced to carry out the conditional mutual information tests. With real world data this may make the indirect discovery of interactions much less reliable than the direct discovery used here.

Although Kononenko [1991] is only looking at attribute values he makes use of the quantity (taken from equation 5.14):

$$P(CL|A1; A2) - \frac{P(CL|A1)P(CL|A2)}{P(CL)} \quad (5.32)$$

which can be rewritten as:

$$\frac{P(CL; A1; A2)}{P(A1; A2)} - \frac{P(CL; A1)P(CL; A2)}{P(CL)P(A1)P(A2)} \quad (5.33)$$

if instead of linear quantities Kononenko had used logarithms as:

$$\log \left( \frac{P(CL; A1; A2)}{P(A1; A2)} \right) - \log \left( \frac{P(CL; A1)P(CL; A2)}{P(CL)P(A1)P(A2)} \right) \quad (5.34)$$



or rewritten as:

$$\log \left( \frac{P(CL; A1; A2)P(CL)P(A1)P(A2)}{P(A1; A2)P(CL; A1)P(CL; A2)} \right) \quad (5.35)$$

Equation 5.35 is closely related to the interaction information. All of the joint probability terms in the interaction information definition are included and Kononenko demonstrates that his method will learn a 2-XOR function which is not possible for a naive Bayesian network or a network using mutual information alone.

The direct application of interaction information to classifiers has been described (seemingly alone) by Jakulin [2005] although his derivation of why it is important differs from the one given here. His algorithm for incorporating interaction information into ‘Bayesian like’ classifiers has some similarities to the method developed for this study and is discussed further in section 5.5.2. Jakulin’s study focusses on comparative classifier performance using standard test sets from the University of California, Irvine (UCI) database (Hettich and Bay [1999]). He finds that his ‘Kikuchi-Bayes’ classifiers do not perform as well as state of the art support vector machines, but tend to out-perform naive and tree augmented naive Bayesian classifiers. Both Jakulin [2005] and Strumbelj et al. [2009] use interaction information to diagnose the performance of arbitrary classifiers. It seems that this is motivated by a belief (which deserves more attention) that the principle of interaction is a fundamental concept underlying that of classification and possibly all statistical learning.

As noted above a typical learning scenario will contain variables with a direct influence on the classifier. Not all the variables will be able to be grouped with others on the basis of high interaction information between them and the classifier. Variables that have high mutual information with the classifier<sup>1</sup> and low mutual information with each other (low redundancy) will also produce variable clusters with reasonably low entropy. This is the basis for the mRMR (minimum redundancy maximum relevance) feature selection technique described by Peng et al. [2005]. The mRMR feature selection technique is an approximation for the maximum dependency criterion for the case where features are selected one at a time. Maximum dependency maximises the mutual information between the classifier and the composite variable formed from all those in the selected attribute set  $A1, A2 \dots An$ . This is written  $I((A1, A2, \dots An); CL)$  by Peng et al. and is not the same as multi-variable interaction information  $I(A1; A2; \dots An; CL)$ . The definitions of relevance,  $R$  and redundancy,  $D$  among the incrementally selected feature sets are given as:

$$D(A1, A2, \dots Am; C) = \frac{1}{m} \sum_{i=1}^m I(Ai; C) \quad (5.36)$$

$$R(A1, A2, \dots Am) = \frac{1}{m^2} \sum_{i=1}^m \sum_{j=1}^m I(Ai; Aj) \quad (5.37)$$

---

<sup>1</sup>Mutual information is of course two-way interaction information and it might be more appropriate to drop the term ‘mutual information’, however it is widely used. Jakulin mixes the terminology and in this study the term ‘mutual information’ is also retained.

Given  $(A_1, A_2, \dots, A_m)$  Peng et al then add the attribute  $A_n$  which maximises  $D - R$  for the attribute set  $(A_1, A_2, \dots, A_m, A_n)$ .

The ‘discrimination’ of a classifier with parents selected on this basis is illustrated in figure 5.13. An example of variables from this study that could be clustered in this way might be a distance variable and an ambiguity variable, the mutual information between the two is low but they both have high mutual information with the classifier. As can be seen in figure 5.13 it is possible that a reference that was at a good distance from the target but was ambiguous might be a borderline case and subsequently the high entropy of the classifier reflects the true situation.

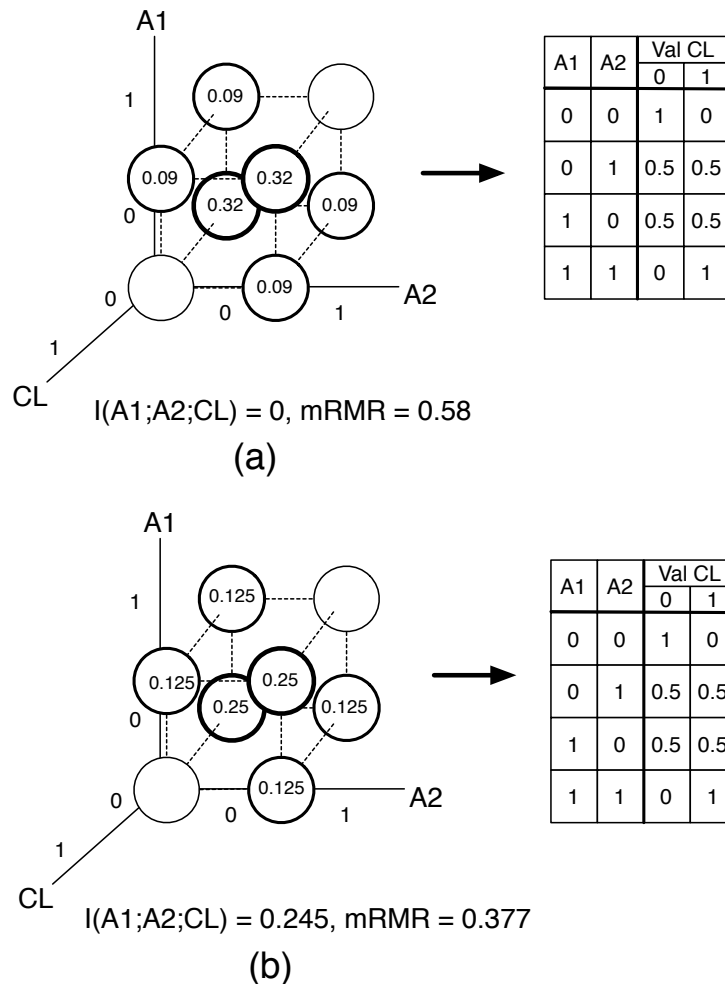


Figure 5.13: Class discrimination with parents chosen using mRMR, the numbers in the balls represent  $p(x, y, z)$  values.

## 5.5 A practical Bayesian classifier construction algorithm

### 5.5.1 Forming feature groups

From the discussion in section 5.4 it can be seen that two distinct criteria are being used to select variables. They should either have a high mutual information with the classifier or have a high interaction information value with the classifier and a third variable. The aim of an algorithm to construct a practical classifier must be to include all significant variables and interacting variable pairs which influence the classifier while rejecting, as far as possible redundant variables. An algorithm for selecting and ranking variables by these criteria can be illustrated with reference to table 5.5 which shows the interaction information values for a selection of variables (a subset of the variables used which are defined in section 5.6.1). The values are scaled to facilitate comparison between variable pairs with different numbers of values. The maximum value of interaction information possible in the table is 10. Some negative values occur due to noise in the data set and these are discarded by the algorithm. Also a few ‘genuine’ negative values of interaction information occur in this study for situations where the  $I(A1, A2)$  is larger than  $I(A1, A2|CL)$ . As noted, high negative interaction values will tend to indicate high values of mutual information between the classifier and the individual attribute variables, which will be identified in their own right. For this reason negative interaction information values are not used in the algorithm at present.

Table 5.5: Interaction information and mutual information values for a variable subset. The classifier  $CL$  is ‘reference suitability’ for all the variable pairs  $A_i, A_j$  for the interaction informations  $I(A_i; A_j; CL)$ . shown. The mutual information values are  $I(A_j; CL)$

		Interaction information values												
Variable No. ( $A_i$ )	Variable Name	Variable number ( $A_j$ )												
		1	2	3	4	5	6	7	8	9	10	11	12	13
1	MobilityTarget													
2	MobilityRef	0.17												
3	Ambiguity	0.02	-0.46											
4	TopologyHull	0.18	0.17	-0.08										
5	ProxDist	0.11	0.31	0.41	-1.13									
6	MinDimTarget	0.04	0.41	-0.02	0.26	0.74								
7	MinDimRef	0.25	0.32	-0.29	0.21	1.37	0.91							
8	MaxDimTarget	0.06	0.34	0.04	0.13	0.73	-0.01	1.05						
9	MaxDimRef	0.12	-0.02	0.02	0.13	0.55	0.97	0.53	1.20					
10	BbVolTarget	0.03	0.42	-0.03	0.13	0.72	-0.12	0.97	-0.05	1.12				
11	BbVolRef	0.30	0.63	-0.14	0.12	0.70	0.71	-0.08	0.84	-0.32	0.76			
12	MaterialVolTarget	0.03	0.34	-0.03	0.15	0.75	-0.15	0.79	-0.05	1.21	-0.09	0.76		
13	MaterialVolRef	0.14	0.11	-0.29	0.09	0.82	0.79	0.17	0.91	0.55	0.81	-0.78	0.84	

		Mutual information values												
		Variable number ( $A_j$ )												
		1	2	3	4	5	6	7	8	9	10	11	12	13
	refSuitability	0.00	0.96	1.98	1.86	3.96	0.20	1.45	0.20	1.07	0.18	1.57	0.25	1.45

The algorithm proceeds in two stages. Firstly the table of interaction information values is searched for the highest values relating to each of the variables. This can be a direct mutual information between a single variable and the classifier or a high interaction between two variables and the classifier. Using table 5.5 as an example this results in the

ranked list of variables and variable pairs shown in table 5.6. This stage of the algorithm is shown in algorithm 5.5.1.

The second stage of the algorithm combines the variables and variable pairs into groups using the following rules:

1. The process starts using the lowest ranked variable or variable pair, attempting to improve the useful information contributed by these variables in combination with others.
2. Single variables are combined with other single variables or groups of single variables if, for all variables, the mutual information between the variables is less than the mutual information between the variables and the classifier. Also the lower ranked variable must have significant mutual information with the classifier. (This is effectively an mRMR feature clustering.)
3. Single variables are combined with variable pairs if, for either member of the variable pair, the single variable has a higher interaction information with that variable than it has mutual information with the other. This prevents variables being attached to variable pairs when there is a redundancy.
4. Variable pairs are combined with variable pairs using the rule for attaching single variables but treating one of the pairs as two separate variables.
5. For any of the combinations the resultant conditional probability table must be smaller than a predetermined maximum practical size.
6. After all variables and variable pairs have been considered any single variables are eliminated as these must be redundant (except possibly in the case of a very small number of variables).
7. Any variables or variable pairs not contributing significant information are eliminated. What constitutes ‘significant’ information is defined in section 5.5.3

After the first stage the variables are arranged as shown in table 5.6. The restricted number of variables used in order to make the illustration tractable has resulted in there being only one interacting variable pair (13, 10) i.e., MaterialVolTarget and MaxDimRef, the variable numbers are from table 5.5). The other variables all have higher mutual information with the classifier (variable number 1, referenceSuitability) than interaction informations with the classifier and a third variable.

The second stage of the algorithm seeks to find further useful information contributions and interactions by combing the variable pairs and single variables, while restricting combination of redundant variables. This results in the variable groups shown in table 5.7. The variables from table 5.6 not included in the groups have been omitted because they could not on their own or in combination with other variables contribute significant information

---

**Algorithm 1** Form a list of the highest interactions and mutual informations with the classifier

---

```

/* Variable definitions */
n /* Number of attribute variables */
I1[n]/* List of mutual informations  $I(A_i; Classifier)$  */
I2[n][n]/* Table of mutual informations  $I(A_i; A_j)$  */
I3[n][n]/* Table of interaction informations  $I(A_i; A_j; Classifier)$  */
InteractionValues[n]  $\leftarrow$  0 /* List of high interactions */
Groups[maxGroups][groupSize][2]  $\leftarrow$  0/* List of variable groups */
groupNumber  $\leftarrow$  0 /* Variable group count */

GenerateHighestInteractionsList (n, I2, I3)

Count  $\leftarrow$  0 /* Count variable */
(Used[n])  $\leftarrow$  0 /* List of included attributes */

while Count < n do
  for all  $i < n, j < n$  such that  $i \neq j, Used[i] \neq 1, Used[j] \neq 1$  do
    /* A is the highest mutual information and B the highest interaction information
    among the variables so far not included in a group */
    (A, i)  $\leftarrow$  argmax(I1[i])
    (B, i, j)  $\leftarrow$  argmax(I3[i][j])

  end for
  if A > B then /* Mutual information is higher */
    InteractionValues[groupNumber]  $\leftarrow$  A /* record the interaction value for the
    group */
    Groups[groupNumber][0][0]  $\leftarrow$  0
    Groups[groupNumber][0][1]  $\leftarrow$  i /* Add the variable to the group */
    groupNumber  $\leftarrow$  groupNumber + 1 /* Increment group number count */
    Count  $\leftarrow$  Count + 1 /* Only one variable is included in a group */
    Used[i]  $\leftarrow$  1 /* Mark the variable as included */
  end if
  if B > A then /* Interaction information is higher */
    InteractionValues[groupNumber]  $\leftarrow$  B /* record the information value for the
    group */
    Groups[Count][0][0]  $\leftarrow$  i /* Add the first variable to the group */
    Groups[Count][0][1]  $\leftarrow$  j /* Add the second variable to the group */
    groupNumber  $\leftarrow$  groupNumber + 1 /* Increment group number count */
    Count  $\leftarrow$  Count + 2 /* two variables are included in a group */
    Used[i]  $\leftarrow$  1 /* Mark the variable as included */
    Used[j]  $\leftarrow$  1 /* Mark the variable as included */
  end if
end while
return InteractionValues, Groups, groupNumber

```

---

Table 5.6: Ranked interacting variable pairs and informative single variables from the 1st stage of the algorithm

Variable number	6	4	5	12	14	8	13	3	7	9	11	2
Variable number	-	-	-	-	-	-	10	-	-	-	-	-
II value	3.95	1.98	1.86	1.57	1.49	1.45	1.21	0.96	0.20	0.20	0.18	0.00

to the classifier (see section 5.5.3). The information values leading to the formation of two of the groups are illustrated in figure 5.14. Note that while it may not be clear that any meaning can be attached to the groups chosen by the algorithm the inclusion of potentially redundant variables such as the bounding box volume and material volume of objects has been avoided. This algorithm is presented more formally in algorithm 5.5.1.

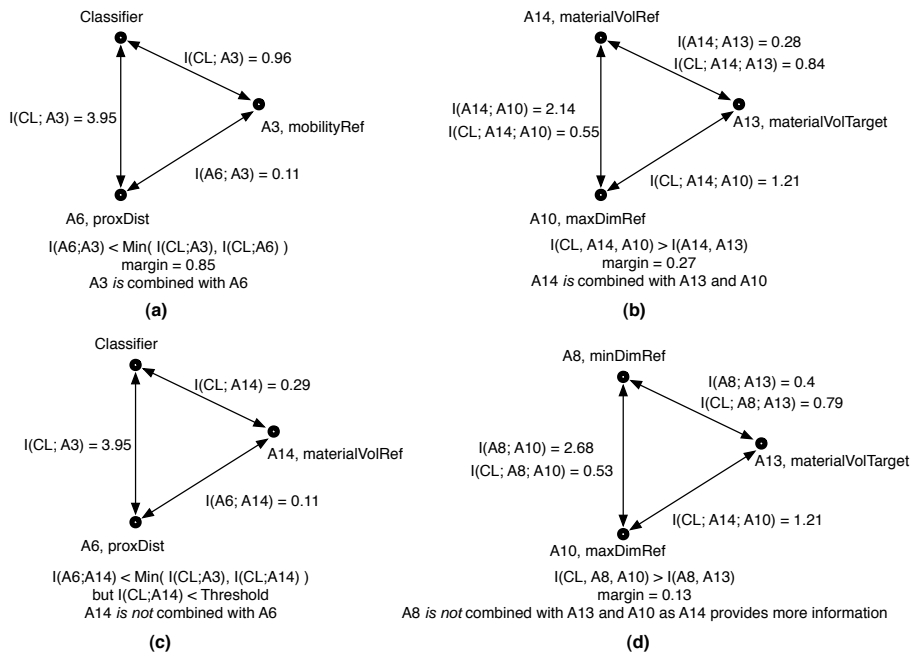


Figure 5.14: Combining interacting variables and variable pairs in to groups; (a) Successful combination of single variables, (b) Successful combination of a single variable with a pair, (c) Unsuccessful combination of single variables, (d) Unsuccessful combination of a single variable with a pair.

Table 5.7: Selected clusters with corresponding information gains (normalised)

Variable number	3	13	8	12
Variable number	6	10	5	4
Variable number	-	14	-	-
Information increase	0.85	0.27	1.17	0.83

This second step could be repeated several times if necessary. The algorithm will not

---

**Algorithm 2** Combine the interacting variable pairs and variables with high mutual information with the classifier into groups. The subroutine **getInformationMargin** is defined in Algorithm 3

---

```

n /* Number of attribute variables */
I1[n]/* List of mutual informations  $I(A_i; Classifier)$  */
I2[n][n]/* Table of mutual informations  $I(A_i; A_j)$  */
I3[n][n]/* Table of interaction informations  $I(A_i; A_j; Classifier)$  */
InteractionValues[n] ← 0 /* List of high interactions from first step */
Groups[maxGroups][groupSize][2] /* List of variable groups from first step */
(GroupSizes[maxGroups] ← 1 /* Size of variable groups from first step */
Ithreshold /* Value below which information is insignificant */
IgroupNumber /* Number of groups from first step */

CombineInteractingVariables (n, I, groupNumber, Groups)

for (i = groupNumber to 1 step -1) do
  for all (j < i, j > 0) do
    (A, j) ← argmaxj getInformationMargin(Groups, i, j)
  end for
  if groupSizes[j] + groupSizes[i] < maxGroupSize then /* If the resulting group is
  practical */
    Groups[j][groupSizes[j]][0] ← Groups[i][0][0] /* add group i to group j */
    Groups[j][groupSizes[j]][1] ← Groups[i][0][1]
    groupSizes[j] ← groupSizes[j] + 1 /* Increment size of group j */
    InteractionValues[j] ← InteractionValues[j] + A /* add new information contri-
    bution */
  end if
  /* Remove the now combined group from the list */
  removeGroup(Groups, i) /* straightforward - not described */
end for
/* Remove variables or groups not contributing significant information */
removeInsignificantGroups(Groups, Ithreshold) /* straightforward - not described */
removeSingleVariableGroups(Groups) /* straightforward - not described */
return (InteractionValues, Groups)

```

---

---

**Algorithm 3** Find the lowest information gain between the variables in two groups (this must be above a threshold value)

---

```

getInformationMargin (Groups, i, j)

/* A little shorthand in this definition:  $V_i$  means 'the variable in question from Group[i]',
 $V_{i1}, V_{i2}$  means 'the variable pair in question from Group[i]' similarly for Group[j] */
/* Try to combine single variables in the first group.... */
for all ( $V_i$  in Groups[i]) do
  /* ...with single variables in the second group */
  for all ( $V_j$  in Groups[j]) do
    /* Check variables are not more redundant with each other than they are relevant
    to the classifier */
     $margin1 \leftarrow \underset{i,j}{\operatorname{argmin}}(\max(I1(V_i) - I2(V_i, V_j), I1(V_j) - I2(V_i, V_j)))$ 
  end for
  /* ...or variable pairs in the second group */
  for all ( $V_{i1}, V_{i2}$  in Group[i]) do
    /* Check, for the single variable, that it is less redundant with one variable of the
    pair than it is interactive with the other */
     $margin2 \leftarrow \underset{i,j}{\operatorname{argmax}}(\max(I3(V_i, V_{j1}) - I2(V_i, V_{j2}), I3(V_i, V_{j2}) - I2(V_i, V_{j1})))$ 
  end for
end for
/* Try to combine variable pairs in the first group.... */
for all ( $V_{i1}, V_{i2}$  in Groups[i]) do
  /* ...with single variables in the second group */
  for all ( $V_j$  in Groups[j]) do
    /* Check, for the single variable, that it is less redundant with one variable of the
    pair than it is interactive with the other */
     $margin3 \leftarrow \underset{i,j}{\operatorname{argmax}}(\max(I3(V_{i1}, V_j) - I2(V_{i2}, V_j), I3(V_{i2}, V_j) - I2(V_{i1}, V_j)))$ 
  end for
  /* ...or variable pairs in the second group */
  for all ( $V_{j1}, V_{j2}$  in Group[j]) do
    /* Check, for each variable of the first pair, that it is less redundant with one variable
    of the second pair than it is interactive with the other */
     $margin4 \leftarrow \underset{i,j}{\operatorname{argmax}}(\max(I3(V_{i1}, V_{j1}) - I2(V_{i1}, V_{j2}), I3(V_{i1}, V_{j2}) - I2(V_{i1}, V_{j1}),$ 
       $I3(V_{i2}, V_{j1}) - I2(V_{i2}, V_{j2}), I3(V_{i2}, V_{j2}) - I2(V_{i2}, V_{j1})))$ 
  end for
end for
/* Return the lowest margin - that is the most redundant combination */
 $minMargin \leftarrow \min(margin1, margin2, margin3, margin4)$ 
return ( $minMargin$ )

```

---



remove all redundancy from the resulting models but at each stage there is a gain in useful information over any increase in redundancy. It could also be extended to include higher orders of interaction information that would reveal cases where combinations of three (or more) variables have a significant influence on the classifier. These interactions are invisible to algorithms using lower orders of interaction information. The penalty for trying to find higher order interactions lies in the computational complexity which could be as high as:

$$O\left(\binom{Vars}{I_{order}} Values^{I_{order}}\right) \quad (5.38)$$

where  $I_{order}$  is the order of interaction information being considered,  $Vars$  is the number of variables being considered for inclusion in the model and  $Values$  is the number of values each variable takes. No significant effort has been made to understand if this complexity can be reduced.

### 5.5.2 Combining feature groups

Depending on the exact algorithm chosen for adding variables the result is effectively a collection of variable clusters which need to be combined to form a complete model. If each hidden variable is made equivalent to the classifier variable then each variable cluster is in effect an incomplete model of the entire problem. These models could be combined using Bayesian model averaging, however a complete model can also be approximated by the scheme shown in figure 5.8b. which simply creates a Bayesian network from the sub-models.

Clearly it is possible to say that the mutual information between the hidden variables and the classifier variable will be high and that therefore most standard mutual information based construction techniques (Cheng et al. [2002] for example) might arrive at a similar configuration. It is also possible to view this clustering method as an extension to that of Kononenko [1991] and it would certainly be possible to combine the models in a naive Bayesian network configuration as shown in fig 5.8a. The benefit of the arrangement shown in figure 5.8b is that any dependencies between the variable clusters can be expressed to a certain extent.

Jakulin [2005] takes an approach which he suggests is similar to ‘boosting’. He performs a greedy hill climb, adding interacting variable clusters one at a time until no further performance advantage is gained. At each stage the actual classifier is the Bayesian model average of all the so far added clusters. He appears to restrict himself to using interactions of order 2 first (i.e.,  $I(A_i; C)$ ) before considering higher orders ( $I(A_i; A_j; C)$  then  $I(A_i; A_j; A_k; C)$ ). Although he applies no arbitrary limit to the interaction order he notes that instances of interactions of higher than order 4 are ‘relatively rare’ in real world data sets and does not report any results using higher order interactions.

The practical considerations of maximum conditional probability table size still apply and if the number of variable clusters is too great and result in an impractical conditional probability table then an extended scheme will be required. Ultimately, in any configuration

except naive Bayes, there is an issue with conditional probability table size if a large number of evidence variables must be within the Markov blanket of a classifier. An obvious extension to the current scheme would be to cluster the models on the same basis as the original evidence variables were clustered and form a new (second) layer of hidden variables. A second layer of hidden variables has been used in the hand constructed networks in chapter 6 and these perform well at the classification task, however the machine learned structures only have a single layer of hidden variables.

### 5.5.3 Terminating the construction process

The process of adding variables to a Bayesian classifier needs to be stopped before any of the following occur:

1. Excessively redundant variables are added. Variables which are duplicates, or near duplicates, of others in the classifier reduce its performance, particularly if the data set is noisy (Langley and Sage [1994]).
2. Irrelevant variables are added. Variables making insignificant contribution to the classification reduce the performance of a classifier in the same manner as redundant variables but need a different measure to prevent their inclusion.
3. The practical limits on network representation (memory requirement) or evaluation are exceeded.

The first of these conditions is dealt with in the algorithm above, which limits redundancy while improving the chances of including variables which have an influence on classification. The second and third conditions are likely to be mutually exclusive, if a maximum practical size is reached while all variables are contributing significant information then a simple process of discarding the least significant (however significant they are) will be required. If a practical size has not been reached then a measure of what constitutes significant information is required to terminate the construction of the classifier.

Examination of the distribution of interaction information values can provide an estimate of where to set a threshold below which information supplied by variables or variable pairs can be thought to be insignificant. Figure 5.15 shows a distribution from an interaction information table similar to that in table 5.5 although using more variables, and also a Gaussian distribution fitted to the first two points in the distribution of interaction information values.

Making the assumption that noise in the data set will cause the majority of interaction information values that are effectively zero to be distributed as a Gaussian centred on zero a confidence measure for the interaction information being non-zero can be derived. Variables that are contributing below this level can then be discarded. In the graph shown interaction information values above 0.5 are not due to noise with 95% confidence, those above 0.6 with 99% confidence.

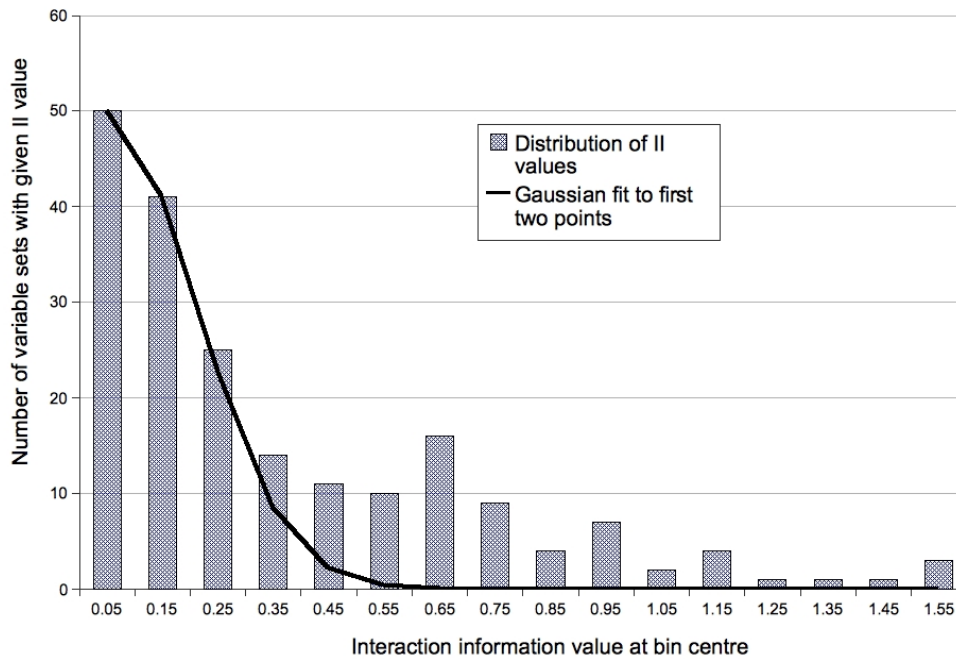


Figure 5.15: Distribution of interaction information variables for a complex network

The graph reflects the procedure as currently implemented, with absolute values of interaction information being used. More generally it might be preferable to use the signed value of interaction information. A slightly better technique would be to use the Gaussian whose variance is the variance of the entire set of interaction values. This will lead to a more pessimistic, but probably more reliable, estimate of confidence for the significance of an interaction.

This procedure works better with interaction information than with conditional or conditional mutual information because there are typically fewer genuine interactions: redundant values for mutual information do not appear. A benefit to using a measure of this kind is that it includes an assessment of the quality of the training data-set. Further time needs to be spent looking in to this technique but it appears to be promising.

## 5.6 Data handling for the reference choice experiments

### 5.6.1 Variable derivations

The Bayesian network models used in the study use evidence variables derived from the test data set (test scenes) by fixed computational models. Details of the variables and their derivations are given here. The variables are discretised although many are continuous in nature (for instance the distance between a target and a candidate reference object). Combining continuous variables typically requires some assumption to be made about the *form* of the joint probability distribution with weighting values then being learned. The nature of the variables in this study suggested that making any such assumption would be

unsafe (see the example of target and reference object sizes in section 3.2). The discretisation process also introduces limitations and potential errors in the model and these are investigated in section 5.6.2. The abbreviations of the variable names have been included to make the graph and figure legends used in chapter 6 a manageable size.

### Classification variable

The classification or ranking variable is referred to as ‘Reference suitability’ (refSuitability). During training the value is given by the assessment of the human participants in the validation exercise or by the author. As described in section 4.5.1, up to three objects in a given test case are decided to be suitable. This is because there may well be more than one effective reference object for a given target. In the case of assessment of reference suitability by human participants in the validation exercise a ‘suitable’ reference must be chosen by more than one participant. Although some references may be ‘more suitable’ or ‘more effective’ than others at locating the target, no account is made for this during training as yet.

During testing the value of this variable is determined for each candidate reference object by the network and the reference with the highest suitability value is returned as the model’s choice of reference.

### Variables related to the distance between objects

All objects are given ‘life-like’ dimensions in the scene definitions (a person is likely to be about 1.8m tall for instance) and ‘life-like’ object separations largely follow from this. However the machine model uses scaled versions of these dimensions to enable scenes of very different scales (from table top to street scale) to be conveniently combined in a single model. Although this seems intuitive there is an implicit assumption that humans scale distance with respect to overall distances in a scene rather than (say) with respect to significant object sizes. Whether this assumption is valid is investigated in section 6.6

Five logarithmically spaced bins are used for discretisation of the distance measures. The bin boundaries are:  $0.02S, 0.05S, 0.1S, 0.2S$  where  $S$  is the length of the diagonal of the scene bounding box (in metres).

Three variants of distance measure are illustrated, in 2-dimensional form, in figure 5.16. They are defined as follows:

1. Proximal distance (ProxDist). The distance between the closest points on the candidate reference object and target object.
2. Centroid distance (CentDist). The distance between the centroids of the candidate reference and target objects.
3. Asymmetric distance (AsymmDist). The distance between the centroid of the target object and the closest point to the target on the candidate reference object, as proposed by Gapp [1995a].

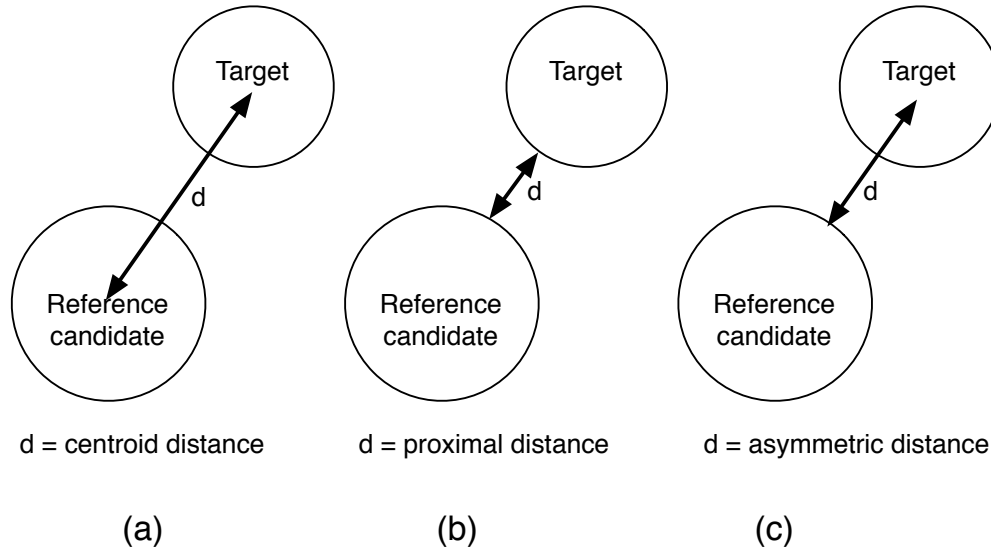


Figure 5.16: 2-dimensional representation of the different distance measures

### Variables related to the angular relationship between objects

1. Proximal Latitude (ProxLat) The angle of the proximal vector to the horizontal plane, range  $[-\frac{\pi}{2}, \frac{\pi}{2}]$ . If the target and reference objects are touching and the proximal vector degenerates to a point the angle of the vector between the point of contact and the centroid of the target object is used. The variable is defined for all values of longitude. There are 9 values with the bin boundaries placed as follows:  $-\frac{7\pi}{16}, -\frac{5\pi}{16}, -\frac{3\pi}{16}, -\frac{\pi}{16}, \frac{\pi}{16}, \frac{53\pi}{16}, \frac{5\pi}{16}, \frac{7\pi}{16}$ . Thus the bins are not all the same size, however single identifiable bins exist for the directions of the cardinal axes.
2. Proximal Longitude (ProxLong) The angle of the proximal vector in the horizontal plane relative to the scene ‘camera’ position. Range  $\frac{\pi}{8}, \frac{3\pi}{8}, \frac{5\pi}{8}, \frac{7\pi}{8}, \frac{9\pi}{8}, \frac{11\pi}{8}, \frac{13\pi}{8}$ . Although this should make no difference to the system, zero longitude (the centre of the bin  $[\frac{15\pi}{8}, \frac{\pi}{8})$  is equivalent to the target object being directly left of the candidate reference object, and  $\frac{\pi}{2}$  has the target in front of the reference. The variable is defined for all values of latitude, if the target is directly above or below the reference a random joggle is applied to the horizontal position to define an arbitrary longitude. This variable does not take into account the possibility of intrinsic reference frames and its definition is illustrated in figure 5.18a. The camera position is located at the centre of the OpenGL viewport which is the centre pixel of the window as viewed on the screen. This should equate to the speakers viewpoint.
3. Reference frame aware proximal longitude (ProxLongRFA) This variable has the same bin organisation as Proximal longitude, however for candidate reference objects with an intrinsic reference frame (for instance a car or a person) the angle is referenced to the orientation of the candidate reference object not the camera position. The

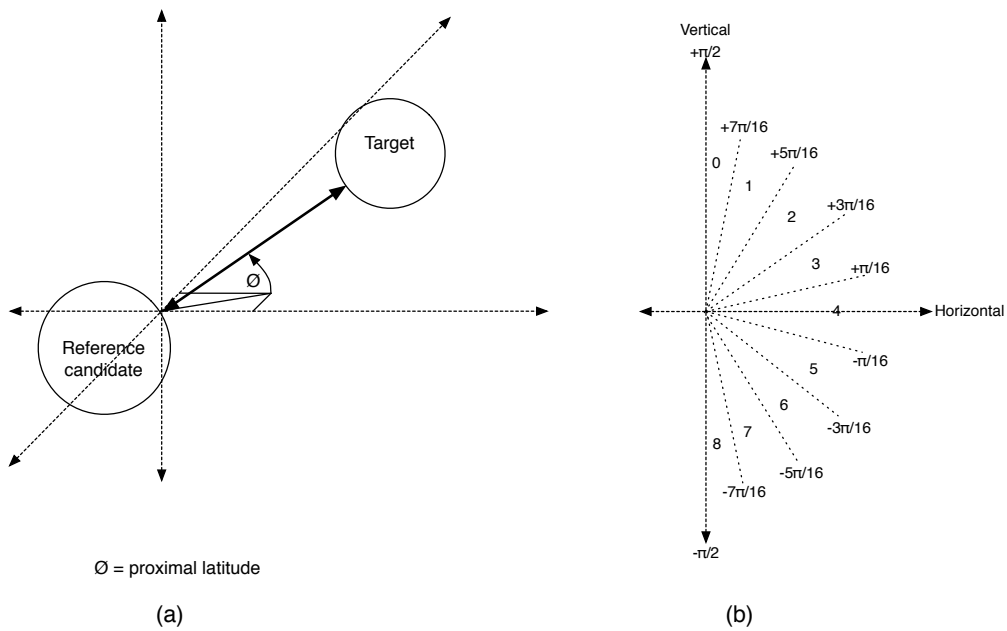


Figure 5.17: Schematic of proximal latitude definition: (a) The angle is measured from the horizontal plane to the proximal vector, (b) The arrangement of bins for discretisation.

presence of an intrinsic reference frame is given in the object definition in the test data set, it is not learned by the system. For objects without an intrinsic reference frame this variable reverts to the non-reference frame aware value. The derivation of longitude for an object with an intrinsic reference frame is shown in figure 5.18b.

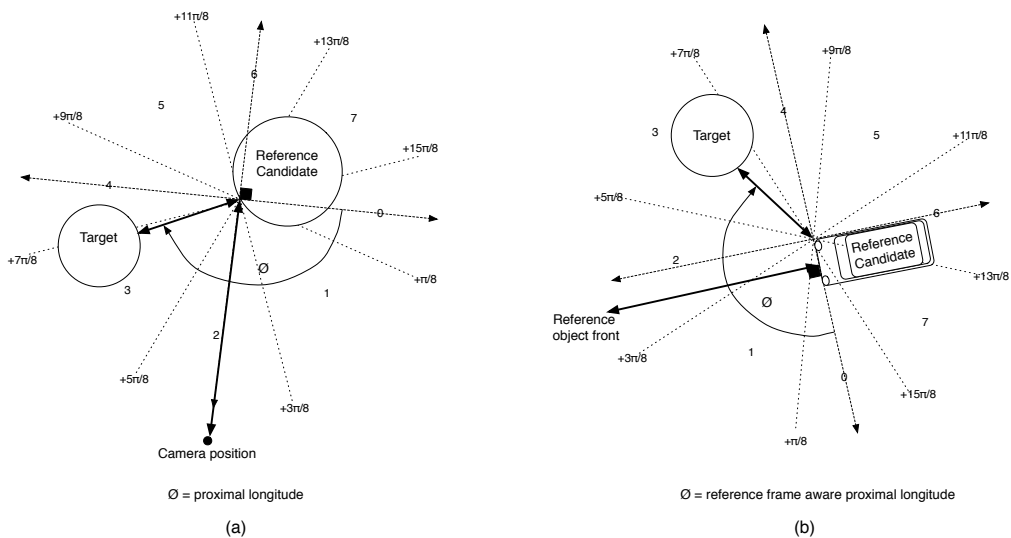


Figure 5.18: Schematic of proximal longitude definitions with bin arrangements: (a) If the intrinsic reference frame is ignored the vector to the camera position and the proximal vector define the angle. (b) If an intrinsic reference frame is available this determines the angle and the camera position is ignored.

### Variables related to object size

1. Bounding box volume (BbVolRef, BbVolTarget). For both the target and candidate reference objects, the volume of the bounding box of the object. The bounding box is minimal and does not vary with the object's orientation with respect to the  $x, y, z$  axes of the overall scene. On this measure a fruit-bowl will be bigger than the apple inside it. 8 values are used with bin boundaries as follows:  $0.025k, 0.1k, 0.4k, 1.6k, 6.4k, 25.6k, 124.0k$ , where  $k$  is  $\frac{\text{sceneVolume}}{16,000}$ . The sceneVolume measure is the cube of the scene bounding box diagonal (*metre* 3), the 16,000 constant simply gives more or less manageable numbers for the bin boundaries. As an illustration an object with a volume of about a third of a litre (a teacup say) is about on the lowest bin boundary in a room-scale scene measuring 4 metres by 4 metres by 2.5 metres.
2. Convex hull volume (HullVolRef, HullVolTarget). For both the target and candidate reference objects, the volume of the convex hull of the object. On this measure a fruit-bowl will usually be bigger than the apple inside it. The bin organisation is the same as for Bounding box volume.
3. Material volume (MaterialVolRef, MaterialVolTarget). For both the target and candidate reference objects, the volume of 'material' in the object. On this measure a thin walled fruit-bowl might be smaller than the apple inside it. The bin organisation is the same as for Bounding box volume.
4. Maximum dimension (MaxDimRef, MaxDimTarget). This is the maximum dimension of the object's minimum bounding box. As with BBVol it is independent of orientation. 8 values are used with bin boundaries as follows:  $0.005S, 0.01S, 0.025S, 0.05S, 0.1S, 0.25S, 0.5S$  where  $S$  is the length of the diagonal of the scene bounding box (metre).
5. Minimum dimension (MinDimRef, MinDimTarget). The minimum dimension of the object's minimum bounding box. The bin organisation is the same as for Maximum dimension.

### Variables related to object topological relationships

1. Bounding box topological relationship (BbTopology). This is an RCC type (see Randell et al. [1992]) measure of the topological relationship between bounding boxes. Using this variable an apple in a bowl would overlap or be a proper part of the bowl. It is 6 valued as follows;
  - (a) Objects are Separate (DC).
  - (b) Objects are touching (EC).
  - (c) Objects are overlapping (PO).

- (d) Reference proper part of target (PP).
  - (e) Target proper part of reference (PPI).
  - (f) Objects are identical (bounding box vertices) (EQ).
2. Convex hull topological relationship (HullTopology). This is the same as for the bounding box topology except that the boundary of the convex hull of the object is used. The 3-dimensional convex hull is calculated using the Qhull algorithm due to Barber et al. [1996]. Using this variable an apple in a bowl would overlap or be a proper part of the bowl.
  3. Material topological relationship (MaterialTopology). This is the same as for the bounding box topology except that the boundary of the ‘material’ of the object is used. Using this variable an apple in a bowl would be touching the bowl or separate if it were on top of other apples.

### Variables related to object characteristics

1. Mobility (MobilityRef, MobilityTarget). This is a measure of the number of times an object’s location changes within a scene series, as a fraction of the number of times it is seen. An object’s appearance in or disappearance from the scene is not included in the measure. It records ‘likelihood of the object to move’, the distance moved, either in absolute terms or relative to the size of the scene or the object, is not recorded. 5-valued with bin boundaries: 0.02, 0.05, 0.1, 0.2. So for example an object that appeared in 10 scenes and changed its position 3 times would have *mobility* = 0.3 and would be in the highest mobility category.
2. Ambiguity (Ambiguity). Currently this is simply a count of the number of objects in the scene with the same name. Some potential references are disambiguated with a simple single adjective (usually size or colour) so for instance a ‘blue house’ and a ‘green house’ are not 2 instances of ‘house’. The variable takes 3 values: 1 instance, 2 instances, more than 2 instances.
3. Reference Name Length (NameLength). The number of characters in the object’s name, only applies to candidate reference objects. This may not be the best representative of the communication cost of an object due to the time taken to enunciate it, but it will have a strong correlation with any such variable.

### Variables derived from ray casting

In the current study 10,000 rays are ‘cast’ in to each scene at random. Each ray is effectively a vector from the camera position which as noted is ‘at the centre’ of the window as seen on the computer screen. This is illustrated in figure 5.19, where the cartesian components of the ray direction  $R = (x, y, z)$  are given by:



$$x = randX * \sqrt{(\tan \phi)^2 - \left(\frac{viewPortY}{2}\right)^2} \quad (5.39)$$

$$y = randY * \sqrt{(\tan \phi)^2 - \left(\frac{viewPortX}{2}\right)^2} \quad (5.40)$$

$$z = 1 \quad (5.41)$$

where  $randX, randY$  are random numbers uniformly distributed between 1 and  $-1$ . Each ray being generated with a new  $randX, randY$ .

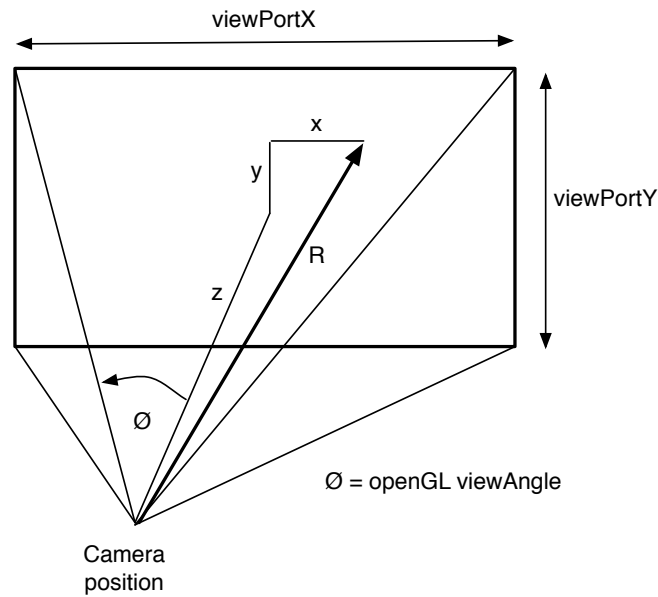


Figure 5.19: Derivation of the components of a ray used in generating ray casting variables

The point at which a ray intersects the first object in the scene along its length progressing from the camera position is recorded. If the ray does not intersect any objects in the scene it is not used, but is not replaced, so for a sparse scene there will be fewer than 10,000 intersections.

1. Viewability (Viewability). This is the number of times an object is intersected by a ray. As noted large but mostly obscured objects and objects only partially in the field of view will have low viewability. The 8 bins are approximately logarithmically spaced with boundaries: 5, 10, 25, 50, 100, 250, 500.
2. Sight-line salience measure (salienceSL). This is an analogue rather than a direct equivalent of Kelleher's salience measure [Kelleher and Costello, 2009] because in this study the centroid of the target object is not always in the centre of the field of view (in fact very rarely, the target can be anywhere in the scene). So Kelleher's weighting

of pixels (or, in this study, ray intersections) by the fraction of their distance between the centre and the edge of the field of view is not possible.

Instead using  $\overrightarrow{VH}$  as the vector from the viewpoint to the intersection point between the sight-line and an object ('hitpoint'), and  $\overrightarrow{VT}$  as the vector from the viewpoint to the target centroid then;

$$hitAngle = \arccos \frac{\overrightarrow{VH}}{\|\overrightarrow{VH}\|} \cdot \frac{\overrightarrow{VT}}{\|\overrightarrow{VT}\|} \quad (5.42)$$

defining  $F(hitAngle)$  as:

$$F(hitAngle) = \begin{cases} 1 & \text{if } \sin(hitAngle) < \sin(viewAngle) \\ 0 & \text{otherwise} \end{cases} \quad (5.43)$$

which is necessary as target objects are not always in the centre of the scene. The 'viewAngle' is the OpenGL view angle (i.e., the projection angle). The sight-line salience (*salienceSL*) is then ;

$$salienceSL(Obj) = \sum_{hitpoint \in Obj} F(hitAngle) \left( 1 - \frac{\sin(hitAngle)}{\sin(viewAngle)} \right) \quad (5.44)$$

hopefully without needing a complex formal definition for  $hitpoint \in Obj$ , this is simply summing over all intersections between rays and the given object.

3. Proximal salience (proxSalience). An issue with the usefulness of Kelleher's salience measure is that it is a measure derived from the projection of a 3-dimensional scene onto a 2-dimensional surface, whereas humans have a depth perception faculty and an ability to judge distances in a 3-dimensional space. A version of the salience measure which combines a candidate reference object's viewability and the distance between the candidate reference and target objects in 3-dimensions is defined as follows:

$$F(hitDistance) = \begin{cases} 1 & \text{if } \|\overrightarrow{HT}\| < sceneBBdiag/2 \\ 0 & \text{otherwise} \end{cases} \quad (5.45)$$

Where  $\overrightarrow{HT}$  as the vector from the hitpoint to the closest point on the target. This is necessary as target objects are not always in the centre of the scene. The 3-dimensional proximal salience (*proxSalience*) is then;

$$proxSalience(Obj) = \sum_{hitpoint \in Obj} F(hitDist) \left( 1 - \frac{2 \cdot \|\overrightarrow{HT}\|}{sceneBBdiag} \right) \quad (5.46)$$

4. Centroid salience (centSalience). This measure simply substitutes the distance between the hitpoint and the centroid of the target for  $\overrightarrow{HT}$  in the definition for proximal salience.

5. Proximal salience Squared (proxSalienceSqr). A problem with compound variables such as the salience variables described is that there is an implicit assumption of the nature of the combining function. To assess whether this might be a significant issue another measure can be defined as follows:

$$proxSalienceSqr(Obj) = \sum_{hitpoint \in Obj} F(hitDist) \left( 1 - \left( \frac{2 \cdot \|\vec{HT}\|}{sceneBBdiag} \right)^2 \right) \quad (5.47)$$

Here the distance is squared meaning that the value of the measure will fall off more quickly as the distance between the objects increases. This version uses the distance between the hitpoint and the closest point on the target object.

6. Centroid salience squared (centSalienceSqr). The same as proximal salience squared except the distance to the centroid of the target is used.
7. Search Distance (searchDist). This is the average distance of a ray intersection on the candidate reference object to the centroid of the target object. The bin organisation is;  $1.0k, 2.0k, 4.0k, 8.0k, 16.0k, 32.0k, 64.0k$  where  $k$  is  $\frac{512}{S}$  and  $S$  is the diameter of the scene bounding box.
8. Weighted Search Distance (wgtSearchDist). This is the same as the search distance except that the distance to each ray intersection on the candidate reference object is weighted by the ratio of; the distance from the intersection to the camera point divided by the distance of the target centroid to the camera point. This variable tries to capture the effect of a search being more time consuming as the area being searched moves further away. The target will appear smaller the further away it is and it might be expected that a human visual search process will acknowledge this. The bin organisation is  $1.0k, 2.0k, 4.0k, 8.0k, 16.0k, 32.0k, 64.0k$  where  $k$  is  $\frac{512W}{S}$  and  $W$  is  $\frac{distanceHitpointCamera}{distanceTargetCamera}$ .
9. Weighted Search Distance Squared (wgtSearchDistSqr) This is the same as the weighted search distance but takes into account the fact that the visible area of the target object will fall as the square of the distance from the camera position. The bin organisation is the same as for weighted search distance except that  $W^2$  is used instead of  $W$ .

### Variables related to the listener in a scene

1. Listener is potential reference (listenerRef). This is a Boolean variable which is 1 if the listener is the candidate reference object (that is, the reference in a sentence such as “The bowl is behind you”) and 0 otherwise.
2. Listener distance to reference (listenerTargetDist). This is the distance between the centroid of the listener and the centroid of the target. The scaling and bin organisation are the same as for the target to reference distance measures, except that a ‘switch’ value is added which is set to 1 if there is no listener in the scene.

3. Listener distance to target (`listenerRefDist`). Again this is the centroid distance between the listener and the candidate reference object. The scaling and bin organisation are the same as for the listener distance to the target measure.
4. Listener to target longitude (`listenerTargetLong`). This is the angle from the listener to the target using the listener's intrinsic reference frame. It is calculated in the same manner and has the same bin organisation as reference frame aware longitude. Again, although it should not matter, a listener to target longitude of 0 means the reference is to the listener's left and  $\frac{\pi}{2}$  has the target in front of the listener. The switch value is again present.
5. Listener to reference longitude (`listenerRefLong`). Identical to the listener target longitude except that the candidate reference object is the object in question.

### Variables related to scene scale

1. Scene scale (`sceneScale`) This is currently a binary variable whose value is 1 if the scene bounding box diagonal is  $> 20m$  and 0 otherwise.

### 5.6.2 Discretisation

The discretisation of the variables into value ranges is undertaken manually. Although machine learning of bin values is possible it was thought that, in this study, this might lead to over-fitting to the test data and therefore to spurious results. As an example in the test data set there are likely to be some objects that are quite often useful as references, houses perhaps, that are similar in size. A bin learning technique might place bin boundaries to bracket the typical size range of the houses in the data set. This arrangement might not translate well to an extended data set containing fewer houses.

The discretisation process for the continuous variables was undertaken by testing for an optimum or near optimum number of bins and then organising the bins logarithmically or linearly as appropriate so as to conform reasonably closely to a uniform distribution. The placement of the bins was then tested to ensure that reasonably robust results were being obtained.

As an example figure 5.20 shows the number of bins for the proximal distance variable being varied while testing the best performing network from figure 6.8. Although it might be expected that more bins would always lead to better results, this is not necessarily the case with a finite and noisy data set, as can be seen. The range of the proximal distance variable is 0 – 1 (see section 5.6.1) and the bins are centred so as to give a close to uniform distribution. In this case 5 bins were used and the system is reasonably robust to the number of bins used. The variation in the results as the number of bins is changed from 4 to 7 is less than  $\pm 0.5\%$ .

Although logarithmically spaced bins seem intuitive (at least for the distance and size variables) there is no absolute reason why this should be so. Figure 5.21 shows the results

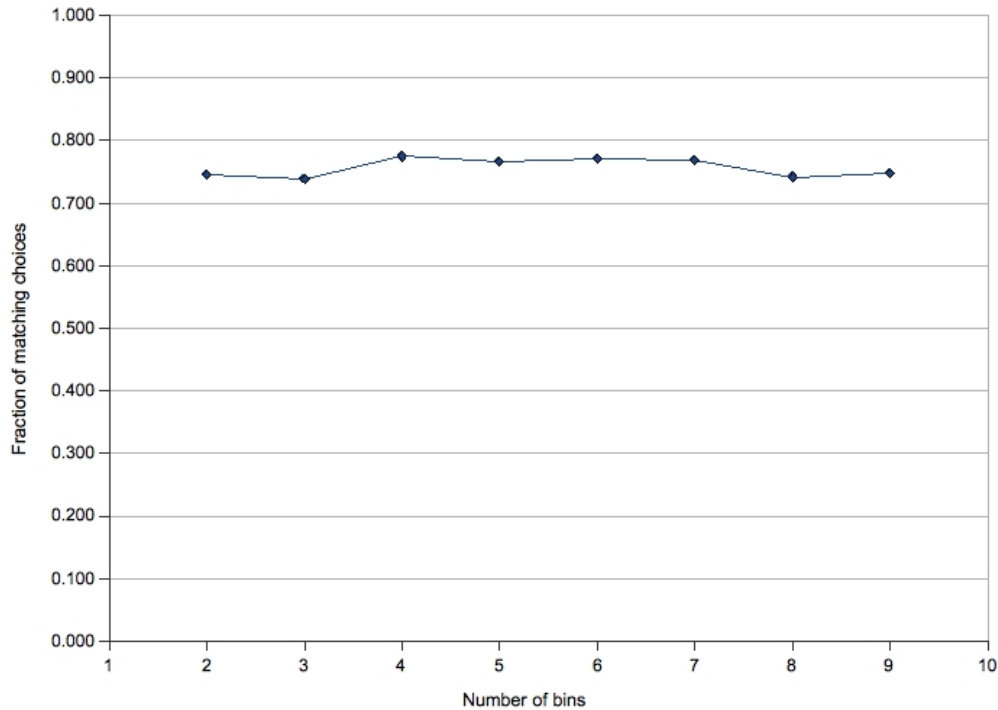


Figure 5.20: Variation in test results for the best performing network from figure 6.10 for varying the number of bins (logarithmically spaced bins)

for different numbers of arithmetically spaced bins using the same test conditions as figure 5.20. Although the bins are arithmetically spaced relative to each other they are not centred over the interval  $0 - 1$  and the distribution is not close to uniform. Figure 5.21 rather illustrates the best performance available from arithmetically spaced bins. The results for a low number of bins are lower than that for logarithmically spaced bins as might be expected however there seems to be a stable region for relatively high numbers of bins where performance is as good as for logarithmically spaced bins. The fact that the performance does not drop away as quickly for high bin numbers in this case is probably due to there being a number of genuinely redundant bins that attract a very low number of significant evidence states. The use of logarithmically spaced bins is retained in this study since good performance is obtained with fewer bins, with a resultant reduction in computational load.

Figure 5.22 shows the effect of off-setting the bin positions from the position which gives a near uniform distribution. This also illustrates the robustness of the results to the exact positioning of the bin boundaries. Although there are no apparent fluctuations for small scale movements in bin boundaries for the distance variable, the ‘plateau’ area of the graph in figure 5.22 is not as flat as might have been hoped. In this case (the distance variables) the lowest bin boundary is set at 0.02 The variation in the results is about  $\pm 2\%$ . The variation for the reference volume variable is much less and other variables should be lower still as they have less effect on the models overall.

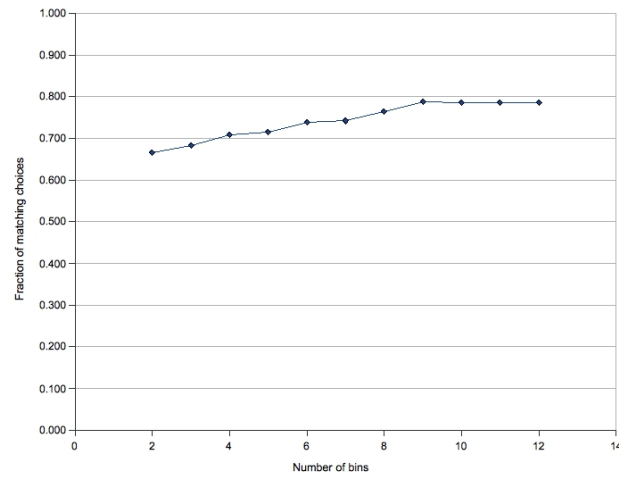


Figure 5.21: Variation in test results for the best performing network from figure 6.10 for varying the number of bins (arithmetically spaced bins)

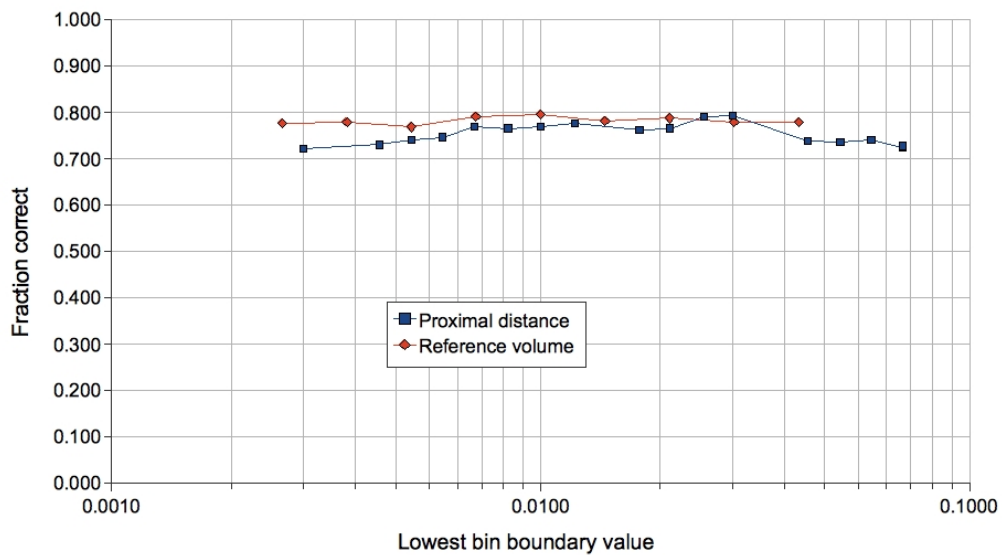


Figure 5.22: Variation in test results for the best performing network from figure 6.10 for varying bin boundary positions for the two most influential variables, reference target distance and reference volume

Although an exhaustive analysis of the manner in which discretization errors for all relevant variables combine has not been carried out, the distance variable is the single most influential variable in most networks and errors in the distance variable will tend to dominate other errors. In the light of these investigations it is thought that an error value of  $\pm 4\%$  overall should be applied to the absolute results presented in chapter 6 due to the discretization process. These errors should not effect the comparative results between different networks, which all use the same bin boundaries.

### 5.6.3 Network training

The networks constructed as described in section 5.5 or those constructed ‘by hand’ have one or more layers of nodes which are not direct evidence nodes. As noted their value for any evidence state is the same as the value of the classifier. But the weighted positions in the conditional probability table to which this value must be assigned are not necessarily known a-priori depending on the order of evaluation of the network.

The network is initialised so that all of the classifier nodes have a likelihood slightly less than the expected likelihood of any given reference being suitable (approximately 3 divided by the average number of objects in a scene). Binomial priors or other smoothing techniques are not employed at present. Examination of the results indicates that sufficient numbers of evidence cases are available for all the significant classification conditions.

The maximum number of evidence states, given the data set size of 529 test cases with an average of 26 candidate reference objects for each case is approximately 14000. If a network of 20 variables is taken as representative of a complex case then 280000 integers would be needed to store all the evidence to train the network. This is clearly well within the capacity of even a small computer. Notwithstanding this, an incremental approach to training the network has been employed with the network being updated after every scene has been processed. This has been done in anticipation of future application areas for the system in virtual environments (as part of an artificial agent in, for instance, second-life) where the data set is not bounded.

For the case of a finite data set iteration over the data set will overcome both the effect of the prior values and the order in which the data is presented, as the values are simply propagated through the network from the evidence variables to the classifier variable. This is illustrated in figure 5.23, where the number of individual results (out of a total of 369) which change from iteration to iteration is plotted as the number of iterations increases. This is for a network with two layers of ‘hidden’ variable which is the maximum used in this study.

## 5.7 Summary

The criterion for choosing the machine learning system for this study was not that it should be thought to give the absolute highest performance in terms of being able to model the reference selection problem but that it should provide satisfactory performance but be able

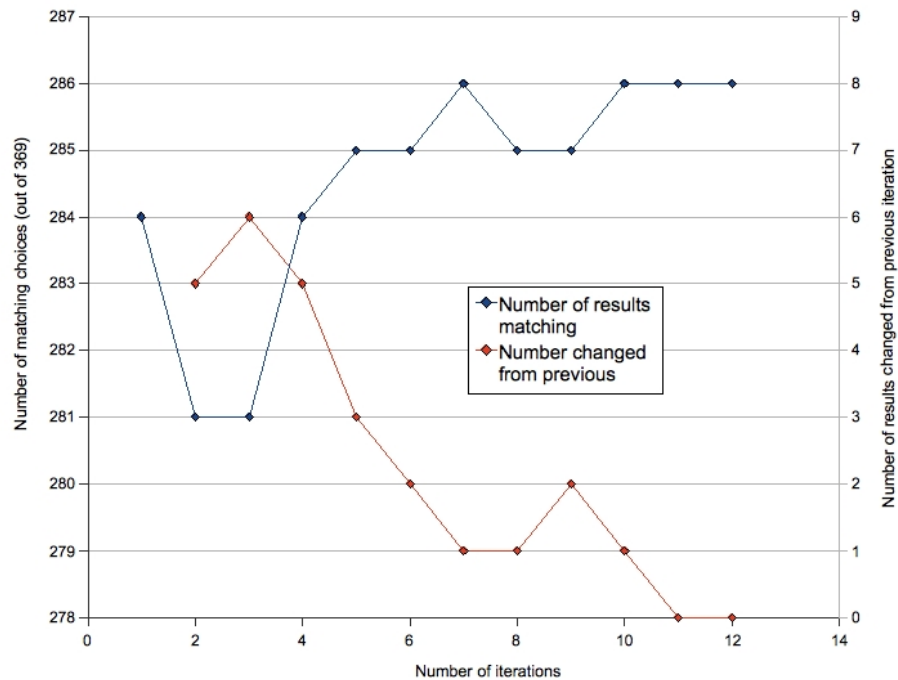


Figure 5.23: Progress of training as number of iterations of training data set increases for the best performing network from figure 6.10

to provide very good illustration of the factors and combinations of factors involved in the reference selection problem. For this reason Bayesian network based classifiers were chosen.

The development of Bayesian network classifiers from the basic ‘naive’ Bayes classifier through various forms of augmented naive Bayes classifiers, to less restrictive forms has been described, highlighting the major practical problems of Bayesian classifiers namely:

1. Maintaining computational feasibility through having reasonably sized probability tables and hence a limited number of parent variables for any given variable.
2. Given this, having all relevant attribute variables in the Markov blanket of the classifier variable, or separated from the classifier only by a hidden variable (which must somehow be defined).
3. Eliminating, as far as possible, irrelevant or redundant variables which would degrade classifier performance.
4. Capturing, as far as possible, dependencies and interactions between attribute variables that have an influence on the classifier
5. Ensuring that the size of the probability tables and detected interactions are properly supported by the size and quality of the data set.

The nature of the variables in the reference choice problem readily highlights the need to capture variable interactions and eliminate redundancy. It is anticipated that classifiers



that do not satisfactorily address this will perform badly. Although some of the currently employed Bayesian classifier construction techniques go some way towards capturing variable interactions, only that due to Jakulin [2005] does so directly using the interaction information function. A Bayesian classifier construction technique is proposed here which clusters variables according to their mutual information with the classifier, or their interaction information with the classifier and a second attribute. The clusters form partial models for reference suitability and these are combined as parents of the classifier variable which should make the resultant network more robust to redundancy in the partial models.

The list of variables, extracted from the raw geometric data in the test set, for presentation to the machine learning system, contains some 32 variants on size, distance, angular relationship, topological relationship, projected area and other measures. Most of these have been suggested by the literature reviewed in chapter 3, others have been added that seem likely to have some bearing on the reference choice issue. The (mostly continuous) variables are discretised without recourse to a machine learned or entropy based process as it was considered that this might lead to over-fitting. The discretisation process introduces some performance variability in to the machine models but this should not dramatically effect comparison between the models which all share the same discretisation.

## Chapter 6

# Results from models of reference object choice

### 6.1 Introduction

The range of variables used in the models, and the test data set (the scene corpus), have both developed during the course of the study and this complicates the presentation of the results in this chapter to some degree. To retain compatibility with already submitted papers and because some insights into factors affecting reference choice were gained during this process, the results are presented reflecting this progression, rather than simply providing a snapshot of the end-point of the study. Thus in section 6.3 only geometric variables are used in the models which are trained and tested on the first 93 scenes (series 1 scenes) only. Later sections add variables derived from ray casting and the models are trained and tested on all 133 scenes (series 1 and 2 scenes). Some of the models from section 6.3 are retested on all 133 scenes for comparison as necessary but the process of developing the models is not repeated in full.

One factor that was considered important and which is addressed through all stages of the study is the necessary degree of complexity of the model and the interactions (or dependencies) between the variables that must be captured in order to match human performance. To facilitate this models based on three different Bayesian network types are compared. These Bayesian network types have inherently different capabilities with regard to capturing variable dependencies:

1. Naive Bayesian networks cannot capture any dependencies between variables.
2. Tree augmented naive Bayesian networks can capture dependencies between pairs of variables.
3. What are here termed ‘combined clustered’ Bayesian networks and which can capture dependencies between up to four or five variables. The limit on the number of variables being a practical one, rather than theoretical one, dependent on the size of the probability tables involved and the quantity of test data.

The reasons for these different behaviours have been described in chapter 5, where it was also noted that the tree augmented naive Bayesian networks tend to capture redundancy between variables as well as genuine interactions. The combined clustered Bayesian networks, used in the first investigations, are structured by hand and tend to avoid redundancy; they follow the hypothesis model of chapter 3 as far as possible. In section 6.7 a version of combined clustered Bayesian networks with machine learned structure as described in section 5.5 is tested and compared with results from previous models.

This chapter is structured as follows:

**Section 6.2.** Describes how the machine model tests are run, how the results are presented and some notes on how they should be interpreted.

**Section 6.3.** Contains results from machine model tests using only ‘geometric’ variables, that is variables derived directly from object volume and dimension as well as variables relating to the geometric and topological relationships between objects. No variables derived from ray casting (‘sight-line’ variables) are used. These results are obtained from the initial data-set of 369 test cases and are the same as those found in Barclay and Galton [2010]

**Section 6.4.** Describes the reason for, and effect of, extending the data-set by a further 160 test cases to 529 in total. Results in this section illustrate the relationship between scene complexity and difficulty of reference object choice.

**Section 6.5.** Contains results from models incorporating sight-line variables, on their own and in conjunction with the geometric variables used in section 6.3. The results in this and following sections use the full data set of 529 test cases.

**Section 6.6.** Looks at the effect of the scale of scenes on the reference selection task. The data set is separated into large scale, exterior scenes (street and vista scale) and small scale scenes (table top and room scale).

**Section 6.7.** Introduces the learned structure version of the combined clustered Bayesian networks and compares the performance of models produced using this technique with the models already described.

**Section 6.8.** Contains results of models using variables relating to a listener, who is present in a scene, in addition to those used in earlier models.

## 6.2 Experimental conditions and analysis

Ten-fold cross validation is employed to maximise use of the training/test data-set. The data set is traversed ten times and each time a different 10% of the data-set is reserved for testing. The networks are trained on the remaining 90% of the data-set. This also yields 10 partial result values which are used to assess the statistical significance of the difference between results from different networks. Wilcoxon’s signed rank test (Wilcoxon [1945]) is used in this study as no assumptions need to be made about the distribution of the result values. Unlike the standard analysis of variance there is no requirement for the data to be normally distributed. With only ten samples in each case the assessment of normality

is difficult and a test based on the ranking of the samples rather than their values should be more robust and less likely to generate spurious significance values. Unlike the Mann-Witney-Wilcoxon rank-sum test (which is generalised in the Kruskal-Wallis test), or the related Friedman test, the Wilcoxon signed rank test makes no assumption of independence between the paired samples in the two sample groups. In this study the ten-fold cross validation uses the same groups of test cases in each of the ten sub-divisions, for all models, so independence cannot be assumed.

The calculation of the relevant statistics is illustrated in table 6.1. The two models being compared (A and B) have the ten paired results from the cross-validation test listed in the top two rows of the table. The difference between the paired values is taken and any identical results ( $A - B = 0$ ) are discarded, in this case reducing the number of samples ( $N$ ) to 9. The absolute values of the remaining differences are ranked, (from 1 to 9 in this case) with paired ranks being assigned the average rank value for the pair. In the fourth row of the table it can be seen that the three cases of  $A - B = 5$  have all been assigned rank 8. In the last row of the table the signs of the differences are restored to the ranks and the sum of this row generates the ‘ $W$ ’ number. For high values of  $N$  (greater than 10 say) a statistic is generated and compared for significance to the normal distribution. For low values of  $N$ , as in this study where the maximum sample size is 10, the probabilities are calculated directly from the sampling distribution of  $W$ . This is simply finding how many permutations of ranks, from the possible  $2^N$ , would equate to more than a given value of  $W$ , and hence the probability of this  $W$  occurring by chance. Critical values of  $W$  for the usual significance levels are shown in table 6.2, note that sometimes tables based on the smaller of the positive or negative ranks, rather than the total as in this study, are used. In the case illustrated in table 6.1, for  $W = 34$ ,  $N = 9$  there is a significant difference between the models at the 0.05 level.

Table 6.1: Illustration of Wilcoxon signed rank test calculation

	Results (number correct) from each of the 10 cross validation tests										Note
model B	27	22	24	24	26	27	24	25	21	22	<b>total 238</b>
model A	22	24	24	22	25	22	23	20	20	21	<b>total 223</b>
B - A	5	-2	0	2	1	5	1	5	1	1	<b>N = 9</b>
rank	8	5.5	0	5.5	2.5	8	2.5	8	2.5	2.5	
signed rank	8	-5.5	0	5.5	2.5	8	2.5	8	2.5	2.5	<b>W = 34</b>

The use of Wilcoxon’s signed rank test can lead to an apparent discrepancy between the figures given for significance and the performance values given in the graphs. It is possible that even if model ‘B’ out-performs model ‘A’, model ‘A’ may be significantly better than a third model ‘C’, while ‘B’ is not. This is illustrated in table 6.3 where model ‘A’ consistently but marginally outperforms model ‘C’, while model ‘B’ sometimes performs much better than model ‘C’ and sometimes a little worse.

Table 6.2: Critical values for the Wilcoxon signed rank test used in this study

N	significance levels for a directional (one tail) test				
	0.05	0.025	0.01	0.005	0.001
5	15	-	-	-	-
6	17	21	-	-	-
7	22	24	28	-	-
8	26	30	34	36	-
9	29	35	39	43	-
10	35	39	45	49	55

Table 6.3: Illustration of apparent discrepancy between significance (using Wilcoxon’s signed rank test) and model performance. Model ‘A’ outperforms model ‘C’ and model ‘B’ outperforms model ‘A’ but only ‘A’ is *significantly* better than ‘C’

Model	Results (number correct) from each of the 10 cross validation tests										Total	W	N	Significance Level	
A	31	33	33	23	23	29	26	29	28	27	<b>282</b>				
B	34	36	36	26	26	27	24	27	27	25	<b>287</b>				
C	30	32	32	22	22	28	25	28	27	26	<b>272</b>				
A - C	1	1	1	1	1	1	1	1	1	1	<b>10</b>	55	10	0.005	
B - C	3	3	3	3	3	-1	-1	-1	-1	-1	<b>15</b>	25	10	< 0.05	

For each of the cross validation passes the network is trained on 12 iterations through the data set (see section 5.6.3) which ensures that networks with hidden variables have converged satisfactorily.

Cross validation is not used for learning the model structures. The interaction information, conditional mutual information and mutual information values are calculated from the complete data set and the structure derived prior to the training using cross validation. This prevents different structures being present in the same test which would complicate comparison between structures. A possible disadvantage to this might be any over-fitting which might occur. It is difficult to envisage how this could occur in the calculation of the information values, which are integrated over the full range of the probability mass functions of the variables in question. This is to say that the data set is being effectively ‘smoothed’ in this step and specific, possibly spurious, features in the data set cannot be selected and (over)fitted to. Whether the overall model is over-fitting is discussed in section 6.9.

Object mobility is also learned from the complete data set prior to structure learning or network training. This should be taken to represent a process of acquiring object knowledge before acquisition of a model of reference choice. Whether this happens in practice is an open question, it is unlikely to make any significant difference to the results, but the models of reference suitability should be regarded as ‘models learned with prior knowledge of object

characteristics’.

In interpreting what is meant by a statistically significant result some care must be taken. What is actually true is the following:

There is an apparent significant *difference* between two machine models in matching the judgements of a small number of humans as to the suitability of reference objects, which were made for a relatively small number of schematically represented scenes, displayed on a computer screen in a clearly experimental setting.

In the following sections a short hand is used which could be read as:

This model is significantly better than others for determining reference suitability

The caveats concerning the reality of the data set and the process of obtaining the human judgements as to reference suitability cannot be repeated everywhere but must be remembered. Whether we are looking at better and worse models, or just different models, is further discussed in section 7.1.1. It should be noted that, while the machine models are matching considerably fewer group choices of reference than the median human in the group, it is probably reasonable to suggest that an increase in the number of matches represents an improvement. The median human matches about 90% of one of the top three group choices of reference.

## 6.3 Results from geometric variable models

### 6.3.1 Single variables

Initially each variable was tested singly for its influence on reference suitability, in a simple network as shown in figure 6.1. Note that when a variable is described as a good (or bad) predictor of reference suitability, this is the reference suitability as determined by human usage. Note that the results for all three topologies shown are identical when a single variable is considered. The results are shown in figure 6.2. The random baseline figure is derived by replacing the values in the network with random numbers, thus effectively selecting a reference at random. The figure is larger than three (there are usually three ‘good’ reference objects for each test case) divided by the average number of objects in a scene, which would be  $\frac{24.9}{3} = 0.120$ , due to the different number of objects in the scenes. The expected value of the random baseline, given the actual distribution of numbers of objects in each scene, would be 0.136 which is not significantly different from 0.15. Except for a few of the single variable models shown in figure 6.2, all the models perform significantly better than this chance level. It can also be seen from figure 6.14 that, for higher performing models, there is no correlation between performance and the number of objects in a scene.

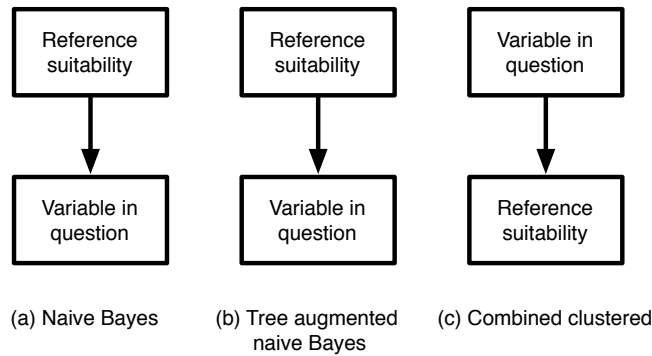


Figure 6.1: Network topology for the assessment of single variables

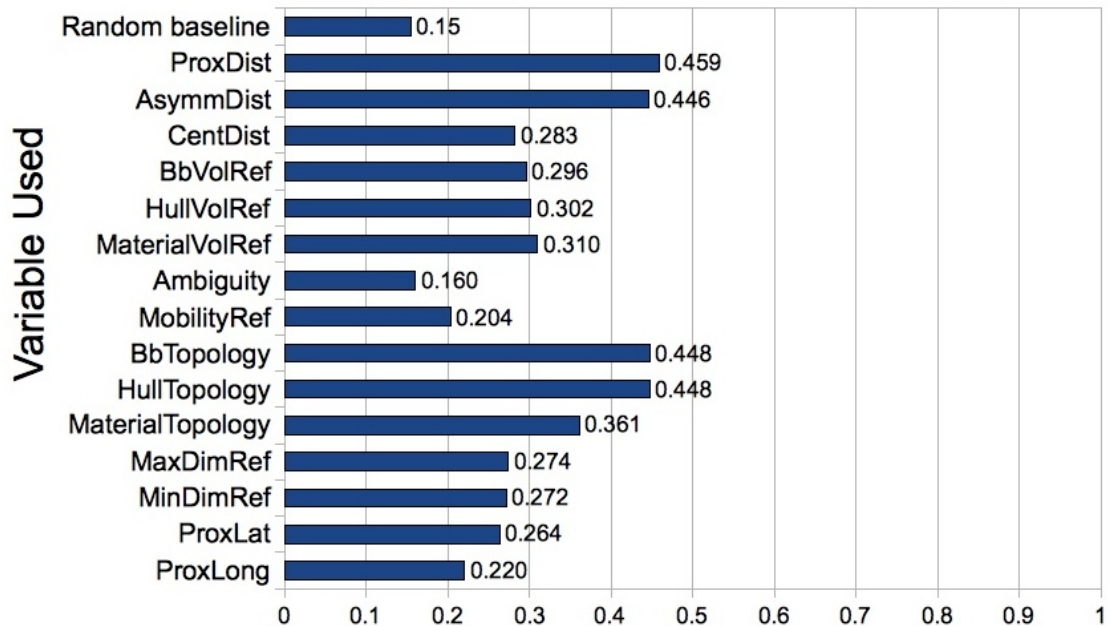


Figure 6.2: Fraction of machine reference choices matching one of the top three human reference choices when a single variable is considered

No variables relating solely to the target object (target volume measure or target mobility for instance) are relevant in this exercise. All candidate reference objects would score the same given only a parameter related solely to the target to discriminate between them. As might be expected, the best single predictor of suitability in a reference is distance from the target object, however there are significant differences between the distance measures. The difference between the predictive power of both the proximal distance (ProxDist) measure, over the centroid distance (CentDist) measure, is significant at the 0.005 level ( $W = 55$ ,  $N = 10$ ). The same is true for the asymmetric distance (AsymmDist) over the centroid distance (CentDist) measure. The asymmetric distance is apparently worse than the proximal distance but there is insufficient data to establish this as a significant result. Because they

are very similar and to reduce clutter in the results only the proximal distance measure is used in the remaining experiments.

The topological relationship between the target and the reference is also a good predictor of reference suitability. This seems intuitive given the topological nature of the prepositions ‘in’ and ‘on’, however it may also be because, in a single variable model, it has a correlation with distance between reference and target and is acting as a ‘proxy’ for distance. This would also explain why the ‘looser’ topological relationships relating to convex hulls (HullTopology) and bounding boxes (BbTopology) which can express disjunction, touching, overlap or containment are a better predictor than the material topological relationship which in this model can only be disjoint or touching.

No statistically significant differences exist between the different measures of reference object volume (materialVolRef, hullVolRef, BbVolRef) when considered as single variables.

The fact that a variable does not predict reference suitability well in this exercise does not mean that it will not be relevant in a more expressive model of course. The key example of this will be seen to be the ambiguity measure.

### 6.3.2 Relative size and distance models

From the results of various studies noted in sections 3.3.3 and 3.4.3. as well as intuition, it seems natural to look at combinations of size and distance measures as predictors of reference suitability. To restrict the combinations of variables tested only equivalent volume measures for the target and candidate reference objects are used. The role of the minimum and maximum dimension measures, which give indications of the geometric extension of objects, is covered in the following section.

The network topologies are shown in figure 6.3. The results for the different topologies are shown in figure 6.4 and as can be seen the different topologies now give different results.

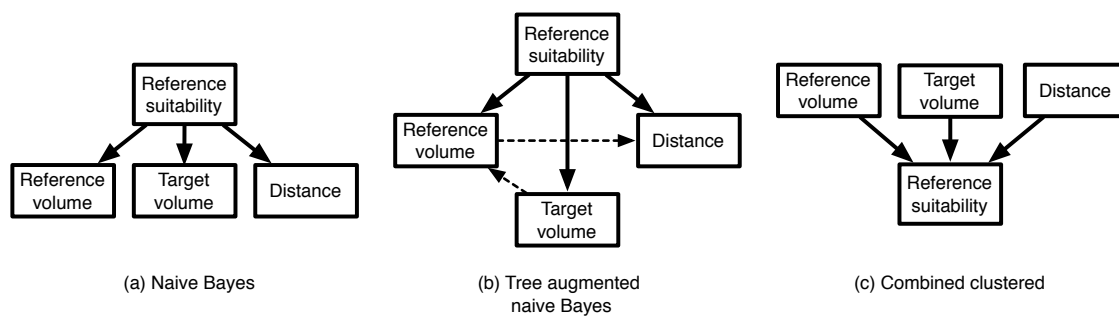


Figure 6.3: Network topologies for the assessment of size and distance variables

At this level of model complexity using ‘relative volume’, that is target as well as reference volume measures, does not offer any statistically significant improvement in prediction of reference suitability over using reference volume alone. Combining relative volume and the proximal distance measure results in an improvement significant at the 0.005 level ( $W = 55$ ,  $N = 10$ ) over using distance alone.



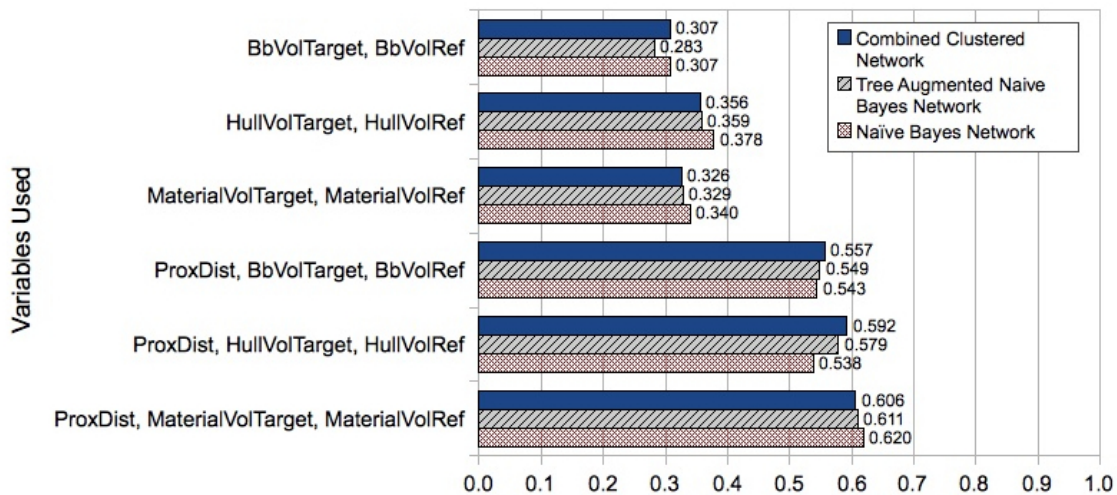


Figure 6.4: Fraction of machine reference choices matching one of the top three human reference choices when reference and target size and target/reference distance are considered

There appears to be no significant overall performance difference between the network topologies and this suggests that the variables used are statistically independent. In this simple case the combined clustered network model should at least equal the performance of the other network types as it contains the full joint probability table. In simple models, however, there is an increased chance of several references being equally suitable. The ‘best’ reference is selected at random from those of equal suitability if this is the case, leading to the possibility of the full joint probability table not giving the best performance. The significance test will account for this random variability. Also there is the possibility that there are some errors due to insufficient test data. The combined clustered model contains the largest conditional probability table and the test data may not have addressed all entries in the table satisfactorily.

Note that although more than 60% of test cases can be correctly matched by the machine to the human reference choice, using these three variables, this is still worse than the most ‘idiosyncratic’ human participant in the validation tests.

### 6.3.3 BaseSet plus one (four variable) models

The use of size and distance variables to identify a good reference is intuitive and accepted by most commentators, so as a step towards building a more complex model this is taken as a baseline variable set (BaseSet) and further variables are added to see if they significantly improve prediction of reference suitability. To give the strictest test of significance the best performing variable set from figure 6.4 is used as the BaseSet, namely proximal distance (proxDist), target material volume (materialVolTarget) and reference material volume (materialVolRef). The other variables are each added to this set to give a series of 4-variable models. The BaseSet variables should ‘filter out’ many obviously poor reference candidates and allow the remaining variable a role in distinguishing between the references

that remain.

The Bayesian network topologies for the different network types are shown in figure 6.5. The full joint probability is still contained in the combined clustered model. The naive and tree augmented naive Bayes networks are simple extensions of the three variable case.

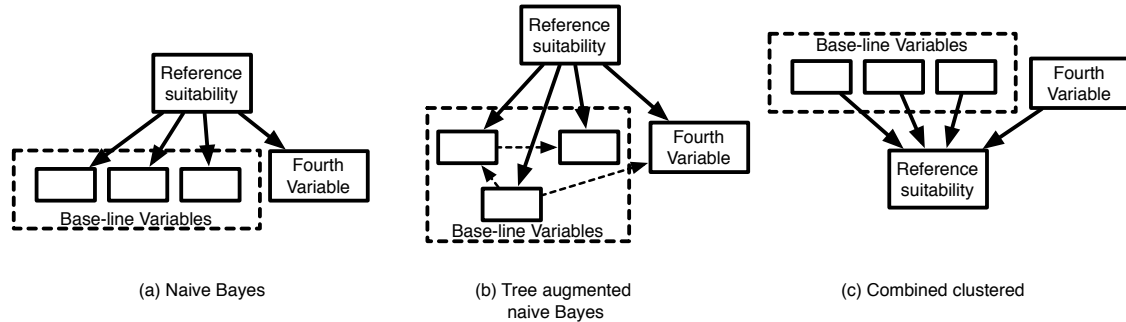


Figure 6.5: Network topologies used for assessing four variable models

The results for all four variable models are shown in figure 6.6. The performance of the different network topologies appears to be diverging at this point. For the variables in which there is a significant improvement in performance over the BaseSet only case, the combined clustered and tree augmented naive Bayes networks perform considerably better than the naive Bayes network. In all but one case the combined clustered network performs slightly better than the tree augmented naive Bayes network. This suggests that the ability to represent statistical dependence between the variables is important at this level of model complexity. For cases where a variable does not improve the performance of the model in predicting reference suitability the combined clustered model often under-performs the naive Bayes variants. This seems to confirm the expected effect that performance of a network will degrade as the size of the joint probability table approaches the size of the data-set.

The variables which, when combined with the BaseSet variables, give a significant improvement in prediction of reference suitability over the BaseSet are listed in table 6.4. It is clear that reference ambiguity and measures of reference geometric extension are playing an important part in reference selection. Measures of target extension, reference mobility and topological relationship between reference and target also seem important. It is perhaps surprising that the ‘intuitively’ important topological relationship, between the reference and target convex hulls (HullTopology), which most accurately models containment and support relationships, is not as significant as the other topological variables. Note that although convex hull topology appears to perform better than the material topology in figure 6.6 this is not reflected in the significance test due to the distribution of the cross-validation results (see section 6.2). The angle between the target and reference objects is not a significant influence in the four variable model.

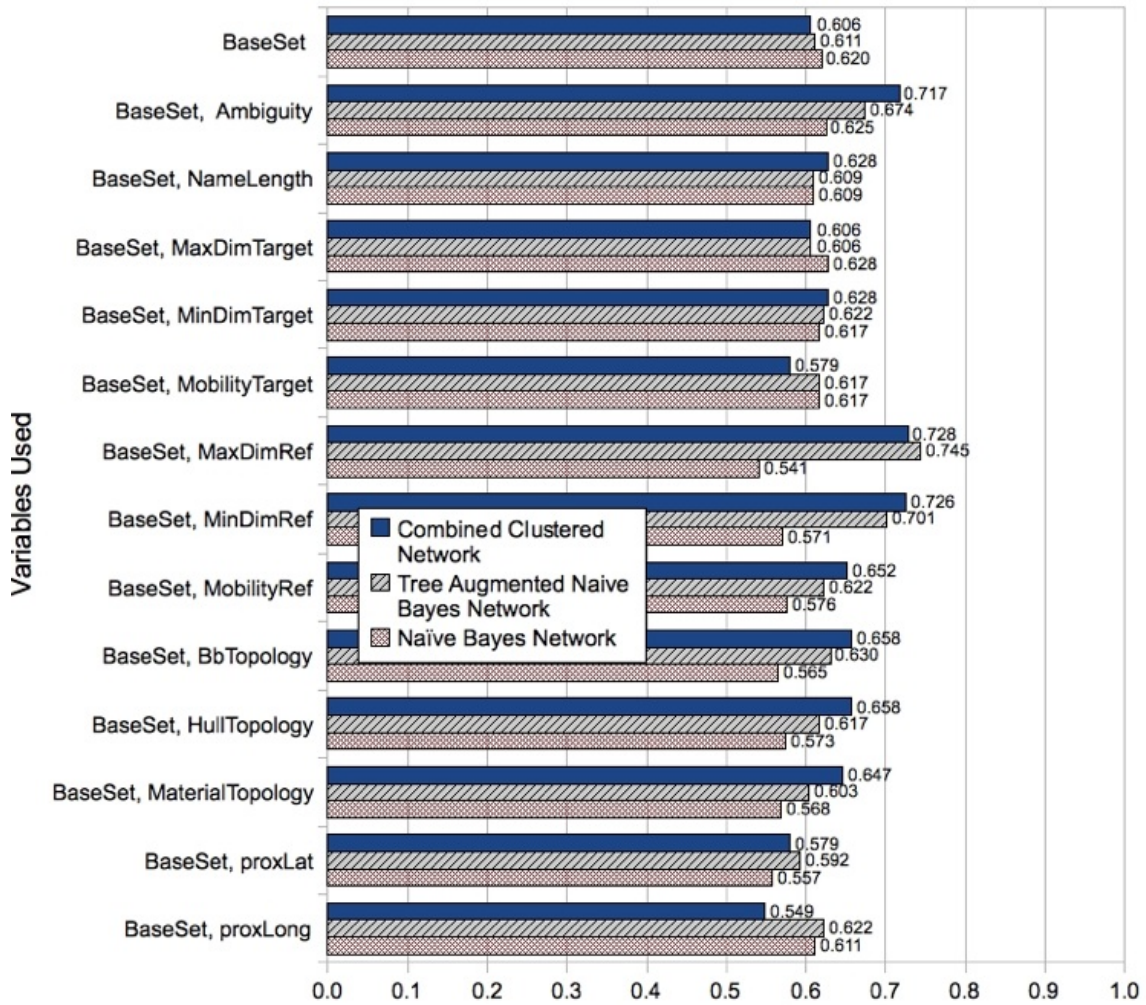


Figure 6.6: Fraction of machine reference choices matching one of the top three human reference choices when relative size, distance and a further variable are considered

Table 6.4: Significance of variable to improvement in reference choice prediction over BaseSet variables

Variable	Abbreviation	Significance Level	W	N
Ambiguity	Ambiguity	0.005	52	10
Target Minimum Dimension	MinDimTarget	0.025	32	8
Reference Maximum Dimension	MaxDimRef	0.025	39	9
Reference Minimum Dimension	MinDimRef	0.010	36	8
Reference Mobility	MobilityRef	0.025	36	9
Bounding Box Topology	BbTopology	0.025	44	10
Material Topology	MaterialTopology	0.025	32	8

### 6.3.4 More complex models

The next step in the development of the model is to combine all of the variables that were significant in the four variable models and combine them into a more complex model.

This is straightforward for the naive and tree augmented naive networks which are simple extensions of their four variable topologies. For the combined clustered model however the network topology needs to be altered to prevent the size of the joint probability tables becoming too large with respect to the data-set, or too computationally intensive.

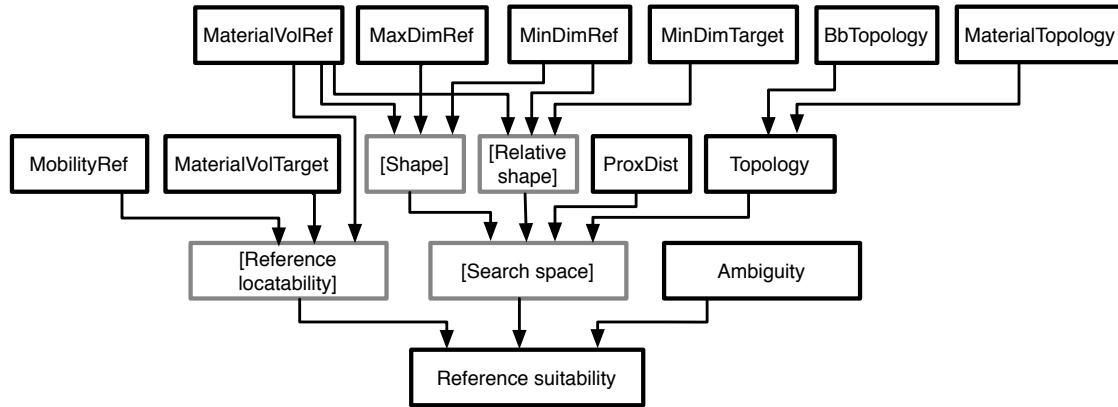


Figure 6.7: The combined clustered network topology for a more complex model

The network topology, which is designed, not machine derived, is shown in figure 6.7. It attempts to follow the rationale given in chapter 3 as far as possible, given the variables which appear significant in the four variable model. The nodes in grey are not variables derived from the test scenes but are ‘hidden’ variables with the suggested names as given in square brackets. Thus the particular arrangement of the model assumes that the locatability of a potential reference is determined by the volume of the target and reference objects along with the mobility of the reference. The space in which the listener must search for the target is determined by the shape of the reference and the ‘relative shape’ of the reference and target, the distance between the reference and target and the topological relationship between them. The ‘meaning’ of the [shape] variable could be described as ‘for a given size of reference there are optimum and less optimum geometric extensions for locating a target’. Similarly the meaning of the [relative shape] measure could be described as ‘for a given size of reference there are optimum or less optimum relative geometric extensions for the target and reference objects’. An example of this would be an extended target object such as a pencil not being well referenced by an eraser, but possibly being well referenced by a ruler of the same volume as the eraser. The ambiguity variable is the sole contributor to communication cost.

Results are shown in Figure 6.8 along with the result for the best 4 variable network (BaseSet maxDimRef) The results show the case where all the significant variables from table 6.4 are combined with the baseline variable set and for the cases where each of them are individually removed from the model. None of the results from the tree augmented naive Bayes or combined clustered networks are significantly better or worse than the best four variable model.

The naive Bayes network significantly under-performs the others at this level of model

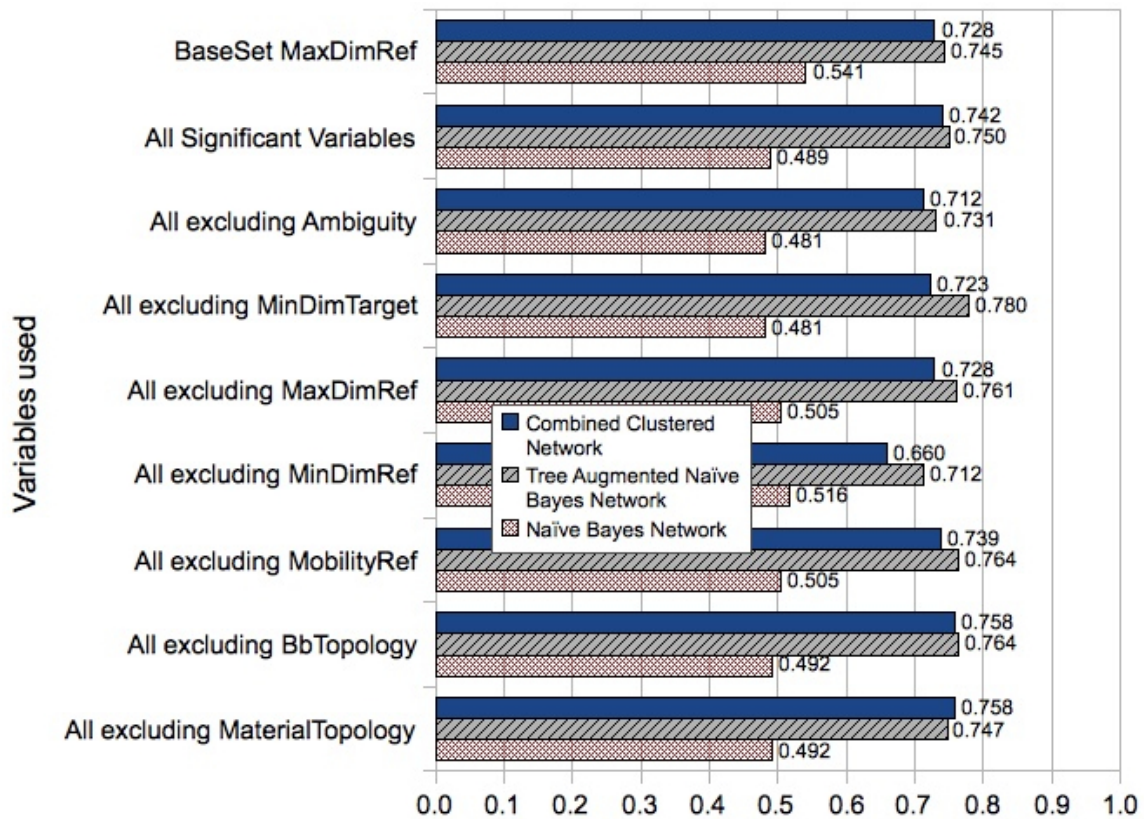


Figure 6.8: Fraction of machine reference choices matching one of the top three human reference choices when a complex model is used

complexity. Although as more near redundant variables are added a drop in performance would be expected, this reduction seems quite marked. The only satisfactory explanation is that dependencies between at least two feature variables and the classifier are important to the process.

The picture regarding individual variables is confused; for instance, the exclusion of the target minimum dimension and reference maximum dimension variables degrades the performance of the combined cluster networks but improves the performance of the tree augmented naive Bayes networks. This suggests that, with the test data and variables used the limit of what can be learned is being approached.

### 6.3.5 Further models

As a final step a few features are added to the models from the last subsection which allow them to address relevant factors even though these are not supported by variables that appeared significant in themselves. These are:

1. Relative mobility of target and candidate reference object. This is the intuitively important factor even though target mobility as an individual variable has not seemed to influence model performance. A hidden node is added to the combined clustered

model which has the variables reference mobility (MobilityRef) and target mobility (MobilityTarget) as parents to model this influence. Both of these variables are included in the naive and tree augmented naive Bayes models.

2. Reference object ‘density’. This is modelled by adding reference object bounding box volume as well as reference material volume. ‘Spiky’ objects such as chairs or taps can be differentiated from ‘blocky’ objects such as cupboards and books.
3. The communication cost is modelled using name length as well as ambiguity.
4. The topology variables are removed as they appear to degrade the performance of the combined clustered model of figure 6.7.

The resulting combined clustered Bayes network is termed the ‘final’ model and is shown in figure 6.9. This model includes a node labelled test variable. This variable allows for direct dependencies between the distance measure and another variable to be tested (no direct dependence with the distance variable and any other feature variable existed in the combined clustered model of figure 6.7). The results for the ‘final’ network with various variables in the test variable position are shown in figure 6.10. For the naive Bayes and tree augmented naive Bayes networks the test variable is included in the network but there is no ‘forced’ statistical dependence check with the distance measure.

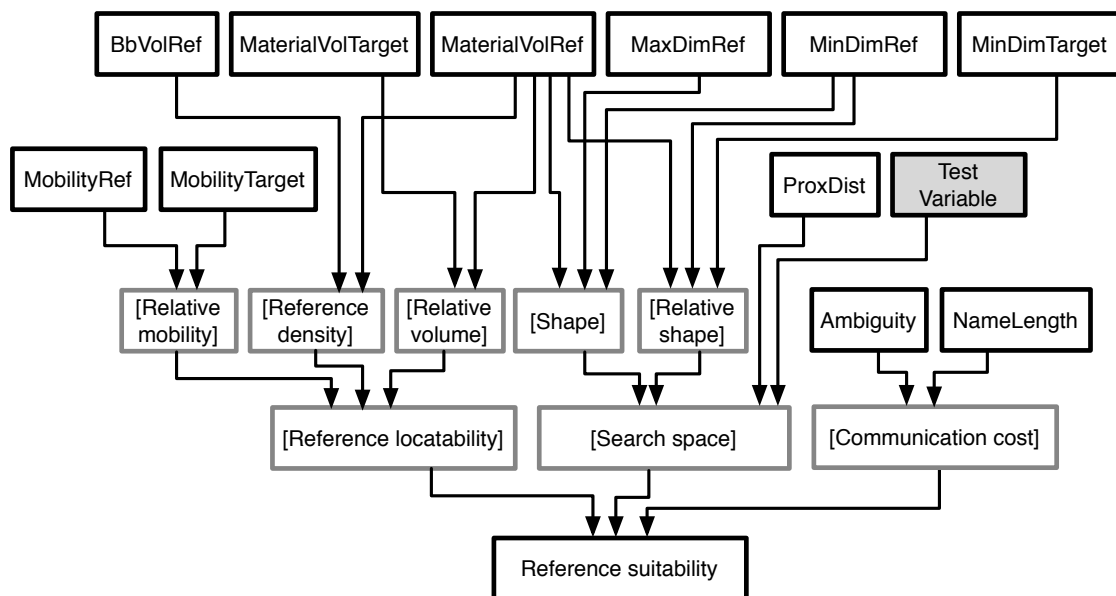


Figure 6.9: The combined clustered network topology for the final models analysed

No statistically significant differences arise with any of the test variables. The model, with or without the target volume variable in the test variable position, is a significantly better predictor of reference suitability at the 0.05 level ( $W = 27$   $N = 8$ ) than the best four variable model. It is not significantly better than either the best combined clustered model or the best tree augmented naive Bayes model from figure 6.8.

It should be noted that addition of redundant variables, or links between variables, to a Bayesian model trained on finite and noisy data typically degrades performance rather than having neutral influence. This depends on the structure of the model but can be clearly seen in the results from the simple naive Bayes structure. It also suggests that the improvements in performance noted as significant as the models have been developed are, if anything, understated.

These are the best models yet found using the geometric variables (as opposed to the sight-line variables used in section 6.5). They do not incorporate the topological relationship between target and reference or the angular relationship between them. Both of these might be considered surprising and possible reasons for this are discussed in section 7.2.2.

Although there are many networks that appear to outperform the best four variable network, achieving further statistically significant results is difficult with the test data set available. As the performance of the networks improves and the number of references correctly predicted increases, the number of test cases available in the data set to discriminate between networks diminishes.

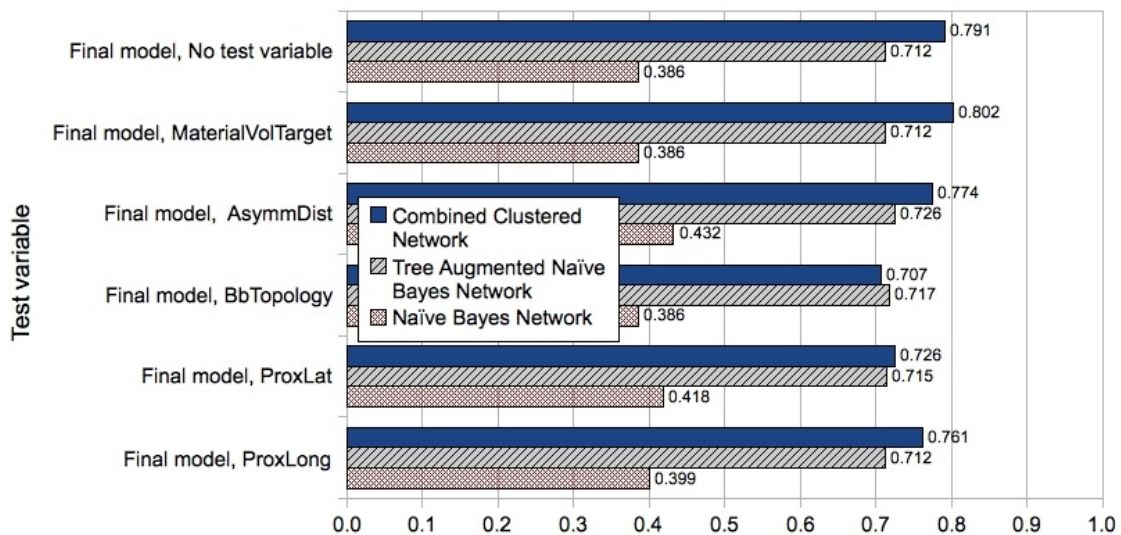


Figure 6.10: Fraction of machine reference choices matching one of the top three human reference choices for the final models shown in Figure 6.9

## 6.4 Extended data set

As mentioned the data set size initially used did not seem to be sufficient to produce statistically significant discrimination between the best machine models, suggesting that the size of the data set could be usefully increased. Examination of the performance of the network as the training set size is varied (figure 6.11, original (series 1 scenes)) also suggests that the amount of training data might not be sufficient to train the models to their fullest potential. In other words the graph of performance against training set size

has not definitely levelled off at the maximum training set size. The data for figure 6.11 is produced using 10 fold cross validation as usual but for each cross validation run a proportion of the training data is discarded. The maximum training set size is 90% of the full data set size and is decreased from this value, the networks are tested on 100% of the data set in question.

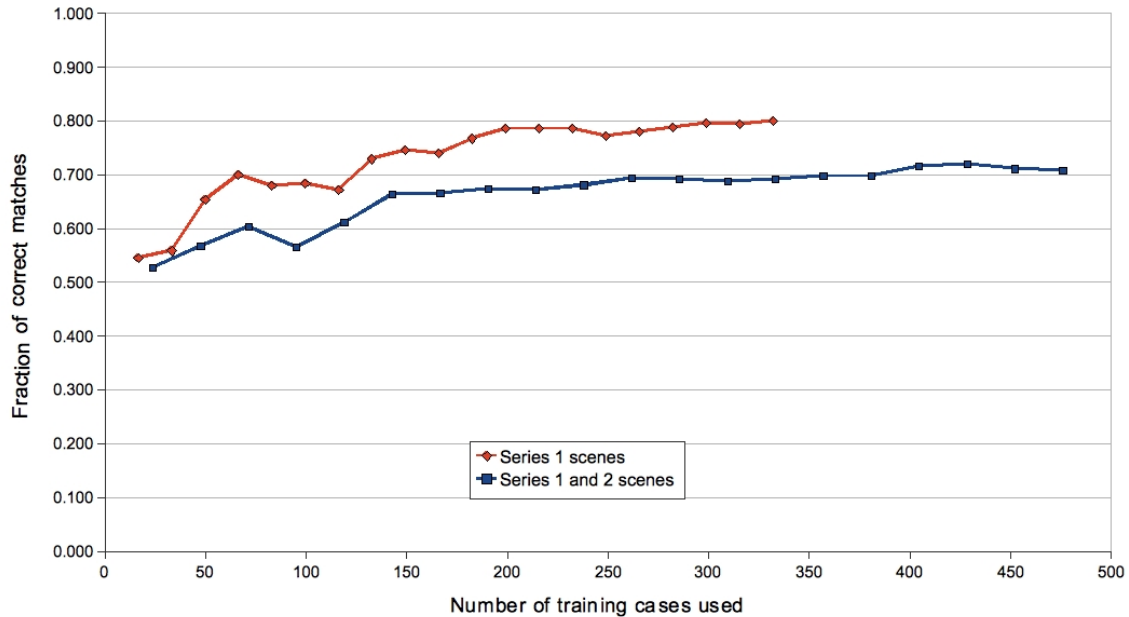


Figure 6.11: Performance of ‘best’ network on reduced training set sizes for original and extended data sets

For these reasons the data set size was increased by the addition of 40 new scenes giving an extra 160 test cases, taking the total from 369 to 529. The original data set is termed ‘series 1’ and the new scenes termed ‘series 2’. The new (series 2) scene set contained 30 scenes at street or vista scale and 10 scenes in a non-domestic interior environment. The number of interior (small scale) test cases is 277 (from 237) and the number of exterior (large scale) test cases is 252 (from 132) after this adjustment. This more equal distribution will facilitate an investigation into the effect of scene scale on reference object selection (the results from which are presented in section 6.6).

The effect of the increase in data set size can be seen in figure 6.11. It is still not entirely clear that sufficient training data is available, however the amount of extra data required to make a significant difference in performance would represent, at present, a prohibitive amount of work to derive. What is immediately apparent is that the performance of the network is distinctly worse on the full data set (series 1 and 2 scenes) than on the series 1 scenes alone. Figure 6.12 confirms that this is the case for a variety of key models of varying complexity from section 6.3.

One possible reason for this is the difference in the average number of objects between the series 1 and series 2 scenes. This is illustrated in figure 6.13. The series 2 scenes are



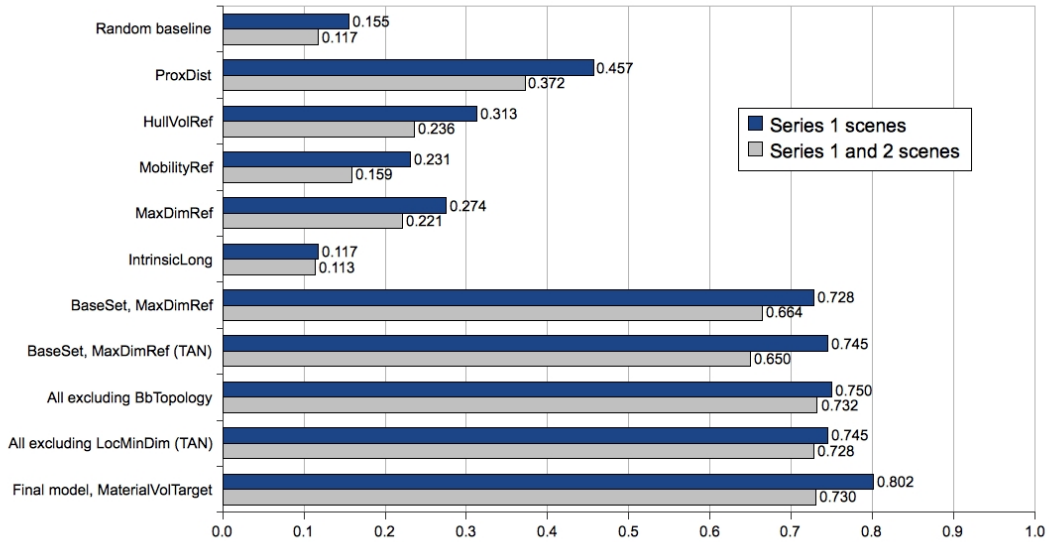


Figure 6.12: Fraction of machine reference choices matching one of the top three human reference choices for the original (series 1) scenes and the full data set of series 1 and 2 scenes

deliberately more complex both in terms of the number of objects in a scene (a minimum of 25), and in terms of the object representation to facilitate future work. It might be expected that a scene with more objects in would present more possibilities for a machine model to choose an incorrect reference and that therefore the percentage of correct matches would be less in more complex scenes. Plotting the fraction of correct matches against the number of objects in a scene shows that this is not the case however. Figure 6.14 shows this is true for both the series 1 and series 2 scenes on their own and for the combined data set.

The correlation coefficients are shown in table 6.5. There is a very slight tendency for the machine model performance to improve in more complex scenes, but this is only significant for the case of the series 1 scenes alone. This cannot explain the drop in performance when the series 2 scenes are added to the data set with the consequent increase in the average number of objects in a scene. Note that the lower overall accuracy on the series 2 scenes alone is probably due in part to the fact that the data set size is small.

The second difference between the two data sets, the fact that the series 2 scenes are biased towards exterior scenes with larger scales, seems to explain some of the discrepancy.

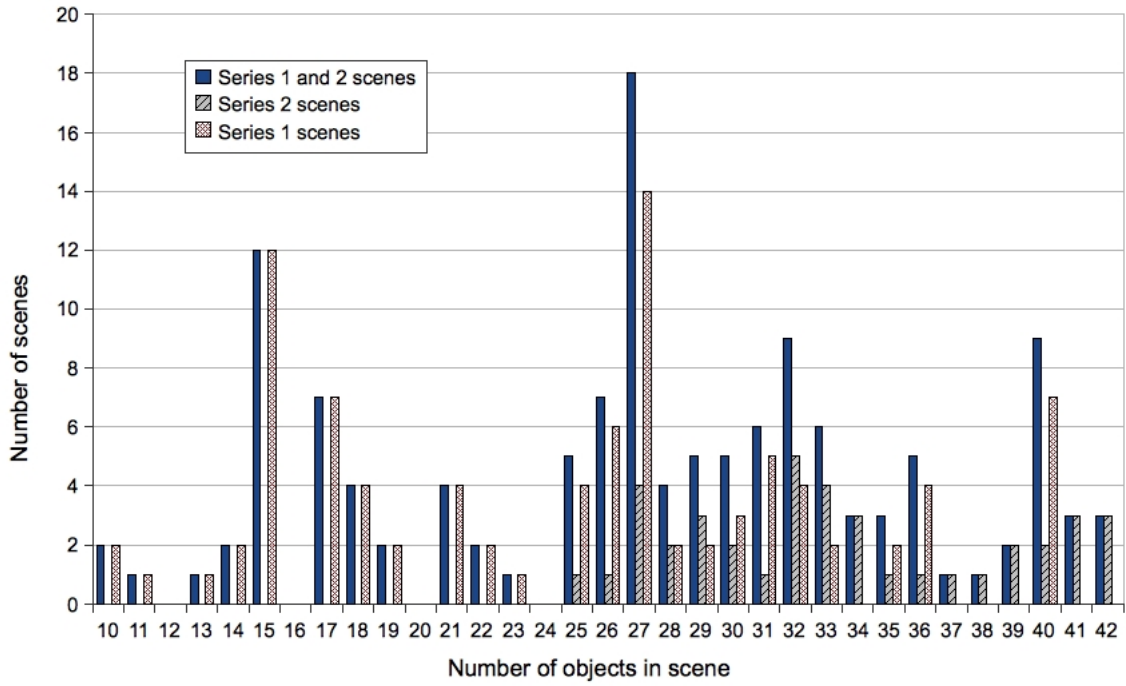


Figure 6.13: Distribution of number of objects in scenes

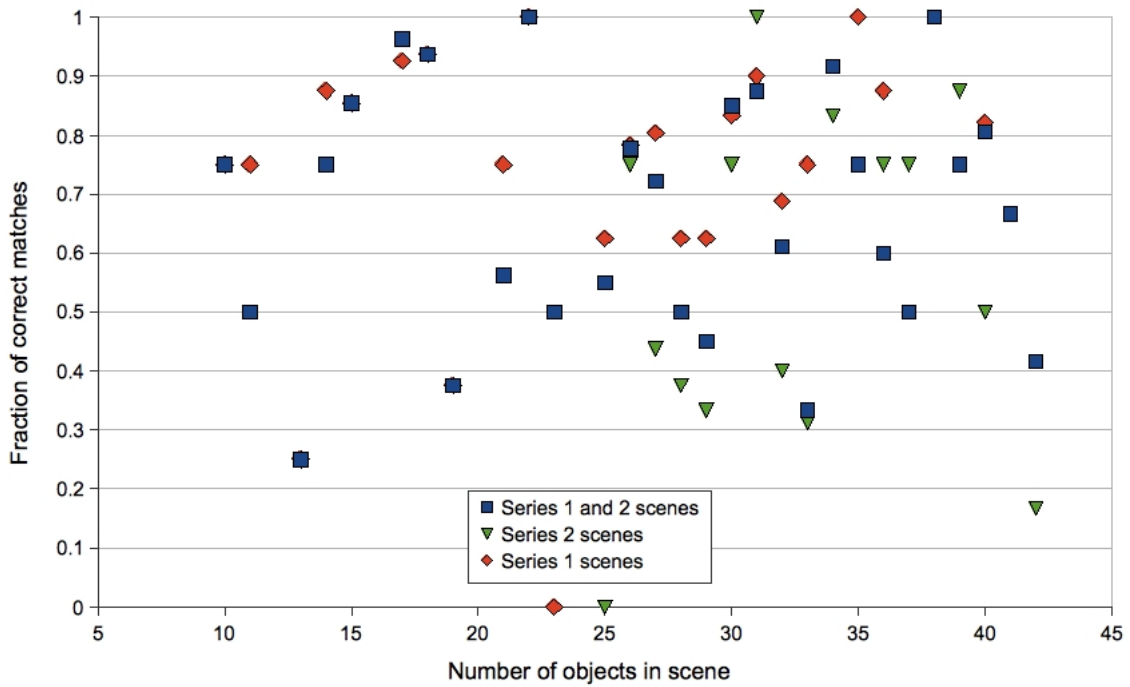


Figure 6.14: Fraction of machine reference choices matching one of the top three human reference choices, for the best model from figure 6.10, as a function of the number of objects in a scene. Models are *trained and tested* on the indicated data-sets

Table 6.5: Correlation of machine model performance with number of objects in a scene for the best model from figure 6.10. (Product moment correlation coefficient).

Data Set	Number of test cases	Average number of objects	Fraction of correct matches	Correlation coefficient	P value
Series 1	369	24.9	0.80	0.12	0.02
Series 2	160	33.5	0.53	0.09	0.26
Series 1 and 2	529	27.5	0.71	0.02	0.64

The models in section 6.3 have no variable that can account for scene scale, and results from models that are aware of scene scale are given in section 6.6.

The other aim of the extended data set, to provide more statistically significant results for better performing models, has been achieved to some extent. Although the previously best performing model is still not significantly better than the other more complex models, all three more complex models from figure 6.12 are significantly better (at the 0.025 level) than the best performing 4-variable model.

## 6.5 Results from sight-line variable models

The sight-line variables should address some issues that the geometric variables cannot, in particular object occlusion, treatment of surface objects, such as roads, and the limited way in which variables describing geometric extension can address the concept of search space. As before individual variables are first considered and then successively more complex models are developed. Naive Bayes networks are not further considered in this study and for brevity tree augmented naive Bayes networks are only compared with combined clustered networks for the most complex models. The extended data set of series 1 and series 2 scenes is used throughout the rest of the study.

It should be remembered from section 5.6.1 that there are effectively two groups of sight-line variables:

1. ‘Saliency’ variables (Viewability, SaliencySL, ProxSaliency, CentSaliency, ProxSaliencySqr, CentSaliencySqr) which relate to the search for the reference object. All contain a measure of the reference ‘visibility’ derived from the number of sight-lines which intersect the reference object. With the exception of the ‘viewability’ variable they also all contain some representation of the distance between the target and reference.
2. ‘Search distance’ variables (SearchDist, WgtSearchDist, WgtSearchDistSqr) which relate to the space in which the listener must search for the target object. These contain a measure of the average distance between all points on the reference, intersected by a sight-line, and the target.

Figure 6.15 shows the results for the sight-line variables individually, in the network topology of figure 6.1c. The immediate point to note is that none of these variables, and in particular the composite saliency variables, perform as well as the proxDist variable on the

full data set as shown in figure 6.12. It might have been expected that these variables would perform better than the simple distance measure as they include a measure of reference size, which leads to a significant improvement in model performance over using distance alone, when reference volume is used to represent reference size. The best performing single variables are the ‘saliency squared’ measures which combine the square of the target-to-reference distance in 3-dimensions along with reference viewability. These could be thought of as giving the greatest weight to the distance measure and hence being most analogous to proxDist.

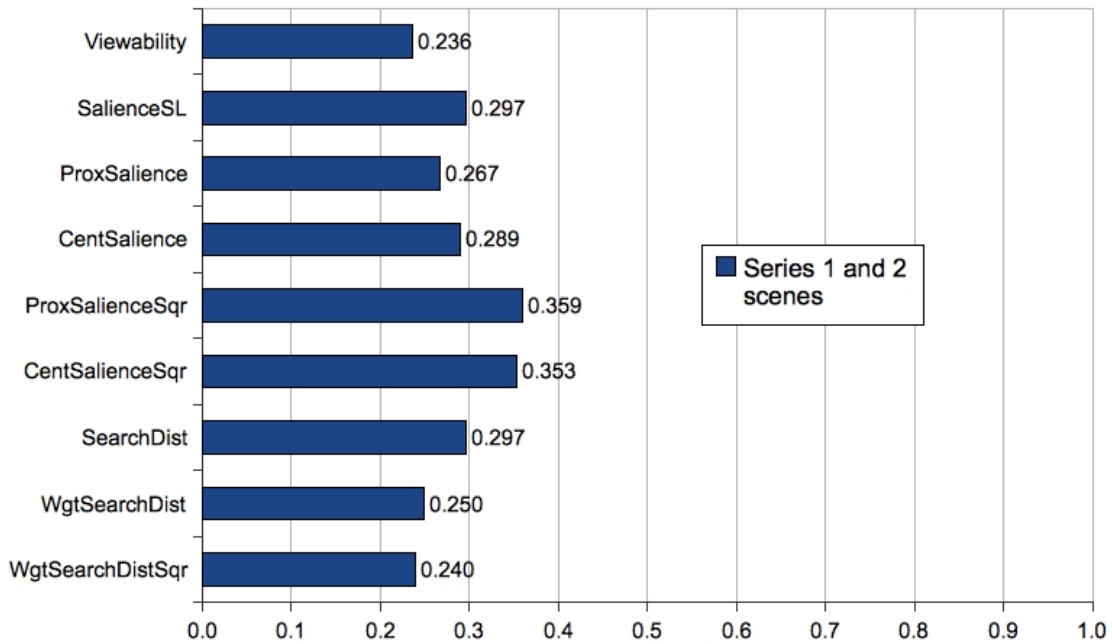


Figure 6.15: Fraction of machine reference choices matching one of the top three human reference choices when a single sight-line variable is considered

Importantly there is no representation of the actual size of the target or reference objects in these single variable models. By combining the individual sight-line variables with the ‘baseline’ variable set (proxDist, materialVolRef, materialVolTarget) from section 6.3.3 these factors are included, although there is some possible duplication of distance representation. The results from these models, which are four variable models with the topology shown in figure 6.5c, are given in figure 6.16. Again the immediate point to note is that none of these models perform as well as the baseline model with the simple addition of the reference geometric extension variable, maxDimRef (figure 6.12). The best performing model is the combination of baseline plus search distance which suggests again that the combination of reference apparency and search space may be a good basic model for reference selection.

The addition of some of the saliency variables leads to better performance than the use of the baseline variable set alone, in particular the 2-dimensional saliency measure, saliencySL. This suggests that there are aspects which might be termed ‘projected view-

bility' that are important in reference selection. Interestingly, viewability alone does not make a significant difference to model performance. The slight improvement when the centroid salience measures are used considered alongside the lack of improvement when proximal salience measures are used might suggest that at some target/reference spacings centroid distance is an important measure.

The weighted search distance squared ( $\text{wgtSearchDistSqr}$ ) measure had been expected to best represent the search space concept but as can be seen this is in fact the worst performing of the three measures of search space. Why this might be so is discussed in section 7.2.2.

The intrinsic longitude measure is also introduced at this point. In common with the simple longitude measure shown in figure 6.6 it slightly degrades the performance of the baseline network. It is not further considered.

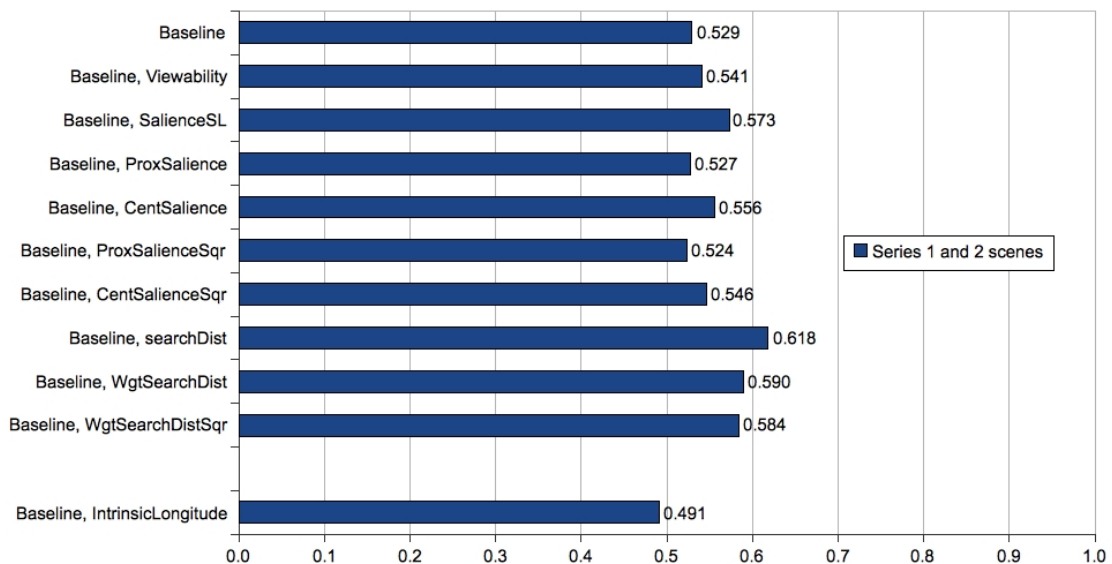


Figure 6.16: Fraction of machine reference choices matching one of the top three human reference choices when a single sight-line variable is considered along with the baseline variable set from section 6.3.3

Although the performance of the sight-line variables is not as good as that of the geometric variable representations in these first models it might be thought that, since they are alternative representations of the measures of reference apperency and search space, they may produce good models for reference selection when combined with other measures such as object mobility and ambiguity. Results for models of this type, again with the network topology shown in figure 6.5c, are shown in figure 6.17. Again none of these perform as well as the best four variable model from figure 6.12. The best performing model is in fact a five variable model containing viewability and distance, again suggesting that a simple measure of distance is a requirement for a good model of reference selection, whether or not a composite variable containing a representation of distance is also included. This further suggests that humans must be using reference/target distance, in combination

with other variables, in a manner that cannot be captured sufficiently by the salience type measures. In other words the probability table in the Bayesian network can model the relationship between target-reference distance and reference object size, as used by humans, in a way that the defined relationships between these quantities in the salience variables, cannot.

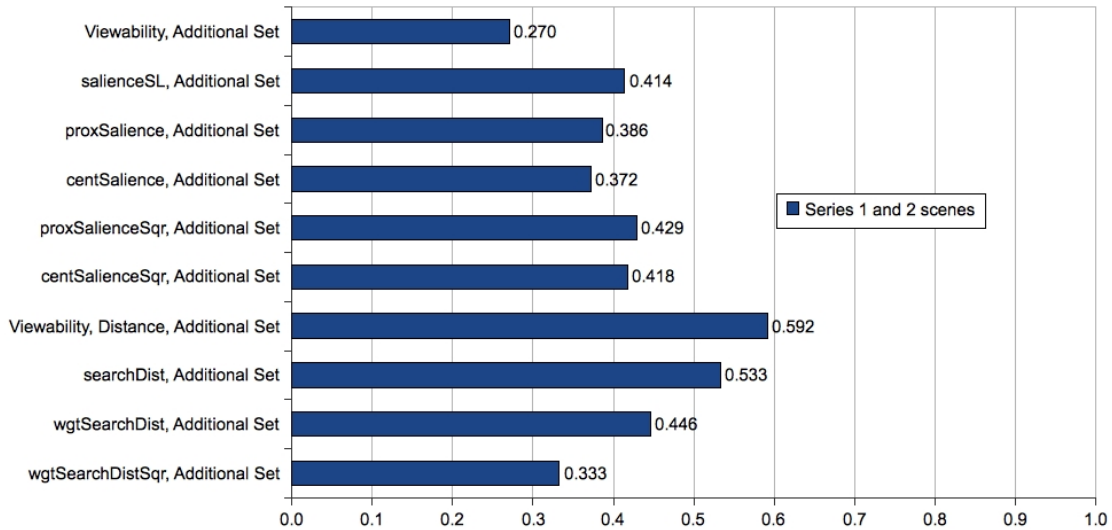


Figure 6.17: Fraction of machine reference choices matching one of the top three human reference choices when a single sight-line variable is considered along with an additional variable set consisting of: maxDimRef, mobilityRef and Ambiguity

What has not been considered so far is a combination of the sight-line variables. In particular it might be expected that a sight-line variable representing reference apparency combined with a sight-line variable representing search space might perform well. Figure 6.18 shows results from models of this type. For brevity the centroid salience measures have been excluded. All combinations of the three search space variables with the four remaining salience measures are considered along with models combining viewability proxDist and the search space variables. In all the models target object volume (materialVolTarget) and ambiguity are also included, as these are the factors not included in the representation of salience and search space that are likely to be most important. The network topology is again that of figure 6.5c.

There is still no network in figure 6.18 that performs as well as the best four variable model (using the geometric variables), on the full data set, as shown in figure 6.12. This is discussed in section 7.3.1, it is a strong indication that humans must be using knowledge of the actual size, or likely actual size, of an object, along with its geometric extension, when making a choice of reference object. The apparent size of an object, given what can be seen of it as expressed by the viewability measure, does not supply the information needed to give the same performance in the reference choice task.

Within these results it seems that there is little difference between any of the salience measures when combined with a search space measure. Of the search space measures the

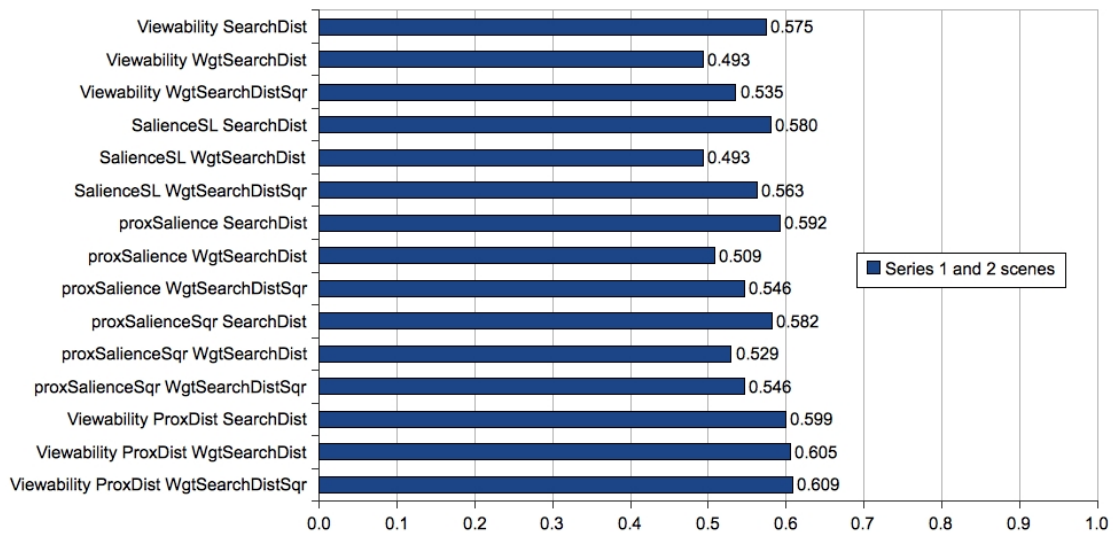


Figure 6.18: Fraction of machine reference choices matching one of the top three human reference choices when a saliency variable and a search space variable is considered along with target volume and ambiguity

simple search distance appears to give the best results but this may simply because it is the best substitute, in this case, for target/reference distance. This view is further reinforced looking at the results for viewability plus distance as the saliency measure. In this case all the three search space measures give similar results with weighted search distance squared (WgtSearchDistSqr) being marginally superior.

Having strong indications that representations of actual geometric size and extensions of the candidate reference objects and a non-compound representation of the distance between them (that is to say distance on its own, not a saliency type measure) are required the question of whether sight-line variables can contribute anything to reference object selection must be asked. To test this sight-line variables are added to a network similar to the best performing network of section 6.3.5. The resultant network is shown in figure 6.19. A saliency measure is added to the reference locatability variable and a new hidden variable is added to the search space variable. Search distance along with target material volume and reference minimum dimension are parent variables of this new variable.

The results for different combinations of saliency variable and search distance variable are shown in figure 6.20. For brevity both the proxSaliency, proxSaliencySqr variables and centSaliencySqr have been omitted as they appear to give similar results to the centSaliency measure used.

The best of these networks now gives an improvement in performance over the best network using only geometric variables (on the full data set). None of the networks is significantly better at the 0.05 level although the network using the viewability and searchDist variables with  $W = 31$ ,  $N = 10$  is only fractionally outside. Note that no tree augmented naive Bayes network gives improved performance over the models using only geometric variables. The performance of the tree augmented naive Bayes networks is discussed in

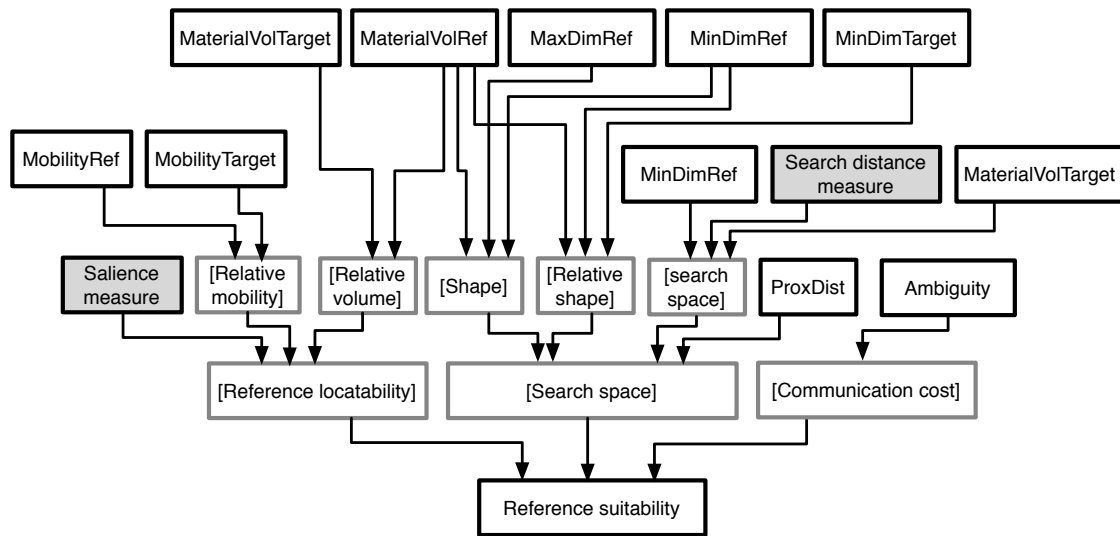


Figure 6.19: Network topology for testing sight-line variables in conjunction with the best performing network using geometric variables

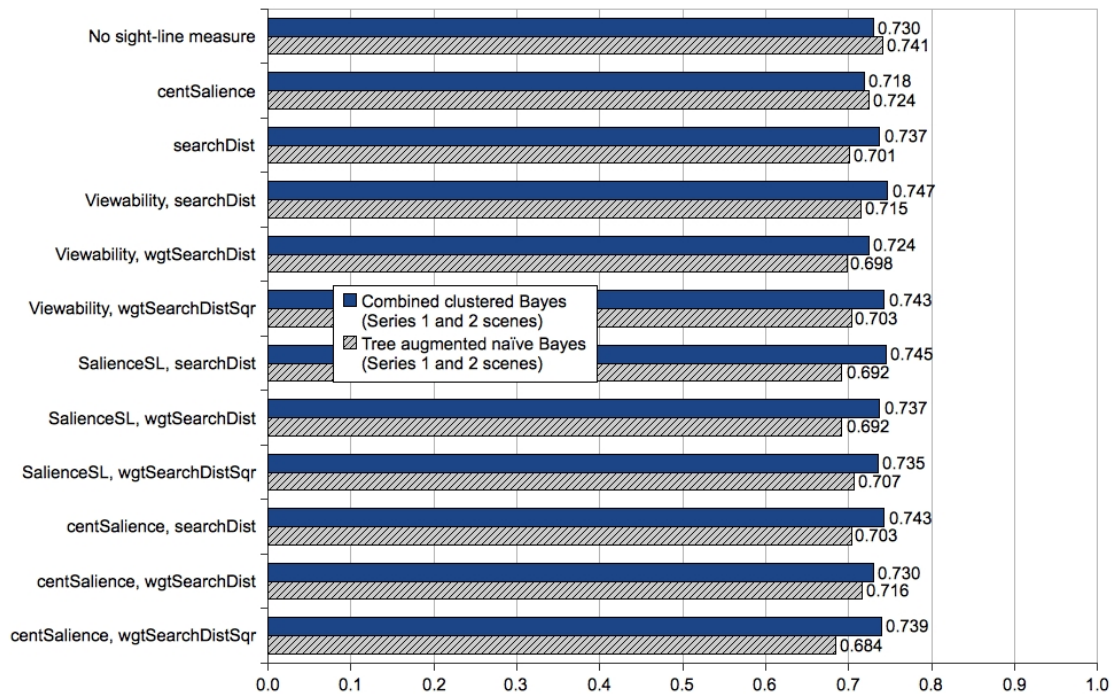


Figure 6.20: Fraction of machine reference choices matching one of the top three human reference choices for the complex networks combining sight-line and geometric variables

section 7.2.2, but there is an indication here that the introduction of the sight-line variables has impaired the ability of the tree augmented networks to represent the necessary interactions in the variable set. When paired with each of the saliency measures the simple searchDist measure is the best performing of the search distance measures. There is no



consistency in the performance of the salience measures.

## 6.6 Effect of scene scale

It was noted in section 6.4 that the inclusion of the series 2 scenes had reduced the overall performance of the networks from section 6.3. The different balance of interior and exterior scenes resulting from this inclusion might be a reason for this. The change in the balance of scenes had been intended to enable better comparison of model performance on interior and exterior scenes but the investigation was given extra significance by the performance discrepancy.

The best network from figure 6.20 was trained and tested on internal and external scenes separately in the combinations, and with the results, shown in table 6.6.

Table 6.6: Performance of the best network from figure 6.20 when trained and tested separately on interior and exterior scenes

Training scenes (number)	Test scenes (number)	Fraction correct
Interior (277)	Interior (277)	0.82
Interior (277)	Exterior (252)	0.44
Exterior (252)	Interior (252)	0.64
Exterior (252)	Exterior (277)	0.61

This is a slightly confusing picture. It suggests that Exterior scenes are inherently more difficult than interior scenes but also that there is a benefit (as might be expected) in training on the same scene scale as testing. Note though that cross validation is used when training and testing on the same scene set but is not used when testing on a different set to training.

An examination of the numbers of objects assigned to different bins and in particular the bins for reference object volume variables shows a possible explanation for the results. Although the initial bin allocation for object volume variables had been reasonably balanced, the addition of the series 2 scenes has skewed the allocation, and this is due to a different distribution of object sizes in the internal and external scenes, as can be seen from table 6.7. This was not anticipated to be the case, but it seems that the volume of objects in larger scale scenes is on average a smaller fraction of the overall scene volume than it is for those in smaller scale scenes.

To try and correct this the volume variables are scaled by  $S^2$  instead of the initial  $S^3$ , where  $S$  is the length of the diagonal of the scene bounding box, with the resulting distribution of bin values as shown in table 6.8. The skewness has been substantially reduced and the distribution of bin allocations for the exterior scenes improved. Repeating the tests on the interior and exterior scene sets leads to the results shown in table 6.9. Although there has been an improvement in the overall performance of the model on

Table 6.7: Bin allocations for reference object volume (materialVolRef) for interior, exterior and combined scene sets using  $S^3$  scale factor

Scene set	Fraction of objects in bin:								Std. dev.
	1	2	3	4	5	6	7	8	
Interior	0.101	0.066	0.128	0.164	0.220	0.105	0.132	0.083	<b>0.056</b>
Exterior	0.329	0.039	0.164	0.115	0.113	0.123	0.115	0.001	<b>0.049</b>
Combined	0.213	0.052	0.147	0.140	0.163	0.116	0.127	0.042	<b>0.097</b>

Exterior scenes there is still a distinct difference between interior and exterior scenes.

Table 6.8: Bin allocations for reference object volume (materialVolRef) for interior, exterior and combined scene sets using  $S^2$  scale factor

Scene set	Fraction of objects in bin:								Std. dev.
	1	2	3	4	5	6	7	8	
Interior	0.144	0.129	0.160	0.191	0.138	0.116	0.098	0.023	<b>0.050</b>
Exterior	0.198	0.118	0.034	0.092	0.174	0.111	0.128	0.146	<b>0.050</b>
Combined	0.168	0.122	0.095	0.139	0.157	0.115	0.116	0.087	<b>0.028</b>

Table 6.9: Performance of the best network from figure 6.20 when trained and tested separately on interior and exterior scenes using  $S^2$  scale factor

Training scenes (number)	Test scenes (number)	Fraction correct
Interior (277)	Interior (277)	0.82
Interior (277)	Exterior (252)	0.54
Exterior (252)	Interior (252)	0.60
Exterior (252)	Exterior (277)	0.63

To assess how much of the difference in performance between the initial data set (series 1 scenes) and the full data set (series 1 and 2 scenes) is due to the effect of scene scale a 2-valued variable is introduced that is 0 if the scene bounding box diagonal is less than  $20m$  and 1 if it is greater than  $20m$ . This also separates the scenes into interior and exterior groups. This variable is applied as a ‘switch’ in various locations in the hitherto best performing network as shown in figure 6.21. As can be seen from figure 6.21 this network uses both geometric and sight-line variables. When added as a parent to the relative volume variable there is a slight drop in performance. When added progressively to other variables alongside reference size measures there is a small, though statistically insignificant ( $W = 14$ ,  $N = 9$ ,  $p = 0.19$ ), improvement in model performance. Adding the variable as a parent alongside the classifier, as a ‘global’ scale switch results in no improvement. Adding the scene scale variable alongside proxDist or viewability also reduces network performance

suggesting that the distance and salience measures are independent of scene scale.

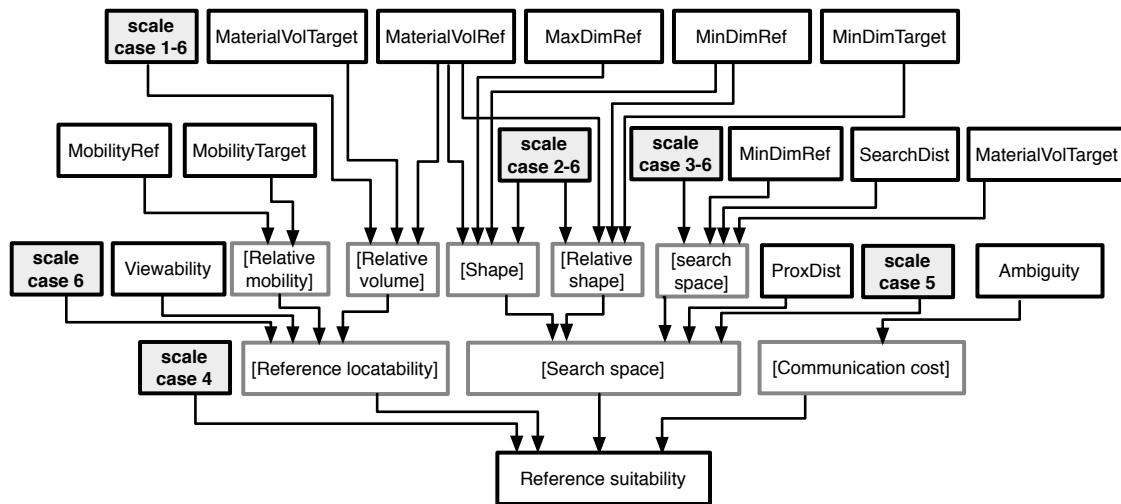


Figure 6.21: Networks including a scene scale variable in positions corresponding to the result cases shown in figure 6.22

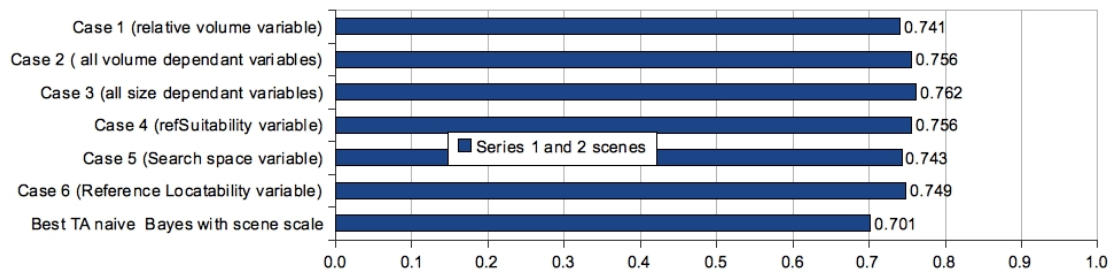


Figure 6.22: Fraction of machine reference choices matching one of the top three human reference choices for the network in figure 6.21 including a scene scale variable

The slight improvement in performance still leaves a substantial gap between the performance of the best networks on the initial and full data-sets which is as yet unexplained.

The best network including the scene scale variable is significantly better (at the 0.025 level,  $W = 37$ ,  $N = 9$ ) than the best network containing only geometric variables. So although it is not possible to say that sight-line variables make a significant difference or that the scene scale variable makes a significant difference, the combination of the two is significant.

The introduction of a scene scale variable to the best performing variable set in a tree augmented Bayesian network causes a small reduction in performance. The best network using scene scale and sight-line variables was tested on the series 1 scenes only where it performed identically to the best geometric variable model from section 6.3.

Note that the experiments in the following sections all use the  $S^2$  scale factor for object volumes.

## 6.7 Results from learned structure models

So far models derived from the interaction information based structure learning algorithm have not been used. The algorithm as described in section 5.5 does not, unlike the tree augmented network algorithm, use all of the variables given in its initial set. So direct comparison of the effects of particular variables on the reference choice task is not possible.

However the algorithm should capture the necessary variable interactions by design rather than by chance as is the case with the tree augmented Bayesian networks. The results from the networks generated by the algorithm when given different groups of starting variables are shown in figure 6.23. The performance of the tree augmented naive Bayesian network is given for the same variable sets and the target to reference distance performance of the best combined clustered network where relevant.

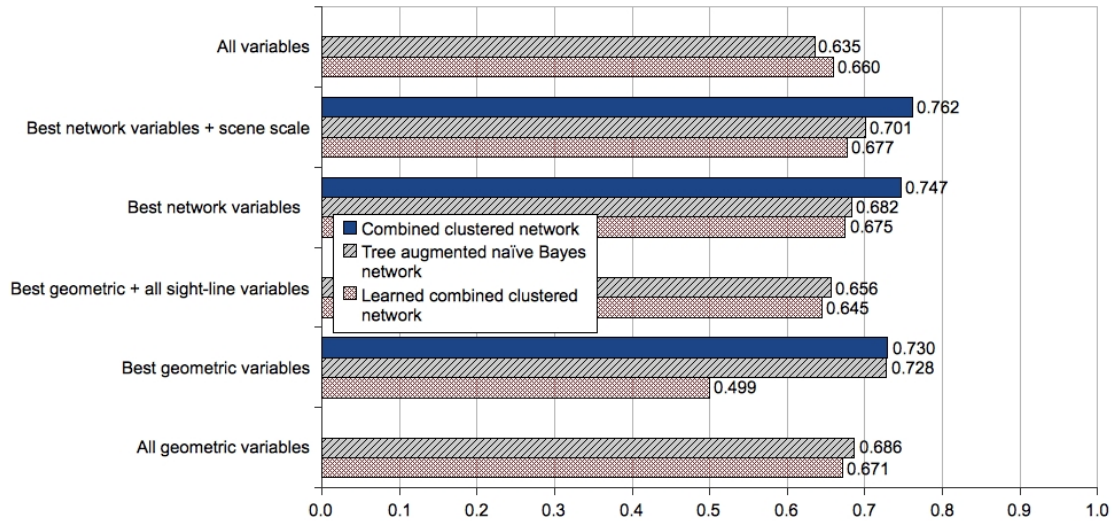


Figure 6.23: Fraction of machine reference choices matching one of the top three human reference choices for learned structure networks operating on various starting variable sets

In most cases the tree augmented Bayes network slightly outperforms the learned combined clustered network although the difference is not significant except in the case of the variables comprising the best performing geometric network. This is a small initial variable set (11 variables) and has been further reduced (to 7 variables) by the combined clustered algorithm in a possibly over zealous attempt to remove redundancy. The learned network structure for this case is shown in figure 6.24. The network for the best learned case is shown in figure 6.25.

Clearly the structure learning algorithm is some way from capturing all the interactions required. The best performing hypothesis model networks are significantly better. However the algorithm has paired reference and target size measures and avoided constructing clusters from redundant variables.

The tree augmented networks perform better than they should, given that they cannot, except by accident, capture interactions between any two variables and the classifier.

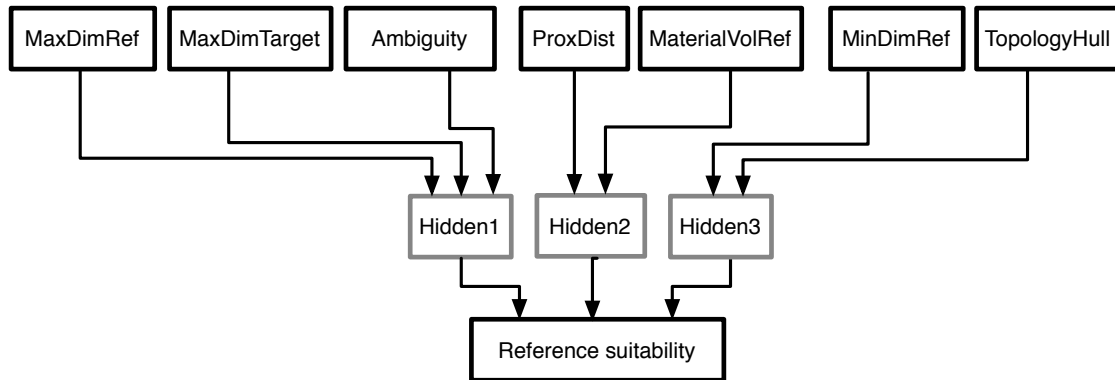


Figure 6.24: Bayesian network produced by the combined clustered structure learning algorithm from the (small) variable set of the best geometric network

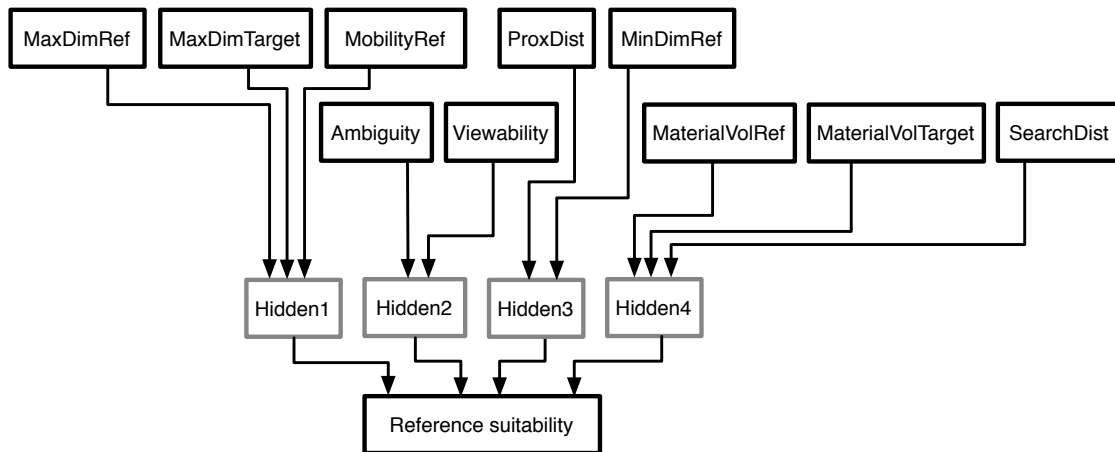


Figure 6.25: Bayesian network produced by the combined clustered structure learning algorithm from the variable set of the best performing network

Possible reasons for this are discussed in section 7.2.2.

## 6.8 Results from listener present models

The effect of having a listener present in a scene is shown in table 6.10 where the best performing network from figure 6.20 was trained and tested on scenes with and without a listener. The improved performance in the train and test with no-listener case (over 6.20) is due to the training and testing being performed on the whole of the indicated data-set with no cross validation. The results indicate that the presence of a listener has a significant effect on the model performance although no cross validation comparison is available. The interesting point to note is that the difference in results when the training scene set is changed is much greater than the difference in results when the test scene set is changed. The machine model is simply treating the listener as another object in the scene so when

used as a training source the presence of the listener makes little difference. However when used in testing the comparison against the human judgements with a listener present leads to a larger change in results. It would appear from this that a listener in the scene has made some difference to human judgement on reference suitability.

Table 6.10: Performance of the best network from figure 6.20 when trained and tested separately on scenes with a listener present and not present. Note, no cross validation has been used

Training scenes (number)	Test scenes (number)	Fraction correct
No Listener	No Listener	0.79
Listener	No Listener	0.77
No Listener	Listener	0.66
Listener	Listener	0.70

The impact of some of the listener related variables in simple single and four variable models is shown in figure 6.26. The normal baseline variable set of proxDist, materialVolTarget and materialVolRef is used. The only variable that improves model performance over the baseline case is the distance between the listener and the candidate reference. While the improvement appears marked it is not statistically significant. It might have been expected that the distance between the listener and the candidate reference would have more influence than the distance between the listener and the target. The reference must be immediately apparent to the listener, not the target. If this is so the reference will define the search area for the target irrespective of the listener's distance from the target. The lack of discernible influence of the angle measures is in line with earlier results. If any influence exists the data-set is unable to illustrate it and no further consideration is given to the angle measures.

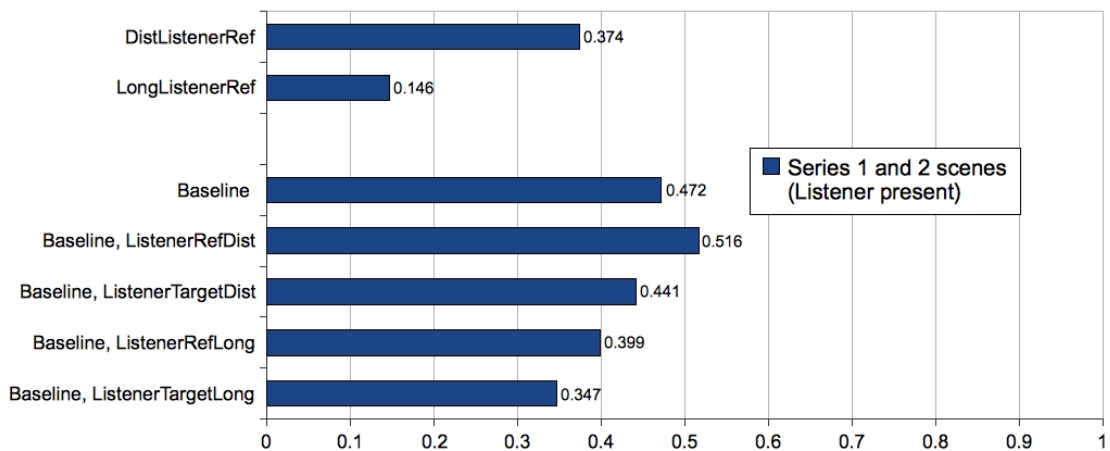


Figure 6.26: Effect of listener related variables in scenes with a listener present

A network for assessing the effect of listener present variables on a more realistic model

is shown in figure 6.27. The two locations shown for the listener related variable were chosen as being the most likely places for a significant influence to be visible. Position 1 in figure 6.27 allows the presence of the listener to influence the appropriate reference size relative to the target size. Position 2 allows the presence of the listener to influence the requirement for search space.

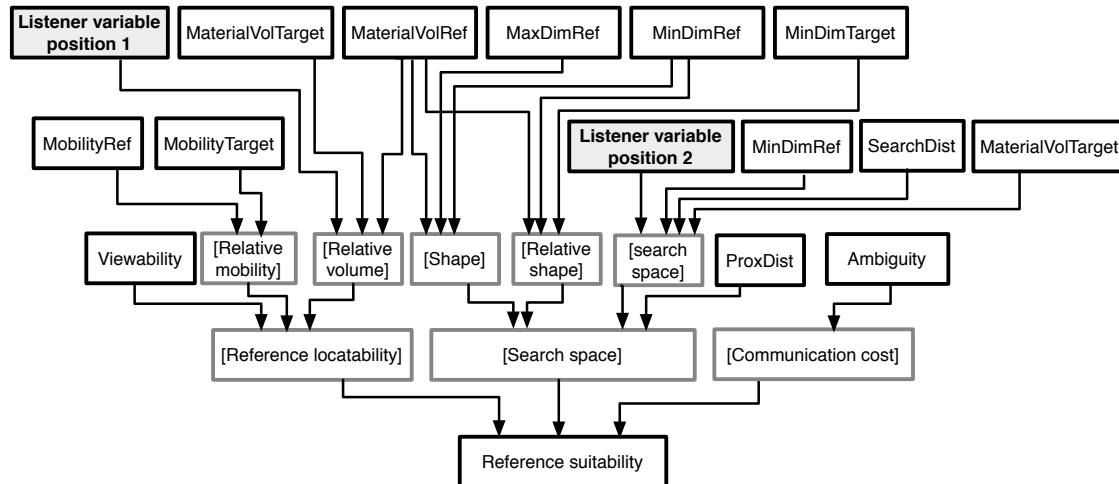


Figure 6.27: Addition of listener related variables to the best network using geometric and sight-line variables. The variables used in positions 1 and 2 are shown in figure 6.28

The results are shown in figure 6.28 for the case where only the listener present scenes are used (that is every scene has a listener in) and the case where all the scenes are used so half have a listener present. Two variables are used, the listenerRefDist variable also contains a ‘listener present’ value which allows the variable to model the absence of a listener or its distance to the candidate reference if a listener is present. The listenerRef variable is set to 1 when the listener is being considered as the candidate reference object (that is, in a “the mug is behind you” type phrase) and 0 otherwise, including when there is no listener present in the scene.

For the case of only the listener present scenes both of the variables produce significant improvement over the model with no listener variables. In this case the listenerRefDist variable is only modelling the distance of the listener from the candidate reference object. The listenerRefDist variable is significant in position 1 at the 0.025 level ( $W = 31, N = 8$ ). The listenerRef variable is significant when in positions 1 and 2 (not position 1 only) at the 0.01 level ( $W = 42, N = 9$ ). For the case of all scenes with and without a listener only the listenerRef variable is significant in positions 1 and 2 at the 0.025 level ( $W = 37, N = 9$ ). This suggests that the presence of a listener has a marginal effect on the way speakers perceive the requirement for reference object size given a size of target, but suggests quite strongly that when considering the listener as a possible candidate reference the requirement for the search area is changed.

Examination of the results suggests that without the listenerRef variable the machine

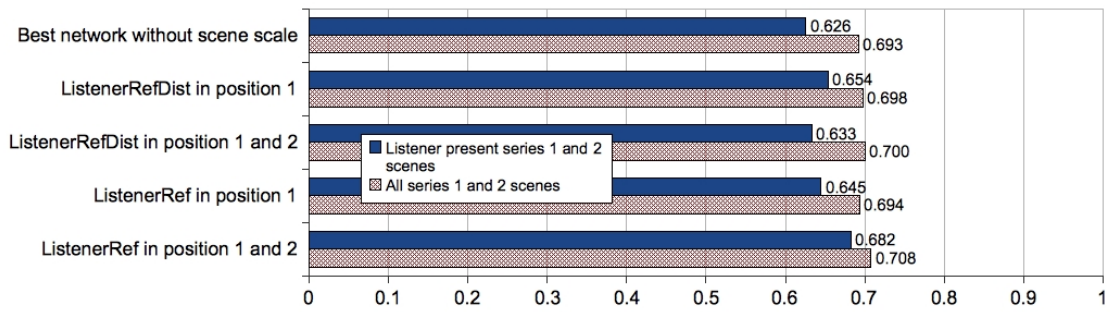


Figure 6.28: Fraction of machine reference choices matching one of the top three human reference choices for the network in figure 6.27 with the variables in the noted positions. The best network from figure 6.22 is the baseline case with no listener related variables

model tends to over-select the listener as a reference, compared to the human participants in the validation exercise. In many cases, particularly considering the constraints on listener placement (see section 4.6), the listener is a reasonably sized and fairly apparent object, so this is perhaps not surprising. The propensity of human participants to use the listener as a candidate reference less often than a computer model may be to do with reference frame confusion issues as discussed in section 4.6. The computer model is ‘unaware’ of the possibility of reference frame confusion. However, since the largest improvement in results is seen when a variable relating to the presence of a listener is added to the search space variable it may indicate that listeners are perceived as not very good at searching in their own vicinity. For this reason listeners should not themselves be used as the reference object, even if, when considered as an ‘object’, the listener is the most suitable reference from the point of view of determining search space.

## 6.9 Assessment of over fitting

Although cross validation has been used throughout the study for training the Bayesian networks, the possibility remains that the networks are ‘over-fitting’ to the training data and that this is affecting the results. It should be noted that over-fitting is a problem of the model in conjunction with the data-set, and though sometimes one may be considered more at fault than the other, there are cases where it may be difficult to judge which is principally to blame. If the data-set is not representative of the whole population, through being too small, or biased in some way, then a model trained on this data may not fit the population as a whole, even though its structure is theoretically correct for the whole population. However if the training data is representative but the model is over-complex it may result in a misfit to the data which may more genuinely be blamed on the model.

To assess whether either of these is happening to a significant extent a selection of networks were trained in the normal way on 90% of the data-set but then tested on a portion of that training data rather than on the excluded 10%. The whole of the data-set is used for testing over the 10 cross validation slices as before. The results are plotted,



alongside the networks being trained and tested in the normal way, in figure 6.29.

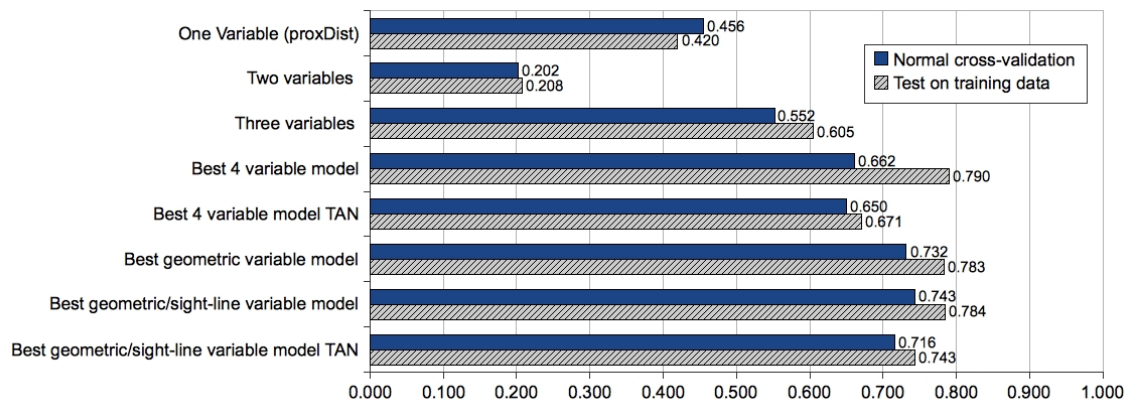


Figure 6.29: Fraction of machine reference choices matching one of the top three human reference choices for a variety of networks tested on a portion of the training data and compared with performance under normal cross validation

There is evidently an influence of conditional probability table size on the results here. The ‘best 4 variable model’, which has a conditional probability table containing 1280 parameters, shows by far the biggest discrepancy (increase in performance) when tested on training data. This will be because, with only about 1500 examples of good references, some training cases will be uniquely entered as examples of good references and when replicated in the test data will be automatically selected as good. If the model is unnecessarily complex this would be a case of over-fitting, if the model is necessarily of this form then the issue is one of insufficient data.

The size of the conditional probability table is a function of the number of variables and the number of values into which the variables are discretised. It is not necessarily the case (see section 5.6.2) that the number of values can be reduced without reducing the performance of the model. With too few values a variable cannot necessarily discriminate between different cases. The number of variables for which dependency needs to be modelled is a parameter of interest in this study and certainly groups of 4 variables should not be ruled out. So it is not clear that the model complexity is unnecessary but equally there is no further data available.

The use of cross validation, in parameter learning, is effective in cases where there is insufficient data, as opposed to cases where training data has a systematic bias (see Jakulin [2005] p74). Cases which are frequently seen in training will tend to be classified correctly when seen in the test data. Infrequently seen cases in training data are also unlikely to be seen in the test data. Use of a binomial prior might help the situation but, as already noted, also complicates the situation in other ways. Currently the study proceeds under the assumption that using cross validation eliminates, to a satisfactory degree, the tendency of (necessarily) large probability tables to over fit to rarely seen cases in the training data.

## 6.10 Summary

The objective of the study was to look at the multiple factors influencing reference choice in a realistic environment, because it seemed that the problem is inherently multi-factorial and cannot only be examined in a synthetic experimental setting containing a very restricted number of objects. This leads however to complexity in the test data set which is compounded with the complexity of the models and the derivation of the variables on which the models operate. Learning anything in this situation has proved more difficult than anticipated.

The unexpected discrepancy between interior and exterior scenes, in particular, is a complication which makes the presentation and interpretation of the results difficult; a model for reference choice based on interior scenes does not work as well when applied to exterior scenes. The reason why this should be so is still not fully understood. Firstly it is not clear that a simple scaling of variables describing an object's size relative to the size of the scene (object volume scaling with scene volume and object linear dimension scaling with scene linear dimension) reflects the way humans judge object size in reference selection. A more complex relationship may be involved and further work will be required to investigate this. Secondly it is also possible that other factors, such as the differences in the way objects are arranged (see section 4.5.4) in interior and exterior scenes may play a part in the reference choice process. In either case a scale, or environment, independent model for reference choice is invalid. The inclusion of scene scale 'switches' does not yet provide a satisfactory answer to the problem.

The sight-line variables offer another way of scaling variables relative to the scene as well as potentially overcoming other complexities such as the different perception of objects and object occlusion and it is not fully understood why these did not perform better. It may be that prior knowledge of actual object size is more important than perceived object size. However it might also be that the correct representation of sight-line variables and the best combination of them and the geometric object variables has not yet been found.

Although the rationale for learning a practical Bayesian classifier based on interacting variables seems sound the implementation is not yet satisfactory. In particular the limitation of the current algorithm in only using each variable once is likely to be reducing its performance. There are indications that the variable interactions represented in the manually defined 'combined clustered' models will lead to performance improvements over the tree augmented naive Bayesian networks as model complexity increases.

It seems that the sufficiency of the test data-set can not be regarded as a binary variable. Rather the data-set should be regarded as sufficient to support different conclusions to differing degrees of certainty. Further extensions to the data-set may lead to more consistent results and allow more significant results to be obtained. It is also possible, with different variable derivations or models, either of which more accurately match human reference choice, that the current data-set would appear far more satisfactory.

The inclusion of the listener figure further complicates the picture and in retrospect

should probably have been left for a future study. Little concrete has been learned about whether the presence of a listener influences reference choice. The results have been documented here in case they are useful for comparison with a future study.

Even accepting the above issues it has been possible to learn a reasonable amount about reference choice. Also, by making the attempt to mimic human behaviour in near real world situations, and expose some of the difficulties involved some basis for further progress has been established. The next chapter considers this in more detail.

# Chapter 7

## System performance, limitations and findings

### 7.1 System performance

#### 7.1.1 Performance of machine models

Assessing the performance of the machine models is not a straightforward task. Even though there are clearly good and bad references in most situations, there is no definitive right and wrong answer, and several borderline cases where two humans might not agree whether a given reference was suitable. Figure 7.1 shows the performance of four machine models plotted on the same graph as the human participants in the experiments from section 4.4.

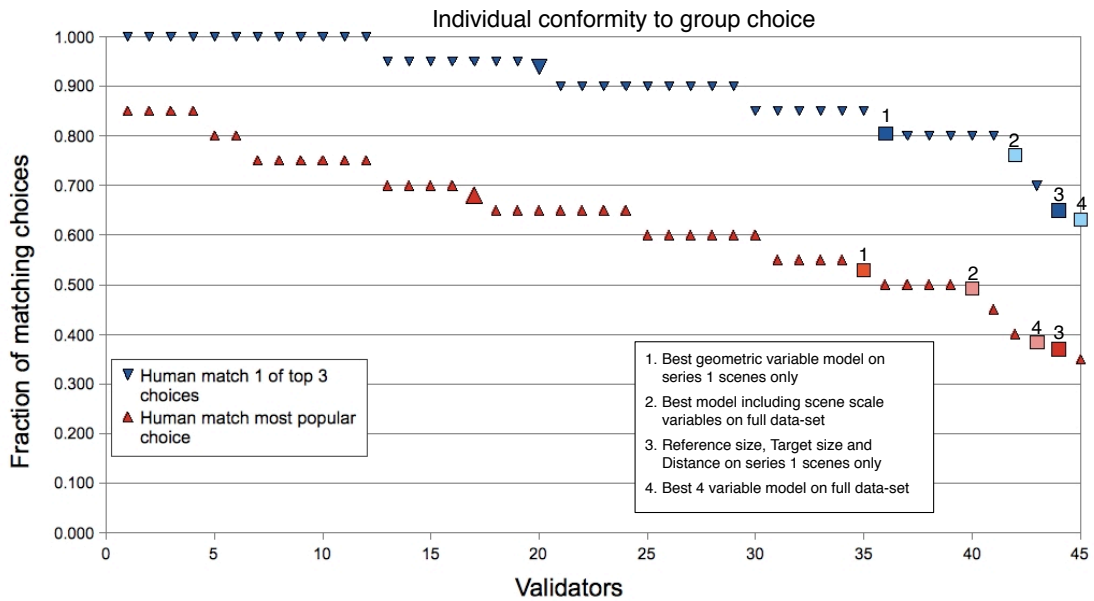


Figure 7.1: Comparison of human and machine performance in the reference choice task

The validation exercise from which the graph is derived took place on the series 1 scenes only, so the best geometric model operating on series 1 scenes only is shown along with the best model using geometric and sight-line variables on series 1 and 2 scenes, which performs less well. The best geometric variable model is still in the ‘bottom’ quartile of conformity to the group view of reference suitability, but if a Turing-type test is considered, it would probably not be distinguishable from a human, although perhaps it might be considered slightly idiosyncratic. This is of course speculative, but it should be remembered that five of the humans who were less conformist to the group opinion than this model were at least competent and fluent English speakers. The comparison of the geometric plus sight-line model with this group of humans is not entirely justified. On the series 1 scenes only this model performs identically to the best geometric variable model and the human participants showed lower conformity when validating series 1 and 2 scenes together although in that case they were also validating listener present scenes. The four variable models and size and distance only models probably would not be considered human on any test data set.

A Turing-type test would certainly not be passed if, although making good reference choices 80% of the time, the machine gave, not just non-conforming but completely outlandish references for the remaining cases. (Remembering that in some scenes there may be more than three reasonable candidate references and the measure being used is to match one of the most popular three choices of the group). The cases where the best performing machine model, using only geometric variables, differed from the group of humans are shown in figure 7.2. Note that there were actually five such cases but the case shown in figure 7.2(a) has a near duplicate. These are all taken from the series 1 scenes that were used in the first validation exercise. None of the machine choices seems entirely outlandish. Possible reasons for the machine model making the choice it did are given in the discussions below. It could be argued that three of the four cases were due to representational issues in the data set rather than simple deficiencies of the model.

The second validation exercise, carried out on the full data set provides further examples of mismatches between the machine choice of reference and that of the human validators taken as a group. These are shown in figure 7.3. The model used this time is the best model using the sight-line variables and the scene scale variable as well as the geometric variables.

The performance of this model on three of the scenes in figure 7.2 was the same as that of the best model using only geometric variables, however in case (b) the machine now matches the group choice. Of the four cases shown in figure 7.3 only one, (b), is possibly a representational issue and is discussed in section 7.2.2. The other three cases are plainly less than optimum references chosen by the machine model, illustrating different sorts of mistakes. The choice of the wooden spoon in (a) is perhaps the most un-human-like, being too small and too mobile. Even though the draining basket, the overwhelming choice of the human validators, might be difficult for a machine it is certainly not clear why the cutlery holder wasn’t chosen instead of the wooden spoon within it. In (c) and (d) the machine has chosen references that are too big, that, in particular in the case of the road in (c), fail

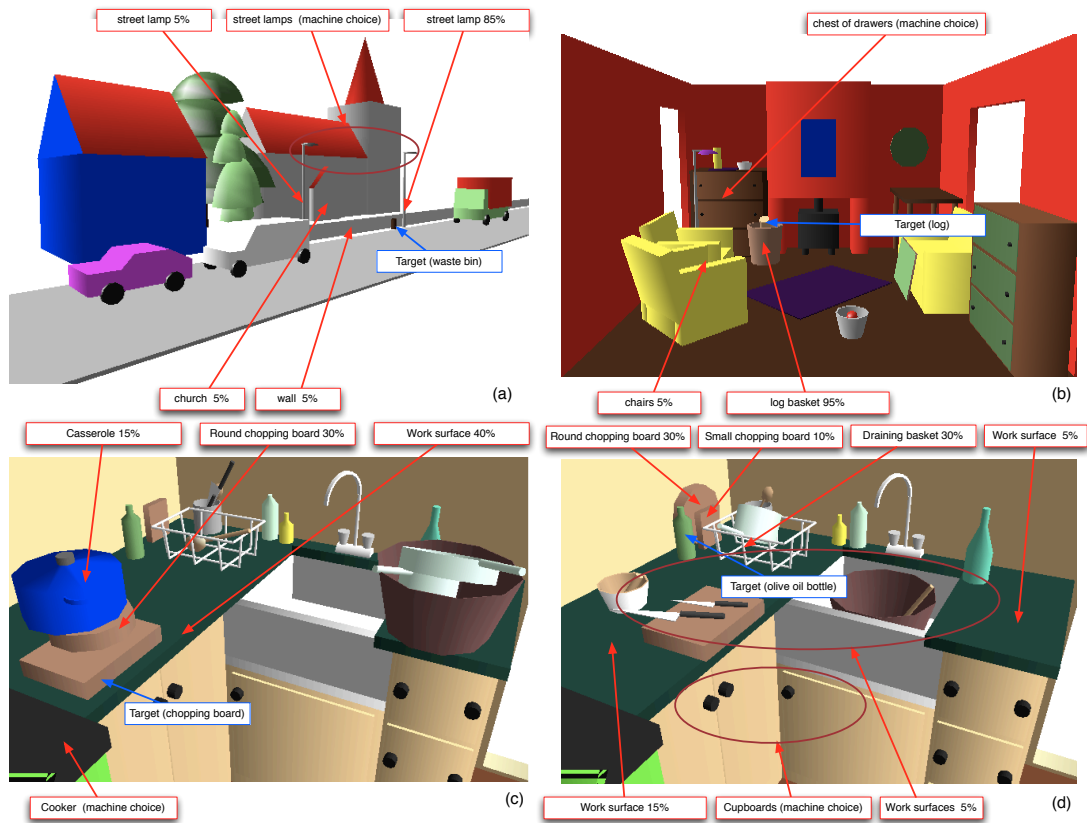


Figure 7.2: Cases where machine model and human (group) choices differ for models using geometric variables, series 1 scenes only

to define the search space for the target satisfactorily. Inspection of the results for most of the better models shows this pattern, that the machine model does not err in a single direction, that is, always choosing over-large, or always over-small, references.

The models using listener present variables can be compared against the human participants shown in figure 4.20. Even with the lower overall conformance of the humans in this case the best model using listener present variables is only just as good as the most idiosyncratic humans. Further investigation into the influence of the listener on the reference choice process is needed and it is possible that there are complexities that are beyond the current data set to extract. An example would be that for a target behind the listener there is a tendency to start by saying “it’s behind you” before qualifying with a second locative expression, or possibly using a hierarchical reference including ‘you’.

The performance of the system as a whole can certainly be improved, but both representational issues in the data-set and the models themselves need to be tackled. Two important questions are whether the system is sufficient to enable anything to be learnt about spatial language and whether it could be used in any practical applications. The answer to both of these questions is, as might be expected, yes and no.

Of the application areas outlined in section 1.5.1 the system as it stands could be used in computer games and training simulators, where the odd sub-optimal reference could be

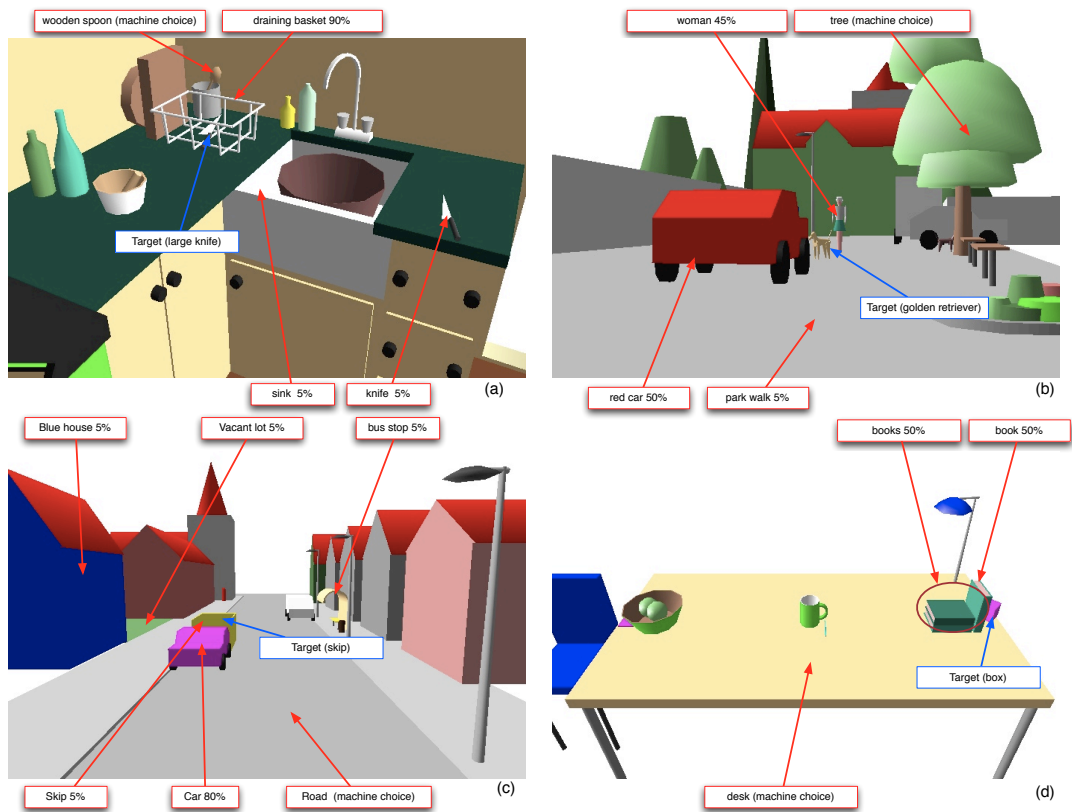


Figure 7.3: Cases where machine model and human (group) choices differ for models using geometric and sight-line variables, series 1 and 2 scenes

tolerated. For mapping and navigation applications more work is almost certainly required, both to reduce the number of sub-optimal references and, particularly, to add capability for reference disambiguation. The “third set of traffic lights” is, in many cases, an effective reference.

The specific areas in which findings have been made about the reference choice task are discussed in section 7.3. There are a lot of areas in which the study provides supporting evidence without being able to decide an issue decisively and many in which the system has failed to address an issue it was hoped to shed light on. The reasons for these are variously discussed in the remainder of this section.

Note also that the performance of the machine models has been achieved without explicitly choosing a preposition although some variables which would also help determine a likely preposition have been used to select the reference.

### 7.1.2 Relative performance of Bayesian network variants

The naive Bayesian networks in which all the variables are statistically independent are clearly inadequate. It therefore also seems unlikely that simple models based on sum-of-weighted-factors, or Euclidian-distance-between-vectors-of-weighted-factors (Gapp [1995a]),

would be usable in a general reference choice situation, with the variables used here. These models are typically less expressive than the naive Bayesian model in which there is no dependence between the variables but in which non-linear weighting factors can be modelled. If composite variables such as the salience variables which combine ‘size’ and ‘distance’ could be used then a linear combination of such variables might be more successful. However the results suggest that constructing and using composite variables such as the salience variables is difficult in practice (without, perhaps, using machine learning in this process as well).

The tree augmented naive Bayesian networks perform much better than expected given that, in theory, they cannot determine some of the variable dependencies which should be necessary to a successful model. Looking at the networks however, shows that although they tend to cluster apparently redundant variables, because they also have to link these clusters, some of the important interactions are actually apparent in the model. This is shown in figure 7.4 which is the tree augmented network for the best geometric model. This network is already ‘pruned’ having many of the initial variables removed and so much of the redundancy. However it can be seen that similar variables (relating to the target or the reference) are clustered due to their high mutual information. A key link between the target and reference variables is represented in the model, even though this has relatively low mutual information, because the tree has to be complete. Variants of the tree augmented algorithm that remove this requirement for tree completeness might be expected to perform worse in this application.

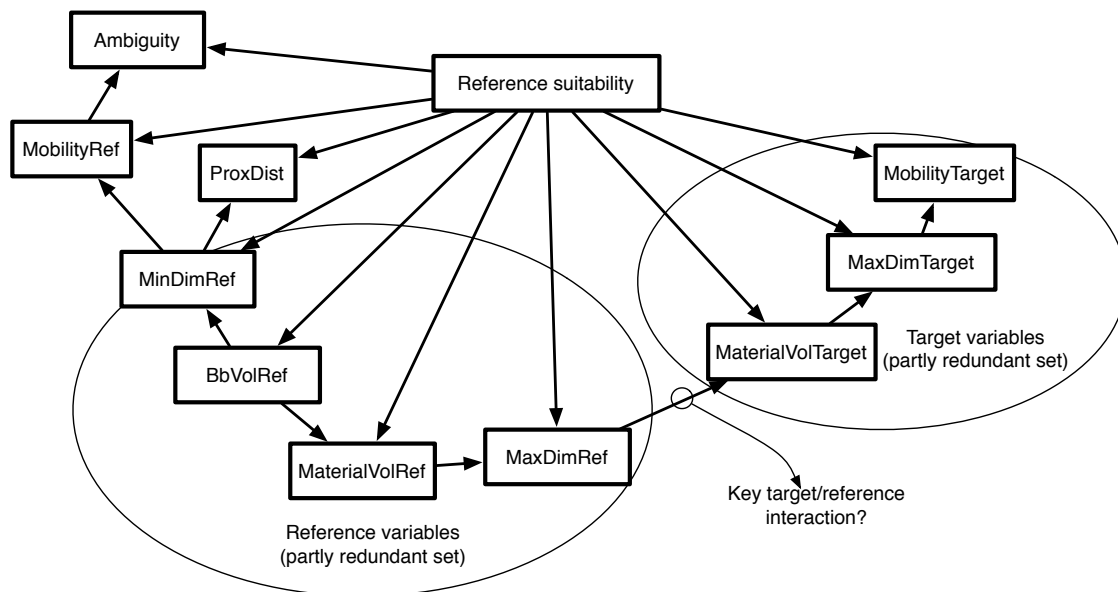


Figure 7.4: An example tree augmented naive Bayes network showing key interactions and redundant clusters

This ‘accidental’ inclusion of links makes the network fragile such that the addition of further variables can lead to the accidental removal of a key dependency. This is probably



the cause of the reduction in performance of the tree augmented networks when the sight-line variables are added (see figures 6.20 and 6.23). Previously the tree augmented networks had performed broadly the same as the best hypothesis model networks.

The performance of the learned structure combined clustered networks relative to the tree augmented networks is slightly disappointing but could probably be improved with further investigation. The results from figure 6.23 suggest that it is consistently if not significantly worse than the tree augmented networks for most starting variable sets. The algorithm appears to be doing well at excluding redundancy as indicated by its performance, relative to the tree augmented networks, when the starting variable set consists of all possible variables. As noted this also results in poor performance when the starting variable set is small, so the algorithm needs some adaptation in this area.

The advantage that both the tree augmented networks and the hypothesis model networks have over the current version of the learned structure combined clustered network is the actual (or effective) multiple instantiation of evidence variables. This is clear in the hypothesis model networks of figures 6.9 and 6.19 where evidence variables such as *materialVolRef* are parent to more than one hidden variable. In figure 7.4 (and remembering that in cases with no missing evidence values the tree augmented network decomposes into pairs of variables with the classifier as a parent to each pair) it can be seen that, for instance, *minDimRef* appears clustered with *BbVolRef*, *MobilityRef* and *ProxDist*. It is thought that developing the combined cluster algorithm to allow this multiple instantiation would significantly improve its performance, although doing this without also incorporating damaging redundancy may be difficult.

### 7.1.3 Were the best models found?

If the measure of best is taken as the highest fraction of reference choices matching the human group choice it is almost certainly true that the best models were not found. This could be either because the best model is not a Bayesian network (as noted in section 5.2) or because the best Bayesian network was not found. A search based technique, possibly a genetic algorithm, could have been employed to locate better Bayesian network structures. However the danger of over-fitting when using a search based on results of previous structures is considerable, particularly when the data-set is limited. The network structures used are either learned from statistics derived from the data in a single step (not also taking account of results of previous structures), or are derived from concepts taken from the hypothesis model, and should be relatively free from over-fitting.

Whether a better network could have been found that enabled anything significant to be learned is more questionable. As discussed below, the limitations may well be in the data set, the variable representations or the use of a single classifier in what may be different classification tasks (or at least scenarios).

## 7.2 System limitations

### 7.2.1 Limitations of the test data set

#### **There may be insufficient training data.**

Figure 6.11 shows the performance with reduced training set sizes while testing on the full data set as well as the initial (series 1) scenes. The likely rate of improvement of the system with further increases in data-set size can be seen to be very slow (the gradient of the curve at the right hand end is low). This suggests that just adding more similar scenes will not significantly improve system performance, but whether this is the whole story is not so clear. Some findings from the experiments, including the lack of influence of angular relationships (discussed below) may suggest that more data would improve performance but it is also possible that what is suggested is that humans use a variety of learning techniques rather than solely digging statistical relationships from complex and noisy reality.

#### **The scenes may not be sufficiently realistic.**

The results from section 6.4 indicate that, in one important respect, this is unlikely to be the case. There is no apparent correlation between the number of objects in the scene and the results (in terms of matching human reference choices) from the machine models. Making the assumption that the human participants are unaffected by the number of objects in a scene suggests that the process of ignoring some of the ‘clutter’ in real world scenes has not affected the results.

The representation of the objects themselves, in terms of detail or realism, does not materially affect the machine models. As long as no significant geometric features are missing and the size of the objects is accurate the machine models will return the same result irrespective of the level of realism. No evidence from the results bears on the question of whether the object representations affect the human participants so we are left with the anecdotal support in that none of the human observers commented on the crudeness of the object representations or the difficulty of interpreting the scenes. Some types of objects are either absent from, or badly represented in the scenes, and these include non-rigid or conformable objects, such as fabrics, and penetrable or partially space filling objects such as foliage. This leads to the omission of some spatial relationships (for instance “the tablecloth is over the table”) and possibly some difficult reference choice cases. Overall it is not expected that object representation has materially affected the results and it should also be noted that in this respect the data-set used in this study is at least as good as in other contemporary studies (for example see Byron et al. [2009]).

#### **The corpus as a whole may not be sufficiently diverse**

There seems to be no way to assess this in the general sense and no target for ‘sufficient diversity’ that can be easily defined. In the limit a highly diverse scene set would contain scenes strange to many humans (who may not have been to the Sahara desert say) but on

the other hand it is true that human ability to describe scenes is not dependent on having prior experience of them.

As noted there is a range of scene scales that might be characterised by large indoor spaces such as an open plan office or an airport check-in hall, or small external spaces such as a suburban garden that are missing from the corpus. These would have enabled a more complete investigation into the effects of scene scale.

Scenes with even larger scales and scenes with significant height extensions (hills) are also missing as they are significantly more difficult and time consuming to produce. It is possible that a range of descriptions and influences on reference choice may have been missed due to the different apparent spatial arrangement of objects on inclined planes. On balance it is felt that lack of diversity is not a major limiting factor on the performance of the system as a whole, and will not be so until some of the issues surrounding variable and object treatment are resolved.

### **The validation process may not match human behaviour in real world situations**

The process of selecting a reference object from a drop down list, for a target in a scene on a computer screen, as described in section 4.4 is clearly not the same as that of forming a verbal description of a real world scene. From observations of the validation process it seems that participants do one of two things. Either they form a description (often verbally) and then use the drop down list to match the reference choice to the description they have made, or they go straight to the list and scroll through the objects until one is highlighted that allows a good description. Although the second of these seems less natural than the first the underlying thought process may be the same, the difference simply being the point in the process at which the computer is used to illustrate the objects under consideration. Experiments could be performed to compare the validation results against scene descriptions in real world scenarios however these would be relatively resource and time consuming and at present reliance is placed on the fact that the vast majority of the descriptions given are reasonable and effective in locating the target. Also note that only using the three most popular reference choices in a scene for training tends to filter out odd references that are either mistakes, or effects of the clumsiness of the process.

Future validation exercises should try to use direct selection of objects from the scene with a pointing device, primarily because this would reduce the effort required rather than because it would produce more realistic results. Use of natural language descriptions is clearly more problematic, in particular because of object naming differences. A desk might be called a table or workbench for instance and the study has already run into minor cultural difficulty by calling a ‘dumpster’ a ‘skip’. Natural language descriptions would either have to be cleaned up by hand or an ontology would have to be employed to reconcile object names. Even so errors due to language use rather than selected reference object quality would almost certainly be present in the training data set.

### **Occluded and partially visible objects**

Occluded objects in scenes are ‘visible’ to the machine model in a way not available to the human validation participants (apart from the author, who ‘knows’ they are there). This is illustrated in 7.2(c) where the machine ‘sees’ the whole of the cooker which it has chosen as a reference although it extends outside the area of the scene visible to the human participants (who did not choose it). This particular case is an example of the scene representation rather than the ‘real’ occlusion of objects causing the discrepancy. Although in practice it does not seem that there are many cases in which this will cause the machine to produce a really bad reference it is an important issue. As with the use of absolute geometric extensions it might have been thought that the use of sight-line variables would significantly improve matters and it is surprising that this is not the case.

### **Screen resolution and rendering**

This is closely related to the issue of the realism of scenes discussed above. The screen resolution and object rendering (particularly the use of lighting within OpenGL) puts a limit on the number of objects that can be incorporated into the scene whilst leaving the scene ‘readable’ by human validators. As noted above this is not thought to be an issue that has materially affected the study. Use could be made of more sophisticated OpenGL lighting techniques in future studies to confirm that this is not significant.

## **7.2.2 Possible deficiencies of machine models**

### **Use of scene scale relative variables**

The use of object size variables scaled relative to the scene they are in has caused some unanticipated problems. These are illustrated in section 6.6 in as far as they are appreciated, however there may be other issues that have not become apparent. The question arises as to whether another system of variables could be devised that would not encounter this problem.

If the decision had been made to use (say) target referenced variables to represent candidate reference volume or linear extension the problem would not entirely have gone away. An initial assumption that the absolute size of the target (or reference) would make no difference could be made, but this was not thought to be satisfactory as a basis for this study. So a system could have been devised that used an absolute measure of target size along with a set of target-relative size measures for candidate reference objects. However the target object still varies in volume by 7 orders of magnitude (in this study, it could obviously be more), so either a coarse representation of target volume, or an impractical number of bins, or a scene scale variable will be required.

It should be noted that there is an implicit scaling in the sight-line variables. However large the space represented in the scene it is effectively projected onto a screen which is randomly sampled 10,000 times. In any case the sight-line variables have been shown to contribute only marginally to effective models of reference object selection.

So is the apparent dependency of reference choice on scene scale an artefact of the variables chosen to represent the relationships between objects in the scene along with the way the model averages scenes over a range of scales or is it a genuine adaptation of human linguistic judgement to different scales? The suspicion remains that the variables used are not a perfect representation of the way in which humans process visual scenes over a range of scales and that if they were improved the need for a scene-scale variable would be removed. However there does seem to be a genuine difference in the distribution of object sizes between scenes of different scales, even if this is just a way of stating that “in general large scale scenes tend to be outdoors and in general there is more air and less solid matter outdoors”.

### **Independent treatment of reference candidates**

Candidate reference objects are considered independently and do not interact in the machine models with other references. Although there is little experimental evidence from the test cases, there are plausible mechanisms (see Gapp [1995a]) by which the choice between two candidate references could be altered by the characteristics or positioning of a third, and this requires further investigation. Inspection of the machine choice in figure 7.2(d) suggests that the machine may not have treated the interposition of the work-surface between the target and the cupboards in the same way as the human participants. It illustrates a situation where the space is divided vertically into two volumes by a planar horizontal object (the work-surface). In cases like this it seems that humans are more reluctant to choose as a reference an object in the space not containing the target object. The machine models used have no means of expressing this preference and to address this issue would require a much more complex model. It is not clear why the machine has not chosen the work-surface as a reference on its own independent merit, and the characteristics of the other references chosen by the human participants are entirely different from the machine choice of the cupboards. Note though that this is a test case that the humans found difficult judging by the lack of consensus displayed.

### **Identical geometric treatment of all objects**

An assumption implicit in the machine models is that all objects are treated the same way by humans. That is to say that humans use the same geometric measures for all objects. It is possible that, for instance, knowing that a bowl is a container a human would assess its size by its convex hull volume but that a chair or table might be assessed by its material volume as this relates to its mass or perhaps ‘presence’, (the assumption being made that these considerations are appropriate for reference object selection). Again it might have been thought that sight-line variables would have been better at addressing this sort of issue, in particular for the surface objects (such as roads), that do not have an easily definable ‘volume’. It is possible that one reason for this apparent failure of the sight-line variables is that they are used by humans for some objects but not for others (or at least are given more weight for some objects) and the current model does not allow for

this. Different geometric treatment of different object types directly would require a full ontological background for classes and perhaps uses of objects.

### **Use of absolute extensions**

Deciding whether a candidate reference object locates the target adequately by considering only the absolute extensions of the objects without considering their orientation will be error prone. Again it is not clear that this happens often in practice but cases where it could cause error can be envisaged. Absolute measures of extension may be good enough to resolve the man on the sidewalk issue (see section 3.4) but a man standing in front of a long train may be adequately referenced by the train whereas a man standing beside it may not. This latter case would require that the extension of the train relative to the position of the target be modelled. In the experiments performed the use of the maximum dimension of the candidate reference object proves a better predictor of reference suitability than the search distance measures derived from ray-casting (sight-line variable) as shown in figures 6.12 and 6.16. Although there are some shortcomings with the sight-line derived variables this is still a surprising result and not yet explained.

### **Treatment of salience and ambiguity**

Salience measures such as colour contrast and brightness are not yet part of the machine model but may be being used by human validators (see sections 3.4.2 and 3.5.1). Salience in the model is treated to a certain extent by the measure of ambiguity. In all discussions of salience there is an implicit acknowledgement that the salient object stands out from an environment that presumably contains other non-salient objects. A red house in a row of grey houses may be easy to spot but this is paralleled by its ease of incorporation in a locative expression, as discussed in section 3.5. A grey house in a row of red houses would presumably be treated similarly. This suggests that a strong correlation can be made between ambiguity and (lack of) ‘visual salience’. In the case of a single red house and a single grey house that are otherwise equally good references the red house may be chosen more frequently by human participants but in practice a lack of this factor this does not seem to affect the performance of the model significantly.

The fairly crude measure by which ambiguity is modelled is an issue which needs to be addressed. Currently objects are ambiguous if they are identical (identically named) and unambiguous if they are not. This polar approach would not translate well to the real world or to other virtual environments.

### **Object aggregation**

Aggregation of candidate reference objects is handled in too crude a fashion. It is included in the model as a way of dealing with ambiguous reference objects but it is clear that it does not cope with the complex way in which humans aggregate objects. Firstly aggregation of non-similar objects is not allowed for, as in the case of the fruit and the bowl aggregating

to ‘fruit bowl’ noted in section 4.5. To solve this problem completely requires a thorough ontological knowledge of object categories and uses. Secondly some aggregated objects, particularly if they produce a composite which is highly dispersed are not well ‘visualised’ by the model. This is the case in figure 7.1(a) where ‘street lamps’ is a distinctly odd reference. In this case though the representation might be thought a little careless and, had there been a row of many street lamps (as world normally be the case), instead of just two, the machine may have chosen differently.

### Angular relationships

The angular relationship between target and reference is not included in any of the best performing models although there is evidence that this does play a significant part in human reference choice. It is clear that it is more costly to say “The bird is above and to the left of the tree” rather than “the bird is above the barn” both for the speaker and listener. The barn is more likely to be chosen as a reference (all other factors being equal) for this reason. This is reflected in the findings of Carlson and Hill [2008] and Carlson and Hill [2009]. It is also true that humans prefer to use projective prepositions other than left or right because of the possibility of reference frame confusion. In the experiments performed no difference was found between the performance of ‘reference frame aware’ and speaker-relative longitude measures. There are other possible complications in the modelling of angular relationships as these can also be affected by dominant objects in a scene fixing a pseudo-global reference frame. Objects lined up on a table may take ‘left of’ and ‘right of’ designations aligned to the table (object-relative) rather than to strictly intrinsic or speaker-relative frames. This is illustrated in a simple (but effective) experiment conducted by Jording and Wachsmuth. [2002]. This effect is not modelled in the study at present and may account for the simple angular relationships not having a visible influence in the complex scenes employed. The vertical angle however should not be affected by reference frame choice and this does not seem to provide any useful prediction of reference suitability either. It is possible that the role of angle does not, in practice, in real world situations, come into play very often and that the data set does not contain sufficient examples to allow the variable to discriminate between otherwise similarly suitable references. The work of Carlson and Hill [2008] used very simple geometrical arrangements containing a target and two potential references, one of which was on a cardinal axis, and the other at 45 degrees. The references were visually similar although one had a ‘functional’ relationship with the target. (A burger, a mustard jar and a tube of toothpaste would be a typical object set). Carlson finds that the reference on the cardinal axis is preferred irrespective of any functional relationships. This is clearly a different situation to the object arrangements in this study where, if the cardinal axis preference is discernible, it will be deeply buried in the ‘noise’ of different object sizes, distances and angular placements. The key question is whether humans arrive at their inclusion of cardinal axis preference through a process of averaging out the noise over a vast array of diverse scene descriptions (as would need to be the case with the pure Bayesian learning in this study) or whether they are learning and applying some hybrid

model in which different preferences and influences are combined from different training sources. The latter seems likely and some such models are discussed in section 7.3.4.

### Topological relationships

The topological relationship is not included in the best performing model. This might seem to account for the case in figure 7.1(b), in which the overwhelming consensus among the human participants is that the log is ‘in’ the basket, but the machine model chose a different reference. However topological relationships will be largely subsumed into a distance measure and would be well modelled by it. It could be argued that the choice of the chest of drawers by the machine is not too bad, considering the basket may move around. It would be interesting to see what reference the humans gave if there had been several logs and these as a group were the target. The tendency to aggregate the logs (plural) with their container (as a vase containing flowers is usually referred to as ‘flowers’) may have produced the chest of drawers as a suitable reference. Clearly it would be expected that the assignment of a preposition (once a reference has been selected) would be strongly influenced by the topological relationship, but that is not what the model is trained for in this experiment.

### Exclusion of preposition choice

Related to the issue of the expression of angular and topological relationships in the model is the omission of the actual choice of preposition from the model. Although, in a *locative* expression, the choice of a suitable preposition will be largely determined by the topological, angular and distance relationships between the reference and target, the listener does not know these relationships before the target is found. The information the listener has to work with is the preposition, and the expectation of where to search given the preposition, once the reference has been located. It follows that the speaker should check that the preposition chosen leads to an appropriately easy search for the target. There would appear to be two, or possibly three, distinct cases where this might not be the case:

1. The preposition fits the spatial relationship between reference and target well but the resulting search is difficult. This would be typified by the case of the “man on the sidewalk”. It is difficult to imagine using a different preposition but the listener may be left with the task of scanning a large extent of space to find the target. The choice here would be to find a better reference or move to a compound locative expression. In this case it is not clear that knowledge of the preposition would supply anything to the model that is not supplied by the reference and target characteristics.
2. The preposition fits the spatial relationship in an acceptable way, leading to a misguided search. This would be typified by the case of “the bird is above the tree” as the bird moves progressively off the vertical axis above the barn. It is also well illustrated by the object placements used by Carlson and Hill [2009]. There will be a



range of locations where ‘above’ is the most suitable (single) preposition to pair with the reference and yet which could lead to an extended search for the target, as the listener will, presumably, commence their search along the cardinal axis denoted by the preposition. In this case not only is the search space extended but the listener has been misguided to an extent. The options for a speaker are, as has been noted above, to accept the cost of qualifying the preposition or to use a different reference. This is a case that is not covered by the representation of search space currently included in the model. A comparison with an ideal spatial template for the chosen preposition would be one way to solve this issue and for this the choice of preposition would be needed in the model. In reality though, the offset of the angular relationship, from the cardinal axis in question, is the key determinant. This is needed only if a directional preposition is under consideration and so could point to a need to model at least the *class* of preposition.

3. An intermediate case exists for proximity prepositions such as ‘near’ in the case of “the man is near the fountain”. As the distance between the man and the fountain increases the search space also increases. The speaker has the option of changing to a directional preposition to reduce the search space at some point in this process and once this decision has been taken this case would appear to devolve into one of the two discussed above.

It does not seem clear from the above cases that consideration of the actual preposition choice, as opposed to the elements of the spatial relationship, is needed in the model for reference choice. However it cannot be considered proven that it is not. As noted in the second point above the use of the angular offset from the cardinal axis is only required if the preposition under consideration is a directional one rather than a proximity or topological preposition. This though could be modelled by a dependence on the distance variable, making the assumption that at larger distances directional prepositions are more likely and that the angular relationship becomes more influential. This would be in line with the findings of Regier and Carlson [2001], although they were not considering a transition between proximity and directional prepositions.

What does seem clear is that the longitude variable used in the model may not be sufficient on its own. It may be necessary to distinguish left and right from front and back (which might be the preferred relations due to reference frame confusion), but is probably too coarse grained, as currently instantiated, to model angular offset from a cardinal axis. A second ‘cardinal axis offset angle’ variable could be included to account for this, that would have a more finely graduated range of  $-\frac{\pi}{4}t + \frac{\pi}{4}$  measured from the relevant cardinal axis.

### **Sight-line variable limitations**

The fact that the sight-line variables made so little difference to the model performance is surprising given that various specific examples had been identified (figure 7.2(c) for

instance) where they should have had an effect. The problem with the salience variables, given the superior performance of the combination of separate viewability and distance variables, would appear to be that the defined relationship between apparent size and distance (effectively the integral of target-to-reference distance over the visible area of the reference) does not well represent the way humans are dealing with these quantities. Use of a salience measure instead of a distinct reference size measure (not already combined with distance) also suppresses the ability of the model to represent the relative size of the target and the reference.

Although there seemed to be little to choose between the different salience measures the case in figure 7.3(d) is interesting. The car, which is the most popular choice for the humans is in fact further from the target (the golden retriever) than the tree which is the choice of the machine. In the scene this is far from clear and the car appears the better reference. Kelleher's salience measure would reflect this better than the 3-dimensional measures postulated in this study. Whether in this case it would translate into true 3-dimensional reality where humans could use binocular vision as well as cues from perspective is open to question. At least one common case, "the moon is rising above the trees", demonstrates that humans will project descriptions, and reference choice, into more or less 2-dimensions.

Although the search distance measures as they stand enable the models to express the required concept they could be improved in two respects. Firstly the search distance measures use the distance from the target to the intersection of the sight-line on the reference. However they could use the perpendicular distance between the proximal vector from the target to the reference and a parallel vector from the intersection point (see figure 7.5(b)). This would make search distance independent of the distance measure and allow the two variables to better define a search volume. Secondly, and again leading on from the need to define a search volume, the search distance measure should really be a search area measure. The current measure models the case of 'the man on the sidewalk' reasonably well but does not completely reflect the more extensive search that would be required for 'the man in the piazza'

The area of the convex hull of the intersection points projected onto a plane perpendicular to the proximal vector might be the most satisfactory measure of search area. (Or possibly the projected area of the object itself if viewability was not thought to be a consideration in defining search space). It was thought that the weighted versions of search distance would improve the representation of the concept and why this has not proved to be the case is not understood.

### **Combined clustered model expressiveness**

It is not clear that the Bayesian representations used are expressive enough at present. Increasing the number of variables in the hidden nodes so that they are no longer trained as if they were equivalent to the classification node would allow linearly inseparable functions to be expressed at this level in the network. Extensions to the training scheme would be

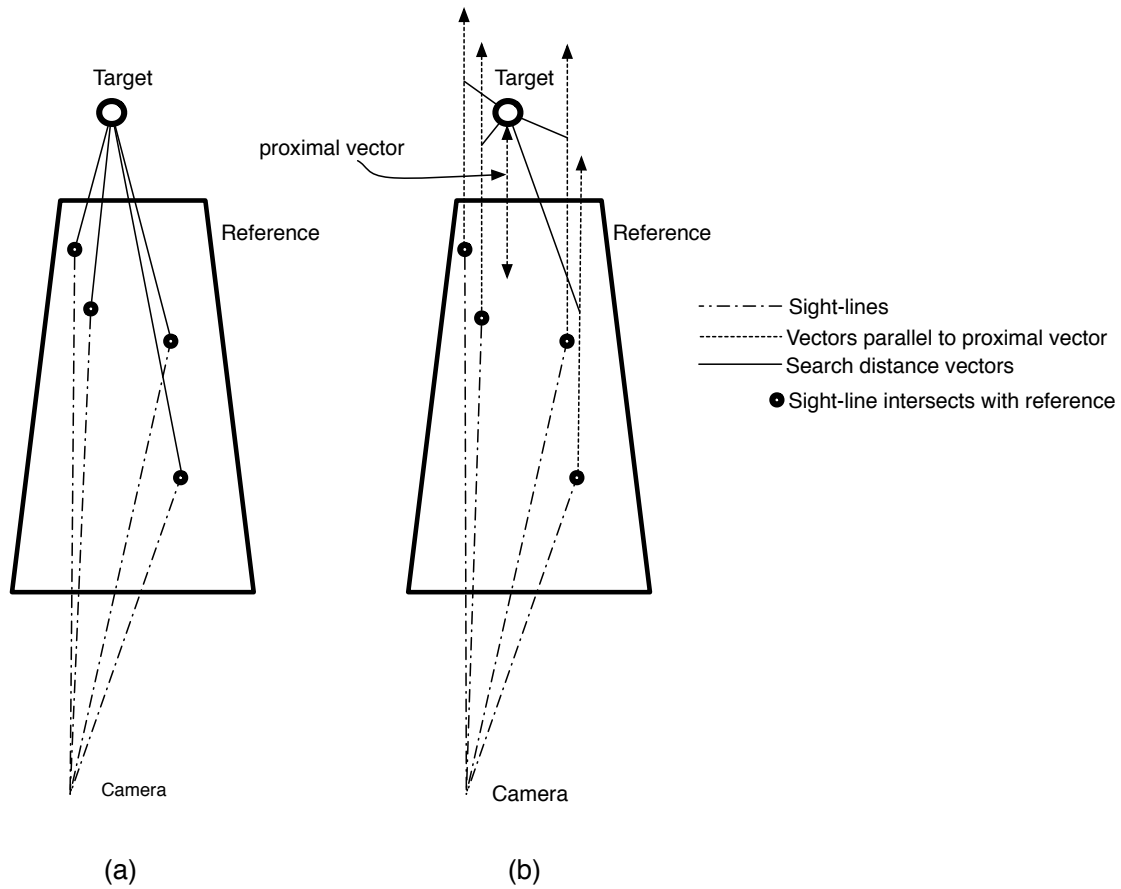


Figure 7.5: Alternative derivations for search distance vectors: (a) as used in the study, (b) from perpendiculars to vectors parallel to the proximal vector

required to make this possible.

## 7.3 What has been learned about spatial language?

### 7.3.1 Influencing factors and variable representations

The results given in chapter 6 lend support to the validity of some of the characteristics of reference objects and landmarks described in chapter 3. In some cases statistically significant evidence is provided for this validity. The findings are summarised here:

1. Perhaps not surprisingly, the idea that the reference object should be 'appropriately located' or 'near' (see Burnett et al. [2001], in the landmark context) the target is confirmed. This is the single most influential factor in determining reference suitability. Further than this it is clear that it is the closest point to the target on the reference, not the reference centroid, that is used as the basis for the distance measure. There is some indication that the closest point to the reference on the target is also used, but no statistically significant evidence.

2. The relative size relationship between the target and reference is given some support [Miller and Johnson-Laird, 1976, Talmy, 2000]. Using a target and reference volume measure is always better than using a reference volume measure alone but only significantly better (at the 0.025 level  $N = 9$   $W = 36$ ) for the convex hull volume measure. The use of any of the reference and target volume measures in conjunction with distance is significantly better than using distance alone. Although bounding box volumes perform consistently worse than either convex hull or material volumes there is no statistically significant evidence for a particular volume measure.
3. There is significant evidence that choosing a reference that could be confusing, such as one of a row of streetlamps, is avoided by humans. In one scene in particular (figure 7.2(a)) with only two streetlamps and a lack of a suitable alternative reference a majority chose the most appropriate streetlamp, but overall this is clearly not the preferred option. It should be remembered that the models used in this study, and the validation exercises given to the human participants, do not currently support disambiguating references with descriptions or other positional or count qualifiers. This might be thought to weaken the evidence since only an all-or-nothing choice between discarding the ambiguous reference altogether or grouping all identical objects is allowed. Thus there is no evidence about how much extra cost in respect of disambiguation would be borne as a trade-off against using a less appropriate but unambiguous reference.
4. The evidence with respect to mobility is mixed. Adding reference mobility alone to a model containing a distance as well as reference and target size measures resulted in a significant improvement. However in more complex models the addition or removal of mobility measures made little difference. The lack of information in the study as to when the locative information was to be used by the listener may have resulted in mobility being assigned little importance. It had been thought that the influence of mobility would be more or less general (see Talmy [2000]. “The chair is beside the dog” seems wrong in any context).
5. The evidence is similarly mixed with respect to topological relationships between reference and target although there was no prior expectation from literature that any given topological relationship would result in an object being a better reference for the target.
6. Variables related to geometric extension are clearly important in selection of reference objects. No other explanation for this is evident except that it relates to the goodness of the reference in defining the space in which the listener has to search for the target. This characteristic of reference objects is largely overlooked by linguistic commentators, but appears more important than the mobility of the objects or any topological relationship between them. The poor performance of the search distance measures which, even given their limitations, should have provided a better measure

of search space, confuses this picture. Further investigation is necessary to establish the right representation of search space as none of the variables used so far, either singly or in combination, seems entirely satisfactory.

7. In spite of the possible shortcomings of some of the salience measures it must be considered highly probable that prior knowledge (or inference) of an object's actual size is an important factor in reference selection. The use of variables relating to an object's apparent size or visibility from the viewpoint of the speaker do not perform as well as measures of actual size. One explanation for this is that use of an object's known or inferred size removes the dependency of the description from any particular point of view. If the listener's point of view is not known this is important. This does not mean that the salience measures are not appropriate for reference resolution (Kelleher and Costello [2009]), which is a different task.
8. There is support for the view that reference choice is independent of preposition choice from the performance of the models and examination of cases where the model chooses wrongly, which do not seem to be related to the lack of a decided preposition. This is not the same as saying that the existence of spatial relationships that could be linked to specific prepositions is not important however (see the discussion on the exclusion of preposition information in section 7.2.2).
9. Perhaps the key finding of the study as a whole relates to the complexity of the model used. As already noted models which treat variables as statistically independent are not satisfactory. There are some indications but no significant evidence that, for more complex models, the tree augmented networks are not performing as well as models which can specifically incorporate interaction information or clusters of more than two variables. Models containing four key variables are significantly less good than more complex models although within the complex models it is not possible to say with certainty what factors are making the significant difference.

This does not amount to comprehensive support for the hypothesis model. In particular absolute measures of reference size, as contributing influences to reference locatability, do not give the best performance. Models in which the reference volume variables are combined with target size measures perform better. As an example the [shape] hidden variable used in several models has as parents only reference size or extension variables but if this is used as a measure of reference suitability instead of the [relative volume] variable performance is significantly reduced.

As noted there is some support for the idea of search space, from the obvious influence of target reference proximity and also from the significant influence of reference geometric extension.

There is no support for the concept of communication cost in spite of the positioning of the ambiguity variable as a parent to a hidden [communication cost] variable in some models. Since in the context of a simple locative phrase there is no scope for reference

disambiguation (with increasing communication cost) no judgements about the effect on reference choice of increasing communication cost can be made. The only real measure of communication cost in a simple locative phrase, that of the utterance length of the reference, has no appreciable effect.

### 7.3.2 Human performance in the reference choice task

From the results presented in section 6 we can infer that humans are using a sophisticated model for choosing reference objects. Certainly it is more sophisticated than simply picking an object of the right size in reasonable proximity to the target. Quantifying the performance of human reference choice models is difficult but an indication is given in figure 4.12 and the significance of the human group conformance is discussed in section 4.5.3.

It is probable that different people are using different model variants and that this leads to different levels of conformity to the reference choices of the group as a whole. However taken as a group the models produce very similar results. This conformity does not necessarily equate to effectiveness but it can be argued that the only plausible driver for this conformity is the need for effective communication. Experiments could be derived to measure the time taken by human participants to locate a target given references produced by human and machine models and work in this direction is under consideration. However given the level of matching between the machine and human models it is thought that to acquire significant data from measurements of search time could be problematic.

It is not possible to say whether humans are using the same variables organised in the same way as the best machine models but it seems unlikely. What can be said is that there is a strong correlation between the variables used by humans and those used in the best machine models, that is they must express the same concepts and processes used to arrive at a judgement of reference suitability.

It seems even more unlikely that humans learn the reference choice task in a manner analogous to the machine models, from correlation of scene and object characteristics with reference choices of other humans. Humans have the possibility, not present in the machine models used here, of monitoring their own visual search task and selecting references that minimise this. This selection method may be augmented with ontological knowledge of the objects involved that might be acquired from observation of the objects directly or from other humans.

The apparent divergence of different human models may be due to the relative performance of humans in different parts of the target location task. The overall task consists, if the hypothesis model is accepted, of an unguided search for the reference object and subsequently a guided search for the target once the reference has been found. A human that performed worse in the unguided search might be expected to choose a more obvious reference even if it made the guided part of the search more difficult as this may minimise the overall difficulty. A human with poorer visual search capability might choose a more 'costly' reference (in the communication cost sense, this might for instance involve more frequent use of compound locative expressions) if this minimised the visual search task.

The machine model has also been evaluated by its ability to ‘match’ human reference choice, that is, to conform to human behaviour, which is not necessarily the same as selecting the most effective reference. If a machine was devised to optimise the overall task its vastly different capabilities might lead to entirely different choices of reference from those of humans. Hence conformity to human behaviour is a more useful concept in this case than some arbitrary measure of machine performance.

### 7.3.3 Extension to compound locative expressions

Various possible algorithms for this can be investigated using the model as described in an iterative fashion. For instance this could be achieved by *conceptually* moving the listener within the scene to a point closer to the target object and selecting an appropriate reference object and then making this the new target object with the listener moved further towards his initial position. Plumert et al. [2001] suggest that in fact the process when direction giving happens in the reverse order. This has the advantage of allowing the listener to start the location process by moving to the region of the most apparent reference before the more detailed parts of the locative expression are uttered, thus saving some time. Both directions of expression formation are illustrated in figure 7.6 which also highlights some of the issues involved. In particular it can be seen that if the references are being identified in descending order of apparentness (that is starting with the ‘finance office’ in figure 7.6), there is no obvious way of stopping the model choosing a long series of references that are all very apparent or unambiguous. The next target is not yet defined when the reference is chosen so a relevant search space is difficult to define. As shown some form of cumulative cost could be obtained and used to ensure that reasonable progress was being made to locating the eventual target.

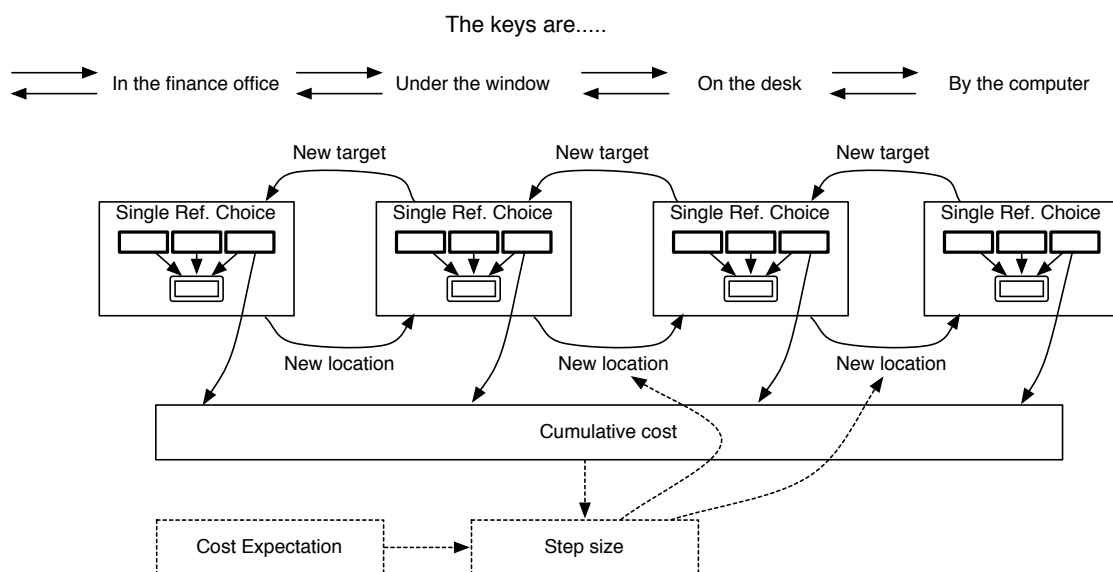


Figure 7.6: An extension to the model to generate hierarchical references

Working outwards from the target (that is, conceptually moving the listener to a point near the target to define the first reference, the ‘computer’ in figure 7.6), is a more obvious process. Targets are always defined, the last reference becoming the target at each iteration of the reference choice model. The tendency to string together too many references could be countered either by monitoring cumulative cost as with the descending reference case or by conceptually moving the listener ‘away’ from the target until only one or two references met some pre-determined threshold of suitability.

In many cases there are recognised steps in the process of hierarchical reference formation; buildings are divided into floors and rooms; towns are divided into districts, streets and buildings; this simplifies the process to a large degree. A complicating factor is that the process often doesn’t proceed in a single direction, in particular when disambiguating references, or references that are parts of objects, is involved. Extending figure 7.6 illustrates this case; the description “the keys are in Harrison building, in the finance office on the second floor, on the desk under the window, by the computer”, which contains several changes of direction, is lengthy but does not sound odd.

In practice learning the conventional sub-divisions of space (streets, buildings, floors, rooms etc.), which simplify the hierarchical reference choice process for humans, may be a more difficult task for a machine, than generating the reference string represented by “on the desk under the window, by the computer”. This would certainly be true if the machine model is allowed to always operate an ascending reference choice and learn other rules to post-fix an order of utterance.

### 7.3.4 Better forms of model

There are two clear pointers from the results to areas where the structure of the model, a simple, single Bayesian network may be inadequate. The first is the use of what might be called ‘context switches’ to improve results across scene scales and with a listener present. The second is the treatment of different types of objects, in particular ‘surface’ objects (such as roads) in the same way as objects with defined volume.

Although only two ‘context switches’ have been used it is easy to envisage others that might be needed, perhaps when the target is visible/not visible or when the listener is on foot/in a car. It is also possible that influences such as mobility and where angular relationships are important, which are difficult to learn in the context of a large general model, should be dealt with in this way. Figure 7.7 illustrates a possible combination, rule based and statistical model. Note that this model will not necessarily be much smaller than a single large network with context switches, the total size of the conditional probability tables may even be larger, however it will be much quicker to compute than a single large network. Also the requirement for training data will be no lower. Most importantly the need to learn exactly which variables or variable groups are dependent on the context switches has been removed and the variable parameters in each statistical model will only be trained on relevant cases.

The treatment of different types of object is less easy to achieve with separate models



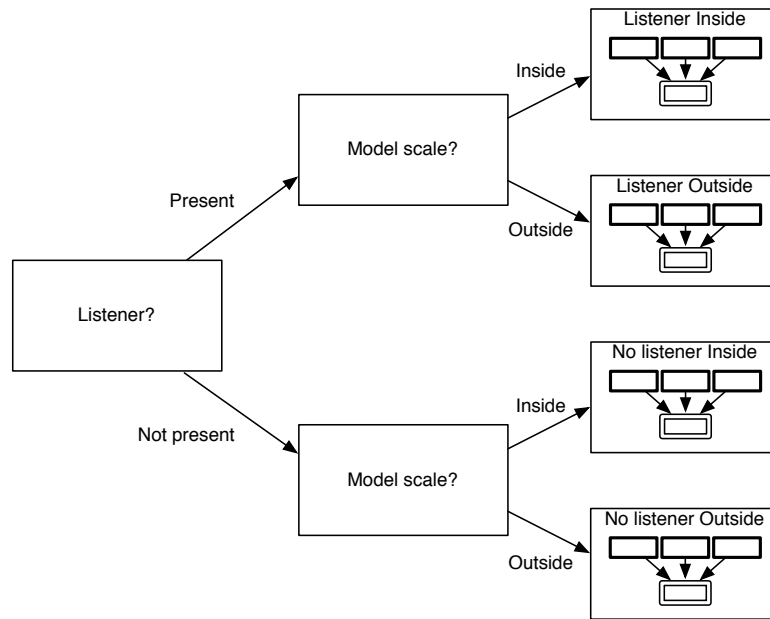


Figure 7.7: Using a decision tree to select among Bayesian network models for reference choice

since this would require comparing the outputs of different models, rather than selecting and using a single model for a single reference choice task. A better method might be to include 'not applicable' states, for instance for object volume in the case of a road.

## Chapter 8

# Conclusions and directions for further work

### 8.1 Nature and context of the study

This is not a neat study. Having taken the decision to look at multiple influences on a complex area of language use and working with a data-set that was deliberately as complex as possible, loose ends are always going to be present and conclusions are always going to carry caveats. Also the study treats an under-researched area of spatial cognition in a way which is still unusual in spatial cognition; the application of machine learning techniques. There are no comparable studies to build on. It has been argued that this is a necessary step in spatial language understanding and particularly in the case of reference object selection. Preposition assignment and reference frame selection do not require such complex environments. Applications such as the ‘GIVE challenge’ also suggest that near real world environments will become an accepted part of computational language systems testing. At the moment the approach is in its infancy and as has been illustrated here a lot needs to be learned about how to represent the environments faithfully to human and machine observers and in particular what is an allowable departure from reality and what may turn out to have a significant effect on the system being trained or tested. It will probably turn out to be true that a perfect spatial language system will need an effectively perfect environment for training and testing, if we accept as perfect something that will pass an exhaustive Turing test.

Another issue is the large number of ways information can be represented to the models and the vast number of possibilities for testing that immediately present themselves when near real world data is available. Just to give a few examples, the vertical extension of objects has not been considered even though in a landscape context (but perhaps not indoors) it seems intuitively attractive. This in itself is just one instance of a variable, which may be specific to different classes of objects, which might be termed their characteristic dimension (width would be the characteristic dimension for a road, Gapp [1995b]). The reinforcement mechanism of Tezuka and Tanaka [2005] was briefly investigated but has not

been included because there was simply no time to investigate it fully and initial results were confusing.

A multitude of further extensions to the machine learning system could have been tried including the use of binomial priors to ‘smooth’ the test data. Kononenko type networks could have been used in combination with the clustering technique developed and varying the threshold for significant interaction information in the learned structure models might have led to improved performance.

So the results of the study must be judged in the light of this. If there had been an accepted scene corpus for spatial language testing this would have been used, but there is not.

## 8.2 Achievements of the study

The principle thesis of the study is that machines can mimic human spatial language use. The best results on the full data set, along with the only implication that can be drawn from the reduction in performance as the data set is extended, that larger data-sets would still further degrade performance, do not support this thesis when taken on their own.

This said there are clearly many ways in which the system as a whole can be improved and with the best performing network no more idiosyncratic than the least conformist humans it is clearly reasonable to hope that performance indistinguishable from humans could be achieved at some point. It is also worth noting that in the cases where the machine disagrees with human reference choices it does not produce entirely outlandish references. Set against this there is the possibility that to get to something like median human performance in a less constrained linguistic setting may require more nuanced judgements about factors such as object aggregation and some more complex local geometries. Although the model could certainly be improved it is probably usable as it stands in highly cluttered, ‘near real world’, virtual environments, at least in applications such as computer games.

Not as much has been learned about the reference choice task as was hoped at the outset but some interesting results have been outlined in section 7.3.1. It is worth re-stating that this is the first time a platform has been developed that is able to illustrate the complexity of the reference choice task. Statistically significant results have been produced that show that reference and target object size and proximity are not sufficient to model reference choice in humans. A machine model needs to incorporate further influences, particularly related to an object’s geometric extension to mimic human behaviour. It is also clear that models must incorporate statistical dependence between variables (or possibly a functional equivalent). One of the more surprising findings is that the sight-line (ray cast) variables on their own perform very poorly. In combination with geometric variables they make a small but statistically insignificant contribution. This indicates that humans use an assessment of, or prior knowledge of, an objects geometry, more than the immediate perception of an object’s visible size, in the reference choice task.

Although some evidence in support of the hypothesis model has been obtained, in

particular indications that search space is an important consideration, it has not been possible to establish anything more comprehensive. The hypothesis model remains a useful structure for organising ideas about reference choice until further evidence is available.

The Bayesian network structure learning algorithm using interaction information clustering needs further development. Interaction information deserves attention from the Bayesian network community as it is capable of identifying relevant influences in a classifier that are missed by conditional mutual information.

### 8.3 Future directions for research

The data set and machine learning platform used in this study open up a wide range of further topics for research some of which have been mentioned as they have arisen. Some of these were anticipated, and have support already in the data set, and some, to a greater or lesser extent, will need additions to the data set and further validation exercises.

An obvious extension would be to add a parallel system to learn the preposition for the simple locative phrase. Given that this would initially be a purely locative preposition assignment it is not thought this will be a difficult task. Validation data has already been collected that will enable meaningful testing of machine learned systems. At this point the addition of some simple syntax will enable a system to produce basic locative expressions from scenes with potential applications in computer games and training simulators.

The ability of a system to produce compound locative phrases using hierarchical references and/or parts of objects or regions associated with objects would allow for far more realistic locative phrases, but is not a simple matter to achieve. Objects in scenes have all their parts available as named geometric entities but the regions that can be associated with objects and the volumes of space associated with them would be the subject of another machine learning exercise. This would probably be a two step exercise in itself requiring firstly some translation from the geometric characteristics of an object to the regions that are appropriate and secondly an exercise analogous to learning the volume of space associated with a projective preposition. Hierarchical references can be used because a single reference does not satisfactorily limit the search area or because an otherwise appropriate reference is ambiguous. In this second case the addition of a referring expression generator is required. For all of these it would also be necessary to devise a system to make a judgement between the increased communication cost of using a hierarchical reference and the benefits to the listener in terms of search time.

There could be some advantage to linking the system to an ontology in the manner of Lockwood et al. [2006]. This would certainly help in classifying objects and helping to learn further object characteristics such as animacy and function (container (for a bowl), support (for a table), for instance). However it is not clear that an existing ontology such as Cyc would have the needed information in a manner that would be immediately useful, in particular for object aggregation and possibly for regions associated with objects. It might be better, as noted, to learn these from grounded examples.

Incorporating a better model for communication cost would also be informative. This would need to be linked with better measures for ambiguity of references (and hence again a referring expression generator) and possibly a mechanism for quantifying prepositions.

Various enhancements can be made to the test data-set and further validation will increase the robustness of the results. In particular adding animation to the scenes will allow the generation of path descriptions and learning of motion prepositions. To assist in the learning of hierarchical references scenes that are ‘cross-scale’ would be useful. These might include buildings with several rooms, at least some of which were ‘furnished’.

There is a lot of scope for work improving the models used. Investigation of multiple variable instantiation in the learned structure combined clustered models is important as is looking at the effect of the threshold for minimum interaction information. Before the learned structure combined clustered technique can be accepted further analysis of its properties is needed along with testing on a range of standard data sets from the UCI repository.

More complex models, possibly based on combined decision trees and Bayesian networks, to cope with the different contexts for reference selection (inside/outside, listener present/absent) look appealing. The major question which would need to be answered would be how the difference between a context switch and a simple influence (like object size) on the reference choice task, could be learned.

# Appendix A

## Object types used

### A.1 Primitive object types

**Triangular prism.** This is composed of distinct quads and triangular end caps, rather than a quad-strip and triangular end caps, to cope with OpenGL lighting. For the triangular face a base, height and horizontal displacement of the top vertex from the center of the base are specified.

**Rectangular prism.** Similar to the triangular prism, this uses distinct quads for the prism sides. For the end face, width and height are specified.

**Truncated cone.** This shape defines regular prisms with an arbitrary number of sides. A quad-strip is used for the sides so the lighting angles change incrementally. The end caps are triangle fans and specified by a radius. The radii at the top and bottom of the prism can be separately specified, hence this shape can approximate a truncated cone as well as a cylinder.

**Cone.** This is a special case of the truncated cone for a zero-radius end cap.

**Straight pipe.** This is a hollow truncated cone. The wall thickness can be specified.

**Pipe section.** By specifying an angle between 0 and 360 degrees, this produces a section of a straight pipe, The exposed sides as well as the ends are capped

**Apertured prism.** This is principally used for producing walls with doors or windows in. The prism and the aperture are both four sided. The aperture is of arbitrary size and position but must always be contained in the prism and cannot directly abut one edge.

**Spheroid.** This has separately specifiable radii. It is a sphere when  $R_x = R_y = R_z$ . It is composed of triangle fans 'top and bottom' and quad-strips in between.

**Spheroid cap.** This is a spheroid segment. The angle subtended at the center of the spheroid by the edge of the cap is specified as well as the radius.

**Warp cylinder.** This is a development of the truncated cone so that different x and y radii can be specified at the top and bottom of the cylinder. It is principally used for human body parts and does not work well for number of sides less than 6.

## A.2 List of named objects

Table A.1: List of distinct selectable object types

lorry	jeep	van	car	open van
ambulance	bus	hand cart	push chair	
man	woman	girl	boy	dog
green grocer	travel agent	shop	cafe	office
advert	advertising frame	menu	blackboards	price list
circular bench	bush	flower bed	plank	
vinegar	mayonaise	ketchup	glass	wine glass
wire basket	box	rucsac	bag	
espresso machine	saucer	cup	mug	milk bottle
bar	till	wine cooler	ice bucket	hand pumps
licquer bottles	lager bottles	wine bottles	champagne	
telephone box	sign post	plant tub	bollard	
pub	town hall	bank	market stall	
bus shelter	bus stop	street lamp	postbox	
road	pavement	vacant lot	skip	pallet
church	terrace	house	living room	kitchen
fountain bowl	circular path	pond	bench	
field	park	hedge	tree	conifer
steps	path	gate	wall	
vase	bowl	waste bin	standard lamp	table lamp
flue	woodburner	log basket	log	chimney breast
chopping board	wooden spoon	knife	mortar	pestle
washing-up liquid	soap	cloth	table mat	ball
cutlery holder	tap	washing-up bowl	draining basket	
casserole lid	casserole	saucepan	fryingpan	
olive oil	sugar bowl	orange	apple	lemon
baking tray	tray	kettle	can	
table	chair	sofa	bureau	
hob	rayburn	oven	sink	
work surface	wall cupboard	cupboard	chest of drawers	shelf unit
book	report	pen	clock	picture

## Appendix B

### Validation results

The full results from the validation exercises are given here. Each of the test cases is shown along with the locations of the target and all chosen reference objects. For each chosen reference object the percentage of subjects who chose the reference is shown. Also shown for each reference are the different prepositions selected to accompany it along with the number of subjects who chose each preposition.

The scenes were chosen in 2 groups of 20, but with 2 of the scenes appearing in each group in order to allow a check to be made for any differences between the groups of subjects. Hence 38 scenes are shown in the figures following. The first group of scenes were given to the first 20 subjects and the second group of scenes to the second 20 subjects. No notable differences between the groups were apparent.

The first group of scenes are shown first, then the two scenes from both groups then the second group of scenes. Within the groups the scenes are shown (broadly) in order of increasing scale, they were shown to the validators in random order.

The choices of the author, for both reference and preposition, for each scene are shown in green text and underlined. The author is not included in the count or percentage of validators, so in the first scene shown (desk1\_04), 55% of validators *and* the author chose the bowl as the reference object and 5 validators *and* the author described the coaster as right of the bowl.



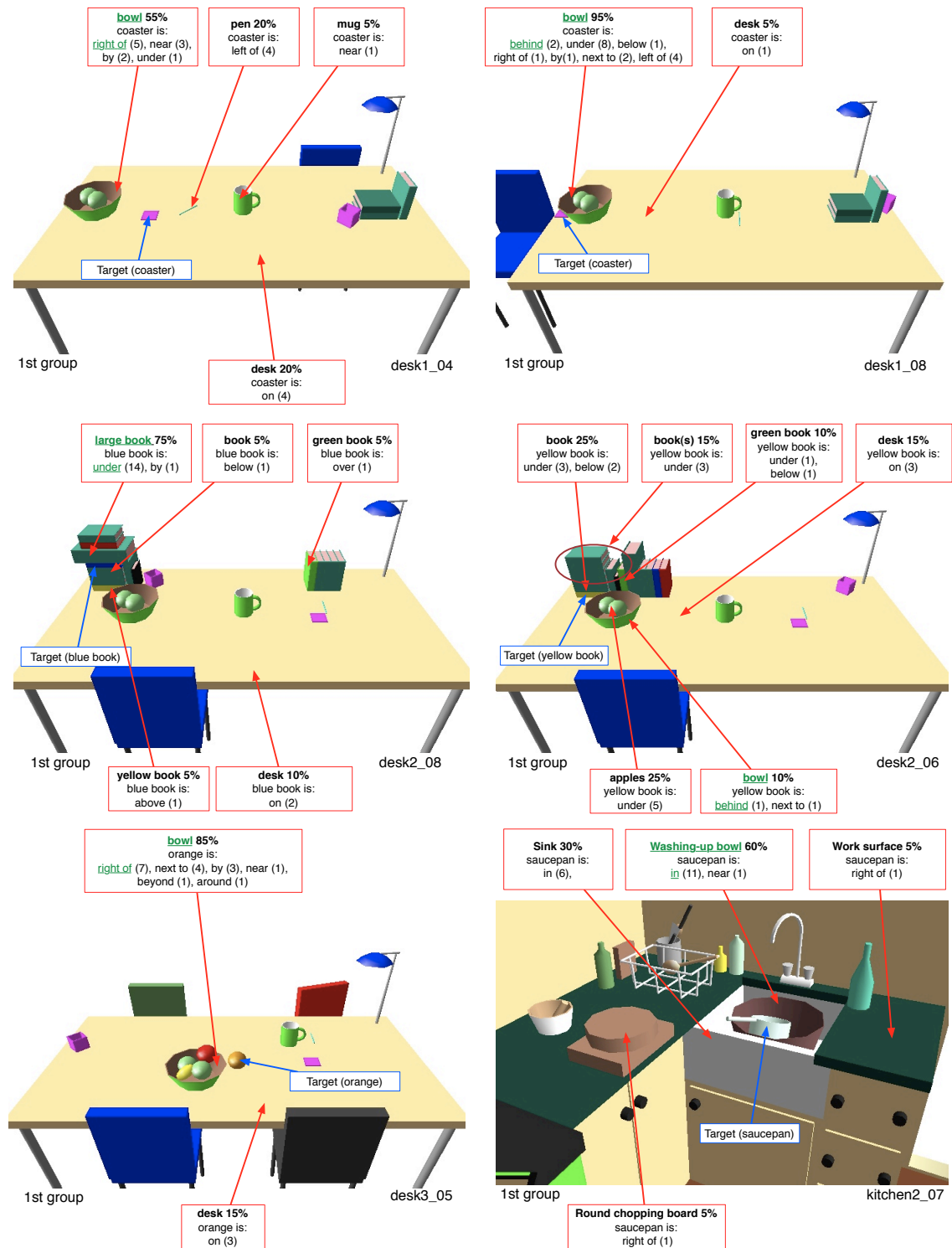


Figure B.1: Validation results 1 of 7

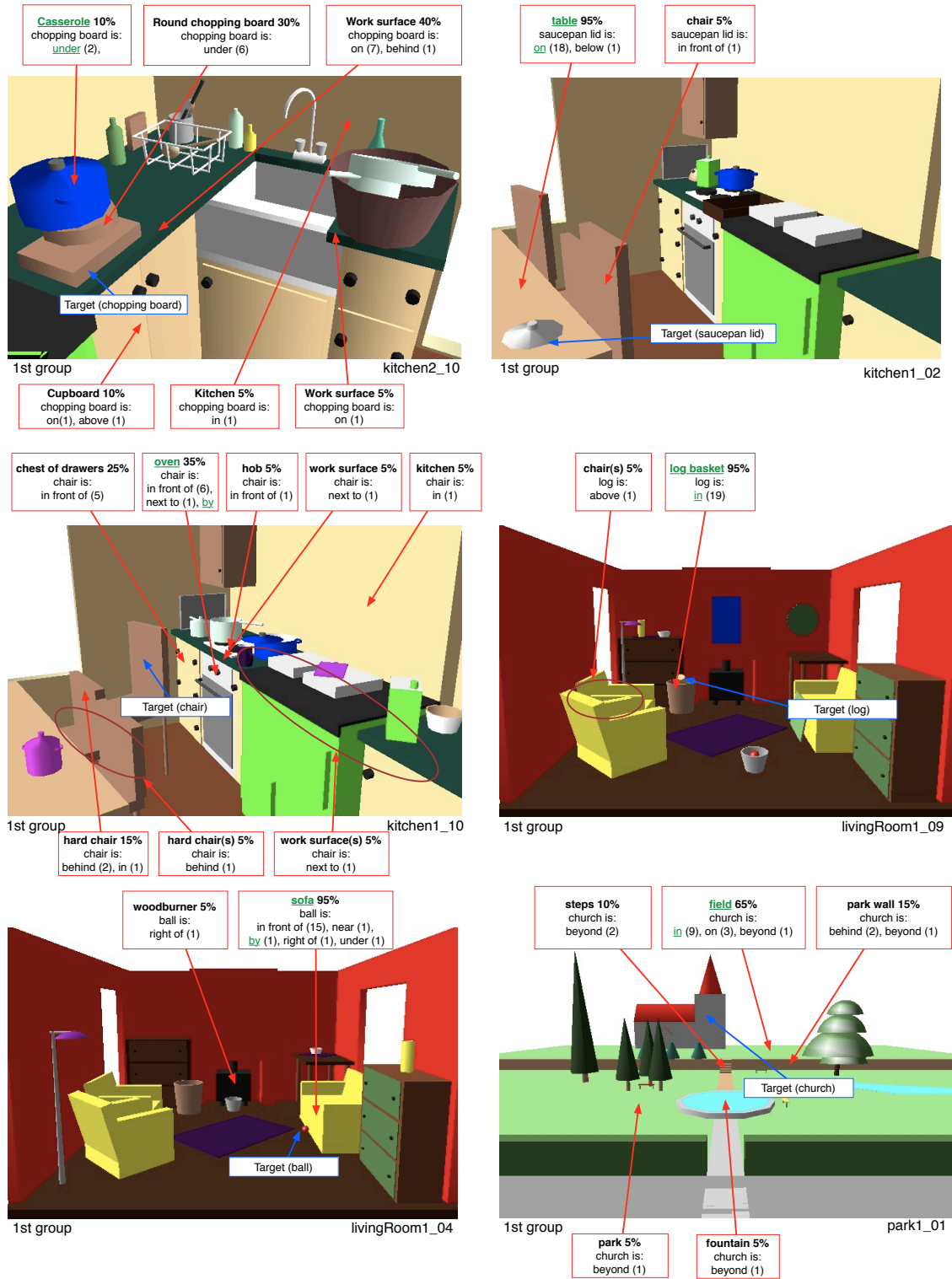


Figure B.2: Validation results 2 of 7

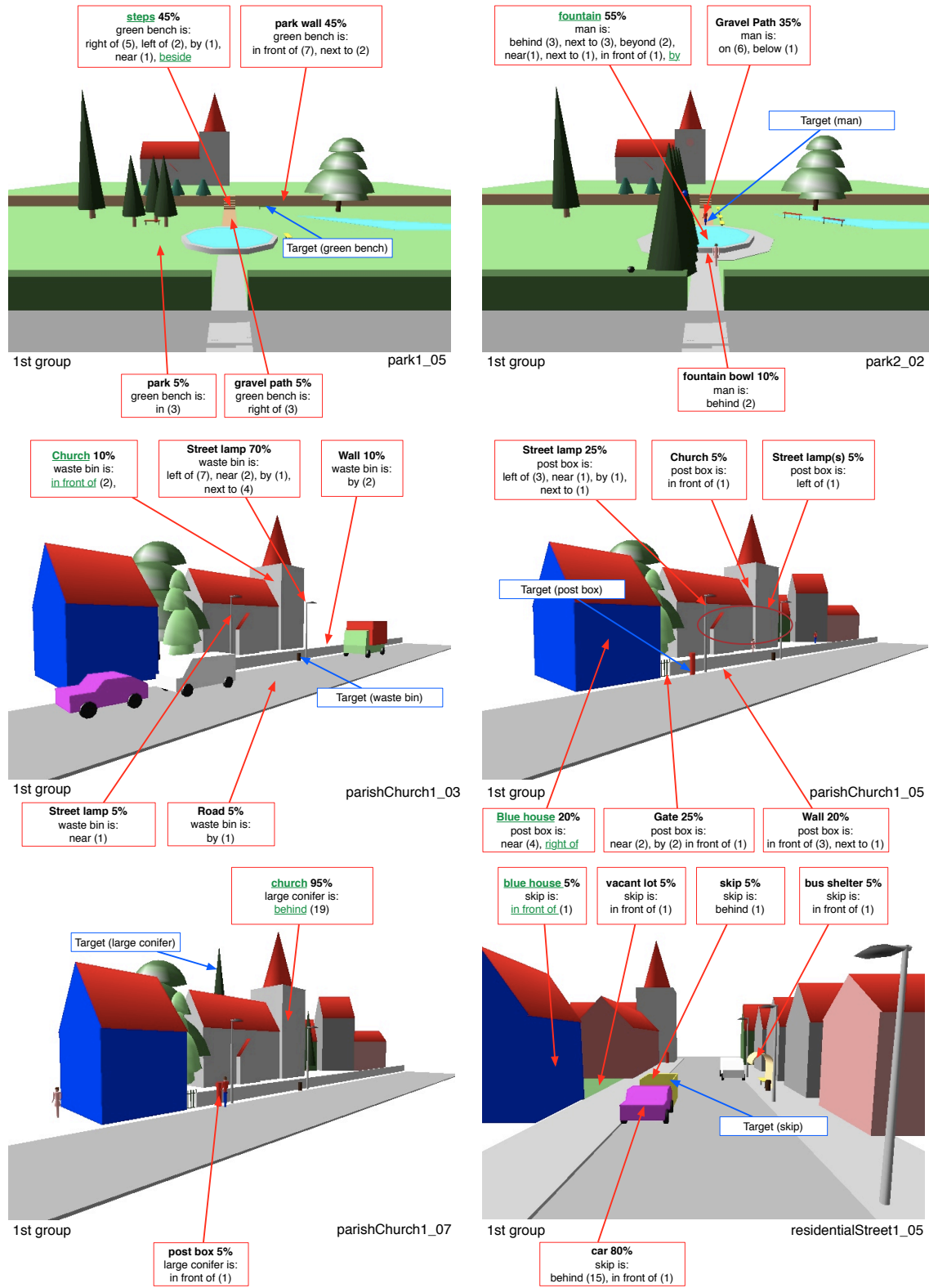


Figure B.3: Validation results 3 of 7

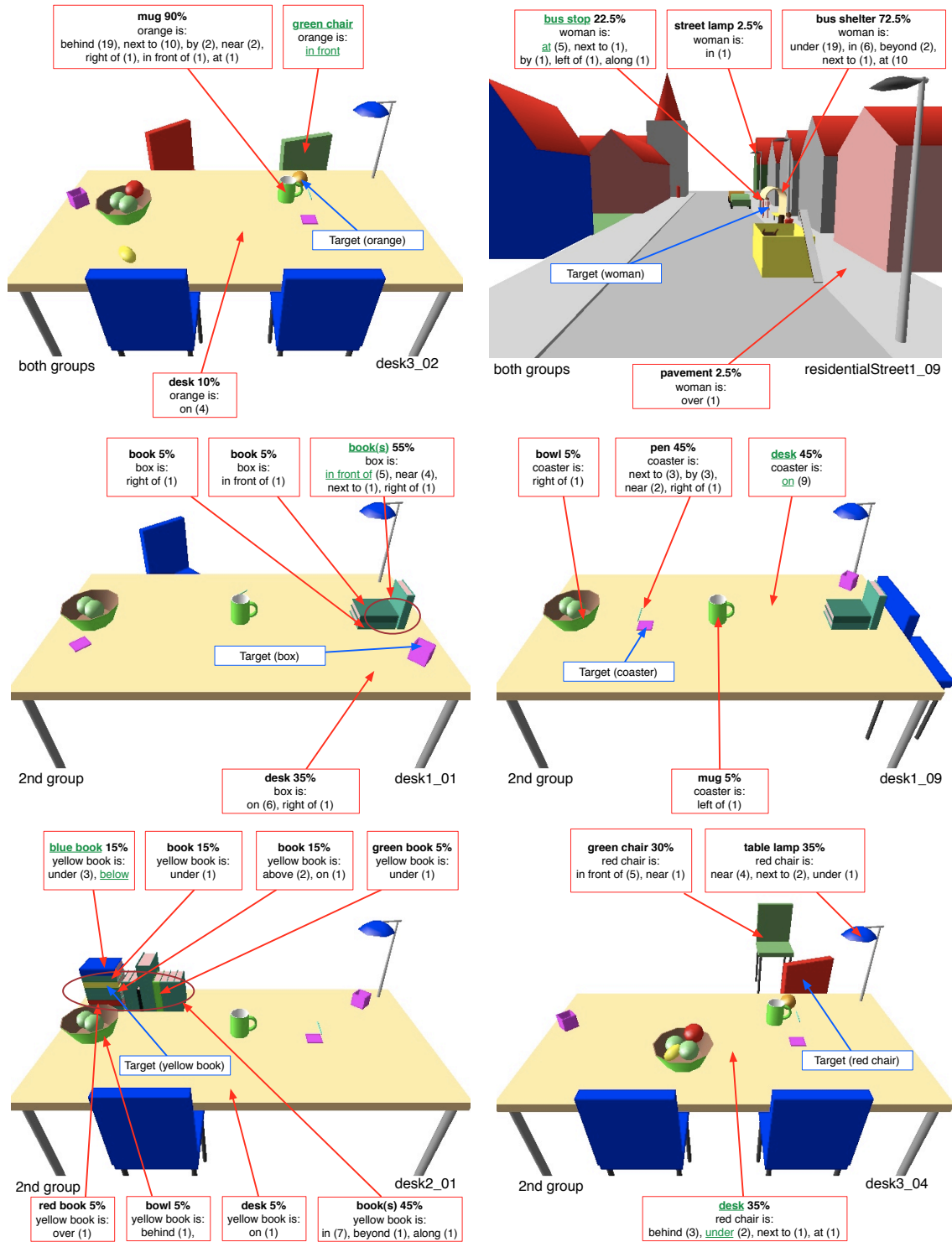


Figure B.4: Validation results 4 of 7

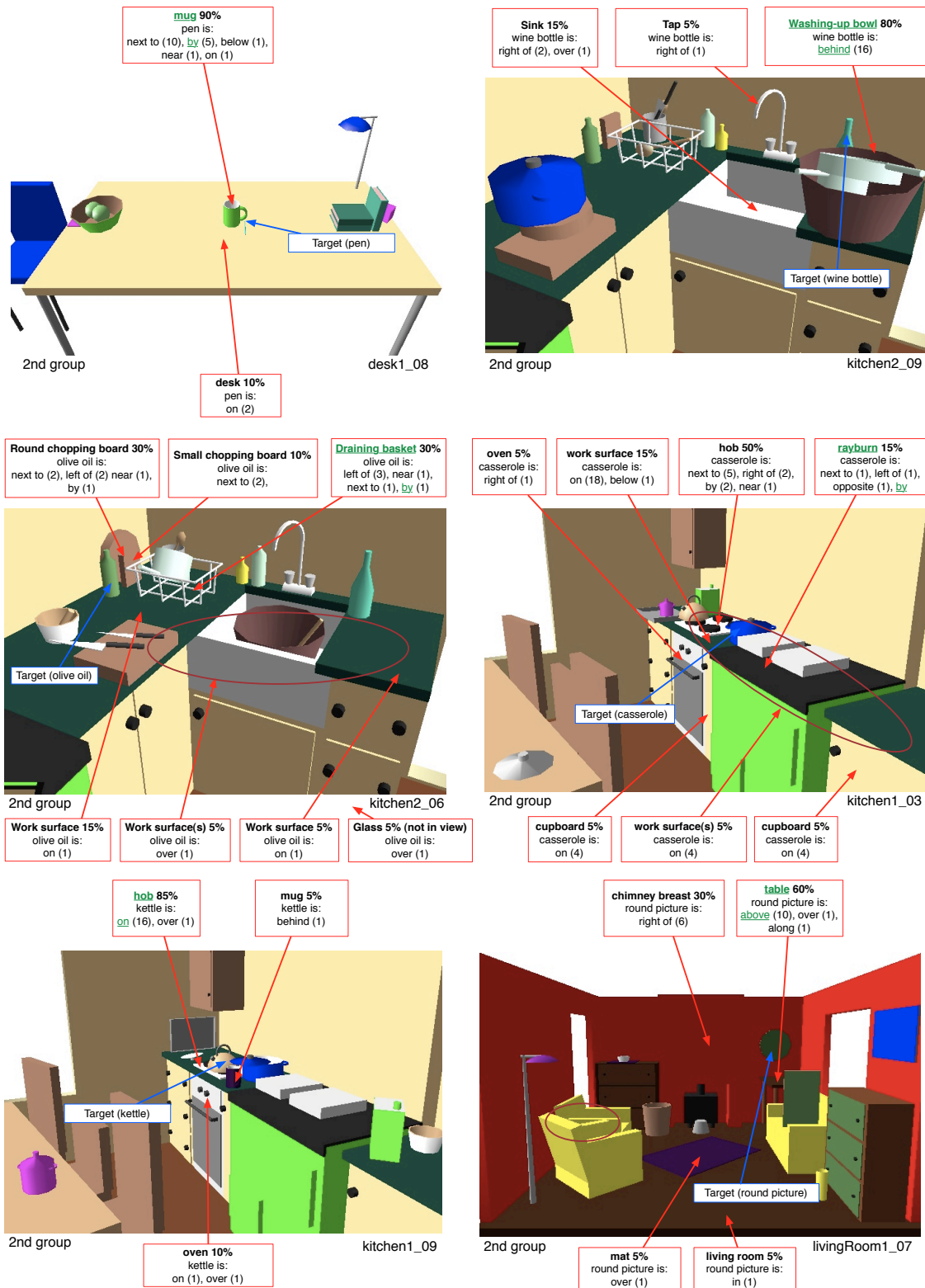


Figure B.5: Validation results 5 of 7

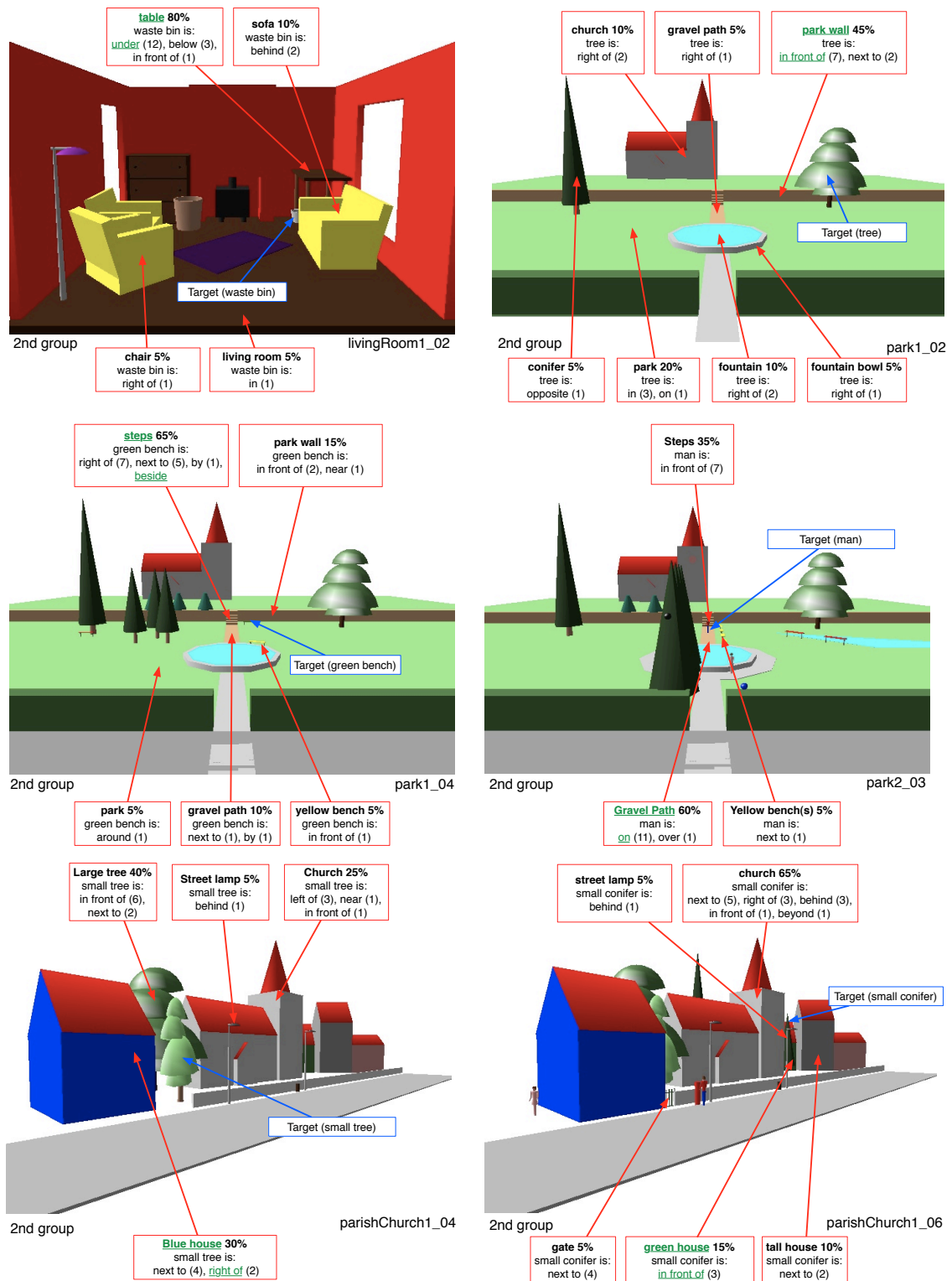


Figure B.6: Validation results 6 of 7

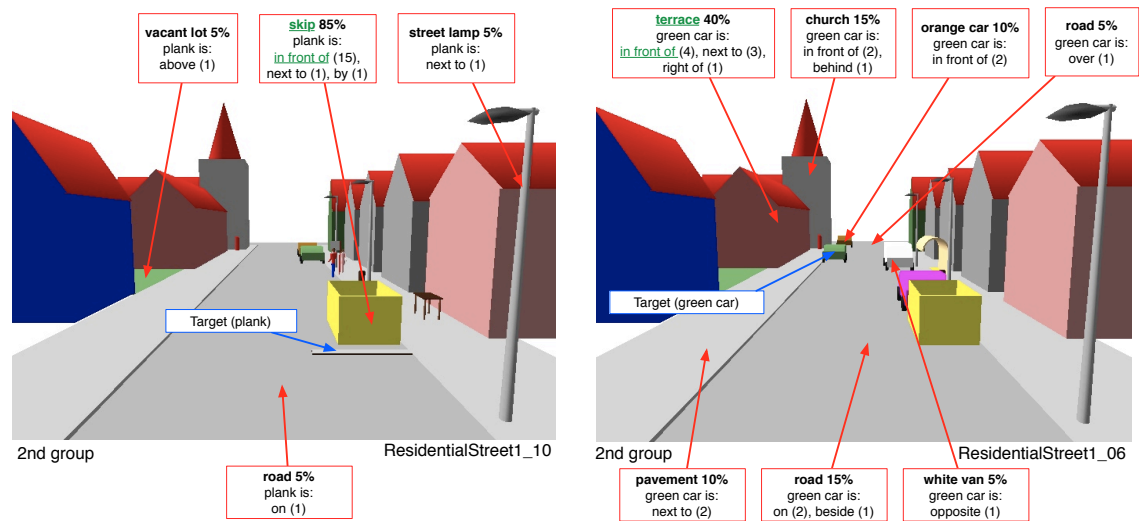


Figure B.7: Validation results 7 of 7

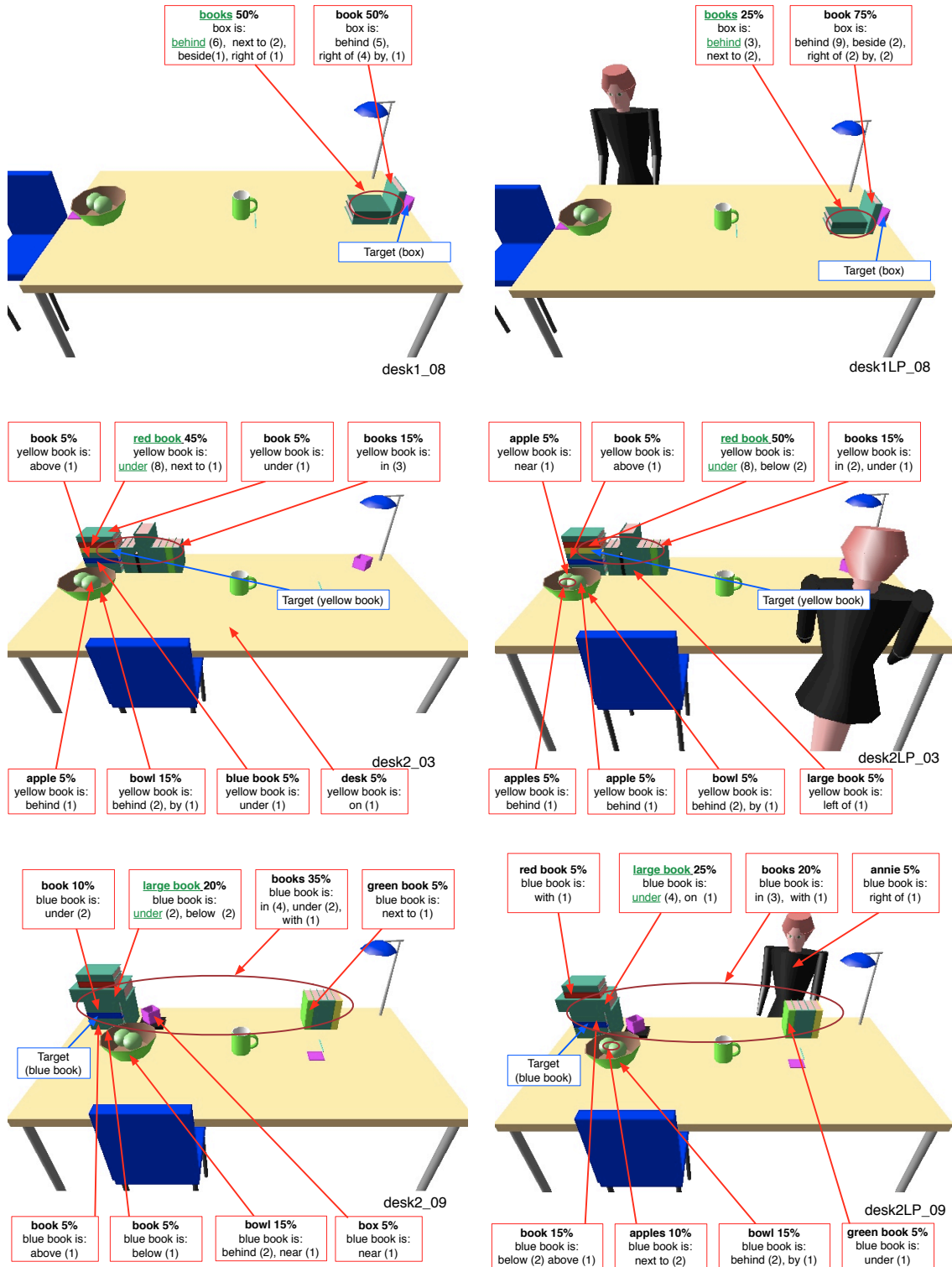


Figure B.8: Listener present validation results 1 of 6



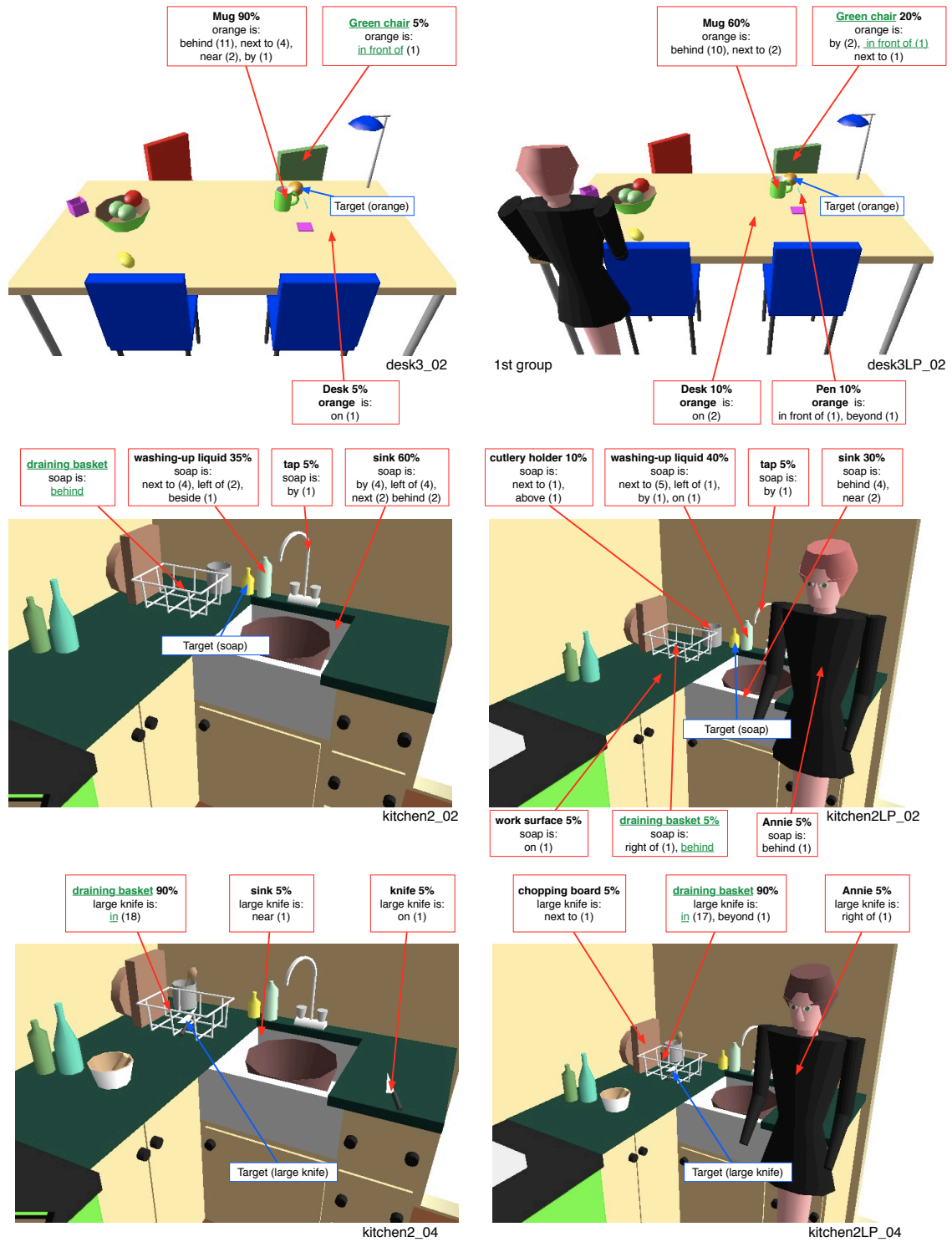


Figure B.9: Listener present validation results 2 of 6

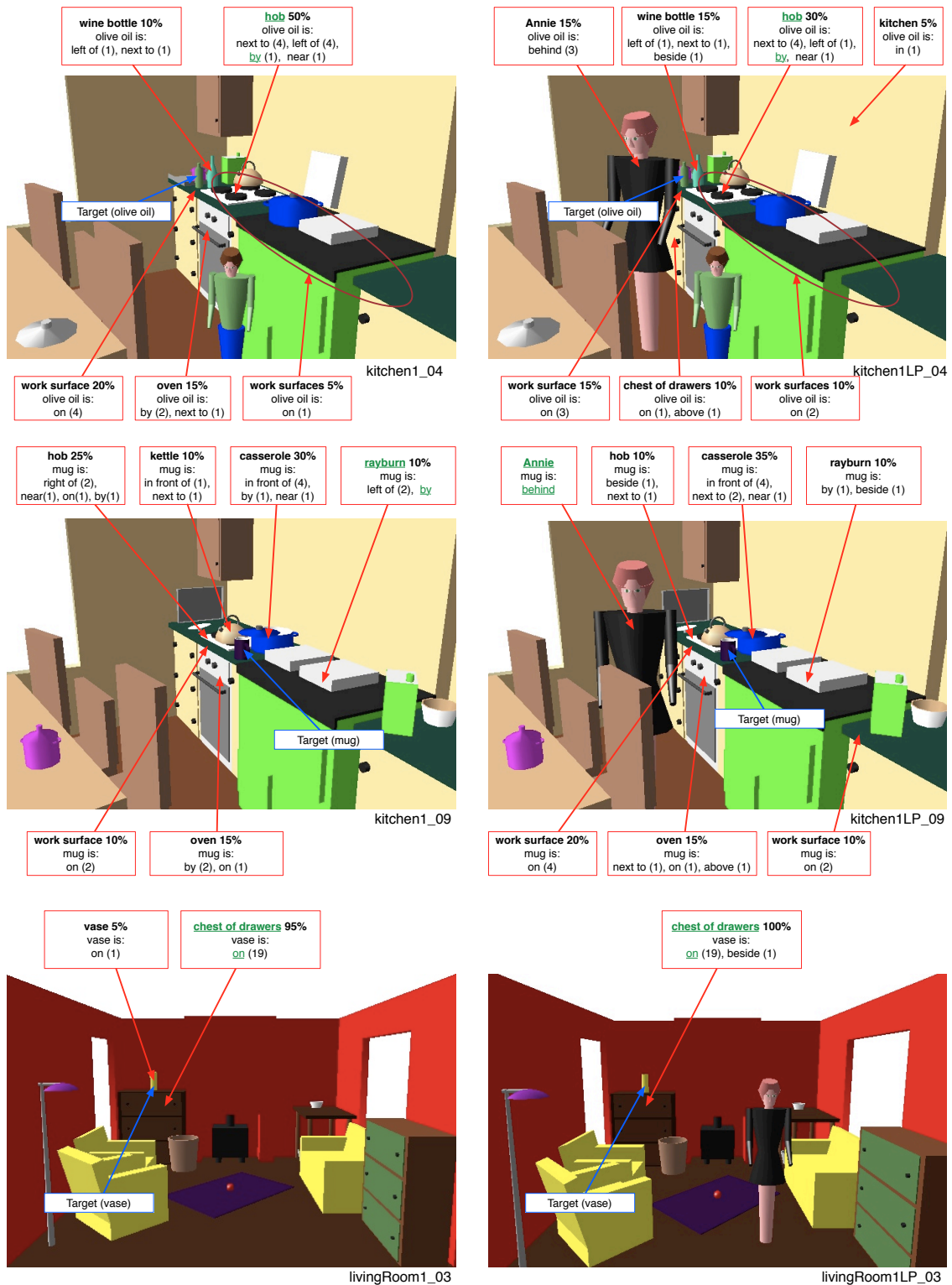


Figure B.10: Listener present validation results 3 of 6

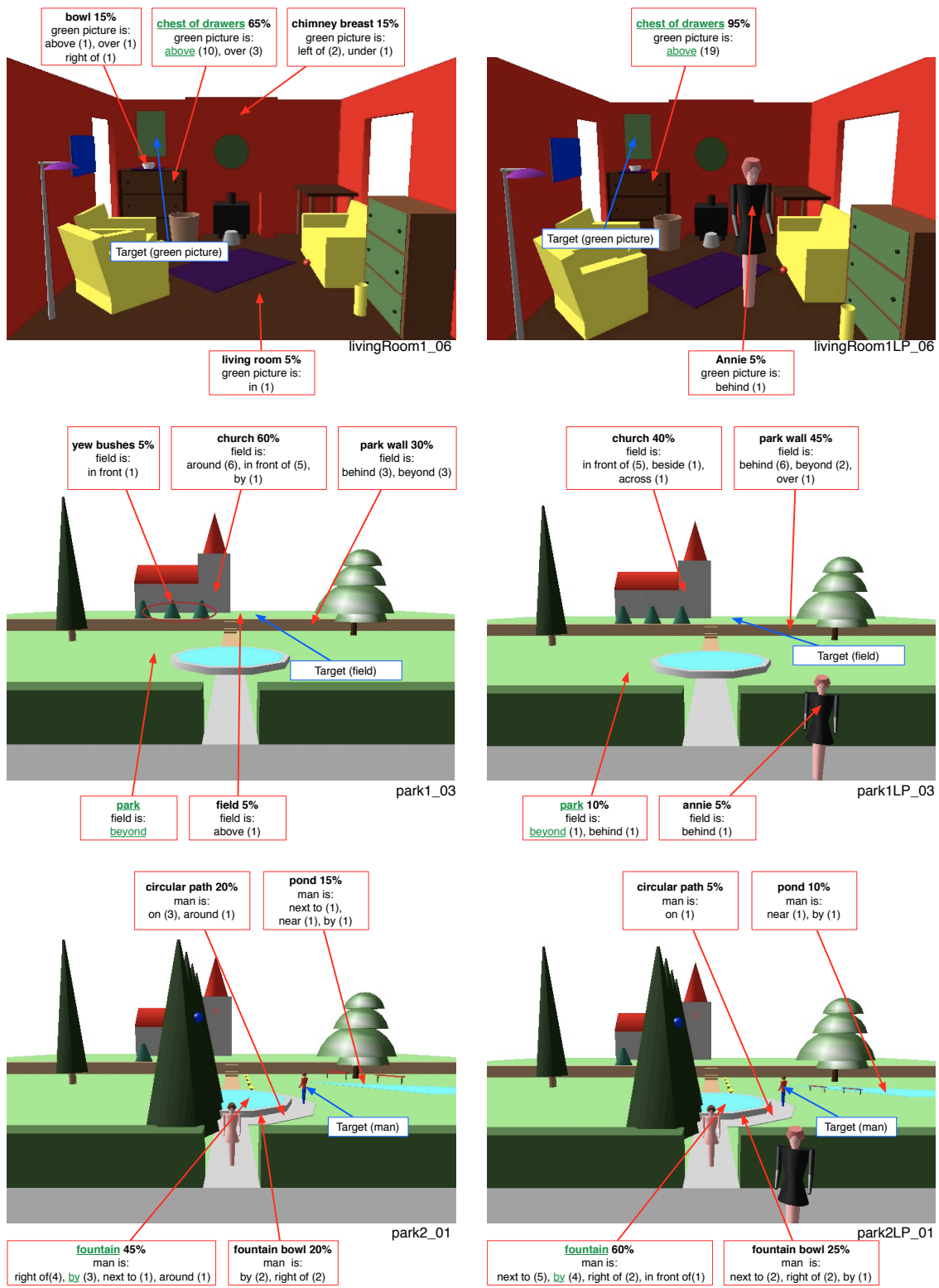


Figure B.11: Listener present validation results 4 of 6

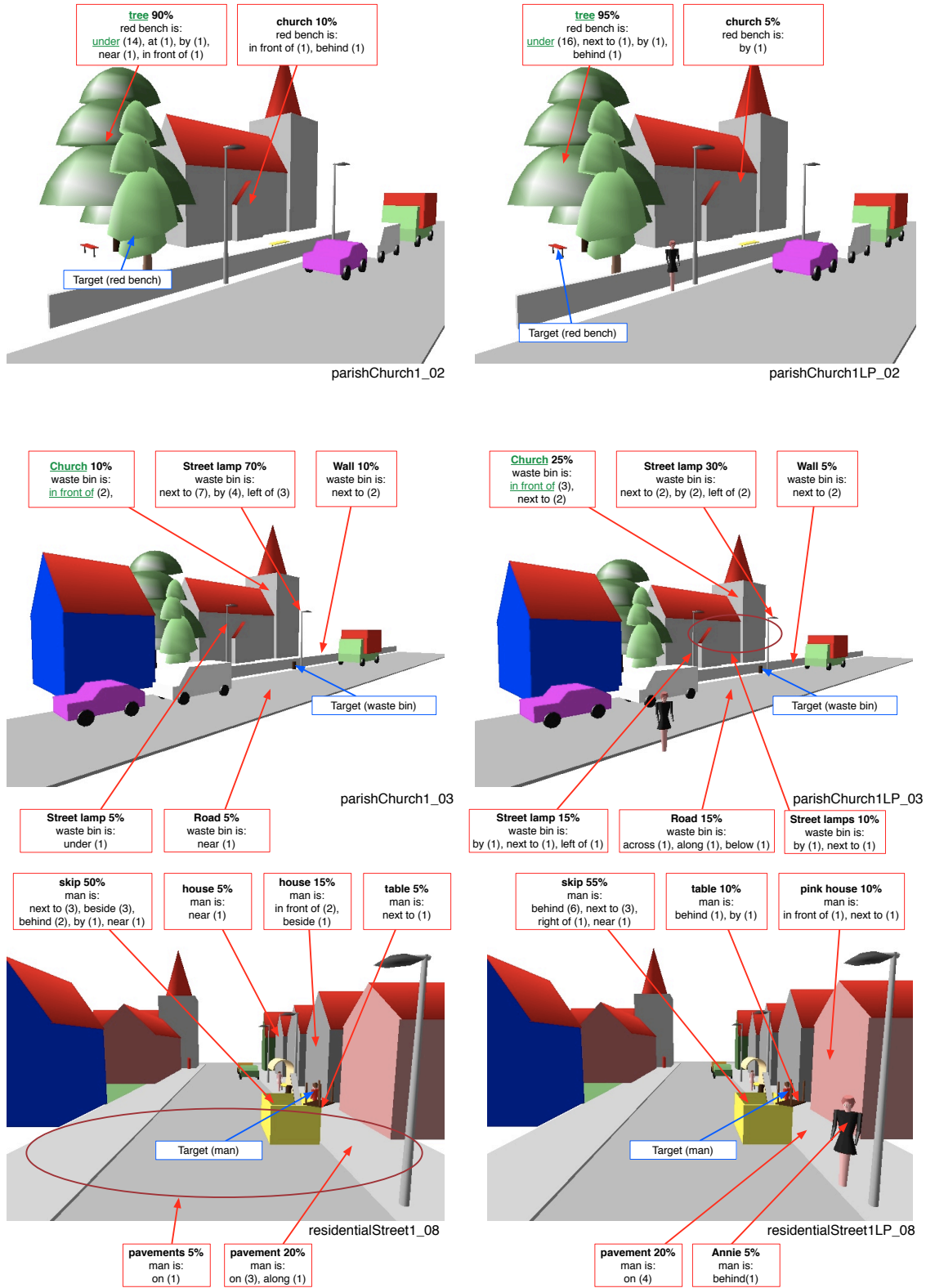


Figure B.12: Listener present validation results 5 of 6

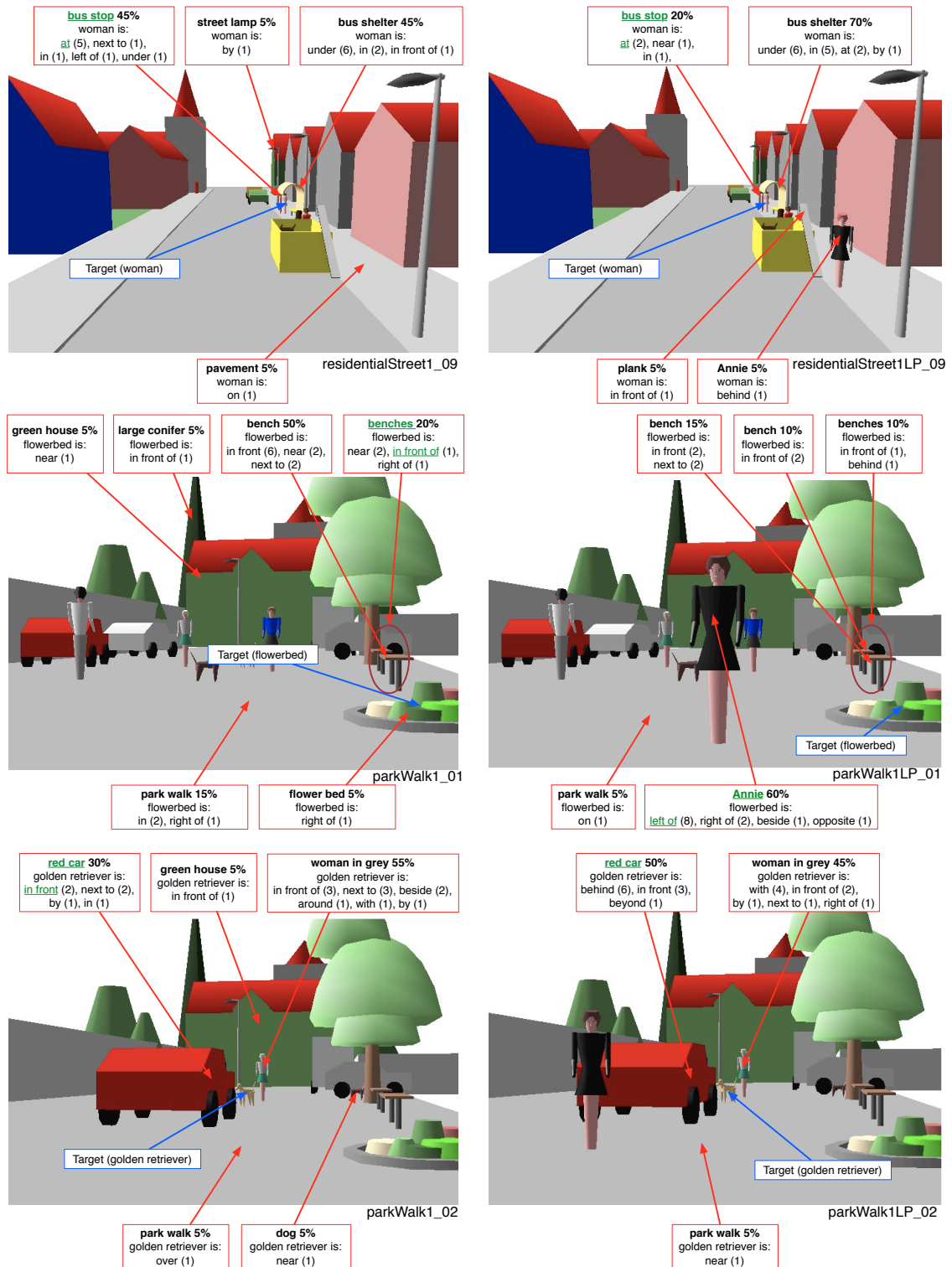


Figure B.13: Listener present validation results 6 of 6



Figure B.14: Scenes from series produced after the validation exercises, showing author's reference choices for a random test case

# Appendix C

## Sample file listings

### **C.1 Sample scene files**

These can be found on the attached CD

### **C.2 Sample network files**

These can be found on the attached CD

### **C.3 Sample validation files**

These can be found on the attached CD

# Bibliography

- Alicia Abella and John R. Kender. From images to sentences via spatial relations. In *SPELMG '99: Proceedings of the Integration of Speech and Image Understanding*. IEEE Computer Society, 1999.
- E. Andre, G. Herzog, and Rist T. Characterizing trajectories of moving objects using natural language path descriptions. *Proceedings of the 7th ECAI, Brighton, England*, pages 1–8, 1986.
- E. Andre, G. Herzog, and Rist T. On the simultaneous interpretation of real world image sequences and their natural language description: The system soccer. *Proceedings of the 8th ECAI*, pages 449–454, 1988.
- A. Aydemir, Sjöo K., and P. Jensfelt. Object search on a mobile robot using relational spatial information. In *Proceedings of the 11th International Conference on Intelligent Autonomous Systems (IAS'10)*, Ottawa, Canada, September 2010.
- C. B. Barber, D. P. Dobkin, and H. Huhdanpaa. The quickhull algorithm for convex hull. *ACM Trans. on Mathematical Software*, 22(4), 1996.
- M. J. Barclay and A. P. Galton. A scene corpus for training and testing spatial communication systems. *Proceedings of the AISB convention: Communication, Interaction and Social Intelligence*, 10:26–29, 2008.
- M. Barclay and A. Galton. Human and machine models for reference object selection in spatially locative phrases. 2010.
- A. J. Bell. The co-information lattice. In *ICA 2003, Nara, Japan*, 2003.
- Anselm Blocher and Eva Stopp. Time-dependent generation of minimal sets of spatial descriptions. In Patrick Olivier and Klaus-Peter Gapp, editors, *Representation and Processing of Spatial Expressions*, pages 57–72. Laurence Earlbaum Associates, 1998.
- Claudia Brugman. *'The story of over*. PhD thesis, University of California, Berkley, 1981.
- Wray Buntine. Theory refinement on Bayesian networks. In B. D. D'Ambrosio, P. Smets, and P. P. Bonissone, editors, *Proceedings of the Seventh Annual Conference on Uncertainty Artificial Intelligence*, pages 52–60. Morgan Kaufmann, 1991.



- G. E. Burnett, D. Smith, and A. J. May. Supporting the navigation task: characteristics of good landmarks. In *Proceedings of the Annual Conference of the Ergonomics Society*. Taylor and Francis, 2001.
- Donna Byron, Alexander Koller, Kristina Striegnitz, Justine Cassell, Robert Dale, Johanna Moore, and Jon Oberlander. Report on the first nlg challenge on generating instructions in virtual environments (give). In *ENLG '09: Proceedings of the 12th European Workshop on Natural Language Generation*, pages 165–173, Morristown, NJ, USA, 2009. Association for Computational Linguistics.
- Laura A. Carlson-Radvansky. Constructing spatial templates: The influence of reference frame selection. 1996.
- L. A. Carlson-Radvansky and D Irwin. Frames of reference in vision and language: Where is above? *Cognition*, 46:223–224, 1993.
- L. A. Carlson-Radvansky and D Irwin. Reference frame activation during spatial term assignment. *Journal of Memory and Language*, 33:646–671, 1994.
- L. A. Carlson-Radvansky and G. D. Logan. Using spatial terms to select an object. *Memory and Cognition*, 29(6):883–892, 2001.
- L. A. Carlson-Radvansky and G. A. Radvansky. The influence of functional relations on spatial term selection. *Psychological Science*, 7(1):56–60, 1996.
- Laura A. Carlson and Patrick L. Hill. Processing the presence, placement, and properties of a distractor in spatial language tasks. *Memory and Cognition*, 36:240–255, 2008.
- Laura A. Carlson and Patrick L. Hill. Formulating spatial descriptions across various dialogue contexts. In Kenny R. Coventry, Thora A. Tenbrink, and Bateman John A., editors, *Spatial Language and Dialogue*, pages 89–103. Oxford University Press, 2009.
- Laura A. Carlson, Terry Regier, William Lopez, and Bryce Corrigan. Attention unites form and function in spatial language. *Spatial Cognition and Computation*, 6(4):295–308, 2006.
- Jie Cheng, Russell Greiner, Jonathan Kelly, David Bell, and Weiru Liu. Learning Bayesian networks from data: An information-theory based approach. *Artificial Intelligence*, 137:43–90, 2002.
- C. K. Chow and C. N. Liu. Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory*, IT-14(3):462–467, 1968.
- D. Connolly. Constructing hidden variables in Bayesian networks via conceptual clustering. In *In Proceedings of the Tenth International Conference on Machine Learning*, pages 65–72. Morgan Kaufmann, 1993.
- G. F. Cooper and E. Herskovits. A Bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347, 1992.

- G. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.
- F. Costello and J. Kelleher. Spatial prepositions in context: The semantics of near in the presence of distractor objects. *Proceedings of the Third ACL-SIGSEM Workshop on Prepositions*, pages 1–8, 2006.
- Kenny R. Coventry, Dermot Lynott, Angelo Cangelosi, Lynn Monrouxe, Dan Joyce, and Daniel C. Richardson. Spatial language, visual attention, and perceptual simulation. *Brain and Language*, In Press, Corrected Proof:–, 2009.
- Kenny Coventry and Simon Garrod. *Saying, Seeing and Acting: The Psychological Semantics of Spatial Prepositions*. psychology press, 1 edition, 2004.
- K.R. Coventry. Spatial prepositions, functional relations, and lexical specification. In Patrick Olivier and Klaus-Peter Gapp, editors, *Representation and Processing of Spatial Expressions*, pages 1–35. Laurence Earlbaum Associates, 1998.
- K. R. Coventry, M. Prat-Sala, and L Richards. The interplay between geometry and function in the comprehension of over, under, above, and below. *Journal of Memory and Language*, 44(3):376–398, 2001.
- K. R. Coventry, A. Cangelosi, R. Rajapakse, A. Bacon, S. Newstead, D. Joyce, and L. V. Richards. Spatial prepositions and vague quantifiers: Implementing the functional geometric framework. In C. Freksa, B. Knauff, B. Krieg-Bruckner, and B. Nebel, editors, *Spatial Cognition, Volume IV. Reasoning, Action and Interaction*, pages 98–110. Springer-Verlag, 2005.
- R. Dale and E. Reiter. Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19:233–263, 1995.
- Michel Denis, Francesca Pazzaglia, Cesare Cornoldi, and Laura Bertolo. Spatial discourse and navigation: an analysis of route directions in the city of venice. *Applied Cognitive Psychology*, 13(2):145–174, 1999.
- Manuel de Vega, Mara J. Rodrigo, Manuel Ato, Doris M. Dehn, and Beatriz Barquero. How nouns and prepositions fit together: An exploration of the semantics of locative sentences. *Discourse Processes*, 34:117–143, 2002.
- R. Duda and P. Hart. *Pattern classification and scene analysis*. New York: John Wiley & Sons, 1973.
- Ingo Duwe, Klaus Kessler, and Hans Strohner. Resolving ambiguous descriptions through visual information. In K. R. Coventry and P. Olivier, editors, *Spatial Language. Cognitive and Computational Perspectives*, pages 43–67. Dordrecht: Kluwer Academic Publishers, 2002.

- B. Elias and C. Brenner. Automatic generation and application of landmarks in navigation data sets. In Peter F. Fisher, editor, *Developments in Spatial Data Handling: 11th International Symposium on Spatial Data Handling*, pages 469–480. Springer-Verlag, Berlin Heidelberg, 2004.
- M. I. Feist and D. Gentner. An influence of spatial language on recognition memory for spatial scenes. In J.D. Moore and K. Stenning, editors, *Proceedings of the 23rd Annual Conference of the Cognitive Science Society.*, pages 279–284, 2001.
- M. I. Feist and D. Gentner. Factors involved in the use of in and on. In *Proceedings of the Twenty-fifth Annual Meeting of the Cognitive Science Society.*, 2003.
- Douglas H. Fisher. Knowledge acquisition via incremental conceptual clustering. *Machine Learning*, 2(2):139–172, 1987.
- Nancy Franklin and Barbara Tversky. Searching imagined environments. *Journal of Experimental Psychology: General*, 119(1):63 – 76, 1990.
- Nir Friedman and Moises Goldszmidt. Building classifiers using Bayesian networks. In *In Proceedings of the thirteenth national conference on artificial intelligence*, pages 1277–1284. AAAI Press, 1996.
- Nir Friedman. The Bayesian structural EM algorithm. In *Proc. Fourteenth Conference on Uncertainty in Artificial Intelligence (UAI '98)*, pages 129–138, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.
- N. Friedman, D. Geiger, and M. Goldszmidt. Bayesian network classifiers. *Machine Learning*, 29:131–163, 1997.
- Thomas Fuhr, Gudrun Socher, Christian Scheering, and Gerhard Sagerer. A three-dimensional spatial model for the interpretation of image data. In Patrick Olivier and Klaus-Peter Gapp, editors, *Representation and Processing of Spatial Expressions*, pages 103–118. Laurence Earlbaum Associates, 1998.
- Klaus-Peter Gapp. Object localization: Selection of optimal reference objects. In A.U. Frank and W. Kuhn, editors, *Spatial Information Theory: A Theoretical Basis for GIS*, pages 519–535, 1995a.
- K.P. Gapp. An empirically validated model for computing spatial relations. *Künstliche Intelligenz*, pages 245–256, 1995b.
- S. Garrod, G. Ferrier, and S. Campbell. In and on: investigating the functional geometry of spatial prepositions. *Cognition*, 72:167–189, 1999.
- S. Geman and D. Geman. Stochastic relaxation, gibbs distributions, and the Bayesian restoration of images. pages 452–472, 1990.

- P. Gorniak and D. K. Roy. Grounded semantic composition for visual scenes. *Journal of Artificial Intelligence Research*, 21:429–470, 2004.
- H P Grice. Logic and conversation. In P. Cole and J. Morgan, editors, *Syntax and Semantics: Vol 3, Speech Acts*, pages 43–58. New York: Academic Press, 1975.
- Annette Herskovits. Semantics and pragmatics of locative expressions. *Cognitive Science*, 9(3):341–378, 1985.
- A. Herskovits. *Language and spatial cognition: An interdisciplinary study of prepositions in English*. Studies in Natural Language Processing. Cambridge University Press, Cambridge, UK, 1986.
- A Herskovits. Schematization. In Patrick Olivier and Klaus-Peter Gapp, editors, *Representation and Processing of Spatial Expressions*, pages 149–162. Laurence Earlbaum Associates, 1998.
- Gerd Herzog and Peter Wazinski. Visual translator: Linking perceptions and natural language descriptions. *Artificial Intelligence Review*, 8:175–187, 1994.
- Gerd Herzog. From visual input to verbal output in the visual translator. Technical Report TR124, Universitat Des Saarlandes, Saarbrucken, Germany, 1995.
- S. Hettich and S. D Bay. *The UCI KDD archive*. University of California, Department of Information and Computer Science, 1999.
- William H. Ittelson. Environment perception and contemporary perceptual theory. In *Environment and cognition*, pages 1–19, New York, 1973. Seminar press.
- Aleks Jakulin. *Machine Learning Based on Attribute Interactions*. PhD thesis, University of Ljubljana, 2005.
- Tanja Jording and Ipke Wachsmuth. An anthropomorphic agent for the use of spatial language. In K. R. Coventry and P. Olivier, editors, *Spatial Language. Cognitive and Computational Perspectives*, pages 69–85. Dordrecht: Kluwer Academic Publishers, 2002.
- Dan W. Joyce, Lynn V. Richards, Angelo Cangelosi, and Kenny R. Coventry. Object representation-by-fragments in the visual system: A neurocomputational model. In *In L. Wang et al. (Eds), Proceedings of the 9th International Conference on Neural Information Processing (ICONP02) IEEE*. Press, 2002.
- D. Joyce, L. Richards, A. Cangelosi, and K.R. Coventry. On the foundations of perceptual symbol systems: Specifying embodied representations via connectionism. In F. Detje, D. Dorner, and H. Schaub, editors, *The Logic of Cognitive Systems. Proceedings of the Fifth International Conference on Cognitive Modeling*, pages 147–152, Universitatsverlag Bamberg, 2003.

- John D. Kelleher and Fintan J. Costello. Applying computational models of spatial prepositions to visually situated dialog. *Computational Linguistics*, 35(2):271–306, 2009.
- John Kelleher and Fintan Costello. Cognitive representations of projective prepositions. In Valia Kordoni and Aline Valencia, editors, *In Proceedings of the 2nd ACL-Sigsem Workshop on The Linguistic Dimensions of Prepositions and their Use in Computational Linguistics Formalisms and Applications.*, 2005.
- J. D. Kelleher. *A Perceptually Based Computational Framework for the Interpretation of Spatial Language*. PhD thesis, Dublin City University, 2003.
- J. Kelleher and J. van Genabith. Visual salience and reference resolution in simulated 3-d environments. *Artificial Intelligence Review*, 21(3-4):253–267, 2004.
- E. Keogh and M. Pazzani. Learning augmented Bayesian classifiers: A comparison of distribution-based and classification-based approaches. *Uncertainty 99, 7th. Int'l Workshop on AI and Statistics*, pages 225–230, 1999.
- Ron Kohavi and George H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97(1-2):273 – 324, 1997.
- Igor Kononenko. Semi-naive Bayesian classifier. In *Proceedings of the European working session on learning on Machine*, pages 206 – 219, 1991.
- E. Kraemer, A. Verleg, and S van Erk. Graph-based generation of referring expressions. *Computational Linguistics*, 29(1):53–72, 2003.
- Geert-Jan M. Kruijff, John D. Kelleher, and Nick Hawes. Information fusion for visual reference resolution in dynamic situated dialogue. In Elisabeth Andre, Laila Dybkjaer, Wolfgang Minker, Heiko Neumann, and Michael Weber, editors, *Perception and Interactive Technologies: International Tutorial and Research Workshop, PIT 2006*, volume 4021 of *Lecture Notes in Computer Science*, pages 117 – 128. Springer Berlin / Heidelberg, 2006.
- G. Lakoff and M. Johnson. *Metaphors we live by*. University of Chicago Press, 1980.
- G. Lakoff. *Women, Fire and Dangerous Things*. University of Chicago Press, 1987.
- Pat Langley and Stephanie Sage. Induction of selective Bayesian classifiers. In *in Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence*, pages 399–406. Morgan Kaufmann, 1994.
- Pat Langley and Stephanie Sage. Tractable average-case analysis of naive Bayesian classifiers. In *ICML '99: Proceedings of the Sixteenth International Conference on Machine Learning*, pages 220–228, San Francisco, CA, USA, 1999. Morgan Kaufmann Publishers Inc.

- Steffen L. Lauritzen. The em algorithm for graphical association models with missing data. *Computational Statistics and Data Analysis*, 19(2):191–201, 1995.
- Anna-Katharina Lautenschütz, Clare Davies, Martin Raubal, Angela Schwering, and Eric Pederson. The influence of scale, context and spatial preposition in linguistic topology. In *Proceedings of the 2006 international conference on Spatial Cognition V: reasoning, action, interaction*, pages 439–452, Berlin, Heidelberg, 2007. Springer-Verlag.
- S. C. Levinson. Frames of reference and molyneux’s question: Cross-linguistic evidence. In P. Bloom, M. Peterson, L. Nadel, and M. Garrett, editors, *Language and space*, pages 109–169. MIT press, Cambridge, MA, 1996.
- Zhaoyu Li and Bruce D’Ambrosio. Efficient inference in bayes networks as a combinatorial optimization problem. *International Journal of Approximate Reasoning*, 11(1):55 – 81, 1994.
- K. Lockwood, K. Forbus, and J. Usher. Spacecase: A model of spatial preposition use. In *Proceedings of the 27th Annual Conference of the Cognitive Science Society*, 2005.
- K. Lockwood, K. Forbus, D. Halstead, and J. Usher. Automatic categorization of spatial prepositions. In *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, 2006.
- Gordon D. Logan. Linguistic and conceptual control of visual spatial attention. *Cognitive Psychology*, 28(2):103 – 174, 1995.
- Michael G. Madden. Evaluation of the performance of the Markov blanket Bayesian classifier algorithm. Technical Report NUIG-IT-011002, Department of Information Technology, National University of Ireland, Galway, Galway, Ireland, 2002.
- S. D. Mainwaring, B. Tversky, Mohoto Ohgishy, and D. J. Schiano. Descriptions of simple spatial scenes in English and Japanese. *Spatial Cognition and Computation*, 3(1):3–43, 2003.
- W. J. McGill. Multivariate information transmission. *Psychometrika*, 19(2):97–116, 1954.
- G.A. Miller and P.N. Johnson-Laird. *Language and perception*. Harvard University Press, 1976.
- Marvin Minsky. *The society of mind*. Simon & Schuster, Inc., New York, NY, USA, 1986.
- Daniel R. Montello. Scale and multiple psychologies of space. In A. U. Frank and I. Campari, editors, *Spatial Information Theory: A theoretical basis for GIS, proceedings of COSIT ‘93*, volume 1, pages 312–321. Springer-Verlag., 1993.
- Bernard Moulin and Driss Kettani. Route generation and description using the notions of object’s influence area and spatial conceptual map. *Spatial Cognition and Computation*, 1(3):227–259, 1999. ISSN 1387-5868.

- Richard E Neapolitan. *Learning Bayesian Networks*. Pearson Prentice Hall, 2004.
- C. Nothegger, S. Winter, and M Raubal. Computation of the salience of features. *Spatial Cognition and Computation*, 4:113–136, 2004.
- Patrick Olivier. A computational view of the cognitive semantics of spatial prepositions. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 303–309. Association for Computational Linguistics, June 1994.
- M. Pazzani. Searching for dependencies in Bayesian classifiers. In *Artificial Intelligence and Statistics IV, Lecture Notes in Statistics*. Springer-Verlag: New York., 1997.
- J. Pearl. Reverend bayes on inference engines: A distributed hierarchical approach. In *Proceedings of the American Association of Artificial Intelligence National Conference on AI*, pages 133–136, Pittsburgh, PA, 1982.
- J. Pearl. Fusion, propagation, and structuring in belief networks. pages 366–413, 1990.
- J Pearl. Evidential reasoning using stochastic simulation of causal models. *Artificial Intelligence.*, 32:245–257, 1987.
- J. Pearl. *Probabilistic reasoning in intelligent systems*. Morgan Kaufmann, San Mateo, CA, 1988.
- H.C. Peng, F. Long, and C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(8):1226–1238, 2005.
- Jose M. Peàa, Roland Nilsson, Johan Björkegren, and Jesper Tegnér. Towards scalable and data efficient learning of markov boundaries. *International Journal of Approximate Reasoning*, 45:211–232, 2007.
- Jodie. M. Plumert, Thomas L. Spalding, and Penney. Nichols-Whitehead. Preferences for ascending and descending hierarchical organization in spatial communication. *Memory and Cognition*, 29:274–284, 2001.
- J. M. Plumert, C. Carswell, K. DeVet, and D. Ihrig. The content and organization of communication about object locations. *Journal of Memory and Language*, 34:477–498, 1995.
- R. Porzel, M. Jansche, and R Klabunde. The generation of spatial descriptions from a cognitive point of view. In K. R. Coventry and P. Olivier, editors, *Spatial Language. Cognitive and Computational Perspectives*, pages 185–207. Dordrecht: Kluwer Academic Publishers, 2002.
- J. R. Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann. San Francisco, CA, 1993.

- D. A. Randell, Z. Cui, and A. G. Cohn. A spatial logic based on regions and connection. In *Proceedings of the 3rd Int. Conf. on Knowledge Representation and Reasoning*, pages 165–176. Morgan Kaufmann, 1992.
- M. Raubal and S. Winter. Enriching wayfinding instructions with local landmarks. In M. J. Egenhofer and D. M. Mark, editors, *Geographic Information Science Vol. 2478 of Lecture Notes in Computer Science*, pages 243–259. Berlin: Springer, 2002.
- Terry Regier. *The human semantic potential: Spatial language and constrained connectionism*. MIT Press, 1996.
- T Regier and L. Carlson. Grounding spatial language in perception: An empirical and computational investigation. *Journal of Experimental Psychology: General*, 130(2): 273–298, 2001.
- G Retz-Schmidt. Various views on spatial prepositions. *AI Magazine*, 9(2):95–105, 1988.
- Gert Rickheit and Ipke Wachsmuth, editors. *Situated Communication*. de Gruyter, 2006.
- D. K. Roy. Learning visually-grounded words and syntax for a scene description task. *Computer Speech and Language*, 16(3), 2002.
- Stuart Russell, John Binder, Daphne Koller, and Keiji Kanazawa. Local learning in probabilistic networks with hidden variables. In *IJCAI'95: Proceedings of the 14th international joint conference on Artificial intelligence*, pages 1146–1152. Morgan Kaufmann, 1995.
- Edward K. Sadalla, W. Jeffrey Burroughs, and Lorin J. Staplin. Reference points in spatial cognition. *Journal of Experimental Psychology: Human Learning and Memory*, 6(5):516–528, 1980.
- Michael F. Schober. Speakers, addressees and frames of reference: whose effort is minimised in conversations about location? *Discourse processes*, 20(2):219–247, 1995.
- Stephanie Schuldes, Michael Roth, Anette Frank, and Michael Strube. Creating an annotated corpus for generating walking directions. In *UCNLG+Sum '09: Proceedings of the 2009 Workshop on Language Generation and Summarisation*, pages 72–76, Morristown, NJ, USA, 2009. Association for Computational Linguistics.
- John Searle. Minds, brains and programs. *Behavioural and Brain Sciences*, 3(3):417–457, 1981.
- Ross D. Shachter. Evaluating influence diagrams. *OPERATIONS RESEARCH*, 34(6): 871–882, 1986.
- G. Socher, G. Sagerer, and P. Perona. Bayesian reasoning on qualitative descriptions from images and speech. *Image and Vision Computing*, 18(2):155 – 172, 2000.



- Kristoffer Söö, Dorian Galvez-Lopez, Chandana Paul, Patric Jensfelt, and Danica Kragic. Object search and localization for an indoor mobile robot. *Journal of Computing and Information Technology*, 17(1):67–80, 2009.
- M. Sorrows and S. Hirtle. The nature of landmarks for real and electronic spaces. In C. Freska and D. Mark, editors, *Spatial Information Theory: Cognitive and Computational Foundations of GIS*. Springer-Verlag, 1999.
- E. Strumbelj, I. Kononenko, and M. Robnik Sikojca. Explaining instance classifications with interactions of subsets of feature values. *Data and Knowledge Engineering*, 68: 886–904, 2009.
- Leonard Talmy. How language structures space. In Herbert L. Pick and Linda P. Acredolo, editors, *Spatial orientation: theory, research and application*. Springer, 1983.
- Leonard Talmy. *Toward a Cognitive Semantics*. MIT Press, 2000.
- Stefanie Tellex and Deb Roy. Towards surveillance video search by natural language query. In *CIVR '09: Proceeding of the ACM International Conference on Image and Video Retrieval*, pages 1–8, New York, NY, USA, 2009. ACM.
- Thora Tenbrink and Stephan Winter. Variable granularity in route directions. *Spatial Cognition and Computation*, 9(1):64–93, 2009.
- T. Tenbrink. Identifying objects on the basis of spatial contrast: An empirical study. In C. Freksa, M. Knauff, B. Krieg-Bruckner, B. Nebel, and T. Thomas Barkowsky, editors, *Spatial Cognition IV: Reasoning, Action, Interaction. International Conference Spatial Cognition 2004*, pages 124–146. Springer: Berlin, Heidelberg, 2005.
- T. Tezuka and K. Tanaka. Landmark extraction: A web mining approach. In A. G. Cohn and D. M. Mark, editors, *Spatial Information Theory, proceedings of COSIT 2005*. Springer, 2005.
- Martin Tomko and Stephan Winter. Pragmatic construction of destination descriptions for urban wayfinding. *Spatial Cognition and Computation*, 9(1):1–29, 2009.
- Claude Vandeloise. *Spatial Prepositions*. University of Chicago Press, 1991.
- K. van Deemter, I. van der Sluis, and A. Gatt. Building a semantically transparent corpus for the generation of referring expressions. 2006.
- K. van Deemter. Generating referring expressions: boolean extensions of the incremental algorithm. *Computational Linguistics*, 28(1):37–52, 2002.
- Sebastian Varges. Overgenerating referring expressions involving relations. In *Proceedings of the Third International Conference on Natural Language Generation (INLG-04) Brockenhurst, UK.*, pages 171–181. Springer, 2004.

- Ipke Wachsmuth. 'i, Max' - communicating with an artificial agent. In I. Wachsmuth and G Knoblich, editors, *Modeling Communication with Robots and Virtual Humans*, pages 279–295. Springer Berlin, 2008.
- W. Wahlster. One word says more than a thousand pictures. on the automatic verbalization of the results of image sequence analysis systems. *Computers and Artificial Intelligence*, 8(5):479–492, 1989.
- Peter Wazinski. Generating spatial descriptions for cross-modal references. In *Proceedings of the third conference on Applied natural language processing*, pages 56–63, Morristown, NJ, USA, 1992. Association for Computational Linguistics.
- Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics*, 1:80–83, 1945.
- Paul Williams and Risto Miikkulainen. Grounding language in descriptions of scenes. *Proceedings of the 28th Annual Meeting of the Cognitive Science Society*, pages 252–287, 2006.
- T. Winograd. *Procedures as a representation for data in a computer program for understanding natural language*. PhD thesis, Massachusetts Institute of Technology, 1971.
- Stephen Winter. Route adaptive selection of salient features. In W. Kuhn, M. Worboys, and S. Timpf, editors, *Spatial Information Theory, proceedings of COSIT 2003*. Springer, 2003.
- Jeremy M. Wolfe. Visual search. In H. Pashler, editor, *Attention*. University College London Press, London, UK, 1998.
- Atsushi Yamada, Toyooki Nishida, and Shuji Doshita. Figuring out most plausible interpretation from spatial descriptions. In *Proceedings of the 12th conference on Computational linguistics*, pages 764–769, Morristown, NJ, USA, 1988. Association for Computational Linguistics.
- Harry Zhang. The optimality of naive bayes. In Valerie Barr and Zdravko Markov, editors, *FLAIRS Conference*. AAAI Press, 2004.
- Brian D. Ziebart, Anind K. Dey, and James (Drew) Bagnell. Learning selectively conditioned forest structures with applications to DBNs and classification. In *Proceedings of Uncertainty in Artificial Intelligence (UAI 2007)*, July 2007.