The role of language in conceptual coordination

Cyprian Laskowski, B.Sc., M.Sc.

A thesis submitted in fulfilment of requirements for the degree of
Doctor of Philosophy

to

School of Philosophy, Psychology and Language Sciences
University of Edinburgh

January 2011

Declaration

I hereby declare that this thesis is of my own composition, and that it contains no material previously submitted for the award of any other degree. The work reported in this thesis has been executed by myself, except where due acknowledgement is made in the text.

Cyprian Laskowski

Abstract

Although concepts are located within individual minds, while word forms are shared across entire language communities, words and concepts are normally deemed to be tightly bound. But in fact, at least to the extent that concepts vary, the relationship between words and concepts may not be as uniform or stable as is often assumed. Nevertheless, language may itself mediate that relationship, through its entrenchment and use. Psychologists have already investigated language use in referential communication, but they have yet to focus in detail on the role of language in conceptual coordination. One of the obstacles has been the theoretical and methodological challenges that arise from seriously abandoning conceptual universals. To that end, an experimental framework was developed based on sorting tasks in which participants freely partition a set of stimuli into categories and an objective measure for comparing two outputs. Four experiments were then conducted to investigate whether people were conceptually coordinated before, during and after linguistic interaction.

Experiment 1 consisted of a cross-linguistic study looking at default coordination between native speakers. Participants both sorted items into groups and named them individually. There was a relatively high degree of categorisation agreement among speakers of the same language, but not nearly as high as for naming agreement. Experiments 2-4 inquired into conceptual coordination during or immediately after linguistic interaction. Experimental manipulations involved the form of language use (full dialogue or only category labels), as well as the type of feedback (category groupings, labels, both, or neither). In particular, Experiment 2 investigated the effects of categorising a set of objects together, with or without dialogue, on subsequent individual categorisation. The results were inconclusive and revealed specific methodological issues, but yielded interesting data and were encouraging for the general framework. Experiment 3 modified the design while testing and extending the same general hypotheses. Participants carried out a sequence of categorisation tasks in which they tried to coordinate their categories, followed by individual categorisation and similarity tasks. The availability of dialogue and feedback was manipulated in the interactive tasks. During interaction, they also received both kinds of feedback, except in the control condition. Pairs that could talk coordinated much better than the others, but feedback didn't help. Experiment 4 looked into the effects of the four possibilities for feedback during a longer sequence of interactive tasks. In general, conceptual coordination was found to depend on grouping feedback only. However, by the end of the task, pairs who received both kinds of feedback did best. All three interactive experiments also measured lexical convergence between pairs. The results generally revealed a dissociation, with lexical alignment showing more convergence and occurring under a wider variety of conditions.

Together with previous research, these findings show that language can bring about conceptual coordination. However, it appears that the richer the form of language use, the more conceptual convergence occurs, and the closer it gets coupled with lexical convergence. The long-term effects, if any, are much weaker. These studies have implications for the general role of language in cognition and other important issues.

Acknowledgements

It has been a very special four years of my life, and I have learned a lot. One of the most surprising lessons has been just how emotional doing a PhD can be. You imagine it's going to be this intellectual period devoted primarily to some serious analytic thinking, but in the end it's perhaps better characterised as an intense emotional rollercoaster. And it's pretty clear that I would not have gotten through it in one piece if it weren't for a lot of help from different sources.

I have been very lucky to have Martin Pickering as my primary supervisor. Despite his demanding schedule, he always found time to meet me, often on short notice or when he wasn't feeling so well, and even if it meant reading some overly verbose and incoherent draft of mine first. He found an excellent balance between giving constructive criticism and encouragement, and at reminding me not to ignore my data (even when they seemed to be at odds with his own theories). I sometimes wonder how much smoother my path would have been if I was less stubborn and followed his advice more often.

Jim Hurford and Nik Gisborne, my other supervisors, have been really encouraging and helpful as well, and during certain periods their contributions were critical. At other times, though, months would pass without them hearing a word from me, when I got deeply entrenched in my experiments and analysis. But when I needed a reminder that all those little numbers on my screen were actually tied to interesting theoretical questions, there was noone better I could turn to.

On the other hand, when I needed to make statistical sense of all those numbers, Alex Weiss was extremely kind in helping me out. Sometimes I would come to see him in a state of complete confusion and urgency. But usually I left with helpful answers, some of which came from him helping me realise that I wasn't as clueless as I thought.

I am very grateful for all the wonderful people I have met during my study and life here in Ediburgh, and with whom I have had a chance to babel chaotically, take frisbee breaks, have Homeric reading evenings, and more. I will not try to list them all here, but I will feel special nostalgia whenever I recall taking mind-clearing walks or runs in the hills with Hannah, listening to Sebastian's entertaining verbal nonsense after a couple of pints, and seeing Adrianna unable to communicate her ace holding. And thanks a lot Gareth for reading bits of my thesis when you should have been celebrating your flawless viva.

I must also apologise to all the people I have been increasingly neglecting as deadlines approached and my PhD consumed my life more and more. And not only those in Edinburgh, but beyond as well. I used to think I was bad at responding to emails, but I was an angel before compared to what I have become! I have never been very good at managing many things at once, a fact of which my flatmates have had to deal with in my neglect of shared flat cleaning duties during busy periods.

Also, I'd like to thank all the people of Edinburgh for being so considerate and not realising how refreshing Arthur's Seat can be at night. Sometimes when I just needed my own space and time, I could go up there in the middle of the night, and sing out loud without, for once, bothering anyone.

And then of course there's my mom, dad and sibling. They too have been neglected, often for months at a time. But when I needed something, sometimes urgently and in various ways, they have always been there, ready and willing to help immediately. They are simply amazing.

I come last to my most incredible discovery, far outstripping anything I have learnt about language and thought. Who would have thought I would have made it in Slovenia, while trying to get my mind off of presearch for a few weeks? From the moment she took my hand, or perhaps long before, Jana's entry into my life has been so beautiful, inspiring and revolutionising, that all I can say is: *wow*.

Contents

Declaration i

Abstract ii

Acknowledgements iv

Chapter 1 Introduction 1

1.1 Private concepts and public meanings? 1

1.2 Language and conceptual coordination 2

1.3 Theoretical and methodological challenges 4

1.4 Roadmap 7

Chapter 2 Conceptualisation in the individual 11

2.1 Introduction 11

2.2 The nature of concepts 13

2.2.1 Theoretical positions 14

2.2.2 Fodor's (1998) conditions 15

2.2.3 Levels of representation 17
2.3 Conceptual variation 18
2.3.1 Variation at different timescales 18
2.3.2 Abandoning conceptual universals 22
2.4 Concepts, categories and words 25
2.4.1 Concepts and categories 26
2.4.2 Categories and words 30
2.4.3 Concepts and words 32
2.4.3.1 Concepts and word meanings 32
2.4.3.2 Word forms and word meanings 34
2.4.3.3 Concepts and word forms 37
2.4.4 Conclusion 40
2.5 The internal structure of concepts 41
2.5.1 Prototypes, exemplars and theories 42
2.5.2 Hybrid views 45
2.5.3 Conclusion 48
2.6 Categorisation, conceptualisation and words 49
2.6.1 Introduction 49
2.6.2 Categorisation 50
2.6.3 Conceptualisation 53
2.6.4 Categorisation and similarity 56

2.6.5 Conceptualisation and words 60

2.6.6 Conclusion 65

2.7 Summary 67

Chapter 3 Conceptualisation and shared words 69

3.1 Introduction 69

3.2 Word input 70

3.2.1 Infants 70

3.2.2 Children 72

3.2.3 Adults 74

3.2.4 Conclusion 76

3.3 Linguistic relativity 77

3.3.1 Whorf 77

3.3.2 Colour 79

3.3.3 Objects 86

3.3.4 Other domains 88

3.3.4.1 Space 88

3.3.4.2 Number 90

3.3.4.3 Time 90

3.3.5 Theoretical positions 91

3.3.6 Conclusion 93

3.4 Linguistic interaction 94

3.4.1 Joint action and dialogue 95

3.4.2 Common ground 97

3.4.2.1 Theoretical background 97

3.4.2.2 Interactive dialogue experiments 99

3.4.2.3 Comprehension and production experiments 101

3.4.2.4 Summary 103

3.4.3 Alignment 104

3.4.3.1 Theory and scope 104

3.4.3.2 Lexical (and conceptual) alignment 106

3.4.3.3 Preserved conceptual alignment 110

3.4.4 Conclusion 114

3.5 Summary 115

Chapter 4 Developing a framework 118

4.1 Criteria 118

4.2 Categorisation tasks 120

4.2.1 Justification 120

4.2.2 Categorisation paradigms 121

4.2.2.1 Picture naming task 122

4.2.2.2 Category training 123

4.2.2.3 Familiarisation and novelty 125

4.2.2.4 Tasks involving two or three stimuli at a time 126

4.2.2.5 Free classification 127

4.2.2.6 Coordination games 129

4.2.3 Conclusion 131

4.3 Comparing two task outputs 131

4.3.1 The requirements 131

4.3.2 An example 133

4.3.3 Possible measures 134

4.4 Experimental manipulations 139

4.5 Stimulus considerations 140

4.6 Summary 142

Chapter 5 Experiment 1: native language and prior conceptual alignment 143

5.1 Introduction 143

5.2 Malt et al. (1999) 145

5.2.1 Purpose 145

5.2.2 Design 146

5.2.3 Analysis and results 147

5.2.4 Interpretation 149

5.3 Experiment 1 151

5.3.1 Overview 151

5.3.2 Method 154

5.3.2.1 Participants 154

5.3.2.2 Stimuli 155

5.3.2.3 Conditions 156

5.3.2.4 Procedure 156

5.3.3 Results 157

5.3.3.1 Example output 157

5.3.3.2 Comparisons of naming patterns 157

5.3.3.3 Comparisons of sorting patterns 167

5.3.3.4 Correspondences between naming and sorting patterns . 169

5.3.3.5 Questionnaire data 176

5.3.4 Discussion 176

5.3.4.1 Replication and extension of Malt et al. (1999) 176

5.3.4.2 Prior conceptual (and lexical) alignment 177

5.3.4.3 Is language responsible? 179

5.3.4.4 Methodological considerations 181

5.3.5 Conclusion 184

Chapter 6 Experiment 2: conceptual convergence due to dialogue 185

6.1 Introduction 185

6.2 Overview 187

6.3 Methods 190

6.3.1 Participants 190

6.3.2 Conditions 190

6.3.3 Stimuli 190
6.3.3.1 Generation 191
6.3.3.2 Allotment 192
6.3.4 Procedure 194
6.3.5 Program 195
6.4 Results 197
6.4.1 Conceptual convergence 197
6.4.2 Lexical convergence 200
6.4.3 Lexical choices 203
6.4.4 Use of dialogue 205
6.4.5 Framework and experiment feedback 208
6.5 Discussion 211
6.5.1 Summary of results 211
6.5.2 Conceptual alignment, lexical alignment and dialogue 212
6.5.3 Methodological issues 213
6.6 Conclusion 216

Chapter 7 Experiment 3: conceptual alignment during and after dialogue 218
7.1 Introduction 218
7.2 Overview 219
7.3 Methods 222
7.3.1 Participants 222

7.3.2 Conditions 223

7.3.3 Stimuli 223

7.3.3.1 Generation 223

7.3.3.2 Allotment 224

7.3.4 Procedure 226

7.3.4.1 Joint phase 226

7.3.4.1.1 Joint categorisation tasks 226

7.3.4.2 Individual phase 226

7.3.4.2.1 Similarity judgement tasks 227

7.3.4.2.2 Individual categorisation task 227

7.3.4.3 Questionnaire 227

7.3.5 Program 227

7.4 Results 228

7.4.1 Conceptual alignment 229

7.4.1.1 Joint categorisation tasks 229

7.4.1.2 Individual categorisation tasks 232

7.4.2 Lexical alignment 234

7.4.2.1 Joint categorisation tasks 234

7.4.2.2 Individual categorisation tasks 236

7.4.3 Similarity convergence 238

7.4.4 Questionnaire data 239

7.5 Discussion 240

7.5.1 Summary of results 240

7.5.2 Conceptual and lexical alignment during interaction 241

7.5.3 Conceptual, lexical and similarity alignment beyond interaction . . 244

7.5.4 Methodological evaluation 246

7.6 Conclusion 248

Chapter 8 Experiment 4: Category versus word feedback in conceptual and lexical alignment 249

8.1 Introduction 249

8.2 Overview 251

8.3 Methods 254

8.3.1 Participants 254

8.3.2 Stimuli 254

8.3.2.1 Generation 254

8.3.2.2 Allotment 256

8.3.3 Conditions 256

8.3.4 Procedure 258

8.3.5 Program 258

8.4 Results 259

8.4.1 Example task output 259

8.4.2 Conceptual alignment 261

8.4.3 Lexical alignment 269

8.4.4 Correspondence between conceptual and lexical alignment 272

8.4.5 Questionnaire data 276

8.5 Discussion 276

8.5.1 Summary of results 276

8.5.2 Conceptual and lexical alignment 277

8.5.2.1 A dissociation 277

8.5.2.2 Not entirely dissociated 278

8.5.2.3 Relation to previous alignment results 280

8.5.3 Conceptual structure 282

8.5.4 Methodological comments 285

8.6 Conclusion 286

Chapter 9 General discussion 287

9.1 Summary of empirical findings 287

9.2 Evaluation of the empirical framework 289

9.3 Conceptual (and lexical) alignment 294

9.3.1 The unit of alignment 294

9.3.2 Prior, online and preserved alignment 297

9.3.3 Dissociation between lexical and conceptual alignment 301

9.3.4 The role of language in conceptual alignment 302

9.4 Linguistic relativity 306

9.5 Revisiting conceptual issues . . . 308

9.5.1 Concepts and categories . . . 308

9.5.2 Concepts, conceptualisation and variation . . . 309

9.5.3 Concepts and words . . . 311

9.5.4 Internal structure . . . 314

9.6 Conceptual development, language acquisition, and language evolution . 314

9.7 Conclusion . . . 317

Appendix A Experiment 1 materials 319

A.1 Instructions . . . 319

A.2 Questionnaire . . . 321

Appendix B Experiment 2 materials 322

B.1 Instructions . . . 322

B.2 Questionnaire . . . 325

B.3 Program interface . . . 327

Appendix C Experiment 3 materials 329

C.1 Instructions . . . 329

C.2 Questionnaire . . . 333

C.3 Program interface . . . 334

Appendix D Experiment 4 materials 337

D.1 Instructions . . . 337

D.2 Questionnaire 340

D.3 Program interface 341

Appendix E Triangle generation program 343

Appendix F Published papers 348

References 359

List of Tables

5.1 Malt et al.'s (1999) correlations 148

5.2 Experiment 1: dominant names 161

5.3 Experiment 1: naming correlations (by distributions) 162

5.4 Experiment 1: naming correlations (by sorting method) 163

5.5 Experiment 1: sorting correlations 167

5.6 Experiment 1: naming(distributional)-sorting correlations 172

5.7 Experiment 1: naming(sorting-style)-sorting correlations 173

5.8 Experiment 1: naming(distributional)-sorting difference correlations . . . 174

5.9 Experiment 1: naming(sorting-style)-sorting difference correlations 174

6.1 Experiment 2: conceptual means 198

6.2 Experiment 2: lexical means 201

6.3 Experiment 2: lexical means secondary 202

6.4 Experiment 2: common labels 204

7.1 Experiment 3: joint task conceptual means 231

7.2 Experiment 3: joint task conceptual means means 232

7.3 Experiment 3: individual task means 233

7.4 Experiment 3: joint task lexical means 235

7.5 Experiment 3: joint task lexical ceilings and floors 236

7.6 Experiment 3: joint task lexical means means 236

7.7 Experiment 3: individual task lexical means 237

7.8 Experiment 3: similarity task means 238

8.1 Experiment 4: conditions 258

8.2 Experiment 4: conceptual means 264

8.3 Experiment 4: conceptual model selection: random effects terms 264

8.4 Experiment 4: conceptual model selection: groupings terms 265

8.5 Experiment 4: conceptual model selection: labels terms 265

8.6 Experiment 4: conceptual means 266

8.7 Experiment 4: conceptual means (last 5) 267

8.8 Experiment 4: conceptual means (first 5) 268

8.9 Experiment 4: lexical model selection: random effects terms 271

8.10 Experiment 4: lexical model selection: labels 271

8.11 Experiment 4: lexical model selection: groupings 271

8.12 Experiment 4: lexical means 272

8.13 Experiment 4: lexical means (last 5) 273

List of Figures

1.1 Dishes 7

2.1 Simple word-concept-object model 12

2.2 A concept and its category 30

2.3 Near synonyms 37

2.4 A concept and its words 40

2.5 A concept, its category and words 41

2.6 A concept's internal structure 48

2.7 A concept (with internal structure), category, and words 49

2.8 An act of conceptualisation 67

2.9 An act of conceptualisation, with internal structure 67

3.1 Comparing linguistic categorisations 94

3.2 Comparing non-linguistic categorisations 94

3.3 Interactive model 115

4.1 Comparing conceptualisation 120

4.2 Comparing categorisations 121

4.3 Measuring alignment 132

5.1 Malt et al.'s (1999) linguistic and recognition categories 145

5.2 Comparing non-linguistic categorisation 152

5.3 Experiment 1: stimuli 155

5.4 Experiment 1: sorting task example 158

5.5 Experiment 1: naming task example (a) 159

5.6 Experiment 1: naming task example (b) 160

5.7 Experiment 1: English MDS solution 170

5.8 Experiment 1: Polish MDS solution 171

5.9 Experiment 1: Japanese MDS solution 172

6.1 Experiment 2: interactive model 186

6.2 Experiment 2: interactive joint task 187

6.3 Experiment 2: base stimuli 191

6.4 Experiment 2: morphing example 192

6.5 Experiment 2: a stimulus subset 193

6.6 Experiment 2: task sequence 195

6.7 Experiment 2: joint task example 198

6.8 Experiment 2: conceptual distributions 199

6.9 Experiment 2: lexical distributions 201

6.10 Experiment 2: lexical distributions alternative 203

7.1 Experiment 3: *control* condition 220

7.2 Experiment 3: *silent* condition 220

7.3 Experiment 3: *talking* condition 221

7.4 Experiment 3: original base stimuli 224

7.5 Experiment 3: derived base stimuli 224

7.6 Experiment 3: morphing example 225

7.7 Experiment 3: joint categorisation example 229

7.8 Experiment 3: joint task conceptual runs 230

7.9 Experiment 3: joint task conceptual distributions 231

7.10 Experiment 3: individual task distributions 233

7.11 Experiment 3: joint task lexical runs 234

7.12 Experiment 3: individual task lexical distributions 237

7.13 Experiment 3: similarity task distributions 239

8.1 Experiment 4: expanded interactive model 250

8.2 Experiment 4: feedback by condition 252

8.3 Experiment 4: stimulus subset 257

8.4 Experiment 4: example output 1 260

8.5 Experiment 4: example output 2 261

8.6 Experiment 4: example output 3 261

8.7 Experiment 4: conceptual runs 263

8.8 Experiment 4: conceptual interaction plot 268

8.9 Experiment 4: lexical runs 270

8.10 Experiment 4: label agreement and conceptual alignment 274

8.11 Experiment 4: lexical and conceptual alignment 275

8.12 Experiment 4: label stability 275

9.1 Concept and conceptualisation 310

B.1 Experiment 2: initial interface 327

B.2 Experiment 2: mid-task interface 328

C.1 Experiment 3: joint task initial interface 334

C.2 Experiment: joint task mid-task interface 335

C.3 Experiment 3: joint task feedback interface 335

C.4 Experiment 3: similarity interface 336

C.5 Experiment 3: individual categorisation interface 336

D.1 Experiment 4: initial interface 341

D.2 Experiment 4: mid-task interface 342

D.3 Experiment 4: feedback interface 342

CHAPTER 1

Introduction

1.1 Private concepts and public meanings?

A few years ago, over Christmas dinner, my father and brother had a long debate about what constituted an empire. While neither would dispute prototypical cases like ancient Rome, they were deeply entrenched in their positions regarding borderline cases. One of them (I can't remember who) kept insisting that Japan was an empire (since it has an emperor), while the other countered with the United States (due to its wide influence in international affairs). But after their characteristic stubbornness made them drag on for a while, I eventually stopped gnashing my teeth and erupted, "You're both right: it just depends on what you mean!"

This example illustrates a fundamental tension in the relationship between public meanings and private concepts. The problem is that, although concepts are located within individual minds, word forms are shared across entire language communities. As such, we must either claim that concepts are identical for all speakers of a given language, or that the same words may identify slightly different concepts for different people (or even for the same person in different situations). But examples like the one above clearly favour the latter position. Concepts vary, so, assuming concepts are word meanings, word meanings vary as well. So why is it so hard to break the intuition that the meanings of words are fixed and public?

Indeed, the tension persists, and is one of the more difficult problems in cognitive science. Philosophers, psychologists and linguists (among others) have dealt with it in a variety of ways (Levinson 1997). Frege (1892/1948) develops a theory which clearly distinguishes the meanings of words from the objects they refer to in the world. But he also distinguishes fixed public meanings from variable private concepts, and makes no effort to relate the two (Johnson-Laird 1983). Fodor (1998) collapses Frege's latter distinction, claiming that meanings are concepts, and concepts are public. But then of course he is confronted with facts of conceptual variation, although he does his best to dismiss them. Putnam (1975) takes the other extreme, taking meaning out of the mind altogether, and grounding it in truth instead. Under this view, any correspondence between meanings and concepts only reflects our degree of understanding of the true nature of things in the world. On the other hand, Johnson (1987) grounds meaning in our interaction with the world through our bodies. While this allows for potential variation between people, he also argues how commonalities in human biology and experience can explain why similar concepts emerge so that we can still communicate. And Rosch (1978), as a cognitive psychologist, explores cognitive representations underlying meaning. Nevertheless, she still abstracts away from individuals and focuses on the structure of public categories. Finally, Barsalou (1999) proposes that perceptual representations, if properly construed, can play a much more sophisticated role in cognition than is normally acknowledged. Although his theory embraces representational variation both between and within people, he also shows how one of its foundational components can help explain conceptual stability.

While this is just a small sampling of theoretical positions on this controversial subject, it highlights how the treatment of word meanings (and not just word forms) as public resources is pervasive even in the scientific community, and how this has marked our notion of concepts. In fact, as we will see in Section 1.3, there are actually sensible theoretical reasons for trying to hang on to conceptual universals, although I will need to resist them. But first, I raise a possibility that will be the focus of my thesis: that language itself may be responsible for bringing people's concepts closer together.

1.2 Language and conceptual coordination

So far I have been implicitly talking about language as if it mirrored the mind (Chomsky 1968). In this view, words merely encode concepts (Jackendoff 1983). But that is not necessarily the case. Once we acknowledge that word-concept mappings may vary from person to person, however slightly, and that word forms (if not their meanings) are public resources, the possibility arises that people might affect each other's word-concept mappings. Given how quickly and accurately children learn words and their meanings (Bloom 2000), and how we are susceptible to influences from the words of others all the way from infancy (Waxman and Markow 1995) to adulthood (Lupyan, Rakison and McClelland 2007), we could hypothesise more specifically that language leads to conceptual coordination. This will be the general question I will focus on in my thesis:

Does language bring about conceptual coordination between individuals?

There are good empirical reasons for supposing that language might do so. Clark (1996) has argued that language use is intimately involved in most forms of social interaction, and he and his colleagues have demonstrated how dialogue in particular both depends on and builds up mutual knowledge during interaction (Clark and Wilkes-Gibbs 1986; Schober and Clark 1989; Wilkes-Gibbs and Clark 1992). Researchers have also proposed, based on empirical evidence, that conversation partners converge on referential expressions and their underlying conceptualisations (Garrod and Anderson 1987; Brennan and Clark 1996), and that engaging in a joint task involving dialogue can result in conceptual convergence that persists even after interaction (Markman and Makin 1998; Voiklis 2008). These experiments have fuelled theoretical proposals suggesting that interlocutors in dialogue align their mental representations at different levels, including lexical and conceptual ones (Pickering and Garrod 2004).

Moreover, if linguistic interaction effects on conceptual coordination are not evanescent, then they might, through repeated application, result in a more permanent transformation of our concepts (Malt and Sloman 2004). If so, we might expect native speakers of the same language to conceptualise things in relatively similar ways, even when they are not using language. This is the basic idea behind Whorf's (1956) influential proposal of linguistic relativity, which states that our native language determines the way we see the world. Although Whorf's claim is now widely recognised as being too strong, and even in the most studied domain controversy remains (Regier and Kay 2009), weaker versions of Whorf's position have taken over and gathered empirical support (Gumperz and Levinson 1996; Gentner and Goldin-Meadow 2003*b*).

However, it would be premature to conclude that language necessarily drives conceptual coordination, because the results may be deceptive. First of all, in order to investigate the relationship between language and conceptualisation, it is important to separate the two phenomena (Nuyts and Pederson 1997). We cannot assume that common word choice reflects common conceptualisation. Just because two people use the same word to refer to the same thing, doesn't mean that their underlying concepts match (e.g., my brother and father still think of empires differently, even if they both consider ancient Rome to be a good example). Our interlocutor's words may prime us lexically, but that does not automatically imply that we are also lining up conceptually (Schober 2005).

Similarly, we also have to be cautious about concluding that language is the causal factor in conceptual coordination. Just as it is difficult to isolate language from conceptualisation, so it is also hard to separate it from both local interaction and general culture. It is possible that the conceptual coordination that seems to occur in dialogue is actually more due to other properties of interaction rather than language itself. Similarly, we may share similar concepts to other people of our linguistic community due to cultural factors that are confounded with language. Therefore, we need to be careful both in designing experiments and interpreting them. While this is already a substantial challenge, it is further aggravated by theoretical and methodological complications, as I will outline next.

1.3 Theoretical and methodological challenges

Having identified an empirical focus of investigation, it is tempting to run to the experimental lab to conduct the first experiment. However, in this case, doing so would result in theoretical chaos, and our only coping mechanism would be to either ignore theory altogether or make a mess of it.

Why is this the case? I have already suggested how an investigation of the role of language in conceptual coordination presupposes that we take two issues particularly seriously: conceptual variation on one hand, and the separation of language from conceptualisation on the other. We will soon find that while these principles may seem relatively innocuous, they have deep theoretical consequences, and staying loyal to them is no easy task. The fundamental issue is that they leave us without a straightforward way to identify concepts. If we cannot rely on convenient universals or language, then how can we study a single person's concepts, let alone conceptual coordination between people?

Perhaps Fodor (1998) puts the danger of conceptual variation best (following up on a thought experiment from Putnam 1975):

> "If everybody else's concept WATER is different from mine, then it is literally true that only I have ever wanted a drink of water, and that the intentional generalization 'Thirsty people seek water' applies only to me. (And, of course, only I can state the generalization; words express concepts, so if your WATER concept is different from mine, 'Thirsty people seek water' means something different when you say it than when I do.) Prima facie, it would appear that any very thoroughgoing conceptual relativism would preclude intentional generalizations with any very serious explanatory power. This holds in spades if, as seems likely, a coherent conceptual relativist has to claim that conceptual identity can't be maintained even across time slices of the same individual." (p. 29)

I disagree with Fodor (1998) that "conceptual relativism" would be so catastrophic. For many purposes, even social ones, concepts probably don't need to be identical, but just similar enough (Segal 2000). And in any case, we cannot reject conceptual variation just because it is theoretically inconvenient, especially given the variety of evidence for it (as we'll see in Section 2.3.1).

Nevertheless, the point stands that rejecting universal concepts or word meanings leaves us without an external, objective yardstick. Methodologically, this implies that I cannot just conduct experiments where I ask participants to categorise stimuli verbally or according to pre-specified criteria. Without a well-defined metric, it is difficult to do science, and concepts are no exception.

Therefore, in order to address this, I first develop a simple theoretical model incorporating the structures, processes and relationships that concepts enter into. The model is not meant to be particularly sophisticated or to definitively resolve any issues, but it is intended to be grounded in previous work, and to serve the foundational role that is required for the design and interpretation of my experiments. As such, it will require me to synthesise empirical evidence and theoretical perspectives from different disciplines, particularly when it comes to relating what's inside the mind and what's beyond it.

The main point of this process is to reach a reasonable answer to the problem raised by Fodor (1998) above, given my necessary prior commitments. How can I identify concepts, both theoretically and methodologically? To anticipate, my solution draws on three main ideas. First, although it requires some theoretical compromise, I identify categories, which I distinguish from concepts, as our best proxy to concepts. As such, categorisation tasks, provided they are of the right sort, are the way to go (Murphy 2002). Second, I emphasise the dynamic nature of conceptualisation, and how we can take snapshots of different people's concepts at different times, without committing ourselves on certain philosophical issues such as conceptual individuation (Fodor and Lepore 1992). Third, I note that the nature of my problem of investigation actually provides a sort of conceptual yardstick for people's categorisations: the categorisations of other people (Markman and Makin 1998). In particular, for my purposes, I do not need to worry about how participants conceptualise experimental stimuli directly, but only how they compare to others. As long as we have an objective measure for comparing people's categorisations, we have enough of a conceptual metre stick. This does not mean, however, that I take a black box approach to categorisation: categories reflect concepts, and studying them carefully can shed light on the relationships between language, culture and thought (Malt 2006). Indeed, I will attempt to relate some of my experimental results to the internal constituents of concepts, as established by psychological theories of categorisation.

To give a feeling for what my approach will involve, as well as how it will manifest itself experimentally, imagine the following situation. You are moving into a new flat with someone, and you share a kitchen, with a variety of tableware as in Figure 1.1. How will you and your flatmate coordinate your organisation of the cupboards? Will you follow the way people do it in your home country? Will you observe how your flatmate places things, and try to adopt it yourself? Will you place labels on the cupboards to establish the appropriate contents? Or will you discuss it directly, and even arrange the things in the cupboard together at some point? Language may be involved to differing degrees for these different approaches, and may have correspondingly different effects. But notice that if I want to compare how you and your flatmate go about it, I don't need to ask you what you call the dishes, or to understand how you individually make your choices. I just need to look into the cupboard regularly, and have a fair way of comparing the sets of things that I find there at different times.

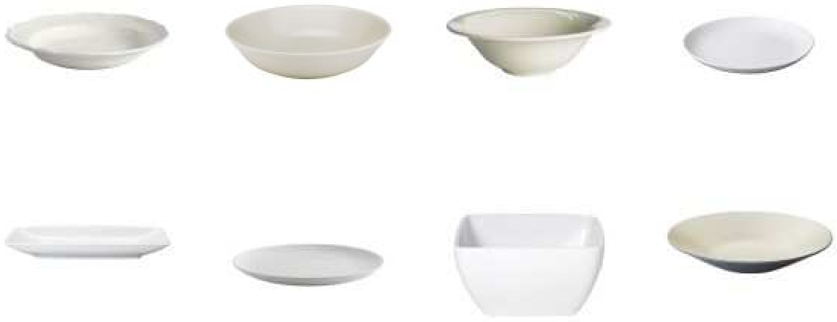


Figure 1.1: Dishes.

1.4 Roadmap

As I have suggested, although largely experimental, my thesis also has a substantial theoretical flavour to it. I develop a simple theoretical model for my purposes, review past work related to the role of language in conceptual coordination, develop an experimental framework for pursuing the issues more directly, present four experiments that did so, and discuss my findings and their theoretical implications.

Chapter 2 lays the theoretical foundations for the rest of the thesis. Although an investigation considering conceptual coordination and language is intrinsically social in nature, we must first look inside the individual in order to figure out the structures that are to be coordinated, the processes that they engage in, and how these structures and processes relate to language. En route, I develop a simple theoretical model which is meant to capture these relationships at an appropriate level of detail. This turns out to be surprisingly challenging, especially if we are to take conceptual variation seriously. However, since allowing for such variation is a logical prerequisite to my study, it is crucial that we stay loyal to it.

Having established my theoretical ground, I move on in Chapter 3 to reviewing mostly empirical work that is of direct significance to my research questions. I begin with what we already know about how lexical input from others can affect our concepts and conceptualisation. Then I turn to two major themes of central importance to my thesis. First, I consider the traditional claim that our native language determines our store of concepts. I review existing literature on this topic, focusing particularly on the widely studied domain of colour. Second, I discuss interactive studies involving dialogue. I converge on two particular sets of studies: those examining conceptual coordination indirectly through lexical coordination, and those testing for conceptual coordination after linguistic interaction. The review in this chapter leads to the identification of more specific empirical targets for my thesis.

Chapter 4 asks what methods can be used to pursue these empirical goals. I first argue that the use of categorisation experiments is the most suitable strategy for my empirical investigations. However, there are many different possible categorisation paradigms, and it is important to adopt one which is appropriate and sensitive to the theoretical issues raised in Chapter 2. To that end, based on criteria that I lay out for my experimental framework, I argue for one particular type of categorisation task, and how it can be embedded in an interactive context involving pairs of individuals and linguistic manipulations. I also mention considerations regarding the choice of stimulus domains, and decide among objective measures for comparing categorisation outputs. I converge on a general experimental framework which I use for all of my experiments.

In Chapter 5, I address the question of how coordinated people's concepts might be even before they interact, by virtue of their being native speakers of the same language. I relate this question to the linguistic relativity hypothesis, and discuss a particularly relevant cross-linguistic study from the literature in some detail. The experiment had examined both linguistic and non-linguistic categorisation, and found that native speakers of different languages tended to diverge in the former but not in the latter. However, although the study had used the same kind of categorisation task as I recruited for my framework, it had not addressed the issue of conceptual coordination directly. Therefore, I then present Experiment 1, which had two aims: to replicate the original study, and to extend its methods and analysis to address my coordination hypothesis. The replication was successful, and my additional analysis complemented but also qualified the original findings.

Chapter 6 then presents the first of three experiments I conducted which involve pairs of participants and focus on the possible effects of language use and interaction on conceptual coordination. Since past work had shown how natural conversation involves a high degree of coordination and development of mutual knowledge, Experiment 2 asked whether carrying out a joint categorisation task while engaging in dialogue led to conceptual convergence between people, as examined by subsequent individual categorisation tasks. I also attempted to isolate the impact of dialogue specifically from interaction more generally by setting up an intermediate condition in which participants categorised jointly together without speaking. Unfortunately, the experiment had methodological flaws and did not resolve the issues I was after, but it did yield other interesting data and showed that the general experimental framework had potential.

Chapter 7 therefore presents a third experiment, in which methodological improvements were made and the scope of inquiry was expanded to consider conceptual coordination not just during but also immediately after interaction. To that end, in Experiment 3, participant pairs also carried out a sequence of joint categorisation tasks, while also modifying the design so that I could isolate their categorisation as individuals during the interaction. After these joint tasks, participants individually carried out both categorisation and similarity judgement tasks. The experiment revealed that dialogue massively facilitated conceptual coordination, but simple interaction without talking did not. However, it also showed how participants coordinated lexical choices, even without full dialogue. On the other hand, in conflict with previous research, there was no evidence of coordination on the subsequent individual part of the experiment.

Chapter 8 presents my final experiment, which delves in more detail into issues raised by Experiment 3. In particular, Experiment 4 looked further into conceptual coordination during dialogue-less interaction, but also places more emphasis on lexical coordination and a comparison between the two. To that end, the experiment consisted exclusively of a (larger) set of joint categorisation tasks, and manipulated whether participants received conceptual and/or lexical feedback after each task. The results suggested that there is a surprising degree of dissociation between lexical and conceptual coordination. Lexical coordination relied exclusively on lexical feedback, and conceptual coordination relied mostly (though not only) on conceptual feedback. This also has implications for the internal structure of concepts.

Finally, Chapter 9 brings the whole thesis together. I first summarise the results of my four experiments. Next I evaluate the experimental framework, discussing its weaknesses and strengths, and showing how it could be applied to many other questions and domains in further research. I then return to the main questions with which I started the thesis, discussing my results in terms of the role of language in conceptual coordination before, during and after interaction. In the process, I also discuss the surprising results concerning lexical coordination and the apparent dissociation between the two. After that, I separately consider the question of linguistic relativity, followed by a reexamination of the foundational theoretical issues raised in Chapter 2. I then relate my results to the otherwise neglected topic of the relationship between conceptual development, language acquisition and language evolution. Finally, I review potential directions for future work, and draw overall conclusions.

CHAPTER 2

Conceptualisation in the individual

2.1 Introduction

The study of coordination of a phenomenon requires first an understanding of the phenomenon itself, especially when the latter is already an elusive subject. As a result, before I tackle conceptual coordination and the role of language therein, it is crucial to first address the underlying process of individual conceptualisation. Nevertheless, words are also stored in the minds of individuals, and thus may already have a role in conceptualisation even before we consider language's communicative function and use. Therefore, this chapter focuses on the question of how a person conceptualises things in the world, and what role words might play in this process.

The reader may question whether the length of this chapter is justified, given that these issues are preliminary to the actual empirical hypotheses I am ultimately addressing. Why delve so far into theoretically murky waters in a primarily experimental psychology thesis concerning categorisation? Why not just adopt a simple and commonly assumed model, as shown in Figure 2.1, with isomorphic relationships between words, concepts and objects?

Unfortunately, as I suggested in Section 1.3, I do not have that luxury here, due to the nature of my research questions. The main problem is the central position that **conceptual variation** occupies in my thesis. In particular, investigating conceptual coordination presupposes the possibility of conceptual variation. Different people may divide the world up in (perhaps slightly) different ways. As a result, I will have to explicitly

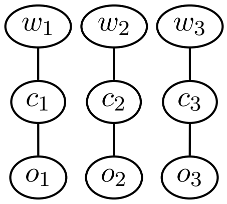


Figure 2.1: An isomorphic relationship between words, concepts, and objects: Each word (e.g., w_2) is associated with a particular concept (e.g., c_2), which in turn is associated with a particular kind of object (e.g., o_2). In such an oversimplified model, the three systems mirror each other perfectly.

reject the claim that different people necessarily have the same concepts (Fodor 1998). Indeed, concepts might also vary intra-personally, so that a single person may conceptualise the same thing in different ways at different times. While this position may not seem terribly controversial from a psychological perspective, we will see that taking the possibility of conceptual variation seriously has surprisingly far-reaching theoretical implications. Moreover, the problem is compounded by the fact that I am specifically exploring the role of **language** in conceptual coordination. As such, any hope of clinging to words as reliable conceptual identifiers is lost, because I will need an experimental way of explicitly separating words and concepts, from both the researcher's and the participant's points of view. In practical terms, I will not be able to identify concepts by relying on tasks with merely linguistic responses (as is often done in psychology experiments; Lloyd-Jones and Humphreys 1997).

Note that these problems are by no means exclusive to this thesis, and are particularly familiar to the developmental psychologist. As Keil (1992) points out, "it is difficult to design and motivate empirical studies on concept acquisition without first committing oneself to a set of assumptions about what concepts are and how they are represented" (p. 25). While my thesis does not focus on conceptual development, some of the theoretical difficulties are analogous. As children learn about the world, the correspondences between their concepts and objects in the world change; and, similarly, as they acquire language, the correspondences between concepts and words changes. As such, it is difficult if not impossible to identify conceptual norms to which their concepts can be reliably related and thereby grounded.

I could still try to downplay these issues, and leave the broader interpretation to those more qualified. This would protect me from launching into dangerous philosophical territory, and allow me to get straight to business. However, it would also make the interpretation of my experiments difficult and limited. Instead, in order to theoretically ground my experiments, I will emphatically drop the convenient idealisation of neat, static and universal isomorphisms between words, concepts, and objects. In its place, I will develop a more dynamic model, bringing together theoretical arguments and empirical evidence.

Therefore, this chapter is necessary groundwork for engaging the actual main topic of this thesis. As such, the review of literature relating to conceptual coordination and its relationship to language is not taken up until Chapter 3.

The chapter is divided as follows. I begin by identifying my position on the nature of concepts within the large interdisciplinary web of perspectives. I then list some evidence for conceptual variation, and commit to taking conceptual variation seriously, whatever the consequences. I then proceed to the relationships between concepts, categories and words, keeping conceptual variation closely in mind along the way. Next, I consider the internal structure of concepts, and how they relate to words and categories. And finally, I consider how concepts are applied in the dynamic process of conceptualisation, and how it relates to words and similarity. As the chapter rolls along, I develop a simple model of conceptualisation which will serve as my guide for the rest of the thesis.

2.2 The nature of concepts

The nature of concepts is a particularly controversial topic with origins in the ancient philosophical debate between empiricism and rationalism (Prinz 2005). Concepts remain a central concern for modern cognitive science in general, where they are approached from an enormous number of different perspectives in psychology, linguistics, philosophy, anthropology, computer science, and neuroscience (Cohen and Lefebvre 2005). However, the actual nature of concepts is well beyond the scope of my thesis. While some of the basic structures, processes and relationships involving concepts will be critical, I do not need to pursue their fundamental nature. Nevertheless, it is important to give some idea of what I understand by concepts.

To that end, in this section I first list a few different theoretical approaches to concepts. Then I describe Fodor's (1998) conditions for a theory of concepts, and use it to lay out my own position. Finally, I mention different levels of mental representation at which we can study concepts, and specify which will constitute my focus.

2.2.1 Theoretical positions

Philosophers, linguists and psychologists have taken many different perspectives on concepts, depending on their purposes and theoretical orientation. Here I briefly list some of these positions, for each of these disciplines in turn. This list is not meant to do justice to this topic, and is unavoidably chaotic. However, my only purpose here is to demonstrate how controversial is the nature of concepts.

I begin with philosophers, since they have the most to say about the theoretical nature of concepts. Fodor (1998) presents a strong nativist theory, according to which concepts are universal, (mostly) innate, and have no internal structure. He acknowledges that such a view may seem radical, but argues against other standard approaches, which he thinks we are destined to abandon. Peacocke (1992) tackles the problem of concept possession. In his view, for a person to have a certain concept, he must satisfy certain objective, necessary and sufficient "possession conditions". Churchland (1989) advocates a connectionist approach to concepts, grounded in neural networks. In this view, concepts are changing distributed representations rather than fixed abstract symbols. Gärdenfors (2000) proposes that concepts occupy regions in an abstract geometric space. His theory attempts to reconcile the gap between symbolic and connectionist representations of concepts. Millikan (2000) argues that humans are remarkably tolerant to large changes in something without changing their conceptualisation of it. Based on this, she challenges the widespread assumption that the set of items subsumed by a concept can be determined by description.

Since language concerns the expression of thoughts, linguists have also taken theoretical positions concerning concepts. Jackendoff (1989) attempts to extend Chomsky's (1986) linguistic theory to concepts. In the process, he emphasises the importance of distinguishing "E-concepts" and "I-concepts", which concern the way the world is and how we grasp it, respectively. Lakoff (1987) claims that concepts are grounded in our bodily experience and emerge through our interaction with the world. For him, concepts are not innate, nor do they exist externally to the mind and body. As such, meaning must be mediated through our minds and bodies. Hurford (2007) is concerned with the cognitive underpinnings of the evolution of language. After comparing the conceptual capabilities of animals with ours, he proposes that human concepts were enhanced by language, but built on pre-existing and simpler "proto-concepts".

While psychologists are generally more interested in empirical evidence rather than abstract theory, they have also made theoretical contributions to the subject. Murphy (2002) treats concepts as mental representations of categories of things in the world. He uses evidence from categorisation experiments to evaluate competing psychological theories of how concepts are represented in the mind. Smith and Samuelson (1997) challenge the common view that concepts are fixed and unchanging representations. They favour a highly dynamic approach, where concepts are built on-line when needed out of memories and knowledge. Barsalou (1999) argues that cognition is grounded in perception, and that concepts are no exception. He proposes a theory of "perceptual symbols", in which perceptual representations in memory can serve many of the functions normally attributed to abstract concepts.

In short, theories of concepts diverge tremendously. A lot of work could be spent on assessing their mutual inter-compatibilities, but that would constitute a separate thesis. So what are we to make of this quagmire here? For this purpose, I recruit help from Fodor (1998).

2.2.2 *Fodor's (1998) conditions*

Although his own position is highly controversial (e.g., Landau 2000), Fodor (1998) also puts forward five weaker conditions which he claims are "not-negotiable" (p .23) for any sensible theory of concepts. I characterise my own position here with respect to his conditions. This will also highlight which aspects of concepts are crucial for my thesis, and which I can choose to remain agnostic about. In this way, I avoid the adoption or development of any particular full theory of concepts.

Fodor's (1998) first condition is that concepts are "mental particulars" (p. 23). From a psychological perspective, this is obvious and important: concepts exist in the mind.

However, it is worth emphasising because, as we have already seen in Section 1.1, concepts are often merged with meaning, and theories of meaning often bypass the mind altogether (see Section 2.4.3). Moreover, people often informally use the term "concept" as if it was an external entity which exists independently of the mind (e.g., "The concept of a fourth dimension is difficult to grasp").

Another condition is that concepts are categories and are used as such. People apply concepts to things in the world, and these things form a category. I agree that concepts are fundamentally used for this purpose, and indeed, this thesis will investigate concepts through categorisation experiments. However, although there is a direct causal link between concepts and categories, they are not the same thing. As I will argue in Section 2.4.1, while concepts exist in the head, and categories may be our best way of studying them, categories themselves are not in the mind. This is a philosophical point, and may in fact be consistent with Fodor's (1998) condition. In any case, it will be important for my purposes.

The next condition is compositionality (p. 25), which relates to the idea of concepts being the constituents of thoughts. Concepts can thus be combined to form more complex concepts by putting them together. This will not be important for my thesis, so I take an agnostic position. However, I do not take this condition as "non-negotiable". While many lexical concepts may be compositional, it is not a priori obvious that this is true of all concepts. In fact, even for lexical concepts, compositionality is not as simple a matter as it first appears, and can be highly sensitive to context (Sweetser 1999).

Fodor's (1998) fourth condition states a lot of concepts must be learned. As he points out, theorists differ widely in what "a lot" means, and his own theory is based on an unusually large base of innate concepts, but this cannot apply to all concepts. I take this condition as obvious. In fact, my work would even perhaps be compatible with the possibility that *all* concepts are learned. However, I do not need to commit to this position, which is fortunate, since it would get me into deep philosophical trouble.

Finally, Fodor (1998)'s last condition is that concepts are public. In particular, people can share the same concepts, and they do so. This point is a critical one for my thesis, and, like Hurford (2007), I must disagree. Concepts may be similar across different people, as is evidenced by how readily we communicate with language. But they are not necessarily identical, and at the very least, cannot be assumed to be uniform a priori. Since I am concerned with conceptual coordination, this point is so crucial that I take it up in more detail in Section 2.3.

In short, I generally agree with Fodor's (1998) characterisation of concepts, subject to some important qualifications. I agree that concepts are mental representations, they pick out categories of things in the world, and many, if not all, are learned. However, I do not take a position on compositionality, and strongly refute the notion that concepts are necessarily public and shared.

2.2.3 *Levels of representation*

Before continuing, there is one more important foundational issue concerning the nature of concepts to settle: the level(s) of conceptual representation of interest. As can be seen from the literature (including some of the positions listed above), there are many different levels of representation which researchers address, even if the boundaries between them are not always clear. At one end, neuroscientists have explored the neural basis for concepts, by studying areas of brain activation during conceptual processing, often involving neuropsychological patients with brain lesions (e.g., Damasio, Grabowski, Tranel, Hichwa and Damasio 1996; Martin 2007). Next, there is a long-standing debate between symbolic and connectionist views of mental representations (Rumelhart and McClelland 1986; Pinker and Prince 1987), although there are increasing efforts to reconcile them (e.g., Gärdenfors 2000; Dale and Spivey 2005; Dale, Dietrich and Chemero 2009). Further, cognitive psychologists study the structures of mental categories, from which they try to infer the mental representations of concepts (e.g., Rosch 1975; Nosofsky and Johansen 2000; Murphy and Kaplan 2000). And finally, philosophers and linguists often abstract away from the internal aspects of concepts, focusing on how they relate to the world and to language (e.g., Chomsky 1986; Peacocke 1992; Fodor 1998; Edwards 2010).

In this thesis, I do not deal with the two lower levels of representation, offering nothing about how concepts are actually physically represented, instantiated or distributed in the brain. Instead, I am attempting to walk the line between the higher two levels. While

I will focus primarily on the relationship between concepts, language and the world, I will also address the internal mental structure of concepts and the interface between the two.

2.3 Conceptual variation

Since the outset of this thesis, I have pointed out that its aims require me to take conceptual variation seriously. But is conceptual variation merely a theoretical possibility, or is it also an empirical fact? In this section, I briefly review some evidence and arguments that strongly suggest the latter, and do so at very different timescales: evolutionary, inter-cultural, intra-cultural, developmental, and online. Combined with theoretical considerations, I then complete the argument that the idealisation of concepts as shared and uniform must be dropped, at least for the purposes of my thesis.

At first glance, it may seem like a discussion of conceptual variation belongs later, once I have developed my theoretical perspective on conceptual relationships, structures, and processes. However, there are three reasons for placing it here. First, we will see at various points in this chapter that developing a theoretical model is strongly affected by a variationist view. Indeed, it is largely because most theories have not taken conceptual variation seriously that I cannot simply adopt a standard position. Second, due to the large variation in the timescales I will discuss, the evidence will draw from a range of different disciplines, including cognitive and linguistic anthropology, comparative and developmental psychology, and psycholinguistics. As such, there is even less consensus on the already controversial nature of concepts than there is within a single discipline. Third, this discussion is not meant to be comprehensive, or even to establish conceptual variation as a fact. The purpose is only to show that no matter what cultural and temporal granularity we look at, there is reason to believe that conceptual representations and content may not be fixed. If so, we must not presuppose that concepts are universal and static, and acknowledge this to be a largely empirical issue.

2.3.1 Variation at different timescales

I begin with the evolutionary timescale, and the differences between animal and human concepts. There is a lot of debate about the nature of animal concepts among philosophers, comparative psychologists and other researchers. Davidson (2004) claims that while animals may behave differently towards different classes of stimuli, that is quite distinct from treating them **as** members of those classes. Bermúdez (2003) argues that each species has its own conceptual ontology of the world, and that animals' concepts have significant limitations relative to our own. Hurford (2007) reviews empirical evidence and suggests that animals have "proto-concepts", which approach but are nevertheless still quite different from human concepts. Deacon (1997) claims that hominin evolution underwent a fundamental cognitive transformation, giving rise to a symbolic capacity in humans which never evolved in other primates. On the other hand, Savage-Rumbaugh, Sevcik, Brakke and Rumbaugh (1990) argue that bonobos have concepts very similar to ours, citing evidence from the language comprehension and production abilities of enculturated animals. Pepperberg (1999) makes similar arguments for an evolutionarily distant species, showing evidence that parrots can go beyond stimulus discrimination and form categorical classes. And Barsalou (2005) argues, on the basis of neuroimaging experiments with monkeys, for an evolutionary continuity between animal and human conceptual capacities. Despite the differences among such views, it appears that there are some kinds of evolutionary precursors to human concepts, however impoverished. Given the fact that there is "large-scale variation" in the human genome (Iafrate et al. 2004), then even if human concepts are largely innate (Fodor 1998), they are unlikely to be identical across the species.

An obvious way in which conceptual variation among humans could manifest itself is at the cultural level: do cultures differ in their concepts of things in the world? In certain domains, such as technology, it is fairly obvious that they do: members of isolated hunting and gathering cultures must surely think of airplanes passing overhead differently than members of modern, technologically developed cultures. In other, more universally relevant domains, the picture is more controversial, in both anthropology and linguistics. For instance, there is substantial anthropological evidence for cross-cultural differences in the folk taxonomies of domains such as plants, animals, kinship and body parts (e.g., Berlin 1992; Ellen 2006). At the same time, there appear to be significant universal patterns and principles in these classification schemes as well (e.g., Woolford 1984; Boster 2005). Similar debate exists amid linguistic investigations of variation in lexically encoded concepts in various domains, including colour (e.g., Regier and Kay 2009), space (e.g., Levinson 2003), body parts (e.g., Enfield, Majid and van Staden 2006) and artifacts (e.g., Malt, Sloman and Gennari 2003). Nevertheless, the controversies do not generally concern whether concepts vary between cultures, but rather the extent of the variation and explanations for it. An influential view on these issues that is still hotly debated and that will be an important locus in this thesis is that cultural differences exist in how the world is conceptually partitioned **and** that these differences are due to language (Whorf 1956).

Having looked across cultures, we can now inquire within cultures: what kind of conceptual differences exist between members of the same culture? However, this is a potentially circular problem, since similarity in concepts could be taken as evidence of similarity in culture. Indeed, this approach has been adopted in cognitive anthropology by A. Kimball Romney and colleagues, as can perhaps best be seen from the title of one of their papers, "Culture as shared cognitive representations" (Romney, Batchelder and Brazill 1995). They have developed analytic tools for measuring intracultural conceptual variation, and these techniques have been applied to verify conceptual coherence within cultures in domains of kinship, animals, emotions and colour (e.g., Romney et al. 1995; Romney, Boyd, Moore, Batchelder and Brazill 1996; Romney, Moore, Batchelder and Hsia 2000; Gravlee 2004). Experimental categorisation studies have also revealed that participants from the same culture tend to have similar typicality judgements and reaction times with respect to different category members (e.g., Rosch 1975; Rosch, Simpson and Miller 1976; Armstrong et al. 1983). However, Barsalou (1987) has criticised some of the statistical methods that have often been used in such studies, pointing out that they depend on sample size and yield misleadingly high values of inter-subject agreement. Indeed, substantial category variation has been documented in more recent studies using sorting tasks (Roberson, Davies, Corbett and Vandervyver 2005; Haslam et al. 2007). Also, conceptual differences have been documented within cultures when comparing laymen to experts in certain domains, such as medicine, trees and birds (Medin, Lynch, Coley and Atran 1997; Bailenson, Shum, Atran, Medin and Coley 2002). For example, experts (e.g., bird-watchers) tend to categorise at subordinate levels (e.g., "sparrow" rather than "bird"). While these cases do not imply that people necessarily have different concepts, they do suggest that a person's level of expertise can impact how they conceptualise something. However, it could be argued that in these cases, the differences are (sub)cultural. In general, there is little clear evidence for conceptual variation within cultures. This doesn't mean, of course, that such variation does not exist. What it does suggest is that such differences might not be very systematic. Indeed, the very fact that psychological experiments usually recruit numerous participants per condition reflects the implicit recognition that minor differences between participants are pervasive.

Until now I have discussed conceptual variation between individuals, but we can look within a person as well. Indeed, conceptual development is a major area of study in developmental psychology (Keil 1992; Nelson 1996; Carey 2004), and such study presupposes that the conceptual repertoire of a child changes over time. There are various controversial issues here, such as the preexisting basis for concept acquisition (e.g., Mandler 2004), whether conceptual systems undergo qualitative changes (Keil 1992), and what the relationship is between conceptual development and language acquisition (e.g., Bowerman and Levinson 2001). Nevertheless, whatever the process may be, even strong nativists acknowledge that some conceptual concepts have to be learned (Fodor 1998). Moreover, it is not only children that undergo conceptual change: we continue learning throughout our lives (otherwise, why have universities?). And this learning involves not only the acquisition of new concepts but also development and refinement of existing ones (Dawson-Tunik 2006). Indeed, the findings mentioned earlier that experts conceptualise differently than novices implicate adult conceptual development, since presumably the experts were once novices.

The final timescale I mention here is online conceptualisation. While the phenomena discussed in the previous paragraph could be the gradual result of years of enculturation, there is also evidence that concepts are created and changed dynamically very quickly on the fly. Children are able to learn novel words for novel objects with minimal exposure, and extend them consistently to other objects of the same type (e.g., Behrend, Scofield and Kleinknecht 2001). More generally, an enormous body of category learning experiments has studied not only that people (both adults and children) acquire concepts online, but also how they do so and the properties of the resulting concepts (see Murphy 2002 for a review). Barsalou (1983) showed that concepts may be more dynamic than we think, since the well-documented properties of concepts like birds and furniture are also found in “ad hoc” concepts like “things to take from a burning house” or “things to have on a picnic”. Finally, Lakoff and Johnson (1980) have catalogued culturally widespread systems of metaphors and shown how the “same” abstract concept can take on different forms in different situations (e.g., love can be conceptualised as war, a plant, a game, etc.).

To summarise, I have briefly presented evidence here from a variety of disciplines for conceptual variation and change at five different timescales. While it is more convincing for some timescales than for others, it reinforces the point that we must treat conceptual variation as a serious empirical issue. As we will see, this has profound theoretical implications.

2.3.2 Abandoning conceptual universals

Concepts seem to exhibit variation at several different timescales. Despite this, as we have seen, some theorists (e.g., Fodor 1998) explicitly insist that concepts are shared and identical across people (see Section 2.2.2). While these may be relatively few in number, we will see in the rest of the chapter how many other researchers implicitly assume similar things as well. The problem is that conceptual variation creates all sorts of theoretical complications, and is thus convenient to ignore. So the question is: are these researchers just being lazy in ignoring conceptual variation, or are they justified in doing so?

Well, it depends. If conceptual variation was rampant, so that there was no consistency between different people's concepts, then theoretical views would need to be modified. However, the evidence from the previous section does not conclusively point to such wide divergence: while conceptual variation does occur, it appears to be limited. It is possible that, to a reasonable degree of approximation, concepts are consistent across individuals (Segal 2000). Therefore, the importance of the variation may depend on one's research questions. For many purposes, an assumption of universal concepts may be justified and appropriate, since it greatly simplifies the theoretical picture. Nevertheless, as I have argued, that does not apply here. Concepts may vary, and this must be incorporated into my theoretical view. It will be important to stick loyally to this position at each step along the way, since it will be tempting to abandon it.

Given how much I am stressing this point, I should also qualify my variationist position. Indeed, my position on conceptual variation is not particularly strong. I am not claiming that people necessarily have different concepts, but only that I must allow for this possibility, and that I am probably wasting my experimental efforts if they don't. It is an empirical question whether and how much conceptual variation actually occurs.

On the other hand, my position goes further than merely rejecting a strong universalist position, such as Fodor's (1998) innate universal concepts. It also rejects the idealisation that there are necessarily universal concepts which people acquire during their lifetimes, whether gradually or suddenly. Hampton (1989) discusses the notion of "normative" concepts, which might be thought of as "the (imagined) end goal of our present scientific advance" (p. 40). Under such a view, with the advance of science, we get closer and closer to true concepts, which are thus reminiscent to Plato's abstract Ideas (Ross 1951). Similarly, the development of an individual concept could be viewed as a convergence in someone's mind towards the true concept. However, as Hampton points out, even if we were to advocate this view, any basis for classification, including science, has a purpose, and as such cannot be said to be objectively true and unique. Moreover, since human acts of categorisation are not performed in a vacuum but are embedded in a context and goal (Ratneshwar, Barsalou, Pechmann and Moore 2001), there is not necessarily just one unique and correct way of dividing up the world. And if people are only converging on concepts, then what do we call those mental entities before they reach them (if ever)?

Nevertheless, as mentioned in Section 1.3, allowing for conceptual variation greatly complicates the theoretical picture. One fundamental complication is that, in contrast to the normative view discussed above, we must now make a strict distinction between how the world is and how we think of it. Since concepts thereby explicitly concern what's in a particular person's mind, which may be different than what's in another's mind, we can no longer take concepts to be simply reflections of how the world really is. For example, we must now seriously appreciate the big difference between claiming that whales are fish and claiming that whales fit a particular person's concept of "fish". The former is a question for biologists, while the latter is one for psychologists. The biological question might have one objective answer, while the psychological one depends on the particular person under study.

It's important to note that it is not just philosophers that tend to collapse this distinction between things in the world and our ideas about them. In fact, even psychology often partly buys into the traditional philosophical view implicitly. For example, a large portion of empirical research on concepts consists of category learning experiments (Murphy 2002). In such studies, category structure and content is normally predetermined by the experimenter, and the participant's job is to learn the category. In other words, rather than studying what concepts people have already or what concepts they form spontaneously, these experiments study how well people learn objective external categories. As such, they frame concepts (at least the concepts that they are studying) as just mental representations of objectively existing categories. This is problematic when the focus is on conceptual variation, and as a result, the relationship between concepts and categories will be reevaluated in Section 2.4.1.

Similarly, philosophers, psychologists and linguists all tend to treat a linguistic label as a useful indicator of a concept. For example, the word "dog" is often taken to identify *the* concept of dogs, as if there were one such concept. In psychology experiments, this implicit assumption manifests itself when participants are asked to categorise stimuli by selecting from a set of labels (Lloyd-Jones and Humphreys 1997). If you and I give the same stimulus the same linguistic label (e.g., "dog"), then it seems natural to infer that we are categorising the stimulus in the same way. But there is a circular problem here, since this presupposes that my concept is identical to yours. If we allow for conceptual variation, then such assumptions must be avoided. The relationship between concepts and words will be revisited in Section 2.4.3.

A related problem concerns conceptual changes that occur within a single person. As a child interacts with dogs and learns more about them, is it a single concept that is being developed and changed, or is it actually being replaced or supplemented by a new concept? And similarly, if a person conceptualises the same thing in a slightly different way from one moment to the next, does that mean that different concepts are being used, or that the concept has changed, or that the concept is being modulated? Some researchers have proposed that concepts are not stable, static entities, but are much more dynamic in nature. Smith and Samuelson (1997) take a particularly strong position, reviewing evidence for online category variation and flexibility, and emphasising that acts of categorisation themselves modify the underlying mental structures. They conclude that a successful theory of categories must "give up timeless abstractions such as concepts" (p. 190). Similarly, Croft and Cruse (2004) argue that the notion of a fixed set of concepts is misguided, and that concepts are actually "created at the moment of use"

(p. 75). Although these positions might be overstating the case, we will see in Section 2.6 that conceptualisation is a highly dynamic process.

Therefore, as outlined in Section 1.3, the rejection of conceptual universals forces us to forgo reliance on objective external categories or on public linguistic labels for conceptual identification. Once we commit to conceptual relativism, we automatically inherit the problem of concept individuation (Fodor and Lepore 1992), which affects both theoretical interpretation and methodological plausibility. Unfortunately, I have no choice but to deal with it, and the rest of this chapter will be largely devoted to exploring the theoretical consequences of this position. However, as we will see by the end of Chapter 4, plausible solutions are available at both theoretical and methodological levels, as long as, given the inherent nature of the problem, we are willing to compromise a little.

Before going on, I offer a disclaimer that I will not be too pedantic in applying the consequences of my variationist position to terminology, unless necessary. In particular, while I am explicitly rejecting conceptual universals and the grounding of concepts in reliable external words or categories, constantly questioning what a plate is and what "plate" means at each step could lead to insanity.

2.4 Concepts, categories and words

In the last section, I emphasised that concepts may vary and pointed out that this has implications for the relationships between concepts, categories and words. Therefore, in this section, I take a closer look at these relationships, merging what is already known about them with a dedication to conceptual variation. I look at each of the three relationships in turn: concepts and categories, concepts and words, and categories and words.

Note that this section focuses on the structural relationships between these notions. How they are used in processing will be addressed in Section 2.6. However, the boundary between structure and processing is not so clean, because, as Barsalou (1990) points out, we cannot empirically study structure without relying on processing. Moreover, the distinction breaks down further if we adopt a dynamic view of structure (e.g., Smith and Samuelson 1997).

2.4.1 Concepts and categories

One of the biggest consequences of taking conceptual variation seriously is the need to fundamentally rethink the relationship between concepts and categories. However, in order to tackle this, we need some more artillery, and for that I turn to an important idea in the philosophy of language.

Frege (1892/1948) made an influential distinction between **sense** and **reference**. Reference concerns the object in the world that an expression refers to, while sense reveals our perspective or attitude to the object. Thus an expression "expresses its sense, refers to or designates its referent" (p. 214). The sense is the "mode of expression" of the reference. The point is that, when we use an expression, we are not just singling out an object, but also framing it in a particular way which reveals our conceptualisation of it. Frege's classic example is that of Venus, which could also be referred to with the expressions of "the morning star" or "the evening star". Both expressions have the same reference, but they identify different senses. More generally, a referent could potentially have any number of senses that single it out, but each sense **determines** its referent.

Frege's (1892/1948) view is not without controversy, and was subsequently attacked by other philosophers, such as Russell (1905) and Kripke (1972). However, the criticisms generally rely on the uniqueness of reference in proper names and certain other kinds of expressions, and are very much centred on a theory of linguistic meaning, not concepts. For instance, Kripke dispenses with sense by arguing that the meaning of a proper name is simply the referent that it is first applied to. As such, these arguments do not concern us here. Moreover, Frege's distinction has been revived and defended by philosophers more recently (Dummett 1981; Evans 1982).

So what does the distinction between sense and reference have to do with concepts? As Hurford (2007) points out, we may be tempted to interpret Frege's (1892/1948) sense as something mental, but "Frege himself might have winced at such an interpretation" (p. 118). Indeed, Frege also distinguished sense from "conception":

> The referent and sense of a sign are to be distinguished from the associated conception. If the referent of a sign is an object perceivable by the senses, my conception of it is an internal image, arising from memories of sense impressions which I have had and activities, both internal and external, which

> I have performed. Such a conception is often saturated with feeling; the clarity of its separate parts varies and oscillates. The same sense is not always connected, even in the same man, with the same conception. The conception is subjective. One man's conception is not that of another. There result, as a matter of course, a variety of differences in the conceptions associated with the same sense. A painter, a horseman, and a zoologist will probably connect different conceptions with the name "Bucephalus". This constitutes an essential distinction between the conception and the sign's sense, which may be the common property of many and therefore is not a part or a mode of the individual mind. For one can hardly deny that mankind has a common store of thoughts which is transmitted from one generation to another. (p. 212)

So, Frege's (1892/1948) conceptions, in contrast to sense or reference, are private, subjective, experiential, and variable. Frege claims we have both, which allows him to put aside people's private conceptions of things and to comfortably ground his theory of meaning in public senses. And due to his distinction, he can still also acknowledge that conceptions can vary between people, and even within people.

However, Johnson-Laird (1983) notes an important problem here from a psychologist's point of view: are senses in the mind or not? If they are not, then how can society pass them down from individual to individual? And if they are, then how are they different from conceptions, and how do they relate to them? It seems more parsimonious, if possible, to wed Frege's (1892/1948) sense and conception somehow, and a commitment to the privacy and variation of concepts would seem to demand it. To that end, why don't we jut throw out the idea of public senses and substitute private conceptions in their place? This would mean that now conceptions (rather than sense) determine reference. Frege himself wouldn't be happy about this, and neither would some contemporary philosophers (e.g., Fodor 1998), as it implies that the meanings of linguistic constituents are unreliable. On the other hand, this perspective is not inconsistent with the idea that initially variable concepts could become (relatively) uniformly conventionalised across a linguistic community. Indeed, some linguists have embraced such an approach in developing theories of linguistic meaning (e.g., Langacker 1987; Taylor 1995; Croft and Cruse 2004).

So where does all this leave concepts and categories? At the outset of his comprehensive review of psychological research on concepts, Murphy (2002) makes the following terminological distinction between concepts and categories:

> In general, I try to use the word *concepts* to talk about mental representations of classes of things, and *categories* to talk about the classes themselves. However, in both everyday speech and the literature in this field, it is often hard to keep track of which of these one is talking about, because the two go together. That is, whatever my concept is, there is a category of things that would be described by it. Thus, when talking about one, I am usually implying a corresponding statement about the other. Writers in this field often say things like "four-year-olds have a category of animals," meaning "four-year-olds have formed a concept that picks out the category of animals." However, being too fussy about saying *concept* and *category* leads to long-winded or repetitious prose (like my example) with little advantage in clarity. (p. 5, author's emphasis)

While these are reasonable working definitions, they are difficult to interpret theoretically. Murphy (2002) defines concepts in terms of categories, so that categories appear to be the more fundamental and independent units. But this raises the question of what is the ontological status of categories.

If categories are universal objective classes of things in the world, then concepts can just be private mental representations of these public classes. As such, any concept can be grounded in a corresponding mind-independent, objective category. On one hand, defining concepts in terms of categories suggests this position. On the other hand, this suffers from the philosophical problems discussed earlier, and indeed, Murphy (2002) doesn't seem to advocate this, since he speaks, for example, of four-year-olds having "a" (rather than "the") category of animals.

Alternatively, then, categories could be in the minds of individuals, in which case different people can have different categories. Thus, there would be a one-to-one correspondence between concepts and categories, but such correspondences would be within the mind of a particular person. However, while this would comfortably allow for conceptual variation between people, it also seems to equate concepts with categories (much as we just saw with Frege's 1892/1948 senses and conceptions): if both are in the mind and both are about things in the world, then what is the difference between them? And this would not be consistent with Murphy's (2002) terms, since he is making a distinction between concepts and categories.

Finally, categories may not exist in either the mind or the world. Categories could just be convenient constructs for psychologists to describe and study people's concepts. For example, when participants carry out categorisation tasks, they produce categories of stimuli which reflect their concepts. However, under this interpretation, it is actually categories that depend on concepts rather than the other way around, which would make Murphy's (2002) definition circular.

How do we resolve this issue, then? Here I return to two key ideas from Frege (1892/1948) that I discussed earlier. First, recall that Frege claimed that sense determines reference, but that, as Johnson-Laird (1983) pointed out, this sense should be construed as private and mental (like Frege's conception rather than sense). Second, Frege pointed out that sense overdetermines reference, such that there are different senses that you could attach to the same reference (as with the morning and evening star). Both of these ideas are relevant to the current dilemma concerning concepts and categories. Much like sense determines reference, concepts determine categories. This means that given a concept, we can study the category that it determines, rather than one that it reflects. It also means that having adopted the position that concepts can vary, we automatically need to allow for the possibility that categories vary as well. At the same time, concepts are more than just uniquely determined mental representations of a class of things, so that different concepts can determine the same category, providing different perspectives on it.

This view of the status of categories and their relationship to concepts fits the mould of the third possibility previously discussed. Categories are neither in the world nor in the mind; instead, they are analytic (and very useful) constructs which allow us to study people's concepts in concrete terms. This characterisation is presented diagrammatically in Figure 2.2. Note that although under this perspective categories do not have an independent real existence, this does not deny that they can still tell us something real about the world and its structure. However, it's important to emphasise that when concepts vary between people, categories vary as well. This diagram must therefore be interpreted as representing a relationship for a particular person, possibly even for a particular moment in time.

So, in contrast to Murphy (2002), I have flipped the direction of the causal relationship, and defined categories in terms of concepts. On the other hand, I have kept what I consider to be the more important parts of his characterisation: that concepts and categories can vary between individuals, and that concepts and categories are not the same thing.

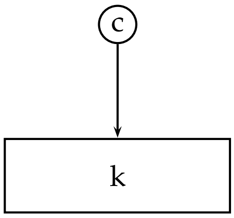


Figure 2.2: The relationship between a concept (c) and its corresponding category (k): the concept determines its category, so both concept and category are specific to an individual, although only the concept is actually in the mind. The category is represented with a box to emphasise that it's a (potentially large or infinite) theoretical set, rather than a particular item.

However, unlike Murphy (2002), I am now left without a clear basis for concepts. If we cannot define concepts in terms of the categories that correspond to them, then how are we to define them? This is a difficult philosophical problem, but fortunately is not critical for my purposes and I do not pursue it further. Instead, like most psychologists investigating concepts, I will, for the most part, rely on the close relationship between categories and concepts throughout this thesis, and will often use the terms interchangeably. However, sometimes the distinction will be crucial.

2.4.2 Categories and words

I have now established that concepts determine categories, and argued that the latter do not have an independent existence. Does that mean that words and categories do not have any direct relationship with each other?

In fact, linguistic and philosophical theories of meaning have traditionally taken the opposite approach. Far from leaving categories out of the semantic picture, they have grounded meaning itself in categories, and bypassed the mind instead. From this perspective, meaning is not concerned with conceptual structures and processes, but rather is defined directly in terms of objective truth and reference (Davidson 2001). In particular, words get meaning by virtue of the categories of things they refer to in the world, and the meanings of sentences are defined in terms of the sets of situations under which they would be true (e.g., Katz and Fodor 1963).

Putnam (1975) advocated a particularly strong version of such a mind-independent view of meaning, and defended it with an influential thought experiment. Imagine that there is another planet just like Earth, with people just like humans, except that the chemical composition of its counterpart for water is XYZ rather than H_20. In that case, even though the contents of our brains and those of our counterparts from this planet might be identical (at least at some earlier point in history), 'water' would refer to different things. Putnam concludes from this that meanings "just ain't in the head" (p. 227).

While such a position has come under considerable attack from philosophers and linguists (e.g., Lakoff 1987), it is particularly problematic from a psychological point of view, especially once we take conceptual variation seriously. Notice that if past humans were to go to this twin planet, they would also conceptualise the substance there in the same way as they did water on Earth. They could be "wrong" that it's the same thing, but that would not change their conceptualisation. Moreover, locating meaning in the world requires standardising it so that the category of water is effectively fixed across the population. But that is in direct conflict with the position of variation in concepts and categories that I have adopted.

Therefore, for my purposes, I need to reject the idea of meaning being grounded directly in categories, much as I abandoned universal concepts earlier. Since concepts can vary among individuals and categories are determined by concepts, word meaning must be conceptually mediated by the individuals. I therefore turn in the next section to the relationship between words and concepts.

However, it is important to point out that although words and categories do not have any direct cognitive relationship, they do manifest themselves together in the world whenever we actively use language to describe or refer to things immediately around us. Every time someone points at a book and says "Pass me that book", a word is co-occurring with a member of some category. Moreover, this does not always occur with one object in isolation, nor is the word's appearance always as ephemeral as it is in speech. For instance, when we go to the supermarket and go to a section labelled "Cereal", we will find different kinds of cereal there. However, this is only because there were people who put them together, and they could only do so with the intervention of their concepts. Moreover the selection that we find in the cereal section varies slightly from store to store and country to country, reflecting the fact that concepts vary, though not rampantly.

The point is thus twofold. On one hand, the connection between words and categories has no direct cognitive reality, since it is about phenomena that are outside of the mind. As such, any relationship between them only exists by virtue of being mediated through the mind. On the other hand, for the same reason, we can sometimes directly observe co-occurring manifestations of the words and categories of others. Since we have no direct access to people's concepts, this may be the best way for us to infer their concepts indirectly. Such considerations will be important for my thesis, especially once we begin to look at conceptualisation in a more social context (see Chapter 3).

2.4.3 Concepts and words

The relationship between concepts and words is a large topic that plays a central role in my thesis. To keep it manageable, I have distributed my discussion of different aspects of the relationship around relevant parts of the thesis, so that this section itself is relatively limited in scope. Here I am only concerned with the private, structural relationships between words and concepts in an individual's mind. The role of words online in the private process of conceptualisation is covered in Section 2.6.5, and effects due to the public nature of words is the main theme of Chapter 3.

In order to address the structural relationship between words and concepts properly, we need to first break words down into their basic constituents. As is captured in de Saussure's (1916/1983) basic model of the linguistic sign, a word is a psychological association between a (phonological) word form and a (semantic) word meaning. These need to be distinguished from their corresponding material manifestations, which are the acoustic sound form and (often) the physical referent in the world, respectively. But how do word forms and word meanings relate to concepts? In the rest of this section, I will explore the relationships between these three mental entities.

2.4.3.1 Concepts and word meanings

In exploring the relationship between concepts and word meanings, we must keep in mind the distinction drawn between concepts and categories raised earlier (see Section 2.4.1). By concepts, I do not mean categories of things in the world, but the mental forms that they correspond to. As such, we are trying to relate here two potentially different mental representations, "semantic" and "conceptual" (Levinson 1997). But what is the difference?

In fact, some theorists, especially those of a universalist persuasion, argue against making such a distinction. Fodor (1975) takes this position, with meanings being conflated with concepts in an underlying and universal language of thought. Jackendoff (1983) is quite explicit, stating that "The terms *semantic structure* and *conceptual structure* denote the same level of representation" (p. 95). If we adopt such a view, then the relationship between words and concepts is simplified, with concepts simply being the meanings of words.

On the other hand, some theorists have argued that this is an oversimplification, insisting on a distinction between semantic and conceptual representations. For instance, Levinson (2003) argues that the vocabularies of languages, being finite and learnable, are necessarily more underspecified than concepts. For example, languages have lexical gaps, without implying that their speakers lack certain concepts. Moreover, Levinson argues that, unlike conceptual representations, semantic ones are supplemented by pragmatics and context during usage to obtain meaning. Bierwisch and Schreuder (1991) offer similar arguments from a psycholinguistic point of view. In developing a version of Levelt's (1989) general theory of lexical production, in which the first step is conceptualisation, they argue that semantic representations are more limited than conceptual ones. In their model, the two forms of representation play distinctly different roles.

Nevertheless, although such arguments insist on a semantic-conceptual distinction, they also maintain a close correspondence between the two representations. Levinson (2003) argues that they are "rather close", with conceptual representations necessarily supporting semantic ones (p. 296), while for Bierwisch and Schreuder (1991) the semantic representation is much like a substructure of a corresponding conceptual one, only "somewhat more flexible"(p. 33). Therefore, unless our goal was to identify different stages of language production, for example, we lose very little by collapsing the distinction.

But what about universalism? At first glance, it seems that equating conceptual and semantic structure means embracing a universalist position, so that not only concepts are universal, but so are word meanings. However, that is only the case if conceptual structure is universal, which I have explicitly rejected. If concepts can vary between people, then so can word meanings, regardless of whether we collapse the distinction between them or not. This is a mirror image of the collapse we encountered in Section 2.4.1, where keeping both the notions of sense and conception from Frege (1892/1948) turned out unnecessary and confusing.

Therefore, since conceptual representations are taken to be the more independent and broad, I will subsume semantic representations within them and will not discuss the latter further. This is also consistent with Murphy (2002), who, while pointing out that psychologists have not addressed word meaning in much depth, argues that it must be concept-based. I will henceforth follow this view. Moreover, I will use the term "word" to refer to word forms specifically.

Collapsing the distinction between concepts and word meanings has another clean advantage beyond simplicity. In particular, it means that we can automatically recruit linguistic perspectives concerning the relationship between word forms and meanings into a psychological understanding of the relationship between word forms and concepts. I therefore turn to this topic next.

However, before ending this section, it is worth pointing out that claiming that word meanings are concepts is not the same as claiming that word meanings and concepts are synonymous. As I will discuss in Section 2.4.3.3, we can have concepts that aren't closely associated with words, and even words that are not associated with well-developed concepts (Bermúdez 2003; Hurford 2007). Therefore, eliminating the distinction does not make the two terms interchangeable. I will use "concept" in general, but will sometimes resort to "word meaning" when explicitly dealing with the concept associated with a particular word.

2.4.3.2 *Word forms and word meanings*

Semantic relations between word forms and word meanings are an elementary and well-established topic in linguistics. The pervasiveness of these relations shows that, even if there were a universal language (Fodor 1975), word forms and meanings do not sit in a simple one-to-one correspondence. However, the picture gets significantly more complicated still when we throw away the idealisation of fixed connections between words and concepts across speakers of a language. In this section, I consider the implications of taking conceptual variation seriously for a couple of basic semantic relations, and converge on a dynamic view of word-concept associations.

One of the most basic semantic relations is synonymy, whereby two words have the same meaning. However, the existence of absolute synonyms with identical meanings is disputed. For instance, "woods" and "forest" could be synonym candidates, but they do differ subtly in terms of size, wildness, and proximity to civilisation, so they are not entirely equivalent (Room 1981). According to Clark's (1987) Principle of Contrast, different forms are always associated with different meanings. Such a meaning contrast can correspond to different categories, but it can just as well be due to stylistic aspects, such as dialect, register or emotional connotation (in Frege's 1892/1948 terms, the contrast may be in reference, or just in sense). Clark claims that such differences can always be found if we look hard enough: "While two terms may be interchangeable in many contexts, they are not so in all, and it is the contexts where they are not equivalent that reveal their often subtle contrasts in meaning" (p. 4). Similarly, Cruse (1986) points out that the meanings of words are constantly changing and adapting, and that "natural languages abhor absolute synonyms just as nature abhors a vacuum" (p. 270).

The issue of whether two words are synonymous or merely nearly synonymous may be a practically intractable one in general, and I do not pursue it further here. However, it does highlight the fact that the meanings of words can substantially overlap without being equivalent. This mirrors the earlier discussion of conceptual variation, and how people's concepts may differ slightly from each other (see Section 2.3). Indeed, considering these issues together raises the question of which differences will tend to be larger. Is there a larger difference between your meaning of "woods" and "forest", or between your meaning of "forest" and mine? In general, this surely will depend on the words, people and contexts involved.

The flip side of synonymy is polysemy, whereby one word has multiple related meanings. For instance, "book" may refer to either a single physical object (e.g., "Pass me that red book") or an abstract cultural product (e.g., "Eco's new book is amazing"). Alternatively, a word may be homonymous, having multiple unrelated meanings (e.g., "crane"), or monosemous, with just one meaning (e.g., "cat"). Importantly, as with synonymy, there may be subtle inter-personal differences in the range of meanings covered by a single word. For instance, my meaning of "book" may be quite similar to yours, but may be a little broader, including things like pamphlets and brochures.

It's important to note that the lines drawn between polysemy, homonymy and monosemy are often blurry, and indeed, one could easily quibble with the examples above as well. Theorists debate about what is the best way a given word's meaning should be analysed or even whether the distinctions above are coherent (e.g., Ravin and Leacock 2000). Moreover, the picture gets further complicated when we consider the role of context in word meaning. Geeraerts (1993) discusses the relation of polysemy to the notion of vagueness. While a polysemous word locates variation in meaning in the lexical item's intrinsic semantics, a word is vague if its meaning (possibly a single meaning) needs to be pragmatically supplemented by context. Geeraerts discusses several tests for distinguishing the two phenomena, but shows that they often give contradictory results. He concludes that "the instability of the distinction between vagueness and polysemy precludes a strict dichotomy between intercategorial and intra-categorial semantic multiplicity" (p. 258). In short, there may be no principled way to draw a fixed box around the meanings of a word. Geeraerts captures this effectively with an alternative theoretical view of word meaning:

> The tremendous flexibility that we observe in lexical semantics suggests a procedural (or perhaps 'processual') rather than a reified conception of meaning; instead of meanings as things, meaning as a process of sense creation would seem to be the primary focus of attention. The image I would like to propose to make this conception more graspable, is that of a floodlight: words are search-lights that highlight, upon each application, a particular subfield of their domain of application. If this domain of application is seen, in model-theoretical fashion, as a set, each time the word is used a particular subset is selected. But while our traditional view of the distinction between vagueness and polysemy entails that the number of subsets that can be lit is fixed and restricted, we now have evidence that the verbal searchlight has much more freedom. The freedom is not absolute, surely, and there will be preferential subsets for each word; even so, the distinction between what can and what cannot be lit up at the same time is not stable. (p. 259)

Notice that Geeraerts's (1993) dynamic view of word meaning can be related with Clark's (1987) Principle of Contrast. If word meanings are highly context-dependent, then under some situations, two near-synonyms may be associated with the same (or nearly the same) range of meanings, while in others, they may diverge. This possibility is visualised in Figure 2.3. Again, to the extent that there is conceptual variation between people, these relationships may also differ. Indeed, Geeraerts's dynamic view may be just what is needed sometimes for people to come to understand each other if they start off with sufficiently different word meanings.

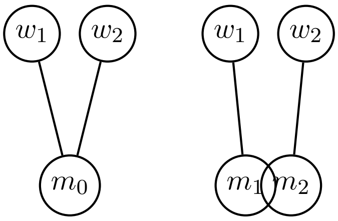


Figure 2.3: Context-dependent view of near-synonyms: under one context (left), two words (w_1,w_2) might have the same meaning (m_0), while in another context (right), they might have different (though overlapping) meanings (m_1,m_2).

2.4.3.3 *Concepts and word forms*

I return now to psychological research concerning concepts and word forms. A close relationship between conceptual and lexical knowledge is supported by neuropsychological experiments. Warrington (1975) studied three patients with impairments in object and word identification. The patients performed poorly on a series of tasks including word recognition, object recognition, and picture-word matching tasks. Moreover, performance was normal on general intelligence and perception tests, suggesting a specific deficit. Hodges, Patterson, Oxbury and Funnell (1992) found similar results with five additional patients. Tests demonstrated impairments with conceptual knowledge and a severe loss of vocabulary. In contrast, other aspects of language, including syntactic and phonological processing, appeared to be preserved. Following Snowden, Goulding and Neary (1989), Hodges et al. called this condition "semantic dementia" (p. 1798). Nakamura, Nakanishi, Hamanaka, Nakaaki and Yoshida (2000) extended these findings through semantic priming experiments. Semantic dementia patients, Alzheimer's patients, and control participants were first primed with a word from one semantic category and then carried out a timed lexical decision task on another word. For all the normal participants and Alzheimer's patients, decision speeds were faster when the word was of the same semantic category as the prime word, whereas no such effect was found with any of the semantic dementia patients.

However, while such studies suggest a tight bond between word forms and concepts, there is also neuropsychological evidence of their decoupling. Indeed, Warrington (1975) had noted that the correspondence between conceptual and verbal deficits was far from perfect. For example, one of her patients had performed relatively well on visual recognition of objects, while another had relatively good performance on verbal recognition. Kay and Ellis (1987) studied one patient in more detail with a more marked contrast in this respect. Their patient didn't seem to have any conceptual difficulties, but was impaired in naming objects, although he was often able to partially access the phonological form of the word. This effect is reminiscent of the well-documented "tip-of-the-tongue" phenomenon, in which normal language users have difficulty accessing the full form of a word they are trying to mentally access (Brown and McNeill 1966). These effects seem to stem from a weak link between word form and concept for rare words, perhaps due to infrequent use and aging (Burke, MacKay, Worthley and Wade 1991). There is also some evidence for dissociations in the opposite direction: impaired cognitive functions with relatively preserved language abilities. Bellugi, Marks, Bihrle and Sabo (1988) found that children with Williams Syndrome had severe deficits in visuospatial processing and reasoning, while maintaining relatively sophisticated linguistic abilities. This could be seen, for instance, in drawing tasks: when asked to draw a bicycle, they could produce the bicycle's different components and label them, but the parts were disjoint and did not integrate into a coherent whole. While subsequent research has detected unusual developmental trajectories concerning certain other aspects of language, most studies that have focused on vocabulary and word fluency in Williams Syndrome patients have confirmed a conceptual (without lexical) deficiency (Martens, Wilson and Reutens 2008).

Therefore, it seems that although word forms and concepts may be intimately related, they are not inextricably locked together. This is consistent with my conclusion in the previous section that word forms and meanings are not in an isomorphic 1-to-1 relationship. Indeed, it is also easy to construct intuitive examples showing how words and concepts appear to be independent of each other, especially if we focus on the relations in the mind of a single person, rather than in the language or population as a whole. For instance, when I go into a bicycle shop to have my bicycle repaired, I often indicate which part is broken by pointing, because I do not have the word for it. Of course, there may or may not be a standard word for it among bike technicians, but that doesn't change the fact that I personally have word-less concepts for these things. More generally, developmental and comparative psychologists have shown that animals and human infants have plenty of concepts, even if they are not as rich as those of human adults (Bermúdez 2003). Conversely, if I flipped through a good English-Thai dictionary, I would surely find some rare English words that I don't know. I may remember some of these words and recognise them as English words in the future, even though I may continue to have no clue what they mean (having absolutely no knowledge of Thai). Again, other people may know these words, but that doesn't change my mental knowledge.

These dissociations between words and concepts should not be taken too far, however. Although we may not have words for some concepts, we are still very capable of expressing them with longer phrases or even coining new words or expressions on the fly (Baayen 1994). Therefore, not having a word for a concept does not prevent lexicalisation of the concept. Conversely, even if we don't know what a word means, storing it in our minds may make a "place" for an associated concept. Indeed, the dependence of semantic on conceptual representations discussed earlier would seem to imply this. Consistent with this, some studies with prelinguistic infants have suggested that words act as "invitations to form categories" (e.g., Waxman and Markow 1995). Moreover, as I mentioned in Section 2.3, acquiring a concept is not necessarily an instantaneous process, and studies of conceptual development often examine how a child's concept gets enriched over time, well after they first learn a word for it (Gopnik and Meltzoff 1997). Thus, we can envision the concepts that are associated with words to lie on a continuum from impoverished placeholders to rich representations.

Putting it together, then, what is the relationship between a concept and words? Figure 2.4 illustrates a concept as not being cleanly associated with any particular word in a 1-to-1 relationship. Rather, a concept can be linked with different words to varying degrees of strength, which reflect their average relative frequency of application for that concept across a range of contexts. Therefore, although during speech we make discrete lexical choices (see Section 2.6.5), that does not imply that the connection is static in general.

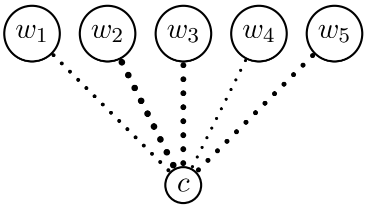


Figure 2.4: The relationship between a particular concept (c) and words (w_1, w_2, w_3, w_4, w_5): Multiple words can be associated with the same concept, with varying degrees of strength (indicated with the relative thickness of the lines). Lines are dotted to indicate that the relationship is not fixed (in contrast to that between concepts and categories). Of course, a polysemous word also maps onto multiple concepts, but that is not the focus here.

2.4.4 Conclusion

To summarise, the relationship between words and concepts is far from simple, being simultaneously quite tight and highly flexible. Words can have multiple meanings, but the meanings are difficult to individuate and may not be stable across time, people, and contexts. Different words map onto different sets of meanings, although in some contexts, they may theoretically coincide. To the extent that word meaning is determined by mental representations in the individual's mind, these representations can, for my purposes, be subsumed under concepts. However, there is good psychological evidence that words and concepts are partially dissociated. Moreover, the relationship is not symmetric: words seem to depend on concepts more than vice versa.

At this point, it is useful to relate these conclusions to the previous sections, in which I discussed the relationships between concepts and categories and between categories and words, respectively. First, I argued for the necessity of distinguishing concepts and categories. Categories do not exist independently of concepts: in fact, concepts determine categories. This also implies that since concepts can vary from person to person, categories can vary as well. Second, I pointed out that influential semantic theories in philosophy and linguistics attempt to bypass the human mind altogether, linking words and categories directly. However, if we are loyal to conceptual variation, this cannot be maintained, even though language use involves co-occurring manifestations of the two.

How do these three relationships fit together then? Recall that the relationship between concepts and categories was captured in Figure 2.2, and that between words and concepts in Figure 2.4. Moreover, as I argued in Section 2.4.2, there is no direct psychological connection between categories and words, so that their indirect relation needs to be mediated by concepts. Figure 2.5 synthesises these three conclusions, effectively merging Figures 2.2 and 2.4, and placing concepts at the centre.[1]

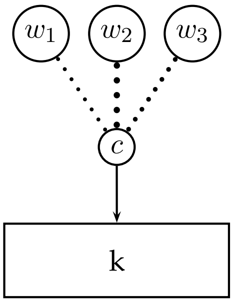


Figure 2.5: The relationship between an individual's concept (c), its category (k) and its associated words (w_1, w_2, w_3).

Notice that this diagram highlights a difference in the structural relationships that concepts take part in. On one hand, concepts and categories are tightly bound in a deterministic causal relationship. On the other hand, comparatively speaking at least, the relationship between concepts and words is relatively loose and malleable. This assymetry reasserts the close bond between concepts and categories, which will be a recurring theme in my thesis, coming up in theoretical discussions (as in Section 2.5), methodological decisions (see Section 4.2), and experimental design and interpretation (especially Experiment 4).

2.5 The internal structure of concepts

So far I have treated concepts as black boxes with no internal structure. However, over the last forty years, psychologists have extensively explored how concepts are represented in the mind. This issue is not as central to my thesis as the relationship between words, concepts, and categories, but it will still be important for grounding the interpretation of my experiments, and Experiment 4 in particular. As a result, in this section, I

[1] The triadic relationship between words, concepts and categories has also been an important focus for semioticians (most famously perhaps, Peirce 1932). Indeed, if a direct connection were made between words and categories, and each concept were associated with only one word, Figure 2.5 would effectively turn into the semiotic triangle (Ogden and Richards 1923).

overview psychological theories of concepts, beginning with the main theories, and then discussing more recent attempts at merging them.

Before discussing the internal structure of concepts, it's important to touch on an important ambiguity. Given the emphasis I have placed on a distinction between concepts and categories (see Section 2.4.1), one might sensibly first ask which of the two is the focus of psychological theories. At first glance, psychologists cannot look directly at people's concepts, but rather conduct experiments in which participants produce or learn categories. Indeed, in her ground-breaking categorisation studies on prototypes (see below), Eleanor Rosch was sometimes reluctant to draw conclusions about concepts themselves: "prototypes themselves do not constitute any particular model of processes, representations, or learning" (e.g., Rosch 1978, p. 40). However, although it is difficult to draw strong conclusions about conceptual representations from categorisation studies, that is what such studies are usually about. Psychologists do not generally care about categories in themselves, but rather what can be inferred about people's concepts (Malt 2006). Indeed, Rosch herself was less careful on other occasions, as in the title of a related paper, "Cognitive representations of semantic categories" (Rosch 1975). I too will take such a perspective, so that, Rosch's caution notwithstanding, I will assume that the structure of "categories" from psychology categorisation studies are actually the structures of mental concepts which determine those categories. Nevertheless, I will follow the field in this section and use the terms relatively interchangeably.

2.5.1 *Prototypes, exemplars and theories*

Until relatively recently, Aristotle's classical view (Ackrill 1963) was ubiquitous. Concepts had the form of definitions, with clear necessary and sufficient conditions. Category membership was a clear black-and-white matter, and boundaries between categories were sharp. This view had many advocates for millenia, including Frege (1903/2004).

However, Wittgenstein (1953) argued that our categories in general can't be given definitions, because we cannot generally come up with necessary and sufficient conditions for what exemplifies a particular category. He gives the example of a game, pointing out that it is extremely difficult, if not impossible, to come up with some specific attribute that all games have in common: some games are played in groups, others alone; some involve physical exercise, others not; some are competitive, while some are cooperative; and the list goes on.

And indeed, in the 1970's, Eleanor Rosch "essentially killed the classical view" (Murphy 2002, p. 16) with prototype theory (Rosch 1978). In contrast to the classical view, categories are characterised by their clear cases rather than their boundaries. Rosch's view is that not all members of a category are equal, with some being more central, more prototypical members than others. The degree of membership of other members depends on their "family resemblance" to the prototype, which is based on their degree of similarity in relevant features. For instance, a robin might be a prototypical bird, while an ostrich, although still a bird, would be more peripheral. Many studies by Rosch and colleagues involving natural categories found evidence for prototypes (e.g., Rosch and Mervis 1975; Rosch, Mervis, Gray, Johnson and Boyes-Braem 1976; Rosch 1978; Mervis and Rosch 1981). Importantly, as Murphy (2002) points out, prototypes are generally interpreted not as actual best examples stored in memory, but as "summary representations" (p. 42), which specify, for example, a weighted set of features. Some researchers adopted this view explicitly, and elaborated such abstract versions of prototypes (e.g., Hampton 1979). Despite little theoretical development over the last 30 years, prototype theory is still very influential and has taken central stage in some linguistic theories of meaning (Lakoff 1987; Taylor 1995; Croft and Cruse 2004).

An alternative view is offered by exemplar theory (Medin and Schaffer 1978). The main idea is that concepts are represented in terms of actual remembered instances. As with prototype theory, similarity still plays a central role, but this time items to be categorised are compared to many previously seen exemplars rather than a single prototype. Thus, one's bird category consists of all the birds one has previously encountered. There is some ambiguity, however, as to what counts as an exemplar. Does my category of dogs consist of all the dogs I have seen, or is each dog encounter counted separately? And how about dog encounters that I mentally imagine or simulate (Barsalou 1999)? These issues notwithstanding, many studies, especially those using artificial and clearly manipulable stimuli, have supported exemplar theory (e.g., Medin and Schaffer 1978; Medin and Schwanenflugel 1981; Ross, Perkins and Tenpenny 1990; Allen and Brooks 1991; Nosofsky and Johansen 2000). Moreover, exemplar theory has been implemented in several incarnations in various rigorous computational models (e.g., Nosofsky, Palmeri and McKinley 1994; Nosofsky and Palmeri 1997), although the behaviour of these models depends strongly on the rule used to calculate similarity (Murphy 2002).

The other major player in psychological theories of concepts is theory theory (Murphy and Medin 1985). This view emphasises that concepts cannot be looked at in isolation, but are deeply embedded in our general knowledge and understanding of the world. Concepts are interrelated with each other, so that learning new concepts happens against a rich existing backdrop of existing ones and can also in turn affect them. Moreover, according to theory theory, such factors are more important and more explanatorily powerful than the notion of similarity (which takes central stage for both prototype and exemplar theories). Our knowledge of birds, then, is not cleanly separate from many other things that relate to them, such as their environment, prey and predators, and general biology; and identifying a bird is not just a matter of comparing its features to previously seen birds. Advocates of theory theory emphasise that many psychological studies, in their attempts at isolating concepts, yield misleading results and implications for real everyday categorisation. And indeed, many empirical investigations which include more background knowledge and coherent inter-category relationships have found support for this view (Murphy and Medin 1985; Pazzani 1991; Heit 1998; Murphy and Kaplan 2000; Palmeri and Blalock 2000; Harris, Murphy and Rehder 2008).

What is the state of our current understanding then? Barsalou (1990) cautions against attempting to settle the debate, at least with regards to prototype and exemplar theories. He points out that empirical evaluations often unfairly characterise one of the theories, and argues that the sophisticated prototype and exemplar models actually carry the same information as each other, so that they cannot in principle be distinguished. Nevertheless, Murphy (2002) reviews a huge body of experimental studies and attempts to evaluate the theories in terms of the findings. His analysis highlights that no one existing theory can account for all the results. He concludes that, in general, theory theory and prototype theory stand up best to the evidence across a wide range of areas. This is particularly true for higher-level aspects, such as the hierarchical structuring of concepts, conceptual combination, induction and word meaning. Exemplar theory, on the other hand, while not accounting for many of these phenomena as well, does have a clear edge which must be acknowledged in category learning experiments. As a result, Murphy cautiously proposes that a complete theory of concepts would need to incorporate elements of all three theories. While the relative importance of these elements is still unclear, it seems that our minds manage and integrate abstracted prototypes, stores of exemplars, and a web of knowledge, and that all of these play central roles in conceptual processes. As such, all three are cognitively real and should find a place in a complete model of conceptual structure.

2.5.2 *Hybrid views*

Given this state of affairs, the last decade or so has seen increasing attempts at developing and testing hybrid theories of concepts. This work is based largely on experiments which have revealed how different theories can complement each other in explaining different aspects of categorisation results.

One solution to the problem is that the type of representation that is used may depend on the type of concept. And this seems to make intuitive sense: we might imagine that our concepts of dogs, height and π could have very different kinds of representations. Along these lines, Atran (1989) claims that there are different types of concepts, so that trying to find some universal method of representation is a lost cause. For instance, he argues that artifact concepts fit the prototype mould better, while natural kinds are better handled by causal theories. Medin, Lynch and Solomon (2000) lay out some considerations for distinguishing types of concepts, based on structural, processing and content criteria. While they consider it premature to draw concrete conclusions, they point out that the evidence does suggest that there are different kinds of concepts, and at the very least, that our theories of concepts should be open to making such theoretical distinctions. Machery (2005) is less conservative, and argues that concepts themselves do not constitute a natural kind. In particular, prototype, exemplar, and theory theories have very little in common, and yet there is solid psychological evidence for each in different domains. Therefore, Machery concludes that “Some concepts are prototypes, some concepts are exemplars, some concepts are theories” (p. 465).

Other researchers have proposed various hybrid models of categorisation theories, often based on direct experimental evidence. Smith and Minda (1998) compared prototype and exemplar models during various stages of category learning. Their starting point was the observation that prior comparisons of prototype and exemplar views in category learning have generally supported exemplar theory, but that this was based largely on experiments which used small, poorly differentiated categories and focused only on performance on the final tasks. As a result, Smith and Minda conducted experiments in which they manipulated the structure of the categories and evaluated different learning models at several stages. As in previous work, the exemplar model dominated for the smaller, less differentiated categories, and at later stages in learning. In contrast, when larger, more differentiated categories were being learned, participants' performance was much more in line with a prototype model. However, when a mixture model was applied (borrowed from Medin, Dewey and Murphy 1983), which incorporated both exemplar and prototype elements, both learning trajectories were satisfied.

Rigorously comparing models and developing hybrids is more difficult when theory theory is also involved. This is largely because while prototype and exemplar theories share substantial common ground in their heavy reliance on similarity, theory theory is more abstract and based on rules and causality. However, a variety of work has suggested that these views are broadly complementary. Sloman (1996) broadly reviews a range of work in the hotly debated topic of rule-based versus similarity-based systems in human reasoning. He argues that both systems are needed if we are to explain all the empirical evidence on different aspects of human behaviour. With regards to conceptual structure in particular, he concludes that the two systems play important, complementary roles. Nosofsky et al. (1994) presented a formal computational "rule-plus-exception" model, in which categorisation is done via a decision tree in which nodes can be either logical rules or checks for exceptional and previously seen exemplars. The model was successfully fitted to a wide range of empirical results, and accounted for many different attested categorisation phenomena. Whittlesea, Brooks and Westcott (1994) set up a range of experiments to study how participants might rely on different levels of knowledge depending on the task and context. They found that when participants carried out a classification task, they tended to use general conceptual knowledge, while when they carried out a recognition task, they relied on item-specific knowledge. Moreover, relative reliance on the two kinds of knowledge was further modulated by a variety of other factors, such as the format of stimulus presentation, the demands of other experimental tasks, social conventions, and the familiarity of cues. Juslin, Jones, Olsson and Winman (2003) conducted two category learning experiments which compared an exemplar model versus a cue abstraction model. The latter model was one in which several property relations are mentally integrated; this has similarities but remains distinct from a prototype model. In the experiments, the quality of feedback was manipulated: it either just identified which category a stimulus belonged to (poor), or also indicated how much of a critical and invisible property the stimulus had (rich). The results showed that participants' performance fit the exemplar model when there was poor feedback, but the cue abstraction model when there was rich feedback. Moreover, these results were modulated when there was time pressure in the tasks. Weiskopf (2009) surveys the body of available psychological evidence and argues that a given category is associated with multiple concepts, each with a different kind of representation. According to this view, for any given category our mind may manage prototypes, sets of exemplars, causal dependencies, words, etc, and these should be treated as distinct concepts. Weiskopf distinguishes such a pluralistic theory of concepts from hybrid theories: in the former, different concepts of cats become activated in different situations, while in the latter, a single complex concept of cat is always activated, even if a particular component is highlighted. However, it seems to me that this issue is an empirical one, and until it is resolved, it is more natural and conventional to use the term "concept" in the hybrid sense.

What are we to make of such attempts at unifying theoretical representations of concepts? Murphy (2002) points out that although no single monolithic theory is sufficient to account for all the known psychological evidence, we must also be suspicious of hybrid models because they are less parsimonious and may be too powerful. However, it is not my goal here to evaluate specific models. Instead, what will be important is that the inner structure of concepts seems to include different forms of representation, which are potentially recruited in different situations and contexts. And these representations can be thought to occur at different levels, possibly along a spectrum. At the bottom, there are exemplars, which encode individual members of categories. Next, we have prototypes, which abstract away from individual members and describe typical features of a concept. Finally, there are theories, which characterise the concepts in still more abstract terms, and relate them to other concepts and knowledge.

2.5.3 Conclusion

In line with the growing consensus that no monolithic theory of concepts will do, I too am adopting a hybrid approach. However, in this thesis, I will not generally have to distinguish between prototype theory and theory theory. Instead, I distinguish only two levels of conceptual representation, drawing a line between a concept's set of exemplars and what I will call its **sense**. The resulting relationship is shown in Figure 2.6. While exemplars are the representations of individual category members, a concept's sense contains the knowledge that is general to it. Thus senses can be seen as encapsulating prototype and theory theories, while exemplars derive straightforwardly from exemplar theory. However, I recognise that this terminology can also be misleading, in light of my earlier discussion of Frege (see Section 2.4.1). For Frege (1892/1948), senses were fixed, public symbols, and determined reference on their own. In contrast, I am proposing that it is the combination of a private concept's sense and exemplars that determines its category.

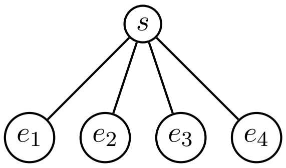


Figure 2.6: The internal structure of a concept, with a sense (s) and a set of four exemplars (e_1, e_2, e_3, e_4). The concept's sense derives from a combination of its prototype and relevant background knowledge.

We can now incorporate internal conceptual structure into the relationship between a concept, its category and words (see Figure 2.5), as shown in Figure 2.7. Since it is the concept as a whole that determines its category, I have causally connected both the sense and the exemplars to the category. On the other hand, since words are general to a concept, and word meaning is more compatible with prototype and theory theories, I have linked the words directly to the concept's sense. Note that although this schematic treats all exemplars of a concept equally, this may be cognitively inaccurate (as suggested by prototype effects, e.g., Rosch 1975) and is done only for visual simplification.

When I develop these relationships and how they are involved in conceptual processing later in this thesis, I will generally work with the simplified version of the conceptual

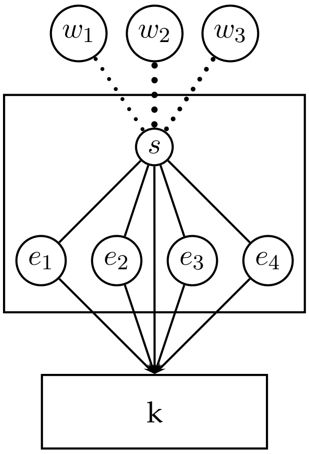


Figure 2.7: The relationship between an individual concept (represented by the larger box), its category (k) and associated words (w_1, w_2, w_3). The concept is represented by its sense (s) and its sets of exemplars (e_1, e_2, e_3, e_4). The word is connected primarily with the sense, while the exemplars are linked more intimately with the category (since they encode actually encountered items from it). However, the concept as a whole determines its category, so all elements of the concept are causally connected to it.

model (Figure 2.5). This allows us to abstract away from internal conceptual structure and is visually cleaner. However, at times I will need to deal with the components of concepts directly and how the inner and outer relationships of concepts interface. In those cases, I will work with the fuller version (Figure 2.7).

2.6 Categorisation, conceptualisation and words

2.6.1 Introduction

Up to this point, I have focused mainly on the mental structures of concepts and their offline relationships to categories and words. But the power of concepts comes from their actual application in important cognitive processes. Although concepts are recruited for a range of processes and it may be ultimately misleading to focus on a single one (Solomon, Medin and Lynch 1999), in my thesis I focus mainly on conceptualisation. However, just as categories are a convenient way to get at concepts, so categorisation is easier to work with than conceptualisation.

Although I have insisted on a clear distinction between concepts and categories (see Section 2.4.1), this does not imply a fundamental split between conceptualisation and categorisation. Since categories are only extensions of concepts (see Section 2.4.1), any act of categorisation has an associated act of conceptualisation, and vice versa. In other words, they identify the same cognitive process. Therefore, the difference between categorisation and conceptualisation lies only in emphasis in the process outcome: are we interested in which concept is activated in the head, or the category of things in the world that it determines? Nevertheless, it is important to remember that since concepts overdetermine categories, conceptualisation overdetermines categorisation as well. For instance, I could group a set of fruit into bananas and kiwi, while you could group the same set into yellow and brown. But although our categorisations would probably be the same, our conceptualisations would be different. Therefore, as with concepts and categories, I will stick to the distinction when relevant; otherwise, I will often use the two terms interchangeably.

In this section, I first converge on a definition for categorisation, driven by methodological considerations which arise out of my rejection of conceptual uniformity. I then discuss the implications for a corresponding definition of conceptualisation, and consider how the internal components of concepts, especially exemplars, are involved in the process. Then I discuss how categorisation is related to similarity, and introduce an important cognitive phenomenon called categorical perception. Finally, I consider how the words individuals know and use affect their conceptualisation processes.

2.6.2 Categorisation

Categorisation is a (or even *the*, Harnad 2005) fundamental cognitive process. Jackendoff (1983) claims that, in fact, without it memory is useless, and that it is central to cognitive psychology. Mervis and Rosch (1981) express something similar: "without any categorisation an organism could not interact profitably with the infinitely distinguishable objects and events it experiences" (p. 94).

Having allowed for the possibility of conceptual variation, the first step towards finding a definition of categorisation is to emphasise that the repertoire of categories (and concepts) is specific to an individual. In other words, when different people categorise the same item, they do so relative to their own set of categories. As a result, strictly speaking, we cannot interpret an experimental participant's categorisation choice as identifying some kind of universal category. For example, at the risk of being overly pedantic, it is not appropriate to claim that an experiment participant puts a stimulus in *the* dog category, but rather in *her* dog category.

Notice that this argument automatically applies to word meaning as well: just because you and I both call something a "dog" does not mean that we categorise it in exactly the same way. This means that it's now theoretically problematic to relate two people's categorisation of the same thing, which was the point Fodor (1998) was making when he attacked conceptual relativity (see Section 1.3). However, we are in too deep now, and have no choice but to accept the resulting difficulties. In particular, we cannot depend on language as a reliable indicator of someone's categories.

How do these considerations relate to available definitions of categorisation? Jackendoff (1983) gives a straight-forward definition, saying that to categorise is "to judge that a particular thing is or is not an instance of a particular category" (p. 77). Thus, when we are exposed to a particular item, we identify, from among our set of categories, the category to which it belongs. If we interpret the category as being local to the categoriser's repertoire, then theoretically, this definition could survive the potential problems above. We would just need to be careful to interpret someone's categorisation within the context of their own system of categories.

However, as it stands, this definition is not much use from a methodological point of view. The problem again lies in the identification of the category. If all we have is a single item matched with a single "identifier", we have no way of actually identifying the category being used, regardless of whether or not the identifier was linguistic.

We therefore need to consider alternative definitions. Note that Jackendoff's (1983) definition embodies a top-down perspective: there is a set of preexisting categories, and when we are exposed to an item, we choose from one of them, possibly using language. However, recall that categories are "classes of things" (Murphy 2002, p. 5). Mervis and Rosch (1981) say that "A category exists whenever two or more distinguishable objects or events are treated equivalently" (p. 89). Similarly, Hahn and Ramscar (2001*a*) state that "Categories allow us to treat different - but in important ways similar - objects equivalently, and hence to communicate about, draw inferences from, reason with, and interpret these objects" (p. 1). These relationships between members of a category suggest an alternative, bottom-up view of categorisation. As Medin and Aguilar (1999) put it, categorisation is "the process by which distinct entities are treated as equivalent" (p. 104). In this view, the categorisation of an item is defined in terms of the other things in its category.

In other words, we need not necessarily have an external identifier of a category. The "particular category" by which an item is categorised can be identified by the set of items it is grouped with. Items that are mentally put together are in the same category, and those that are put separately are in different categories. This view easily accommodates the possibility for variation among people, and no longer relies on language (or any other kind of symbol) for category identification. Indeed, this approach would even allow us to have brand new categories every time items are grouped, consistent with a highly dynamic view of categorisation (Smith and Samuelson 1997).

Although framing categorisation in a top-down way may seem more intuitive and simple, in some ways a bottom-up view is actually more fundamental. In a recent paper which emphasises the central cognitive importance of categorisation (indeed, it is entitled "Cognition is categorization"), Harnad (2005) gives a general bottom-up definition: "most simply and generally, categorization is any *systematic differential interaction between an autonomous, adaptive sensorimotor system and its world*" (p. 21, author's emphasis). In this view, reminiscent of Johnson (1987), categories are grounded in experience, and emerge out of our bodily interaction with the world. Thus, far from being artificial, bottom-up categorisation may in fact be far more natural and evolutionarily primal.

However, we still have a problem: while there may be nothing intrinsically wrong with Harnad's (2005) view, it has taken us a bit too far afield, as it has dissociated categorisation from the human mind. His definition could get by without any mental concepts being associated with the categories at all, like a thermostat responding to temperature changes. As such, Harnad's definition is inconsistent with my definition of categories, which demands that they be associated with underlying human concepts. What I need, then, is a step back from Harnad's position.

Therefore, combining the definitions, insights and considerations discussed above, I define categorisation as **the mental act of assigning items to categories**. The key is the plurality here, allowing one act of categorisation to handle multiple items and categories together. The definition combines Jackendoff's (1983) notion of assignment to categories with Medin and Aguilar's (1999) emphasis on grouping, while appealing to Harnad's (2005) ecological justification. The explicit mention of a mental process also helps link categories back to concepts, and we just need to be careful to interpret the categories as being specific to the categoriser (and perhaps even to the moment of categorisation).

This perspective may be a little unusual, in that it frames an act of categorisation as relating to multiple things together at the same time. In particular, it could lead to the absurd suggestion that we cannot recognise lions (for example) unless we see them in batches. However, the definition is motivated largely on methodological grounds. From a methodological point of view, once we abandon universal and public concepts and the use of words as reliable category identifiers, we can no longer study categorisation of one item in isolation. Although this is not the standard notion of categorisation, I appeal to the intuition that the result of acts of "categorisation" should be "categories". Moreover, the definition is quite flexible and actually subsumes the canonical case involving one item and category (Jackendoff 1983) as a special case.

Before moving on, I should briefly address one potential concern. In reviewing the internal structure of concepts in Section 2.5, one of the clearest conclusions was that concepts are not black-and-white boxes with necessary and sufficient conditions. And yet, the way I have been talking about and defining categorisation makes it sound like that is the view that I am adopting. So what gives? The catch comes from the difference between the structure of concepts and their use in the process of categorisation: the latter is black-and-white, while the former is not. As Croft and Cruse (2004) argue, there is nothing incompatible between taking concepts as having prototypical structure while still requiring discrete decisions for any individual act of categorisation. This can be easily seen if we consider how we use language, where we are required to make lexical choices all the time. An ostrich may be a relatively poor example of a bird, but I still need to decide, when speaking about it, whether to call it a "bird" or not.

2.6.3 Conceptualisation

Now that I have defined categorisation, my definition of conceptualisation can simply piggy-back off of it. In particular, since categorisation is the act of assigning items to categories, and categories are associated with concepts, conceptualisation is **the mental act of assigning items to concepts**. Again, the main difference between conceptualisation and categorisation is in the process result: when we talk about conceptualisation, we are primarily concerned with the concept rather than the category that an object is assigned to.

Note that while the terms "concepts", "categories", and "categorisation" are all quite widespread in the psychological literature, "conceptualisation" is less so. This may be because modelling conceptualisation requires incorporating concept senses and categorisation processes, which typically belong in the purview of different disciplines. In particular, while philosophers and linguists tend to focus on the senses that correspond to words, psychologists conduct experiments with categorisation tasks.

However, this separation is rather artificial. As pointed out at the outset of Section 2.5, categorisation experiments in psychology are not just about groups of things in the world, but about the corresponding concepts in the minds of the categorising humans (Malt 2006) . As such, in the terminology that I've adopted here, psychology categorisation studies are actually largely focused on conceptualisation. When a participant puts an item into a category, psychologists are ultimately interested in the underlying mental assignment of the item to a concept. And psychologists are by no means the only ones concerned with conceptualisation. Indeed, cognitive linguistics has placed conceptualisation at its core (Langacker 1987).

In considering conceptualisation processes, it's worth raising two theoretical issues from earlier in this chapter. First, there may theoretically be variation in conceptualisation without necessarily implying variation in concepts. Rather, differences might come from which concept is selected by two different people for the same thing. For instance, as discussed in Section 2.3.1, I may conceptualise a particular flying thing as a bird, while bird watchers may do so at the more specific level of a robin (Bailenson et al. 2002). More generally, there are likely to be many different concepts that could be assigned to a particular item even from within the same person's conceptual system, so differences between people may sometimes stem solely from which choices they make. Second, shifting the focus from concepts to acts of conceptualisation provides a partial solution to the concept individuation problem (see Section 2.3.2). Rather than worrying about whether conceptual shifts imply a switch to a different concept, change of an existing concept, or creation of a new one, we can ignore this question and simply look at conceptualisation snapshots: how does a person conceptualise something at a particular time? This does not deny that acts of conceptualisation intimately involve underlying concepts. However, it does allow me to take an agnostic stance regarding whether concepts are largely static structures or are generated anew at each moment of use (Smith and Samuelson 1997).

It is also useful here to return to the internal structure of concepts and consider how they participate in conceptualisation. In Section 2.5.3, I settled on a conceptual representation that includes a sense and a set of exemplars. Recall that a sense contains things that are general to a concept, while exemplars encode individual category members. Notice that there is an asymmetry between the relations of these two parts of a concept to an act of conceptualisation. While the impact of conceptualisation on a concept's sense is not clear a priori, the effects on its set of exemplars is relatively straightforward: the conceptualised items become encoded and added to the concept's exemplars (by definition).

One noteworthy consequence of this situation is that we now seem to have conceptualised stimuli in two places: both in the world, among a category's members, and in the mind, among a concept's set of exemplars. But this is not redundant. Dogs really are in the world, and memory representations of dogs I have seen really are in my mind. And if we want a theory which handles both a concept's reference and is psychologically plausible, as is required here, then we need both.

This distinction between the world and our representation of it lies at the core of Johnson-Laird's (1983) influential theory of mental models. Johnson-Laird argues that we construct and develop working models of the world, and that these models are central to our reasoning, perception and thought. Our mental access to the world is therefore indirect: "Human beings, of course, do not apprehend the world directly; they possess only an internal representation of it, because perception is the construction of a model of the world" (p. 156). Mental models can also provide the basis for a theory of linguistic meaning (which was one of Johnson-Laird's objectives), because the truths of propositions can be evaluated relative to a mental model of the world, rather than the world directly. This means that referential aspects of meaning can be taken seriously, without complications being automatically caused by typically problematic phenomena such as hypothetical or counterfactual situations.

Although Johnson-Laird (1983) didn't apply his theory to relate a concept's set of exemplars and a concept's category, it is easy to adopt it for this purpose. Consider, for example, dogs. My set of dog exemplars consists of all the dogs that I have seen, heard or perhaps even imagined. In contrast, my dog category consists of all the things in the world that I would judge to be a dog if they appeared before me. Dropping for the moment the difference between my encoding of a particular dog and that dog itself, we can consider all four set intersection possibilities here. There are dogs that I have both seen and remembered, but there are also dogs that exist although I have never seen them, as well as dogs that I have imagined but which do not exist. Mental models are therefore also useful from an analyst's point of view, as they allow us to both dissociate and relate the otherwise easily confusable notions of exemplars and category members.

2.6.4 Categorisation and similarity

In discussing the inner structure of concepts in Section 2.5, an important theme was that of similarity. Indeed, exemplar and prototype theories both rely heavily on similarity, but the relationship between categorisation and similarity is still highly debatable (Hahn and Ramscar 2001*b*). In particular, on what basis is the similarity of items assessed? This issue turns out to be surprisingly problematic, and has interesting implications for the potential role of language in categorisation. Therefore, in this section, I explore the notion of similarity further, and its relation to categorisation.

For prototype and exemplar theorists, similarity is used to compare stimuli being categorised to prototypes or sets of exemplars, respectively. Similarity is often defined in terms of perceptual features. Features may be categorical (e.g., colour), discrete (e.g., number of legs) or continuous (e.g., height). The similarity of two stimuli can then be assessed by comparing their feature values. If the features are numerical or binary, this notion of similarity can be captured in a formal geometric model: the stimuli are conceived as being points in a multi-dimensional similarity space, where each dimension represents a feature, and the similarity of stimuli is defined as the distance between the points. Although such geometric models are probably the most common, they have been heavily criticised on both theoretical and empirical grounds (e.g., Tversky 1977; Gauker 2007), and psychologists have proposed several other types of similarity models as well (Goldstone 1999).

However, the notion of similarity has a number of problems. In their initial proposal of theory theory, Murphy and Medin's (1985) starting point was an attack on the explanatory usefulness of similarity. In particular, there are no clear and unproblematic constraints on what counts as a feature or why. What list of attributes is used to assess the similarity of two things? In principle, Murphy and Medin (1985) point out, the list of potential features is infinite, so that without some a priori understanding of which features are important, how they should be weighed, and potentially how they co-occur, "any two entities can be arbitrarily similar or dissimilar" (p. 292).

One way to try to get around this problem is to propose that there's a fixed universal set of features and weights. Then there would be a unique, objective degree of similarity for any two stimuli. However, this work-around has several problems of its own. First, the set of features (or at least a set of primitive features) could not be conceptual, but rather would have to be entirely perceptual. Otherwise, we would have a circular argument, and would have to re-adopt conceptual universals, which I have adamantly rejected. But none of the typical features typically listed for birds (e.g., has wings, has feathers, flies), for example, seem to satisfy this criterion. Second, Schyns, Goldstone and Thibaut (1998) argue theoretically and demonstrate experimentally that a fixed set of features cannot easily explain concept learning in cases that involve new kinds of features. Instead, the new features too must be developed, after which then can serve as the basis for the relevant concepts. Third, there is empirical evidence that, in fact, similarity depends on various factors, including context, perspective, and expertise (Tversky 1977; Medin et al. 1993). Together, these considerations strongly suggest that there is no fixed mental similarity space, supporting Murphy and Medin's (1985) arguments of its inadequacy.

Indeed, Rips (1989) neatly demonstrated a dissociation between categorisation and similarity in an influential study. In one experiment, he first chose pairs of categories which differ on a salient dimension, and also in their degree of variability (e.g., pizzas and quarters). Then he asked participants for similarity and categorisation judgements concerning objects whose value on the relevant dimension was between that of the two categories (i.e., 3 inches in diameter for the pizza/quarter example). Participants judged the object to belong to the more variable category (i.e., pizza), but to be more similar to the less variable one (i.e., quarter). In a second experiment, participants read a story in which one animal was somehow artificially transformed from one type to another (e.g., bird → insect). Participants judged the transformed animal to be more similar to the new animal type (i.e., insect), and yet to still be of the old type (i.e., bird). Rips' findings do not imply that similarity has no role in categorisation, but rather that it cannot be the whole explanation. However, if we allow for our knowledge of the world to alter our similarity space (as theory theory could have it), so that, for example, similarity is graded differently in the cases of pizzas and quarters, then Rips' dissociation may well disappear.

But this suggests that we may judge similarity relative to the categories that objects belong to. In other words, rather than similarity wholly underlying categorisation, categorisation may in turn affect similarity. Indeed, this is manifest in a well-established phenomenon called "categorical perception" (Studdert-Kennedy, M., Harris and Cooper 1970). Categorical perception occurs when one is more sensitive to perceptual differences if they cross a category boundary. The most prominent example comes from phonology. In a classic study, Liberman, Harris, Hoffman and Griffith (1957) systematically prepared sequences of speech sounds with equal phonetic distances between adjacent sounds. For instance, */da/* can transition into */ta/* by gradually increasing the voice-onset time. Participants first listened to two sounds, and upon hearing a third sound, had to indicate which of the first two it was identical to. Participants performed more accurately when the two candidate sounds involved different phonemes rather than variants of the same phoneme. This was the case even when the phonetic distance between them was constant.

Although speech sounds are perhaps the best documented domain of categorical perception effects, they have been demonstrated in a variety of cognitive domains (Harnad 1987), including object categorisation. Goldstone (1994) studied how categorical perception effects emerged during category learning. In a series of experiments with visual geometric stimuli varying gradually in size and darkness, he found evidence for increased perceptual discrimination across category boundaries. His analysis also revealed that discrimination improved mostly because of "acquired distinctiveness" across a category boundary rather than “acquired equivalence” within a category, although other experiments have shown evidence for both (Rothbart, Davis-Stitt and Hill 1997). In another category learning experiment involving morphed pictures of human faces, Goldstone, Lippa and Shiffrin (2001) sought to disambiguate between two possible explanations of categorical perception: do participants merely use category membership as an extra criterion when actually judging similarity, or is their initial encoding of objects already affected by the object’s category? Their results, based on similarity ratings of category members to a novel neutral stimulus, found support for the latter hypothesis. This suggests that the way in which we represent exemplars in our minds is significantly influenced by what categories they belong to.

Lupyan (2008*a*) studied the effects of pre-established concepts on visual processing in another way, using a visual search paradigm. His study is particularly important, because it also touches on the role of language. Participants had to look for a target stimulus amongst a set of distractors, and the conceptual heterogeneity of the non-targets was manipulated, while keeping the perceptual differences constant. In particular, the target was a character which looked like a hybrid between a “b” and a “p”, and the distractors were either “B”’s and “b”’s (conceptually homogeneous) or “B”’s and “p”’s (conceptually heterogeneous). The results indicated a “conceptual grouping” effect: participants were faster at locating the target in the conceptually homogeneous case. Lupyan then conducted two follow-up experiments in order to disambiguate between two possible explanations for the results. He first had participants carry out a speeded similar-different judgement task, and examined participants’ reaction times. The results showed that although participants were definitely slower if one of the stimuli was the novel b-p hybrid, there was no speed difference between the B-b and the B-p discriminations. This finding argued against a long-term categorical perception explanation. In the other follow-up, participants heard either “find the target” or “find the B”, and had to look for a “B” among either “b” or “p” distractors. Importantly, the participants already knew that they would be looking for a “B” beforehand, so the difference in instructions did not actually provide any new information. The results revealed that the label significantly speeded up search, but only when looking amongst “p” distractors. Therefore, it appears that hearing the category label prior to the search improved discrimination, provided that it crosses the category boundary. This finding suggested an online effect of conceptualisation on perception, mediated by concept labels. Lupyan concluded that concepts impact on perceptual processing, and do so even more when they are labelled online.

Together, these studies reveal a potential general mechanism through which concepts could become coordinated between people. They suggest that to the extent that concepts align, there may also be a convergent effect on the feature space in terms of which particular items are encoded. In turn, this supports the possibility of a much more interactive view of top-down and bottom-up cognitive processes (Lamme, Super and Spekreijse 1998), so that alignment at conceptual and perceptual levels could mutually reinforce each other. In particular, if alignment in conceptualisation results in convergent similarity spaces, then this may in turn cause people to subsequently conceptualise more similarly. This suggests a way in which language may bring people's conceptualisations together, not only explicitly and consciously, but also through lower-level and more implicit cognitive mechanisms. I explore this possibility when I look into the relationship between words and conceptualisation in the next section.

2.6.5 *Conceptualisation and words*

In the last section, I discussed how conceptualisation could have top-down effects on similarity and perception. In this section, I move up one level further, and discuss how words might affect conceptualisation online. In particular, to the extent that words might affect conceptualisation, do they always have this effect, or only when they are invoked? As suggested by Slobin (1996) (see Section 3.3.5), it's possible that we process things differently depending on whether or not we are using language. Our conceptualisations may be modulated as a result of online lexical processes.

However, it is important to first note that it is highly controversial whether words affect private conceptualisation at all. This is partly because this possibility is incompatible with the traditionally dominant sides of two broad debates in cognitive science. One of these concerns the modularity of the mind, and language in particular. In Fodor's (1983) influential proposal, the mind is organised into different functional modules. Modules are domain-specific, only operating on specialised inputs and yielding restricted outputs, and they are informationally encapsulated, operating independently of each other. language is proposed as one such module. In particular, language can take input from perception, but operates independently of it, and does not make its output available to it. Chomsky's (1988) influential theory of language explicitly takes this position. In his view, language is an innate module, or "mental organ", which operates mostly independently of other cognitive modules. Similarly, Levelt et al.'s (1999) psycholinguistic model of lexical access in speech production feeds forward from conceptualisation through semantic levels to other levels of linguistic representations, but not backwards. Although the model does include self-monitoring which can lead to reconceptualisation, the latter only occurs when a person notices herself making (or about to make) an error. Even more interactive models (e.g., Dell 1986) only incorporate bidirectional flow among the linguistic levels of representation (e.g., phonological, syntactic, semantic).

However, other researchers take issue with such modular views of language, claiming that language is closely integrated with the rest of cognition. In particular, cognitive linguists have argued that language is not domain-specific or encapsulated, and that it is grounded in the rest of cognition (e.g., Lakoff 1987; Langacker 1987). These theories borrow heavily from insights in related disciplines (such as prototype theory, Taylor 1995) and stress the role of conceptualisation and other cognitive processes in linguistic theories (Croft and Cruse 2004). These views are more comfortable with language having a bidirectional relationship with conceptualisation. It is worth noting that criticism of Fodor's (1983) modularity hypothesis has not been confined to language, and includes both strong objections (e.g., Uttal 2003) and reconciling proposals (e.g., Karmiloff-Smith 1992).

The other important issue in cognitive science I want to raise concerns the function of language. Language obviously serves as a tool for human communication: we exchange information and opinions about the world via language. Some researchers emphasise the communicative aspect of language and claim that while words serve to exchange ideas, the concepts underlying them are relatively fixed (Fodor 1975; Pinker 1994). Under this view, there is nothing for words to do other than passively attach to pre-existing concepts and express them. Note also that an emphasis on the communicative function of language seems inevitable if we adopt innate, universal concepts (Fodor 1998), although it does not require them (Pinker and Jackendoff 2005).

However, some researchers have also argued that language also has important cognitive functions. Goldstein (1948) suggests that language is not only used for communication, but also can "support" and "fixate" thinking (p. 115). Similarly, Chomsky (2000) claims that although language can of course be recruited for communication, what primarily distinguishes it is that "it is a system for expressing thought" (p. 75). Carruthers (2002) argues that language helps us integrate information from different otherwise domain-specific cognitive modules. Dennett (1995) claims that language is largely responsible for human consciousness and gives us cognitive abilities which greatly outstrip those of other animals. And Clark (1998) suggests several specific areas in which words might enhance human computation, including memory, attention, representation, and simplification.

I will not engage the modular or functional debates further here. The key point is that the role of words in conceptualisation is theoretically important, and yet has been understudied. Although I have given two important historical reasons for this, another major cause is the methodological challenges behind it. As Nuyts and Pederson (1997) point out, in order to study the relationship between language and conceptualisation, we must first separate them, and that is not an easy task.

Note that investigations of the role of words in conceptualisation generally reflect the view of language as a tool for communication. In most studies, words are provided by the experimenter (e.g., concept learning: Lupyan, Rakison and McClelland 2007) or other participants (e.g., communicative tasks: Markman and Makin 1998), rather than the participant himself. And the majority of the recent studies that have focused on the participant's own pre-existing knowledge have been cross-linguistic. Therefore, while such work is extremely relevant for my thesis, it will be covered in Chapter 3.

In contrast, there are not many studies which investigate how a person's own words affect their conceptualisation online, although a few key recent studies have begun to address this gap. In one series of experiments, Lupyan (2008*b*) explored the effect of linguistic categorisation on item recognition. Participants were first presented with sequences of pictures of furniture and asked to either categorise them linguistically (from among a set of two categories, like "chair" and "lamp") or to indicate preference (whether they liked them or not). Afterwards, participants were tested on both previously seen and unseen items, and had to indicate for each whether it was old or novel. The results showed that old items in the categorisation condition were more likely to be mistaken for novel items than those in the preference condition. After rejecting several alternative explanations via follow-up experiments in which important variables such as stimuli and timing were altered, Lupyan (2008*b*) argued that the results were best explained with a "representational shift account". In this view, when items are linguistically categorised, the memory representations of items are influenced not only by bottom-up perceptual features of the items, but also by top-down conceptual representations activated by the linguistic labels. As such, the representations of items which are linguistically categorised get "shifted" towards the category prototypes, so their memory encoding is less loyal to the original. Thus, words seem to affect conceptualisation, and conceptualisation in turn affects memory encoding.

A couple of other intralingual studies have taken a different approach. People may internally verbalise things during categorisation, even when they don't explicitly produce them: when presented with a chair, we may sometimes think "chair" without actually saying it. This suggests that perhaps the role of words could also be studied by designing experimental conditions in which one manipulated whether or not participants were verbalising internally. While this may be impossible to achieve with certainty, Roberson and Davidoff (2000) used verbal interference tasks for this purpose and obtained fruitful results. Building off of previously documented categorical perception effects in the colour domain (see Section 3.3.2), they conducted several experiments to see whether such effects were robust to verbal interference. Participants were first shown a colour stimulus in the green-blue range, then carried out an interference task for a few seconds, and finally had to identify the original colour from a pair of similar stimuli which included the original. The interference task was either visual (tracking a line through a dot pattern) or verbal (reading words aloud), and there was also a control condition with no interference task. The results confirmed categorical perception effects in the control and visual conditions, but also showed that they disappeared in the verbal conditions. This was the case regardless of whether the words that participants read were colour words (none of which matched the stimuli) or unrelated words. Thus, it seems that if we are prevented from internally labelling something due to other linguistic processing, our words might not affect how we encode things; however, words can affect perception under other conditions, even if we do not explicitly produce words. Together with

Lupyan's (2008*a*) results, this suggests that our words can affect conceptualisation, as long as we are not simultaneously processing other words.

Lupyan (2009) used verbal interference tasks to study another aspect of how words might affect conceptualisation. His starting point was the aphasia studies and reviews of Cohen and colleagues (Cohen, Kelter and Woll 1980; Cohen, Woll and Ehrenstein 1981), from which the authors concluded that aphasics have a defect in analytically isolating single features (such as colour or size) of stimuli, while performing normally on global comparisons. Lupyan conducted two experiments to see whether these patterns would also be found in normal participants under verbal and/or visual interference. Participants were shown triads of stimuli, and had to choose the odd one out, based on either colour, size or thematic (e.g., cake and balloon are both related to the theme of parties) criteria. Stimuli were presented either as pictures or as words. In the verbal interference condition, they were shown a 9-digit number before each trial and were asked to rehearse and remember it, for which they were tested after the trial. In the visual interference condition, a visual grid display was used instead of a number. The results showed that judgements were impaired for the single dimension judgements (colour or size) for both picture and word stimuli, but only under verbal interference. However, this effect did not extend to thematic judgements, nor did it occur under visual interference. These results suggest that words help people conceptualise in terms of a particular dimension, but are less critical for more general assessments.

A final relevant line of evidence here consists of developmental studies with infants. The advantage of using infants is that, in contrast to older children or adults, their vocabulary and object familiarity is relatively limited and can be approximately indicated in advance by their caregivers in questionnaires. This allows us to study how knowing a word for something may affect an infant's processing. Schafer and Plunkett (1999) carried out such a study with 17-month-olds. Each infant was simultaneously shown pairs of object images and their viewing times were recorded to each image. The objects were selected individually for each infant so that they were both approximately equally familiar but the infant only knew one of the object's names. Results showed that infants looked significantly longer at objects for which they knew the name, even though there was no linguistic input in the experiment. Similarly, Gliga, Volein and Csibra (2010) conducted a neuroscientific study to see whether infants' brains processed stimuli more extensively (in terms of a measure that was previously shown to be involved in visual object processing, Tallon-Baudry and Bertrand 1999) if they knew their names. 12-month-olds were shown images of objects which were either unfamiliar, familiar with known names, or familiar with unknown names, and their brain activity was recorded. Results showed greater activation only for objects with known names, and no effect of mere object familiarity. Moreover, the same pattern of results was duplicated when infants were first trained on novel word-object associations. Finally, Rivera and Zawaydeh (2007) had infants carry out object individuation tasks with spatiotemporally ambiguous events. 10- and 11-month-old infants were first shown a pair of different objects coming in and out from behind a screen, after which, in some trials, one object was surreptitiously removed. Then the screen was lifted and the infant's looking time was recorded. The results indicated that infants noticed inconsistencies best when they were familiar with names of both of the objects, even when controlling for object familiarity and general receptive vocabulary. Together, such developmental findings suggest that knowing a word for an object increases attention and processing in young infants, even without verbal input.

Overall, then, there is plenty of evidence that the words we already have in our minds affect how we conceptualise and process things in the world. These effects do not require verbal input, or even explicit lexical production, and they are documented from a very early age. However, they are also not static or permanent, since they can be blocked by conflicting lexical processing.

2.6.6 Conclusion

In this section, I have focused on processes involving concepts. First, I looked for a methodologically motivated definition of categorisation which was compatible with my commitment to conceptual variation. I settled on a hybrid of top-down and bottom-up definitions by which categorisation is defined as the act of assigning items to categories. I then discussed conceptualisation, and how it actually identifies the same process as categorisation, but emphasises the resulting concepts, rather than the resulting categories. I also explored how conceptualisation relates to theoretical issues concerning concepts, and how the internal components of concepts are involved in conceptualisation. I then considered the relationship between categorisation and similarity, showing that it is not as simple as it first appears. After that I introduced a well-documented phenomenon known as categorical perception, in which categorisation can affect the perceived similarity of things. Finally, I moved on to discussing the complex relationship between conceptualisation and words. I discussed how, although words are generally understood as reflecting rather than determining conceptualisations, there is increasing evidence for them also playing a causal role.

We can now put together a model for an act of conceptualisation by an individual at a particular time, as shown in Figure 2.8. A concept is applied to three objects in the world, and is potentially (but perhaps not) lexicalised with a word. Notice the differences between this diagram and that developed earlier for the structural relationship between concepts, categories and words (Figure 2.4). This is because it is modelling something different, and has a different purpose. The previous diagram had focused on the structural relationships. In contrast, here we are taking a dynamic snapshot of a person's concept being applied online. First, since the individual has to make a discrete lexical decision, there is now just one word (assuming the conceptualisation is lexicalised). Also, there is now a bi-directional relationship between words and concepts, since conceptualisation affects the choice of words but the word also affects the conceptualisation. Finally, a specific set of objects (which is a small subset of the concept's category) in the world triggers the conceptualisation, while before I focused on how a concept theoretically determined its whole category (hence the arrow pointing now in the opposite direction). The two models are compatible, however: strictly speaking, we could still have an arrow down from the concept to the category (as a whole), but that relationship is not the focus here and this would only add confusion.

Finally, as shown in Figure 2.9, we can also incorporate what we established about internal conceptual structure into this model (see Figures 2.6 and 2.7). The diagram also captures how categorisation is not only driven by exemplars we encounter but also in turn affects it (i.e., there are both top-down and bottom-up processing influences between a concept's sense and its exemplars), as we saw in Section 2.6.4. However, the actual objects in the world only trigger the conceptualisation: there is no effect in the opposite direction.

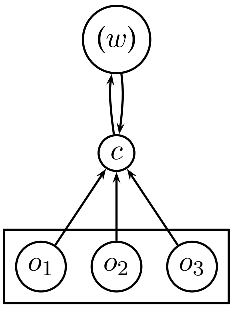


Figure 2.8: An act of conceptualisation, in which an individual's concept (c) is applied to a set of objects (o_1, o_2, o_3). A word (w) may also be applied, but it is optional (indicated with brackets). The arrows between the word and the concept show how the levels influence each other in both directions: conceptualisations trigger words, but words also influence conceptualisations. The direction between objects and concepts is unidirectional: objects trigger conceptualisations, but although conceptualisations can affect the perception and encoding of objects, they do not change the actual objects themselves.

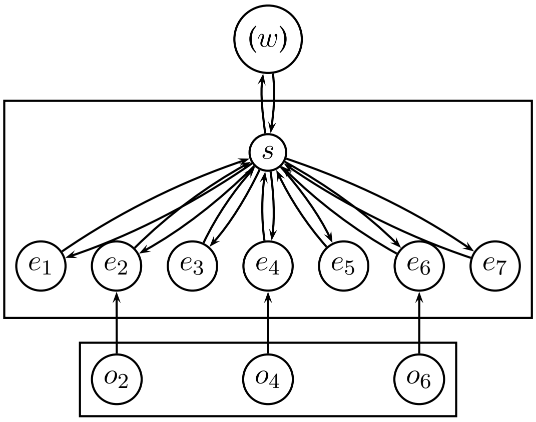


Figure 2.9: An expanded view of the act of conceptualisation, incorporating the internal structure of an individual's concept (represented by the box). The word (w) relates to the sense (s), and objects ($o1_1$, o_2, o_3) become new exemplars (e_2, e_4, e_6), which are thus added to the previous exemplars (e_1, e_3, e_5, e_7). The existence of mutual (top-down and bottom-up) processing influences in categorisation is indicated by the arrows running in both directions between the sense and the exemplars; however, as this will not be the focus, these will generally not be represented in subsequent diagrams.

2.7 Summary

In summary, this chapter focused on individual conceptualisation and its relation to words and objects. Although it is theoretically convenient to posit that concepts are universal, there is good evidence for conceptual variation, so we must abandon a universal position, especially in a thesis about conceptual coordination. Doing so, however, has important theoretical implications. Most fundamentally, we must now define categories and words in terms of concepts, rather than the other way around. Concepts determine categories, rather than reflecting them, and they have no fixed 1-to-1 relationship with words. They also have an internal structure, with both exemplar-specific and concept-general components. The usefulness of concepts, however, comes from their application. In particular, in the process of conceptualisation, the mind assigns items to its concepts. Conceptualisation has tight but non-trivial relationships with both similarity and language. It also offers us a way to access concepts, provided that we keep in mind that the process is dynamic, and may vary from person to person.

The most important result of this chapter is the simple theoretical model that I have developed. The model has two parts, which I have referred to as offline and online. The former (see Figures 2.5 and 2.7) concerns the abstract general relationships between concepts, categories and words, while the latter (Figure 2.8 and 2.9) addresses actual applications of concepts during acts of conceptualisation. In principle, which one we place our emphasis on depends on our theoretical orientation. However, in practice, even when we are interested in the abstract relationships, we can only get at the underlying structures through processing (Barsalou 1990). As such, for most of my thesis, I will primarily use the processing aspect of the model. However, it's good to keep in mind how the two relate, especially when we want to draw conclusions about conceptual structure, or the relationship between language, mind and world.

CHAPTER 3

Conceptualisation and shared words

3.1 Introduction

In the last chapter, I discussed conceptualisation from the point of view of an individual. Even when language was discussed, the focus was on the associations between words and concepts in an individual's mind. However, a fundamental property of words is that they are a shared, public resource. Although two different English speakers may associate slightly different concepts with the word "dog", they nevertheless both store the same word form in their individual minds. Since these words are used in linguistic communication about things in the world, they might serve as a vehicle through which people, consciously or otherwise, influence each other's views of those things. As such, consideration of the shared nature of lexical knowledge and the interactive nature of language use is a critical step in addressing the hypotheses of my thesis.

In this chapter, I review evidence concerning the public aspect of words and how they might coordinate concepts and conceptualisation. First, does having the same set of words as other speakers of the same language result in having similar concepts? And second, does the use of words in linguistic interaction bring together our conceptualisations? These two questions are explored in detail in Sections 3.3 and 3.4, respectively. Having done so, I will also apply the theoretical model I developed in Chapter 2 and incorporate the new social issues. However, first, in Section 3.2, I discuss an important issue that in some ways is presupposed by the other two: whether the words of others affect how we learn and use concepts.

3.2 Word input

In this section, I consider the effects of external lexical input. How does hearing the word label of others while processing a stimulus affect a person's concept and conceptualisation? This is important to my thesis because, if words are to bring about conceptual coordination between individuals, then a logical pre-condition must be that the words of others somehow affect a person's own conceptualisations.

However, I do not attempt to review this issue comprehensively here, as it would take us much too far afield. In particular, the issue is central to the controversial relationship between language acquisition and conceptual development (Bowerman and Levinson 2001; Nelson 1996), and indeed the cognitive processes underlying language comprehension in general. Therefore, here I will only highlight several key findings from three broad age groups: (prelinguistic) infants, children and adults.

I should first point out that, in principle, there are three different but related ways in which hearing another's words during conceptualisation could affect a person's conceptual system: the building of connections between concepts and words, the development of concepts, and the recruitment of concepts online during conceptualisation. Also, the effects might depend on whether the words heard are already familiar, and, if so, how well they match the referent in a person's preexisting conceptual system.

Note that the necessity of linguistic input in the first of the above issues is beyond dispute: although there is plenty of debate concerning how exactly words are learned (Hall and Waxman 2004; Bloom 2000), noone questions that they must be learned. And since word learning is not the focus of this thesis, I will not address it here, focusing instead on work that is relevant to the latter two issues: the effects of word input on conceptual development and conceptualisation. As these two issues are related and I can only address them briefly, I will not explicitly separate them in this discussion, but it is worth keeping them in mind as we go along.

3.2.1 Infants

Although the extent and specificity of an innate human language capacity remains a topic of intense debate (Hauser, Chomsky and Fitch 2002; Pinker and Jackendoff 2005), it is uncontroversial that particular words and their conceptual associations are not innate. Different languages have different lexicons, and children must learn their native lexicons from the people around them. However, there is ample evidence that even for young "prelinguistic" infants, external words do not merely passively map onto preexisting concepts, but rather play an active role in conceptual development.

Xu (2002) studied the role of words in object individuation in 9-month-old infants. Two different objects were revealed to infants from behind a screen, one at a time, and then returned behind the screen. While an object was being visually presented, the infant also heard either a linguistic label (e.g., "Look, a ball"), an auditory tone, other kinds of sounds (e.g., a car alarm), or an emotional expression (e.g., "Ah"). After both objects had been presented in this way, the screen was lifted, revealing either one or both objects. Based on infants' looking times, the results showed that hearing two different linguistic labels for the two objects helped them individuate them. In contrast, a single label for both objects, or two different non-label sounds did not have this effect. These findings have been corroborated by several other studies which have shown that infants individuated objects better if they already knew the words being used (Rivera and Zawaydeh 2006), that they expected the number of objects to correspond to the number of labels even when they were never shown any objects in advance (Xu, Cote and Baker 2005), and that they expectedx different labels to be used for two objects that varied in shape but not in colour (Dewar and Xu 2007). These results suggest that words guide individuation in infants, and that hearing words may therefore affect early conceptual development.

A related line of work has looked more directly at how words may influence the formation of concepts in young infants. Waxman and Markow (1995) first showed 12- to 13-month-old infants a series of toys from the same category (e.g., animals), one at a time. For each toy, the experimenters either labelled the toy (e.g., "Look, an animal") or didn't (e.g., "Look what's here"). After this familiarisation phase, infants were shown two objects simultaneously, one from the same category (e.g., another animal) and one from another category (e.g., a vehicle). The results revealed that infants familiarised to categories faster and showed a greater preference for the object from the novel category. In particular, when hearing a label, infants' attention to new exemplars from the same category decreased faster, and they subsequently showed a greater preference for items from a novel category. Waxman and Markow conclude that even in these earliest stages of language acquisition, labels act as "invitations to form categories" (p. 298). Waxman and colleagues have extended these results to show that these effects are already found in 6- and 9-month-old infants (Balaban and Waxman 1997; Fulkerson and Waxman 2007), that they rely on the use of a consistent word label (Waxman and Braun 2005), and that they do not occur for non-word auditory stimuli such as tones (Fulkerson and Waxman 2007). Moreover, Plunkett, Hu and Cohen (2008) showed that when infants were trained to distinguish two novel categories, they were able to do so without any labels or when the items from the two categories were consistently given two corresponding labels. However, infants failed to learn the two categories if the labels and categories were uncorrelated, or if only one label was used.

It should be noted that although efforts have been made to identify word labels as the specific source of the effects discussed in this section, this does not imply that they are the only source. Indeed, based on carefully coded video data during story-reading with 2-3 year-olds, Gelman, Coley, Rosengran, Hartman and Pappas (1998) have argued that influential maternal input on early conceptual development consists of much "beyond simple labelling routines" (p. v), and includes both non-linguistic cues (e.g., gestures) and subtle linguistic guides (e.g., generic noun phrases). However, as Keil's (1998) commentary points out, it remains unclear how much of this rich information actually reaches and is exploited by the child. Thus, although word labels do appear to play a special role in early development, it is not clear how important they are relative to other factors and kinds of input.

3.2.2 *Children*

While the previous section was concerned with the developmental effects of words in the earliest stages of language acquisition, researchers have also looked at how lexical input affects children who have already acquired substantial language. This period of development is important, because children are already actively using word-concept associations while still being very much in the process of acquiring language and developing concepts.

A key issue here concerns the role of word labels in children's induction: how much do children rely on the labels of others to infer the unseen properties of objects? Since young children have been shown to rely heavily on superficial perceptual properties in cognitive tasks (Flavell 1963), Gelman and Markman (1986) tested four-year-olds by pitting perceptual similarity against category labels in a property-induction task. Participants were first shown pictures of two different objects. Information was given about each object using a sentence which provided both a category label for the object and an unseen property of it (e.g., "This dinosaur has cold blood", "This rhinoceros has warm blood"). Children were then asked to infer the appropriate property of a third object, which was more perceptually similar to one of the preceding objects but shared a label with the other one. Children tended to make their choice based on a shared label, rather than shared properties. Davidson and Gelman (1990) conducted further experiments which qualified these results. They showed that if the category labels were novel, and they showed no correspondence with appearances, children resorted to perceptual properties. On the other hand, if labels were familiar or there was some correspondence between labels and features, four-year-olds still resorted to labels. However, in a study involving three age groups (i.e., 4-5, 7-8, and 11-12) Sloutsky, Lo and Fisher (2001) demonstrated that the effects of labels in category induction increases with age. The youngest children's reliance on labels depended on the degree of perceptual similarity among the items, the oldest children relied entirely on labels, and the 7- to 8-year-olds showed an intermediate transitional pattern. Based on these results, Sloutsky et al. proposed a developmental model in which children shift from treating words as category attributes to category identifiers. Together, these results suggest that the effects of other people's words on a child's conceptualisations can override perceptual similarity, and that children's trust of words as conceptual indicators is not blind and yet increases with age.

While such studies concerning category induction in object domains constitute the bulk of developmental studies of relevance here, it is worth noting that word input has also been shown to enhance children's abilities to perform other cognitive tasks, and not always with object concepts. To illustrate, Loewenstein and Gentner (2005) studied how hearing words for spatial relations might help children perform a relatively difficult mapping task. In a first experiment, children watched as the experimenter put an object either on, in, or under a box. While placing the object, the experimenter either also indicated it overtly with a relation term (e.g., “I’m putting this on the box”) or not (e.g., “I’m putting this here”). The child then closed their eyes while the experimenter placed an object in the corresponding position of another box, and then the child guessed its location. Although the relation term did not provide any new information, children performed better on this task if they had heard the relational term. Follow-up experiments showed that older children also benefited from hearing relational terms if the task was harder, that the effects disappeared if inappropriate spatial terms were used, and that the effects persisted when the children were re-tested two days later.

3.2.3 *Adults*

As shown in the last two sections, developmental evidence suggests that words play a causal role in conceptual development and conceptualisation. Moreover, the increasing reliance with age on word labels in category induction (Sloutsky et al. 2001) suggests how adult conceptualisation is likely to be particularly prone to linguistic effects. Indeed, Gelman and Markman (1986) preceded their developmental induction experiment by confirming that adults relied on category labels to infer hidden object properties. This is not surprising, since compared to children, adults have well-developed lexicons and are accustomed to using language for learning about the world through indirect means (e.g., in university lectures). Indeed, it is important to point out that although adults may arguably not rely on others as much as children do to learn about the world around them, adult conceptualisation is also very much affected by the words of others. In this section, I point out a few studies that have demonstrated that further.

In a classic study, Carmichael, Hogan and Walters (1932) first showed participants a series of ambiguous figures. Presentation of each figure was accompanied by one of two verbal labels (except in a control condition). For instance, a circular figure with loops around its diameter was labelled as either “ship’s wheel” or “sun”. Participants were then asked to redraw the items they had seen, and their drawings were coded blindly for various criteria. The results showed that relative to the original figures, participants’ drawings tended to be transformed to more closely match the labels they had heard. Similar effects have been found more recently in other domains. Billman and Krych (1998) studied the effects of labelling on participants’ recognition of motion events. Participants were shown short videos depicting events and simultaneously heard auditory labels that described either the manner or path of motion (e.g., a child "skipping" across a room to "exit" it). The next day, participants were tested on their memory of these events, by showing them variations where either the path or manner was different. Recognition was affected by the type of verb that had been used: when an event had been labelled with a path verb, errors were more likely for events with changes in manner, and vice versa. Similarly, Feist and Gentner (2007) tested spatial terms' effects on people's memory. Participants were first shown a set of pictures depicting borderline examples of spatial relations (e.g., for "on", a balloon was touching the surface of the table, but was actually hanging from a support above), possibly accompanied with a sentence that depicted it. After a ten-minute filler task, they were then tested on their recognition of the original stimuli, using pictures that either exemplified the spatial relation's prototype or deviated still further from it. The results showed that when the original relations had been labelled linguistically, participants made more recognition errors with the prototypical examples of the spatial relations; this was not the case when participants did not get this linguistic input. Together, such experimental results are consistent with Lupyan's (2008*b*) findings that category labels can actually skew and thus worsen item memory (see Section 2.6.5).

While the experiments above deal with words and concepts that participants are already familiar with, adults too learn new words and concepts throughout their lives. In a recent experiment, Lupyan, Rakison and McClelland (2007) investigated whether redundant category labels enhance category learning in adults. Participants were first trained on two types of alien-like figures. After presentation of each alien, there was feedback as to whether it was friendly or dangerous; moreover, in the label condition, this feedback also included a category label (i.e., "leebish" or "grecious"). Even though the labels were completely redundant, participants performed better in subsequent testing (as to whether aliens were friendly or dangerous) if they had been in the label condition. In a follow-up experiment, Lupyan et al. showed that these effects occurred regardless of whether the labels were presented visually or auditorily, but did not happen when participants were provided instead with a non-linguistic association (i.e., whether the alien lived above or below). These results with normal adults are reminiscent of the infant studies which suggested that words act as "invitations to form categories" (Waxman and Markow 1995). Indeed, even when adults know that they are trying to learn perceptual categories and they get formally equivalent alternative feedback, words still enhance their abilities to learn them.

3.2.4 Conclusion

In this section, I have discussed some studies which have shown that lexical input from other people does not just communicate ideas or drive our word learning, but also affects the development and use of our conceptual system. Moreover, these effects take place throughout our lives, from before we produce our first words ourselves right through our adulthood. Hearing words affects how young infants attend to things and what concepts they form. It drives children's inferences about the properties of things and enhances their cognitive abilities. And it influences how adults encode things and events that they experience and improves adult category learning.

In themselves, these findings show only that hearing words affects our conceptual system, but says nothing directly about conceptual coordination. Just as eating the same food affects people but does so in different ways (due to differences in taste, allergies, medical conditions, etc), it is theoretically possible that words may affect our conceptualisations without necessarily bringing them closer together. For example, there may in principle be more variation in people's linguistic concepts than there is between the way that they conceptualise non-linguistically "by default". Therefore, conceptual coordination does not logically follow and still needs to be tested. Nevertheless, these findings do provide good motivation for thinking that words might bring about conceptual coordination.

In particular, there are two ways in which the words of others may coordinate concepts, corresponding to two different timescales. First, over the course of our lifetimes, and perhaps especially in childhood, repeated exposure to hearing words from the same language may result in us developing concepts that are similar to the people around us. As a result, people with the same native language may end up with similar concepts. Second, when people interact linguistically, they may coordinate their conceptualisations of things in the world. Such online coordination may or may not persist beyond the interaction or affect their long-term conceptual development. Indeed, these two possibilities constitute the focus of this thesis, and provide the empirical hypotheses for my experiments. I therefore review relevant work for these two possibilities in the next two sections, respectively.

3.3 Linguistic relativity

In this section, I first introduce the Sapir-Whorf hypothesis. The bulk of this section then consists of a review of empirical work in the area. Next I briefly discuss recent and more subtle reformulations of the hypothesis and how they match up to the empirical data. Finally, I draw conclusions concerning the effects of language on conceptual coordination and further work that is needed.

3.3.1 Whorf

The question of how our native language may affect the way we think has been of interest for millenia, but nowadays is most commonly associated with the American anthropologist Benjamin Lee Whorf. In the 1940's, Whorf put forth strong views regarding the relationship between language and thought (Whorf 1956). Using cross-linguistic research for support, he expressed ideas that have become known as the "Sapir-Whorf hypothesis", or the "linguistic relativity hypothesis". In its strong form, this is basically the following argument: (1) there are substantial cross-linguistic differences in semantic structure, and (2) linguistic categories determine aspects of non-linguistic thought; therefore (3) speakers of different languages think differently.

Whorf's views were popular for a time, but came under increasing attack during the rise of the cognitive sciences in the 60's and 70's. The general dominant position became that "(1) human conceptual structure is relatively constant in its core features across cultures, and (2) conceptual structure and semantic structure are closely coupled" (Gentner and Goldin-Meadow 2003*a*, p. 5). Added to methodological criticism of Whorf's empirical work and findings of seemingly language-independent cognition in the domain of colour (Berlin and Kay 1969; Heider 1972*a*), the strong Whorf hypothesis was convincingly rejected and work in the area was widely abandoned.

It should be pointed out that the idea of linguistic relativity is not monolithic and does not actually constitute a single well-defined hypothesis. Moreover, as we will see, the body of empirical results makes taking an extreme position (i.e., that language fully determines thought or that it has no effect on it) untenable, as is evident from the positions of most researchers (Gleitman and Papafragou 2005; Carruthers 2002). But in moving away from a black-and-white view, the number of questions quickly expands as we consider the issues in more detail: *how much which* aspects of one's native language affect *which* aspects of thought in *what* ways.

Note that Whorf was primarily interested in grammatical effects on thought. For instance, he noted that Hopi does not encode verb tense, and argued that its speakers must therefore have a different understanding of time than (for example) English speakers (Whorf 1956). Much of linguistic relativity research has looked into cases where one language makes certain formal grammatical distinctions that another does not, including gender (Sera, Elieff, Burch, Forbes and Rodríguez 2002), event structure (Papafragou, Massey and Gleitman 2002), spatial frames of reference (Levinson 2003) and count/mass nouns (Imai and Mazuka 2007).

However, the idea of linguistic relativity could also be interpreted in a lexical sense. Perhaps the words of a language and how they divide up the world into categories has an impact on non-linguistic cognition. Indeed, this possibility seems to be embodied in probably the most frequently cited formulation of Whorf's ideas:

> We dissect nature along lines laid down by our native languages. The categories and types that we isolate from the world of phenomena we do not find there because they stare every observer in the face; on the contrary, the world is presented in a kaleidoscopic flux of impressions which has to be organised by our minds - and this means largely by the linguistic systems of our minds. (Whorf 1956, p. 213)

In this view, the natural world is chaotic and has no predetermined structure, reminiscent of William James' characterisation of the infant's world as "buzzing, blooming confusion" (James 1890/1981, p. 462). According to Whorf (1956), it is our native language that gives it structure and determines concepts and categories. As such, people will conceptualise things differently to the extent that their native languages provide them with different concepts.

Whorf's (1956) claim has a direct connection to my thesis of language-induced conceptual alignment. If a language provides a person with a distinct repertoire of concepts, then speakers of the same language should have more similar concepts and therefore conceptualise things more similarly to each other than speakers of different languages. Importantly, the claim is that language is responsible for such alignment, rather than other cultural or environmental factors. Moreover, it's important to emphasise that in this view, language's role is not just invoked during linguistic categorisation, but applies to conceptualisation in general. Merely noting that speakers of different languages have different lexical concepts is thus not enough to provide support for linguistic relativity.

In the next few sections, I will review empirical investigations of the lexical version of the Sapir-Whorf hypothesis. I focus primarily on two domains: colour and objects. Colour is important because it has been the longest and most intensively studied domain, and demonstrates both how complex the questions and findings are, as well as how much we still do not know. I then focus on objects because I used object stimuli in all of my experiments (see Chapters 5-8). Afterwards, I will also briefly mention some key findings from other domains to round off the picture. Finally, I will consider recent and subtle theoretical formulations of linguistic relativity which have arisen out of these investigations, and how they relate to my hypotheses.

3.3.2 Colour

The first and most vigorously explored domain has been that of colour, and there are good reasons why. Languages vary significantly in how they lexically partition the colour space (Kay, Berlin, Maffi and Merrifield 1997). It is comparatively easy to objectively control and manipulate colour stimuli in psychology experiments, and even to compare colour term systems with the distribution of colours in the perceived world (Yendrikhovskij 2001). Moreover, children get proficient with colour terms relatively slowly (Bornstein 1985), allowing for tracking of the phenomenon at various ages.

An influential early study by Brown and Lenneberg (1954) suggested a relationship between colour naming and recognition. They asked native English speakers to name colour stimuli, and coded their output in various ways, including length of names, reaction times, and agreement between speakers. Another set of participants was first shown colour stimuli and then asked to identify the ones they had been shown among a larger set. Brown and Lenneberg found that certain lexical variables were related to recognition: in particular, colour stimuli which were given shorter names, were named more quickly, and exhibited less variation among the speakers, were also remembered more accurately.

Subsequent studies yielded mixed results (e.g., Burnham and Clark 1955; Lenneberg 1961; Lantz and Stefflre 1964; Stefflre, Castillo and Moreley 1966), but then two papers were published which had a devastating impact on the Whorfian hypothesis in general. First, Berlin and Kay (1969) carried out a large cross-linguistic survey, analysing how different languages divided up the colour space. Their analysis was conducted in terms of "basic" colour terms, which they characterised as words which consisted of a single morpheme, were of general use, and were known and used by all native speakers of a language; this influential notion has formed the foundation for most subsequent work in this area. Berlin and Kay found systematic patterns across languages, on whose basis they proposed a universal evolutionary trajectory for basic colour inventories. The number of colour terms increased along the stages of the trajectory, but at any particular stage the space was more or less fixed, with the colour terms centering on universal "focal points" in the colour space (i.e., points which people largely agreed were the best exemplars of a colour). Thus, the way different languages partitioned the colour space was a product of fixed properties of human perception and biology, rather than arbitrary language-specific conventions.

This universalist view was corroborated by two important empirical papers by Eleanor Rosch (Heider) a few years later. Heider (1972*b*) developed Berlin and Kay's (1969) work on focal points and conducted several cross-linguistic experiments investigating colour naming and memory. In one experiment involving speakers of 23 different languages, she found that focal colours were given shorter names and named faster than non-focal colours. And in two follow-up experiments focusing on speakers of English and Dani (a language with only two basic colour terms), they found that speakers of both languages remembered focal colours better than non-focal colours, and that Dani were better at learning to associate words with focal than non-focal colours. Heider and Olivier (1972) ran another study with English and Dani speakers, which investigated whether there were correspondences between naming and memory within languages. Participants carried out a naming task in which they named colour stimuli, and a memory task in which they looked for a previously seen stimulus among an array. Using multi-dimensional scaling techniques, Heider and Olivier reconstructed a naming space and a memory space for speakers of both languages. They found that there was more divergence between the resultant naming spaces than between the memory spaces, and that there was no correspondence between the naming and memory spaces for either language. Together with the Berlin and Kay survey, these studies strengthened the universalist position, and research in this area was quite limited over the next couple of decades, though not entirely abandoned (e.g., Kay and Kempton 1984; Lucy and Shweder 1979; Bornstein 1985).

However, at the turn of the century, the universalist position was seriously challenged. Focusing on Berinmo, a language with five basic colour terms, Roberson, Davies and Davidoff (2000) (see also Davidoff, Davies and Roberson 1999) conducted a series of experiments which set out to replicate some of the key Dani results (Heider 1972*b*; Heider and Olivier 1972), and explore the issues further. Their results were generally in conflict with Rosch's: they found that Berinmo memory corresponded more closely to Berinmo naming than to English memory, that Berinmo did not remember focal colours better than non-focals (once they accounted for response bias), and that the Berinmo were no better at learning new associations for focal rather than non-focal colours (except for red, but this is also the centre of a Berinmo category). Then, in several experiments with different kinds of tasks, they tested for categorical perception with Berinmo and English speakers, and found that regardless of the paradigm, categorical perception effects did occur and that they were consistent with the speakers' native language. In discussing the results, Roberson et al. also criticised some aspects of Rosch's methodology and interpretation, including the use of a biased stimulus set and unjustified emphasis on only those of her analyses which supported her conclusions. This paper thus gave strong new support for linguistic relativity.

Roberson et al.'s (2000) experiments revitalised linguistic relativity research in the colour domain. Roberson, Davidoff, Davies and Shapiro (2005) extended their findings to Himba, another language with five basic colour terms. They found the same general pattern of results as for Berinmo: Himba memory matched Himba naming better than it did English memory, there was no evidence for a focal colour advantage in memory tasks, and language-specific categorical perception effects were found using various methodologies. Winawer et al. (2007) conducted a categorical perception experiment with Russian and English. Russian makes a category distinction which is absent in English, having separate basic colour terms for (approximately) light and dark blue. Winawer et al. elicited the Russian boundary and then used it as the basis for a simple perceptual matching task, in which they found a clear categorical perception effect for the Russian (but not the English) speakers. Moreover, this effect disappeared when participants were given a verbal interference task, but not when given a spatial interference task. Athanasopoulos (2009) took up a similar case with Greek (another language with a basic distinction between light and dark blue), but focused on bilingual Greek-English speakers. He found that the focal colours shifted for participants with different levels of English. Moreover, although there was no clear difference between bilinguals and monolingual English speakers which would reveal a categorical perception effect, results correlated with two of the other independent variables that Athanasopoulos measured: salience of colour terms and length of time spent in an English-speaking country. While the results of this study are not unequivocal, they do mark the first bilingual study concerning colour and linguistic relativity.

Although the above studies, focusing on particular languages, have generally found evidence for linguistic relativity, a few other studies analysed data from the World Colour Survey (WCS) in various ways which tended to support a universalist view. The WCS was compiled in an attempt to improve on Berlin and Kay's (1969) survey, after criticisms that it had consisted mainly of languages spoken in industrialised societies and that much of the data was collected from bilingual speakers who also spoke English (e.g., Hickerson 1971). The WCS has colour term data from 110 languages spoken in non-industrialised societies, gathered mostly from monolingual speakers. Kay and Regier (2003) used mathematical clustering techniques and a psychologically meaningful colour space to plot the WCS colour term systems together, and found that this resulted in much less variation between languages than there was in hypothetical randomly generated datasets. In addition, when they made an overall comparison between the WCS colour terms and those from Berlin and Kay (1969), they again found a correspondence much greater than expected by chance. Regier, Kay and Cook (2005) followed this up by looking at focal colours across the WCS dataset, and found that they generally lined up well with English focal colours. Moreover, comparing again data from the two surveys, there was a better correspondence in focal colours than there was in category extensions.

Regier et al. argued for a universal tendency towards certain points in the colour space. Similar results were obtained by Lindsey and Brown (2006) by applying additional statistical techniques, such as cluster and concordance analyses, to the WCS data. Their analyses suggested that although there may be differences between languages, there are still universal language-independent tendencies in colour term systems, and these seem to derive from biologically biased points in the colour space. A corroborating conclusion was reached from a cross-linguistic study in which participants both named colour stimuli and freely sorted them into groups (Roberson, Davies, Corbett and Vandervyver 2005). The results showed that although there were substantial differences between the languages in naming, colour sorting exhibited less cross-linguistic variation. Roberson, Davies, Corbett and Vandervyver explained the patterns with a hybrid model, combining both universalist and relativist components.

Some researchers have also investigated developmental issues. Roberson, Davidoff, Davies and Shapiro (2004) carried out a longitudinal study, assessing 3-year-old (at the beginning) English-speaking and Himba-speaking children six times over a period of three years. The children carried out 12 naming, comprehension and memory tasks, which yielded several findings. First of all, prior to learning any colour terms, children tended to perform similarly on the memory tasks, but the patterns appeared to be based on perceptual distance rather than the 11 basic categories of English. Secondly, there was a lot of variation in the order in which children learned colour terms in both populations. Third, a memory advantage for focal colours was absent at first, but increased longitudinally; and this advantage was specific to the focal colours of children's native languages. Fourth, learning colour terms was not an instantaneous process: even after initially learning colour terms, children would only gradually learn their extensions over a couple of years. Roberson et al. argue that these findings are in conflict with a universalist position, according to which children would just need to learn the names for their pre-existing colour categories.

However, the developmental evidence is varied. Franklin and Davies (2004) conducted three experiments with 4-month-old infants, using a looking time technique, and found that they showed categorical perception effects for adult colour categories. Franklin, Clifford, Williamson and Davies (2005) then followed this up with experiments with naming, comprehension and matching tasks, involving young English-speaking and

Himba-speaking children. The results were consistent with Franklin and Davies (2004) but quite different from those of Roberson et al. (2004). Both Himba and English children showed categorical perception effects, even for category boundaries that existed in English but not in Himba. Moreover, the effect was no larger for children with more developed colour vocabularies. Goldstein, Davidoff and Roberson (2009) sought to reconcile the findings of Roberson et al. and Franklin et al. with two more developmental experiments. With English children, they replicated Franklin et al.'s (2005) results when using their methods of analysis, but found that using a stricter criterion for colour term knowledge, the categorical perception effect disappeared for the children with less colour term understanding. With Himba children, they found a categorical perception effect for only one of the two category boundaries they investigated, and argued that even this could be explained in terms of other linguistic and cultural factors. The debate was continued by Franklin, Wright and Davies (2009), who found their original results unchanged when they reanalysed their data using Goldstein et al.'s adapted criterion, and discussed other work which had presented findings in conflict with those of Goldstein et al. They also pointed out theoretical shortcomings in this line of work, and called for new methods to investigate the issues.

A recent study has also shed light on the debate in an intriguing new way. Gilbert, Regier, Kay and Ivry (2006) pioneered an investigation based on a consideration of brain lateralisation. They hypothesised that, due to dominance of the left hemisphere of the brain in most language tasks (Hellige 1993), and the contralateralisation of the primate visual system (Tootell, Silverman, Hamilton, Switkes and De Valois 1988), language may be more implicated in perception in the right visual field (RVF) than the left visual field (LVF). They tested this possibility with two experiments in which participants looked for a target stimulus among a ring of otherwise identical stimuli. The results confirmed the hypothesis. In the RVF, 13 participants were significantly faster at finding between-category rather than within-category targets; no difference was found in the LVF. Moreover, these results were reversed when participants were also given a verbal interference task: in this case, they were actually faster on within-category than between-category trials. A third experiment sought evidence that the source of the effects was indeed brain lateralisation. The same experimental procedure was carried out by a split-brain patient with lesions in the left hemisphere. The results confirmed that the patient was much slower in general when the target was in the RVF, and that only in the RVF was there a between-category advantage.

There have been a couple of follow-ups to Gilbert et al.'s (2006) study. Drivonikou et al. (2007) first examined data from two previous experiments in which a target colour was searched among distractors (Daoutis, Pilling and Davies 2006), and reanalysed it in terms of the LVF and RVF. They also conducted their own two experiments, in which participants saw a single stimulus on the background of a similar colour. Participants indicated on which side of the display the stimulus appeared, and their reaction times were measured. The results from both approaches were similar though not identical to those of Gilbert et al.: in general, the category effect was stronger in the RVF than in the LVF, but there was still an effect in the LVF as well. Drivonikou et al. speculated that the LVF effects could be due to either language affecting both visual fields (albeit, not equally), or that they indicate universal category distinctions. Roberson, Pak and Hanley (2008) decided to test these possibilities. Noting that Korean made a colour distinction which English did not, they conducted an experiment testing for corresponding categorical perception effects with Korean and English participants. They used the same visual search paradigm with a target among a ring of distractors as Gilbert et al. and Drivonikou et al. The results revealed a between-category advantage for the Korean speakers in both visual fields, and in neither for the English speakers. Roberson et al. argued on the basis of these results that the LVF effect is also due to language, and not to universal categories. However, Franklin et al. (2008) extended this line of research to infants, from which they drew different conclusions. They compared performance in the visual fields for both adults and infants, and found that while adults again did better in the RVF, prelinguistic infants actually performed better in the LVF. They argued from this that only the RVF effect was language-driven.

A few other colour experiments have added evidence from neuropsychological patients. Roberson, Davidoff and Braisby (1999) conducted a case study with a patient who had suffered a left hemisphere stroke which left him with intact visual processing but impaired naming. They compared his performance with controls on several kinds of tasks, and found that he had normal performance on some tasks (e.g., recognising a previously seen colour or choosing the "odd" one out of three colours), but heavily impaired in others (e.g., naming colours and freely classifying them into groups). They argued from the pattern of results that he had intact and normal implicit knowledge of colour categories, but was unable to explicitly use it. However, Haslam et al. (2007) carried out a follow-up longitudinal study which looked further into free classification performance in another patient with declining language abilities. They first emphasised the need for an objective measure to compare two people's sets of categories (which they stressed was lacking in Roberson et al.'s study), and found substantial variation among control participants upon applying one. They then tested the patient in both naming and free classification tasks across three different sessions. The results showed that although his performance in naming underwent profound degradation, his categorisation was similar to that of controls, and did not diverge across the sessions. Haslam et al. concluded, contra Roberson et al., that colour categorisation does not depend on language.

In summary, the multi-pronged investigations of linguistic relativity in the colour domain have yielded a wealth of data with varying results. While questions remain, it is becoming increasingly clear that it is oversimplistic to expect a simple black-and-white answer concerning language's role in colour cognition. As Regier and Kay (2009) argue, it is long time to end the battle between relativists and universalists. Language does play a role in colour conceptualisation, but its impact is limited and balanced by other factors, particularly human biology and perception.

3.3.3 Objects

Object stimuli have also been extensively used in linguistic relativity research. However, for the most part the purpose has been to study the potential impacts that different grammatical distinctions might have. For instance, languages vary in whether they require count/mass or gender distinctions, and researchers have examined whether these distinctions have cognitive consequences for how objects are conceptualised (e.g., Sera et al. 2002; Imai and Mazuka 2007; Lucy and Gaskins 2001; Boroditsky, Schmidt and Phillips 2003).

However, as indicated in Section 3.3.1, I focus here on lexical effects, and these have been much less studied. An important exception consists of a series of experiments conducted by Barbara Malt and her colleagues. These experiments have investigated whether differences in how object domains get lexically partitioned correspond with how the objects are perceived and categorised non-linguistically. In a first experiment (Malt, Sloman, Gennari, Shi and Wang 1999), native speakers of English, Spanish and Mandarin carried out two kinds of tasks with sixty pictures of every-day containers (i.e., bottles, jars, etc.). First, they sorted the items into groups based on similarity. The numbers, sizes, and nature of the categories was up to the participants. After that, they were asked to label each item with the way they would normally refer to it. Linguistic categories were then induced by considering what head nouns (e.g., the head noun of "big, red bottle" is "bottle") were used for each item. Upon comparing category patterns, the results showed that the different language groups partitioned the set of items differently in naming, but achieved quite similar groupings in sorting. Moreover, there was no correspondence between the naming and sorting patterns within a language. These results suggested that languages have different lexical concepts, but that these do not affect how we group or perceive things in general. Note that my first experiment in this thesis was largely a replication of Malt et al.'s study, so I will discuss it in more detail in Section 5.2.

Since their original study, Malt and colleagues have pursued several further issues regarding the relationship between linguistic and similarity-based categories. One study (Malt and Sloman 2003) looked at the naming patterns of non-native speakers of English in the United States, and discovered that there were differences between their (English) linguistic categories and those of native English speakers. Moreover, they found that these differences were smaller for speakers with more English experience, and that, in this regard, the length of immersion in an English-speaking environment was more relevant than the amount of formal instruction or age of introduction to English. This study did not, however, involve similarity-based sorting tasks.

In another study, Ameel, Storms, Malt and Sloman (2005) looked at the correspondences between naming and sorting patterns in bilingual speakers of Dutch and French, compared to monolinguals. The procedure was very similar to Malt et al.'s (1999) original experiment, with the main difference being that bilinguals did the naming task in both languages. The analysis first replicated the previous findings: Dutch and French monolinguals exhibited different naming patterns from each other, even though they sorted the items in much the same way. In addition, a closer look at the data, and elaborated further by Ameel, Malt, Storms and Van Assche (2009), showed that the bilinguals' naming categories in the two different languages were very close to each other, and lay in between the monolingual French and Dutch ones.

Another set of experiments (Ameel, Malt and Storms 2008) investigated developmental aspects of these issues. Participants were native Dutch speakers, and included children of several ages between 5 and 14, and monolingual adults. The naming patterns for the different age groups showed a gradual convergence of the children's linguistic categories onto those of adults. Ameel et al. also explored the features which participants were using to label the items, and found that although children initially seemed to categorise in a more holistic way than adults, they gradually learned to attend to the adult feature set. Finally, sorting patterns also suggested that younger children judged similarity between items differently from adults, but again converged on the adult patterns with time. However, Ameel et al. point out that the fact that convergence occurs in both naming and sorting does not imply that naming has a causal role. Indeed, the direction may be reversed, so that sensitivity to more complex features emerging during development may be required to fully learn adult lexical categories. Alternatively, given the dissociation between linguistic and non-linguistic categorisation found by Malt et al. (1999), the two may develop relatively independently.

3.3.4 Other domains

3.3.4.1 Space

After colour, the most studied domain in linguistic relativity, especially in recent years, is probably that of space. This is because spatial relations also exhibit substantial cross-linguistic variation (Talmy 1983), but are less likely to be constrained by biological aspects of human perception, and have the potential for more far-reaching cognitive consequences (Gentner and Goldin-Meadow 2003*b*).

Choi and Bowerman (1991) explored a few basic spatial relations in Korean and English experimentally, and showed that the spatial categories of the two languages cross-cut and overlapped each other in clear and significant ways. This means that the languages required their speakers to make different distinctions in their language use. McDonough, Choi and Mandler (2003) followed this up and showed that adult native speakers of a language were insensitive to some distinctions important in the other language, even though this was not yet the case for prelinguistic infants from the same cultural background. In contrast, Munnich, Landau and Dosher (2001) compared adult speakers of three languages in spatial language and memory tasks, but found that despite certain key lexical differences between the languages, there were no corresponding cross-linguistic differences in memory performance.

In a related line of research, Levinson (2003) has explored the cross-linguistic variation in spatial frames of reference, which concern (at the risk of oversimplifying) how the relative locations and orientations of things are expressed (e.g., "in front of me", "to the north of me"). He conducted experiments involving rotating tables and memory tasks to show that the type of frame of reference typically employed by one's native language has consequences for how one encodes spatial configurations and events. Other experiments have found that the non-linguistic "defaults" of infants and apes coincide on the same general frame of reference, suggesting that 16 languages which use different systems must "override" those defaults (Haun, Call, Janzen and Levinson 2006; Haun, Rapold, Call, Janzen and Levinson 2006). However, there are also findings which argue against Whorfian effects, or at least in favour of weaker ones: for instance, Li and Gleitman (2002) have shown that the kinds of effects found by Levinson are limited, in that they can be overridden by contextual factors, such as the presence of conspicuous landmarks. Thus, overall, findings in the spatial domain have generally supported claims of linguistic relativity, although the interpretation and scope of the results needs to be treated with caution.

There have also been studies concerned specifically with motion events. Gennari, Sloman, Malt and Fitch (2002) investigated whether the conflation pattern differences in linguistic event encoding between English and Spanish (Talmy 1985) had consequences in recognition and similarity tasks. Using triads of short video clips in which the target clip had one alternative which varied in path, and another which varied in manner, they confirmed the linguistic differences via naming tasks, but generally found little correlation between them and performance in non-linguistic tasks. However, they did find that if Spanish speakers desribed the clips linguistically while first viewing them, it affected their performance on subsequent similarity tasks in a way which reflected the Spanish language's emphasis on path over manner. Thus, while these results were generally anti-Whorfian, they do suggest that language may have an effect in the encoding of events online. A similar study was conducted by Papafragou et al. (2002) with Greek and English. In this case, the results were more unequivocal: the languages differed in how they encoded motion, but participants performed identically on non-linguistic tasks.

3.3.4.2 *Number*

In the domain of number, there are languages which do not have full counting systems, but only have a few words for the smallest integers, and then global terms like "few" or "many". For instance, Mundurukú uses count words for 1, 2 and 3, and arguably 4 and 5 (Pica, Lerner, Izard and Dehaene 2004), while Pirahã does not even use 1 and 2 consistently (Gordon 2004). Does this mean that native speakers of these languages have a difficult time dealing with non-trivial numerosities? It has often been argued that language is necessary for the development of number concepts (Hurford 1987; Carey 2004). Empirical results on this question are mixed: some experiments seem to show that numerical cognition is limited in speakers of such languages (Gordon 2004), while others find that there is little or no impact (Gelman and Gallistel 2004). Nevertheless, if there are language-induced limitations, they do not seem to be insurmountable, since there is evidence (for example) that speakers of these languages can easily learn counting words from other languages if there are pressures for doing so (e.g., as in modern commerce, Dixon 1980).

3.3.4.3 *Time*

Some researchers have also looked at the domain of time. Although time is a more abstract domain and difficult to pin down non-linguistically, it has been argued that in many (if not all) cultures, people think of time in terms of space through broad conceptual metaphors (Lakoff and Johnson 1980). In a series of experiments, Boroditsky (2001) tested this proposal with native speakers of Mandarin and English. According to Boroditsky, while English predominantly frames temporal relations horizontally (e.g., "push deadlines back"), Mandarin also systematically frames them vertically (although English does so sporadically, e.g., "the meeting was coming up"). In line with this difference, Mandarin speakers processed temporal relations faster after being exposed to a vertical spatial prime (e.g., one ball above another ball) than a horizontal one, while the opposite was true for English speakers. However, if English speakers were trained on a set of vertical spatial terms, their behaviour on the vertical temporal relations began to mirror that of the Mandarin speakers. Boroditsky concludes that our native language powerfully shapes our thinking about time (and other abstract domains), but does not strictly determine it.

However, the reliability of Boroditsky's (2001) results has come under considerable attack. Both January and Kako (2007) and Chen (2007) failed to replicate her findings. In addition to criticising Boroditsky's (2001) interpretation of her results, January and Kako made six separate replication attempts with English speakers, and despite discussions with Boroditsky to resolve any methodological differences from the original studies, none of the experiments revealed a significant advantage of horizontal primes over vertical primes. Chen first searched the Internet and found that in fact, contra Boroditsky's assumption, Mandarin, like English, uses horizontal expressions for time more frequently than vertical expressions. Moreover, he also failed to replicate Boroditsky's findings despite four attempts with Chinese-English bilinguals.

3.3.5 *Theoretical positions*

In summary, then, the range of empirical studies on linguistic relativity has yielded mixed results. There appear to be universal and non-arbitrary aspects to colour cognition, but language-specific partitionings of the colour space do play a role as well. People are sometimes more sensitive to the spatial distinctions which are encoded by their language, and their memory can be affected by the spatial frame of reference normally employed by their language, but the extent and conditions of these phenomena are still unclear. Numerical cognition does seem to sometimes correlate with the range of number terms in a language, although these can be overcome through cross-cultural interaction. The spatial metaphors that languages use to talk about time might have an impact on thinking about time, but results are inconsistent. Object cognition seems to be uncorrelated with cross-linguistic differences, although there are exceptions, especially those involving obligatory grammatical distinctions.

In light of this growing body of subtle results, advocates of linguistic relativity have generally shifted from the strong claim that language determines thought to the weaker claim that language influences thought. However, as Carruthers (2002) points out, an ultra-weak Whorfian view (i.e., that language affects thought in some way) is as trivially true as an ultra-strong Whorfian view is trivially false. Consequently, the current challenge is to formulate precise weak versions of the Whorf hypothesis, and evaluate them based on empirical research (Hunt and Agnoli 1991). Essentially, the questions now comes down to: to what extent, how, when and why does language affect non-linguistic cognition?

A variety of "neo-Whorfian" views have been put forth. Hunt and Agnoli (1991) argue that in shifting to a weaker and more plausible version of linguistic relativity, the question should shift to how natural and computationally costly expressing the same thought is in different languages. Languages differ in what kinds of structural features they require and encode, so that "at any point in time a language user thinks most efficiently about those topics for which his or her lexicon has provided an efficient code" (p. 378). Levinson (2003) uses evidence mainly from cross-linguistic spatial cognition experiments to argue that, in order for it to be possible to later discuss experiences, it is necessary for people to encode the aspects of those experiences that their native languages emphasise: "Language is an output system. The output must meet the local semantic requirements. Consequently, the input to language production must code for the right distinctions" (p. 301). Slobin (1996) proposes a more conservative but more tacklable version of the Whorfian hypothesis, by shifting from the static notions of 'language' and 'thought' to the more dynamic notions of 'speaking' and 'thinking'. The basic idea is similar to Levinson's in its emphasis on the different distinctions that different languages require of their speakers, except that Slobin restricts his claims partly to the time of speaking: "the expression of experience in linguistic terms constitutes *thinking for speaking* - a special form of thought mobilised for communication" (p. 76). For example, people who speak a language which emphasises path over manner in motion may attend to path aspects of motion events more, but perhaps only if they are witnessing them while using language. Another view is that, when faced with a difficult cognitive task, people can rely on the distinctions of their native language (Kay and Kempton 1984). In this view, lexical classification can be exploited by people if needed as a sort of "naming strategy" when appropriate.

At the moment, it remains difficult to disambiguate between these different views. Although some researchers have made efforts to assess these views based on particular experiments (e.g., Gennari et al. 2002) or larger sets of studies (e.g., Levinson 2003), it is too early to make a single coherent story for all the data available. However, at the risk of speculation, perhaps the various neo-Whorfian proposals can be seen broadly as varying in strength, and which one best applies may depend on the domain and context under consideration. In some cases, the cognitive system seems to be significantly transformed by language in order to provide necessary input for it (e.g., orientation for languages with absolute frames of reference). In other cases, it may be merely more sensitive to distinctions made by the language, but can still handle other distinctions (e.g., categorical perception in colour). Still further, these differences may only occur when language is being used (e.g., processing of motion events). And then there may also be situations where it only comes up when people consciously use language as an explicit strategy (e.g., difficult colour discriminations). Finally, there may of course be cases where language has no impact on non-linguistic cognition at all. Therefore, it may turn out to be fruitless to seek one overarching explanation for all linguistic relativity phenomena.

3.3.6 Conclusion

Linguistic relativity research relates closely to one of the main questions in my thesis. In particular, does having the same native language coordinate the way people conceptualise things? While the evidence suggests that it does, the depth of this coordination is unclear, and seems to depend on the domain and context.

Interestingly, although the object domain plays a primary role in many disciplines concerned with concepts, ranging from experimental cognitive psychology to theoretical philosophy, it has seen little direct attention in investigations of linguistic relativity. The limited studies that have been done suggest that non-linguistic object categorisation is not influenced by one's native language. However, they (and other studies of this sort) have not addressed conceptual coordination.

Nevertheless, provided that we solve certain methodological challenges (see Chapter 4), the kinds of experiments used to study linguistic relativity could be readily accommodated for this purpose, and fit in well with the theoretical model I developed in Chapter 2.

Consider, for example, Malt et al.'s (1999) study. Again, native speakers of three different languages carried out two different categorisation tasks, one linguistic and one non-linguistic. Using data gathered in this way, it would be possible to pair up people (whether with speakers of different languages or the same language) and compare their categorisations (whether linguistic or non-linguistic). Such comparisons are shown in Figures 3.1 and 3.2 for linguistic and non-linguistic comparisons, respectively. Therefore, in Chapter 5, I replicate Malt et al.'s study while also using the data gathered to test for conceptual coordination.

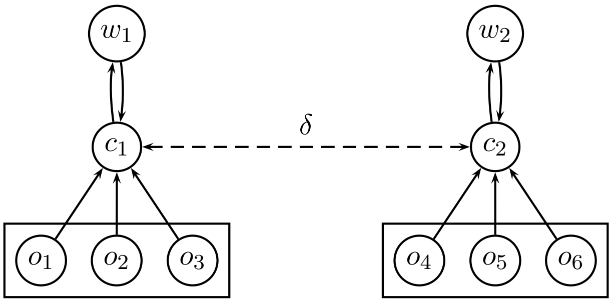


Figure 3.1: Comparison of two people's linguistic categorisations. The difference between their conceptualisations (assuming it can be quantified) is indicated with δ. Note that if the people are native speakers of the same language, then we may have $w_1 = w_2$. Also, there may be overlap in the objects being conceptualised; this is not incorporated in the diagram to avoid clutter and confusion.

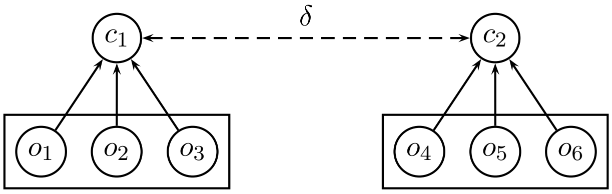


Figure 3.2: Same comparison as in Figure 3.1, except this time the categorisation is non-linguistic, so that concepts are not lexicalised.

3.4 Linguistic interaction

In this section, I move on to the other timescale that I have identified at which language could coordinate people's concepts. I begin by discussing how dialogue, which constitutes the most natural form of language use, is both a paradigmatic example and special component of social interaction. I then discuss how dialogue relies on shared information between interlocutors, and builds off of it. Next, I discuss a recent theory claiming that during dialogue, people align different levels of mental representation with each other. After that, two key sections review empirical evidence on conceptual alignment, first during and then after interaction. Finally, I summarise the discussion and point out important unanswered questions concerning conceptual alignment in dialogue.

3.4.1 *Joint action and dialogue*

Up to this point, we have looked at language and words from the point of view of the individual, and how long-term immersion in the same language might result in some degree of conceptual coordination between people. Even though I have begun to consider the implications of the shared nature of words on an individual's concepts and conceptualisations, I have still, in a sense, been modelling the individual as a mostly isolated input-output organism.

However, since hearing other people's words can affect how a person conceptualises something, and linguistic interaction typically involves a rapid exchange of many words, we need to look beyond this oversimplification. Indeed, Clark (1996) emphasises this issue in developing a theory of language use. He points out that language use has typically been studied from two broadly different perspectives. Cognitive scientists have generally adopted a "product approach", in which the focus is on the individual, treating hearers or speakers in decontextualised isolation. In contrast, social scientists have more often followed an "action approach", which centres on intentions and social actions but neglects the thoughts and actions of individuals. Clark opts strongly for the action approach, arguing that the product approach is fundamentally inadequate and incompatible with the social, interactive and contextual nature of language use. This is especially clear since face-to-face conversation seems to be the most fundamental use of language, as it is universal across human societies, does not require special skills, and is the basic setting of children's language acquisition. However, Clark (1996) also emphasises that a theory of language use must incorporate aspects of both individual and social cognition, since it must take into account both the intentions and actions of individuals and how they are embedded in joint processes.

Clark's (1996) view centres on the idea that language use is a form of "joint action". He identifies joint actions as those in which two or more people coordinate their individual actions. Participants may play different individual roles and carry out correspondingly different actions, but they achieve something together as a unit (e.g., two people playing a duet). Clark argues that linguistic dialogue is a particularly good example of joint action. This is because conversation does not involve an isolated speaker and an isolated listener, but rather the intimate coordination of speaking and listening processes between interlocutors attempting to achieve mutual understanding. Therefore, a theory of language use can benefit significantly from the study of joint action in general.

This is significant, because joint action is pervasive in human social life, and is exemplified in such diverse joint activities as buying food at the supermarket, playing a game of chess, or moving heavy furniture. Moreover, in addition to being itself a form of joint action, language use is also intimately involved as a means of achieving other forms of it. In fact, Clark (1996) argues that "language use and joint activity are inseparable" (p. 29). However, he also emphasises that some activities rely more on language than others, so that the extent to which joint activities rely on language varies along a "discourse continuum" (p. 50), ranging from mostly linguistic activities (e.g., talking on the phone) to mostly non-linguistic ones (e.g., playing a string quartet) .

Systematic study of joint action has been significantly developed in recent years, and it has not been confined to language-based studies. To illustrate, Richardson, Marsh, Isenhower, Goodman and Schmidt (2007) examined the physical coordination between people sitting side-by-side in rocking chairs. They found that participants tended to synchronise their rocking, even when they were asked to rock at their own pace and were given chairs with different natural rocking frequencies. Sebanz, Bekkering and Knoblich (2006) review a range of recent studies in developmental psychology, cognitive psychology and cognitive neuroscience, with the aim of identifying mechanisms which underly successful joint action. They conclude that joint action relies largely on the abilities to share representations and predict and integrate actions, and that these abilities in turn build off of a capacity for joint attention and a close link between perception and action. Moreover, Tomasello, Carpenter, Call, Behne and Moll (2005) argue that the abilities to share behaviours, goals and intentions emerge in stages very early in human development. According to them, human infants are able to interact with others towards a shared goal and with coordinated action plans from around the age of 12-15 months. This ability seems to underly language acquisition and precedes a full-fledged theory of mind by several years. In comparing human infants to non-human primates, Tomasello et al. also argue that these abilities are uniquely human.

To sum up, then, joint action is pervasive in human social life, and involves the coordination of individual actions by two or more people for a common purpose. This coordination seems to concern both mental representations and processes. While joint action is evident in non-linguistic human interaction, language use is intimately associated with joint action in two ways. First, even in activities that are not intrinsically linguistic and could in principle be carried out without language (e.g., playing football), participants often use language in the course of the activity. Second, language use itself, and especially dialogue, its canonical form, is a good example of joint action.

3.4.2 Common ground

3.4.2.1 Theoretical background

Participants in a conversation rely on lots of shared information about things, people and events in the world. Clark and Brennan (1991) emphasise its importance in conversational coordination: “they [interlocutors] cannot even begin to coordinate on content without assuming a vast amount of shared information or common ground” (p. 127). Common ground is a technical notion developed by Stalnaker (1978) (although it has earlier origins), who defined it as the knowledge that speakers believe they share in communication. Clark (1996) broadened the concept to include other types of representations shared by people, and not necessarily restricted to conversation: “Two people’s common ground is, in effect, the sum of their mutual, common, or joint knowledge, beliefs and suppositions” (p. 93). Clark uses the example of a chess game to show how common ground generally consists of different kinds of information, obtained at different temporal stages. At any point during a game of chess, the players share knowledge of how the game works in general, of how to interpret the board and pieces, and of popular strategies. But they also share knowledge of the current state of that particular game, and the steps that took place to get it there. Of course, as Clark points out, there may also be discrepancies between people’s representations of their common ground. When these are discovered, they can either be ignored (especially if they are small), or corrected (as is usually done, according to Clark).

But what is the basis for common ground? Clark and Marshall (1981) argue that it is not enough for people to independently have the same knowledge. For example, if I refer to "Sue's cat" while talking to you, this requires not just me to know Sue and that she has a cat, but also you to know that too, and for me to know that you know. In fact, Clark and Marshall show with an extended example that in general, the common ground that underlies conversation involves an infinite number of recursively defined conditions. In particular, two interlocutors A and B have mutual knowledge of a proposition p if A knows that p, B knows that p, A knows that B knows that p, B knows that A knows that p, A knows that B knows that A knows that p, etc, etc. Of course, people cannot in practice check such a list of conditions when constructing or interpreting an utterance, and Clark and Marshall emphasise that this would be psychologically implausible. Instead, they propose that common ground is checked via "copresence heuristics", by which people check their memory for evidence that they were "openly present together" with the referent in question, either directly or indirectly. Such evidence can be of three general types: community membership (interlocutors belong to the same community and hence can be expected to know certain things), physical co-presence (the referent is or was being attended to by both interlocutors) or linguistic co-presence (the referent has already been established earlier). For this to work, Clark and Marshall argue, our memory must contain information not just about the things we interact with, but also with whom we do so.

Clark and Brennan (1991) point out that, during conversation, interlocutors are not just passively relying on their common ground, but are constantly updating it. This process, which Clark and Brennan refer to as "grounding", occurs not just by a speaker mentioning new things in a conversation, but also by the interlocutors both making sure that the hearer understood them as intended. Clark and Brennan discuss how people resort to different kinds of strategies for achieving this depending on their purpose of communication. For instance, when identifying objects, they tend to provide alternative descriptions or use gestures like pointing, while for communicating something verbatim (e.g., a telephone number), they might resort to spelling. Clark and Brennan also explore how the medium of communication can shape grounding, identifying eight relevant dimensions along which different media differ, such as copresence and simultaneity. Depending on the features of a given medium, grounding may be easier or more difficult, and may require different methods, which can in turn result in additional processing costs. For example, communication via email does not normally involve sharing the same physical environment, and it is asynchronous. As such, it may often require more explicitness than face-to-face conversation.

3.4.2.2 *Interactive dialogue experiments*

Common ground has been experimentally shown to play an important role in communication. However, in order to contextualise some key examples, it's important to mention a paradigm which was originally used for somewhat different questions but which has since been recruited for many studies of common ground. In particular, Krauss and colleagues used a referential communication task with participant pairs in several studies. For instance, in one experiment (Krauss and Weinheimer 1966), the speaker had a card with six unusual figures on it, and the listener had several such cards with the same figures but in different orders. Their task was to get the listener to identify which of his cards had the figures in the same order as the speaker's cards. The results showed that over a series of such tasks, the speaker's referring expressions became shorter. Moreover, the decrease was greater when the speaker got either of two forms of feedback: concurrent feedback from the listener and confirmation of whether or not the listener identified the right card.

Clark and Wilkes-Gibbs (1986) explored how dialogue participants established reference. They point out that traditional theories of reference separate the speaker and hearer, with the speaker being fully in charge of selecting an appropriate noun phrase, and the hearer merely receiving and interpreting it. However, in real dialogue, hearers contribute to the conversation and speakers monitor hearers for understanding. Clark and Wilkes-Gibbs adapted Krauss and Weinheimer's (1966) method to investigate this issue. They found that speakers and hearers collaborated closely and gradually in the construction and interpretation of referential expressions. And indeed, their roles were not discrete: for instance, speakers used hedges or asked for confirmation of understanding, while listeners expressed confusion, made guesses, and proposed their own contributions. Negotiation continued until they reached "mutual acceptance" (p. 9) that understanding had been achieved. Clark and Wilkes-Gibbs thus proposed that speakers and hearers take "mutual responsibility" (p. 33) for the hearer's understanding of the speaker's intended meaning. Moreover, like Krauss and Weinheimer, they also found that referential expressions got shorter during interaction. They proposed that this was because interlocutors sought to "minimize collaborative effort" (p. 26), so that as they converged on a common understanding they could get away with more compact expressions.

Another important issue relating to common ground concerns one's degree of involvement in a conversation. Schober and Clark (1989) conducted an experiment with a referential matching task and a triad of participants: speaker, addressee and overhearer. The participants could hear but not see each other, and had the same set of figures before them. The speaker instructed the addressee on how to rearrange the figures in a certain order and the addressee could respond and ask for clarifications, while the overhearer only listened. Even though overhearers heard everything that speakers said, they rearranged the figures less efficiently than addressees. Wilkes-Gibbs and Clark (1992) looked into this issue further by considering different non-addressee (or "bystander") conditions and having the speaker switch to addressing them part-way through the experiment. They found that the efficiency with which the new pairs carried out the matching tasks depended on the condition that the addressee had been in as bystander. Efficiency was highest when the bystander had sat next to the speaker, intermediate when he had watched and heard everything through a video screen (and the speaker was aware of this), and lowest when the bystander had either sat out of view of the cards or had been absent completely. Moreover, speakers' referential expressions also reflected this pattern, so that the greater the participation of the bystander, the shorter and more concise were the speaker's referential expressions upon switching. Wilkes-Gibbs and Clark concluded that there were various levels of common ground, and that these affected the form and efficiency of subsequent communication.

It's worth emphasising that there was no difference in efficiency in Wilkes-Gibbs and Clark (1992) experiment when the bystander only heard the speaker and addressee communicate and when he was not present at all. This indicates that hearing the referential expressions on their own was useless without seeing how they were coupled with particular figures. Several other experiments have examined how important visual information may be in successful communication in these kinds of tasks. Kraut et al. (2003) conducted two experiments in which participants carried out a physical bicycle repair task either alone or with the help of a bicycle expert. In the latter case, they manipulated the location of the expert and the communication medium. They found that the repairs were more efficient when the workers received help, and most efficient when the experts were in fact in the room with them. However, if the expert was remote, it made no difference if they were only able to communicate through audio, or if the expert also had access to a video display (from a camera mounted on the worker's forehead). Gergle et al. (2004) used a computerised matching task, in which they manipulated whether the director had a view of the matcher's workspace while instructing him how to arrange a set of colour patches in a particular configuration. They found that pairs solved the task faster and communicated more efficiently (i.e., used fewer words) when they shared the visual space. Moreover, the advantage was larger if the patches fluctuated in colour (making them hard to describe verbally), and a three-second delay in updating the director's view removed the speed advantage (but not the efficiency advantage). Clark and Krych (2004) conducted a similar experiment in which directors instructed builders to build lego models. The experimenters manipulated whether the directors could see the matchers' workspace and whether the participants could see each other's faces. They found that visibility of the workspace greatly increased the matchers' building speed and decreased errors. In contrast, although facial visibility was exploited through the use of eye gaze and head gestures, it did not lead to faster model-building.

In considering the usefulness of visual information, it's important to acknowledge that various body signals and gestures can also be very useful in communication and indicative of coordination. In a referential task, Hanna and Brennan (2007) showed that eye gaze can be used faster to disambiguate between items than linguistic expressions, even when matchers and directors had reversed displays. Of course, when people are actively interacting, they can and do also use their hands. Goldin-Meadow and Wagner (2005) review gesture studies and argue that gestures are used both by speakers to help with their thinking and by listeners to infer speakers' intended meanings. Pointing in particular plays an important role in establishing joint attention and building common ground from a very early age (Tomasello, Carpenter and Liszkowski 2007).

3.4.2.3 Comprehension and production experiments

Many researchers have focused on either comprehension or production when investigating the role of common ground in referential communication. Although my own experiments will not exhibit such an asymmetry, it's worth having a quick overview of the range of effects that have been documented.

By conducting experiments in which the speaker and listener had visibility of slightly different subsets of the same grid of objects, it has been shown that listeners seem to consider what speakers can see and thus what is common ground, even if they are also influenced by things that they can only see themselves (Hanna, Tanenhaus and Trueswell 2003). However, Keysar, Barr, Balin and Brauner (2000) showed that listeners' initial interpretations of speakers' expressions were egocentric, not paying attention to common ground. Moreover, Barr and Keysar (2002) showed evidence that listeners expect speakers to adhere to linguistic precedents (rather than common ground), even if they are unnecessarily specific (e.g., "the big apple" when there is only one apple), or when a new speaker takes over. In contrast, Metzing and Brennan (2003) demonstrated a partner-specific effect: listeners comprehended new expressions to old objects more slowly when they were used by the same speaker rather than a new speaker. Kronmüller and Barr (2007) qualified these results by showing that listeners tend to rely on partner-specific associations only later in processing, so that there were two distinct processes at work. Moreover, Shintel and Keysar (2007) demonstrated that listeners expected speakers to use referring expressions they had used previously, even if they were not aware that the listeners had heard them. However, Brown-Schmidt et al. (2008) showed that common ground effects were sensitive to the types of utterances involved. And Brown-Schmidt (2009) showed that if the tasks were more interactive, as in normal dialogue, then partner-specific effects were evident from early processing stages. This family of studies has been complemented by recent eye-tracking experiments: Richardson and Dale (2005) demonstrated that interlocutors' coordination of eye movement (even when they did not see each other) with respect to an array of figures corresponded to better comprehension on the part of the listener, and Richardson, Dale and Kirkham (2007) found that coordination increased if their common ground was first supplemented by giving them the same background information. Nevertheless, the debate about the role of common ground in comprehension continues (Brennan and Hanna 2009; Shintel and Keysar 2009).

The oppose side of the coin, of how speakers tailor their referential expressions to addressees in production (known as "audience design") has also been investigated extensively, especially by William Horton and colleagues. Brown and Dell (1987) showed that speakers, when retelling a story, did not generally take confederate listeners' knowledge into account, and suggested that speakers only did so in monitoring and repair. However, using naive participants as listeners, which may have been important if speakers pick up on subtle cues from listeners, Lockridge and Brennan (2002) found the opposite result. Horton and Keysar (1996) further showed that speakers did not tailor their referring expressions to the common ground if they were under time pressure. Similarly, Bard et al. (2000) found that speakers simplified their expressions over time regardless of whether they changed who they were addressing. They argued that speakers' production was best explained by a combination of fast ego-centric priming based on the speaker's knowledge and slow, optional inferences about the listener's knowledge. However, Galati and Brenna (2010) qualified these findings by showing that speakers' simplifications were more dramatic when they interacted with the same listeners rather than switching listeners. On the other hand, Horton and Gerrig (2002) found that when speakers were aware of the need for audience design, they resisted the tendency to use consistent or increasingly short expressions when switching addressees, while also learning to improve their expressions better with time and experience. Furthermore, Horton and Gerrig (2005*a*,*b*) and Horton (2007) used experiments with changing addressees and a telephone conversation corpus to argue that speaker partner-specific effects are not a product of special processes involving explicit consideration of common ground, but rather are rooted in basic memory traces. As such, people are more likely to use certain terms when speaking with certain interlocutors because they are associated relatively strongly in their memory. Finally, Keysar and Henly (2002) showed that when speakers were asked to say ambiguous sentences to addressees with a certain intended meaning, they overestimated their own effectiveness, even though overhearers hearing the same utterances did not.

3.4.2.4 *Summary*

In summary, in this section I first introduced the notion of common ground. I then reviewed empirical work which explored the role and development of common ground in conversation, covering both more interactive experiments as well as those which focused more on either comprehension or production. Considering the body of studies together, it has been clearly established that common ground plays a role in comprehension and production during linguistic interaction. What is debatable, however, is how much interlocutors take it into consideration, when they use it, and to what extent it leads to partner-specific effects.

3.4.3 Alignment

3.4.3.1 Theory and scope

Building common ground in the course of dialogue involves coordination between speakers and listeners. But what is coordinated at the level of mental representations, and how? This is a crucial question to consider, since I am interested in conceptual coordination. Pickering and Garrod (2004) pick up these issues and propose a mechanistic theory of dialogue which they call the "interactive alignment" account. They point out, following (Clark 1996), that dialogue is the most natural form of language use, and yet has long been neglected by linguistic and psycholinguistic theories. Moreover, contrary to common belief (and as we have seen in the previous section), it is possible to conduct controlled experiments to investigate dialogue phenomena together with cognitive processes and representations. Pickering and Garrod's theory states that dialogue does not just involve general coordination, but specifically brings about **alignment**, in which interlocutors adopt similar cognitive representations at a particular level. Alignment occurs at various levels of representation, and comes about through priming. The process is automatic and does not require understanding or modelling the cognitive states of others. Hence, their position is most compatible with the previously mentioned egocentric views of processing during dialogue, such as those of Horton and Gerrig (2005*a*) and Barr and Keysar (2002). For instance, syntactic alignment would normally occur when dialogue participants unknowingly adopt each other's syntactic structures during conversation. Different levels of representation are linked, so that alignment at one level can bring about alignment at other levels. This kind of coordination greatly simplifies language comprehension and production processes in dialogue processing. Moreover, Pickering and Garrod argue that there is no dichotomous distinction between dialogue and monologue; rather, various forms and uses of language involve interaction to varying degrees, resulting in various degrees of inter-personal alignment.

It's worth pausing for a moment here to address a bit of terminology. Pickering and Garrod (2004) distinguish alignment from the looser notion of **coordination**, whereby people are engaged in a joint activity (Clark 1996) but need not share representations (e.g., when performing a duet). In contrast, alignment is specifically about shared representations. The distinction is a valuable one in general, but is not critical for my purposes, because as we will see, coordination in my experiments will be explicitly concerned with alignment specifically. Therefore, I will use the two terms fairly interchangeably, just with a slightly different emphasis, with alignment being more emphatic about shared representations. In addition, since most of my experiments will look at the timecourse of alignment, I will also use the term **convergence** when I want to emphasise an increase in alignment over time.

Alignment has been demonstrated at various levels of linguistic representation. Brennan and Clark (1996) showed that interlocutors align at a lexical level in a referential coordination task, converging on the same words or phrases. Expressions were implicitly agreed upon together, and became shorter and more stable during interaction as the same figures were referred to repeatedly (I return to this study when I focus specifically on lexical alignment in the next section). Branigan, Pickering and Cleland (2000) studied syntactic alignment instead. The experiment exploited the dative alternation in English, whereby we can say "John gave the book to Mary" or "John gave Mary the book". Participants were shown to align on the same syntactic variant when describing a series of pictures to each other, even though the particular words varied (since the different cards depicted a variety of characters and actions). Finally, Pardo (2006) investigated phonetic alignment in dialogue. Conversation participants' speech was recorded before, during and after a joint task. Other participants then listened to some of the recorded expressions and judged the similarity between pronunciations. The results revealed that interlocutors' pronunciations were more aligned during their interaction than before, and that the alignment persisted beyond the interaction.

But dialogue does not only involve linguistic alignment. Garrod and Pickering (2009) argue that in conversation, interlocutors must coordinate their mental representations and behaviour at various levels, both linguistic and non-linguistic, and these levels can mutually reinforce each other. Indeed, the scope of the coordinating effects of dialogue reaches surprisingly far, including speech rate (Giles, Coupland and Coupland

1992), posture (Shockley, Santana and Fowler 2003), laughter and yawning (Hatfield et al. 1994), gaze (Richardson and Dale 2005) and gait (Murray-Smith, Ramsay, Garrod, Jackson and Musizza 2007). Given the close relationship between words and concepts, this suggests that words, especially when used freely in dialogue, may also coordinate people's conceptualisations.

So what about alignment of conceptualisations? Although Pickering and Garrod (2004) do not address conceptual alignment directly, they start their discussion of alignment at the level of situation models. Situation models are non-linguistic multi-dimensional mental representations of the current situation of interest, and encode such aspects as space, time and causality (Zwaan and Radvansky 1998). As such, they encode the mental picture of the thoughts underlying the discourse at hand. Therefore, they are related to concepts, although the exact relationship or correspondence is unclear. Pickering and Garrod argue that although alignment of situation models might not be required for communication, the alternative would be to maintain two different representations (for oneself and for one's interlocutor), and that this would be "wildly inefficient" (p.172). Moreover, misalignment would raise the question of whether participants really understand each other, even if they communicate successfully. However, they do concede that interlocutors' situation models will generally vary a little, and thus will not be fully aligned. Indeed, several commentators (e.g., Branigan 2004; Schober 2004) argued that interlocutors are less likely to align with as much precision at the level of situation models compared to linguistic levels of representation.

I will not delve any further into situation models and how they may or may not correspond to concepts. Instead, I now move on to studies that shed light on conceptual alignment specifically.

3.4.3.2 *Lexical (and conceptual) alignment*

In this section, I discuss evidence from dialogue experiments which sheds light on how interlocutors might align their conceptualisations. However, given how hard it is to get at people's concepts, how can we assess conceptual alignment? Just as psychologists cannot access individuals' concepts directly and often rely on naming tasks, much of what we know about conceptual alignment comes from people's lexical choices in dialogue. It would therefore be difficult and even counterproductive here to completely filter out studies dealing with the latter, as long as we guard against prematurely equating the two.

Brennan and Clark (1996) focused on how lexical choices are made in conversation depending on the context and history of interaction. They first claimed that "labels reflect conceptualizations" (p. 1482), and noted that the same thing can be conceptualised in many different ways. Perhaps the most obvious manifestation of this lies in different levels of specificity of a category (e.g., animal, dog, terrier, small black terrier). They conducted referential communication experiments in which a target object (e.g., a shoe) had to be distinguished from other objects in a sequence of tasks. The experimenters manipulated whether the context of surrounding objects included other objects from the same category (e.g., other kinds of shoes). The results showed that participants tended to start with whatever level of specificity was required to distinguish the items. But when the context was widened during the course of interaction, participants often continued using the (unnecessarily) specific terms they had adopted (e.g., "loafer"). This effect was reduced, however, if participants changed partners during the experiment. These results suggest that naming choices do not just depend on an object's properties and on informativeness (being as precise but no more precise than needed, Grice 1975), but also on factors that are specific to the local interaction, such as recency and partnership. They also showed that convergence can be quite gradual and involve provisional stages.

Brennan and Clark (1996) concluded that the process of settling on a term for an object was best understood as the development of a "conceptual pact": interlocutors gradually converge on a joint conceptualisation for the purposes of a particular interaction. Notice that, under this view, people who agree on a term for something implicitly achieve conceptual alignment. However, this interpretation relies on assuming an equivalence or direct correspondence between lexical and conceptual alignment, which in turn presupposes that words are reliable conceptual identifiers. But as I argued in Chapter 2, this assumption cannot be made in this thesis. From my perspective, it's necessary to adopt a more conservative interpretation: although converging on the same words suggests some degree of conceptual coordination, it is not clear a priori to what extent. You and I may agree on the term "loafer", but have somewhat different concepts associated with this word (indeed, I had never heard this word until I read this paper).

Garrod and Anderson (1987) used a different methodology to study lexical and conceptual coordination. They developed a computerised game in which participants had to individually navigate from one position to another in a two-dimensional maze composed of nodes and connections. However, the maze had obstacles, and these could only be overcome if one's partner moved to certain special locations in the maze. Much as in the previously discussed experiments, participants could not see each other or each other's mazes, but they were allowed to talk freely during the task. Participants therefore discussed their locations in the maze, providing data on how they referred to locations and thereby how they mentally modelled the maze. The results showed that pairs did align in their descriptions, but did so in different ways. Garrod and Anderson identified four different kinds of schemes from the dialogues, which they called "path" (e.g., "See the bottom right, go two along and two up"), "co-ordinate" (e.g., "I'm on the third row and fourth column"), "line" (e.g., "Third bottom line, third box from the right"), and "figural" (e.g., "See the rectangle at the bottom right, I'm in the top left-hand corner"). The coordination task is challenging, because participants must align both conceptualisations and linguistic ways of describing them, and the correspondence is not trivial (e.g., from where do you count rows to determine which one is "the third row"?). So how do interlocutors achieve this? Garrod and Anderson first point out that it was very rare for participants to explicitly agree on a scheme in advance, and even when they did so, they tended to abandon it once it proved inadequate. Instead, they argue that participants followed a simple "output/input co-ordination" principle, whereby you would formulate your output based on the same interpretative principles as applied to the most recent relevant input. This simple mechanism, they argue, is sufficient to explain how such linguistic and conceptual conventions arise. In any case, this experiment demonstrates how lexical and conceptual alignment are not trivially equivalent and yet how both occur during interaction.

In a follow-up study using the same methodology, Garrod and Doherty (1994) explored how a convention could arise in a small linguistic community. Twenty participants were assigned to one of two conditions: either they worked with the same partner throughout nine different mazes, or they worked with a different partner from the same "community" of ten for each of the nine mazes. This allows one to investigate coordination and stability for isolated pairs versus small communities. Garrod and Doherty found that the isolated pairs used substantial and stable proportions of both matrix and line schemes throughout their games. In contrast, the community group converged rapidly on one particular type of matrix scheme, which was the most common scheme used in the first game across the group. Closer analysis showed that although pairs in the isolated pairs group attained a fairly high degree of lexical alignment quickly, they did not get any higher, while pairs in the community group started low but eventually reached nearly perfect lexical alignment. In a second experiment, participants interacted with new partners for every maze, but these partners never interacted amongst themselves (i.e., it was not a community). In this case, no convention emerged, and participants did not increase their alignment over the course of the study. Garrod and Doherty explained the results by arguing that isolated pairs resorted to precedence and salience constraints, which are sensitive to changes in the immediate context (in this case, in the mazes), while the community group quickly developed a more stable convention, which "overrides precedence and salience" (p. 208).

So far, I have mainly discussed lexical alignment, and implicitly appealed to Brennan and Clark's (1996) claim that these alignments reflect conceptual pacts. However, although they are no doubt related, it is important not to assume equivalence between lexical and conceptual alignment. Schober, Conrad and Fricker (2004) demonstrated this from a different and more applied perspective. They conducted telephone surveys about employment, housing and purchases, in which respondents were given fictional scenarios and asked questions about them. The scenarios were set up so that as long as the participants understood the terms used in the questions in line with the interviewers' standardised definitions (e.g., of terms like "household furniture"), there were objectively correct answers. In this way, respondents' comprehension could be tested. The experimenters manipulated when participants received clarification: only when they asked for it, also when the interviewer thought they needed it, or neither. The results showed that when the scenarios were straight-forward (i.e., any ambiguity was unlikely to affect relevant comprehension), comprehension was very high across all conditions. But when the scenarios were more complicated, respondents' comprehension was a function of how many kinds of clarification they received: those who received no clarification diverged the most, while those who received both kinds of clarification understood best. These results demonstrate how even speakers of the same language presented with the same context and language can differ significantly in their underlying conceptualisations. However, they also suggest that the more collaborative the dialogue is, the closer the alignment. In sum, these results both caution us against equating lexical with conceptual alignment (in Schober's 2005 terms, "linguistic alignment does not guarantee conceptual alignment", p. 249), and support the idea that full dialogue may be able to bridge the gap between them.

In summary, in this section we have seen how interlocutors tend to gradually develop conceptual pacts, reducing the length of expressions and converging on particular terms through two-way interaction and negotiation. However, there is no consensus on the extent to which lexical alignment reflects underlying conceptual alignment. Part of the problem with settling this issue using the methods above is that the data collected during online tasks is primarily lexical, which makes it difficult to separate words from concepts. As a result, in the next section I discuss a few key studies which have decoupled these levels of representation using separate tasks.

3.4.3.3 *Preserved conceptual alignment*

In this section, I discuss research that has tested for conceptual alignment between dialogue participants *after* their linguistic interaction. These studies are important to my purposes for two reasons. The first is methodological, and concerns the challenge of separating the questions of lexical and conceptual alignment. It is difficult to give experiment participants tasks that simultaneously have them communicating linguistically and performing some kind of non-linguistic tasks that probe their concepts. This is evident in all the studies discussed in the previous section: they draw conclusions about conceptual alignment but rely enormously on linguistic evidence. In contrast, it is relatively easy to have participants first perform a joint task together while conversing, and then carry out individual non-linguistic tasks. One can then examine whether partners have convergent responses in the latter tasks. While separating the linguistic and non-linguistic tasks in this way offers clear advantages, it is important to acknowledge that it has limitations too, and asks somewhat different research questions. In particular, conceptual alignment could well occur during an interaction without persisting beyond it. Therefore, evidence for the lack of post-interaction convergence does not necessarily indicate that interlocutors were not aligned conceptually while engaged in dialogue.

The other reason why these kinds of studies are important is that they provide a potential link between the two timescales of potential coordination that I have identified. In particular, to the extent that linguistic relativity effects do occur (see Section 3.3), they must somehow originate in linguistic interaction or acquisition. However, it is not obvious how exactly this occurs, and evidence for preserved post-interaction alignment would provide support for a plausible explanation. If people align conceptually during dialogue *and* stay aligned beyond it, then perhaps repeated linguistic interaction with speakers of the same language throughout one's lifetime could lead to a more permanent kind of conceptual alignment. Of course, such conclusions would still need to be highly provisional, since whatever effects are found immediately after interaction may well dissipate later on.

Despite the important implications of this kind of research, there have been very few studies which have adopted the use of a joint communication task followed by individual non-linguistic categorisation tasks. As a result, I describe the ones that have been carried out in a relatively high amount of detail.

Markman and Makin (1998) investigated the role of communication in category acquisition, structure, and consistency. Their studies used LEGO pieces and involved building models. Dyads were first given a set of pieces and allowed to come up with labels for them together. They then were assigned to two roles, with one participant directing the other how to build a model of either a car or a spaceship. Afterwards, participants were individually given the pieces and asked to sort them into groups. There were also two kinds of control participants: those who built models without communication, and those who did not build models. Although Markman and Makin were interested in three different central questions, I focus on the main one of relevance here: whether linguistic communication draws individuals' categorisation closer together. Their analysis revealed that participants who communicated during model building categorised more closely than those who did not communicate or those who did not build a model. This was the case even when comparing two communicating participants who had been re-paired in analysis with *other* communicating participants, as long as they had both worked on the same kind of model (car or spaceship). In a second experiment, dyads built three models, and were assigned to build either two vehicles and then a car, or two spaceships followed by the (same) car. The sorting task for all participants was based on the pieces used to build the car. The results showed that participants categorised more similarly to other participants who had worked on the same two initial models than those who had worked on different models, even though they all built the car model in the end and sorted only its pieces. These findings suggest that communication aligns not just people's words, but also their underlying conceptual structures, and that this alignment persists beyond the immediate interaction. However, their results also suggest that it is not so important *who* you communicate with, but only *that* you communicate. Communication, and word usage in particular, seem to place certain constraints on category acquisition that are not specific to a given dyad. In connection with this, it's worth noting that in their second experiment, a delay of 2-5 days was imposed after building the first model, but that dyads maintained use of their established names when building the second and third models. Thus, it seems that conceptual pacts (in Brennan and Clark's 1996 sense) can persist over time and can in turn affect concepts and conceptualisation.

Voiklis (2008) combined a category learning paradigm with referential communication tasks. Stimuli consisted of unfamiliar creatures that varied along several perceptual features, which corresponded to two key functional features (i.e., nutritive/non-nutritive and destructive/non-destructive) in a complex way. Participants underwent a large training sequence, in which they judged individual creatures on the two functional features and then received corrective feedback. However, in a dyad condition, one participant saw a creature and described it to her partner while the other judged it based on this description. Before and after these category learning tasks, all participants carried out sorting tasks, in which they put creatures into groups. The results showed that participants in the dyad condition learned the categories better and faster, especially when the category structures were relatively simple. Moreover, in line with Markman and Makin's (1998) findings, there was more convergence in the post-training sorting tasks between dyad participants than between non-dyad participants, regardless of whether the comparisons were made between partners or non-partners. Voiklis noted that these dyad effects could be due both to linguistic factors (e.g., linguistic terms highlighting the relationships between features or serving as a compressed form of concepts and features) or non-linguistic factors (e.g., knowledge diversity, division of labour, increased motivation). In any case, they concluded that "communication may push 'public' conceptualizations and publicly-formed 'private' conceptualizations towards a limited range of widely shareable conceptual structures" (p. 86).

Another relevant study was conducted by Malt and Sloman (2004). In a pair of experiments, they used a coordinated matching task, and used stimuli which could (based on pre-tests) be readily labelled with one of two labels (e.g., bucket/pail, bottle/jar). In a first set of matching tasks, for each picture, confederate directors consistently referred to it with one of the two alternative words. Then the matchers became directors for a second batch of matching tasks. Afterwards, participants individually rated a sequence of pictures in two of three ways. In a naming preference task, participants indicated the appropriateness of the two alternative labels for each item. In a typicality task, they rated how typical items were relative to the named categories. In a similarity task, they judged the similarity of the items to their imagined prototypical category members. Malt and Sloman's analysis revealed that, first of all, participants stuck to the terms that were first used by the directors. This was the case even though the two name alternatives were generally at the same level of abstractness (unlike Brennan and Clark's 1996 study). In addition, the main analyses showed that which of the two alternatives was used had an impact on all three individual post-interaction tasks. However, Malt and Sloman emphasised that the effects were stronger for the naming preference task than the other two. They concluded that naming choices affect further naming access and usage, but had less of an impact on non-linguistic tasks. While the methods used in this experiment had the advantage of getting at lexical and conceptual alignment on relatively equal grounds (since, in a way, both were tested after rather than during the interaction), it should be noted that the naming preference tasks were carried out before the typicality or similarity ratings, which means that the smaller effects in the latter may be due to general dissipation of any influence of interaction.

In summary, the studies discussed in this section suggest that engaging in dialogue does not only align people's lexical representations, but also related cognitive representations as well. Moreover, these effects persist beyond the immediate interaction. However, the degree of alignment that is evident in a given process seems to be a function of how linguistic that process is, although it is not clear how exactly these processes relate to conceptual representations (and thus conceptual alignment) per se. In addition, the studies suggest that alignment with another person only requires linguistic interaction with someone, not necessarily with that person specifically.

3.4.4 Conclusion

People clearly adapt to each other when they interact and engage in joint action. While it is debatable how much and when they shift to other people's perspectives, there is ample evidence that they coordinate their actions with others. During dialogue in particular, evidence suggests that interlocutors align mental representations with each other at various levels, both linguistic and non-linguistic.

Conceptual alignment in particular, however, has not been studied much directly. This is not surprising, given the methodological challenge of not only accessing people's concepts, but also doing so online during dialogue. Indeed, investigations of conceptual alignment have generally been of two types. On one hand, many studies have documented lexical alignment, showing that interlocutors converge and entrench on particular lexicalisations to refer to particular referents. To the extent that words reflect conceptualisations (Brennan and Clark 1996), dialogue can be claimed to also result in conceptual alignment. This is particularly evident in cases where different lexicalisations clearly reflect different schemes and solutions to some kind of joint problem (Garrod and Anderson 1987).

On the other hand, a few studies have looked more directly at conceptual alignment non-linguistically, but only after, rather than during, interaction (Markman and Makin 1998; Voiklis 2008). These studies have shown that people who communicate during joint tasks involving objects then seem to conceptualise those objects more similarly to each other. Interestingly, these effects were not specific to conversational partners: people seemed to align to communicating participants in general, even those that they did not interact with. Moreover, they also aligned more if they had been using the same set of referential terms (Malt and Sloman 2004). To the extent that speakers of the same language do share the same concepts, these results are consistent with studies showing how the use of language can affect conceptualisation online (e.g., Lupyan 2008*b*; see Section 2.6.5).

Despite the progress that has been made on these issues, many important questions remain regarding conceptual alignment and its relationship with language. The first question is primarily methodological: how can we separate investigations of conceptual alignment during interaction from reliance on language? I identified this issue back in

Section 1.2, but it still remains to be solved. Second, does online conceptual alignment occur only during *linguistic* interaction, or during interaction in general? In order to have an empirical basis to claim that it is specific to linguistic interaction, we need to compare interactive conditions which vary in how much language is available. Third, even if lexical and conceptual alignment do turn out to be closely intertwined, this says nothing in itself about the direction of causation. Does dialogue cause both lexical and conceptual alignment, does one of them bring about the other, or is there still another explanation? Although we cannot manipulate these directly, we can control interactional conditions, and in particular, the exchange of lexical and/or conceptual information between participants.

Chapters 6-8 present experiments intended to address these issues. Figure 3.3 adapts my conceptualisation model for the interactive case to give a hint of how this will work. Pairs of participants will carry out joint tasks, and I will measure the conceptual alignment between them while (or after) they do so. Within this set up, I will manipulate the information exchange that takes place between participants, including the availability of dialogue and the exchange of words and subsets of categories. However, before I can get concrete, I need to resolve some methodological issues, which I take up in the next chapter.

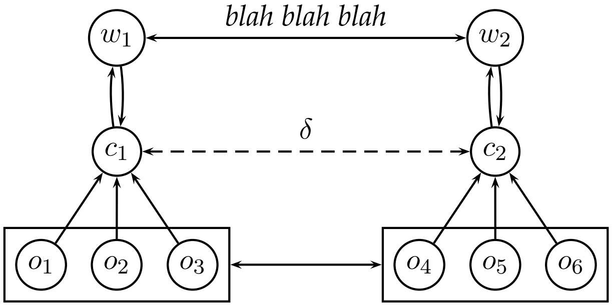


Figure 3.3: Interaction between two people, and potential sources of conceptual coordination. People may be able to speak freely with each other (*blah blah blah*). They might also have access to each other's concepts through their categories and words (indicated by the bi-directional arrows connecting them).

3.5 Summary

Although conceptualisation is a psychological process that takes place in the minds of individuals, the publicity of words and their close associations to concepts implies that our conceptual systems and processes are unlikely to be independent of each other. Indeed, in this chapter, I have reviewed work concerning three broad issues, which revealed evidence that language may draw people's conceptualisations together in various ways.

First, I discussed how the input of words from others affects conceptual development and conceptualisation, starting from early infancy and continuing throughout our lives. Hearing words has diverse effects, such as guiding infants' early conceptual development, influencing children's category induction, and shifting adults' memories. Although the influence of others' words is not the focus of my thesis, it does underly my main hypotheses, because it confirms that people's conceptual systems do influence each other, and suggests how others' use of language could play a causal role in conceptual processing.

Second, I introduced the linguistic relativity hypothesis, according to which our native language determines how we partition the world into categories. I discussed empirical evidence in several domains, especially colour and objects, and considered several theoretical reformulations of the original idea which better fit the body of findings. However, there is no clear winner, since the results are quite complex and subtle, and much further study is needed before we can arrive at a satisfying full story. While this line of work relates to my hypothesis of language coordinating conceptualisation, this question has not been the direct focus of investigation. As a result, the effect of our native language on our "pre-alignment" will be addressed in Experiment 1.

Third, I considered how language use could lead to the coordination of people's conceptualisations. I outlined the view that dialogue in particular is a form of joint action, and involves coordination of behaviours and alignment of cognitive representations. I discussed how dialogue depends on common ground, while also building it up further, but that visual information also plays an important role. I then discussed two lines of evidence which consider conceptual alignment in the context of dialogue. One line takes up conceptual alignment during dialogue, but tends to rely strongly on lexical alignment data and assumptions of how they correspond. The other avoids these assumptions, but is focused on conceptual alignment which persists beyond dialogue, and still doesn't separate dialogue from interaction in general. Consequently, the role of language in conceptual alignment will be taken up in Experiments 2, 3 and 4.

However, before we can dive into empirical investigations, there are methodological challenges to confront. In particular, we need to decide on kinds of experimental tasks and how to quantitatively compare two outputs from them. I turn to this next.

CHAPTER 4

Developing a framework

In Section 1.3, I pointed out how my research question automatically comes in a package with two uncomfortable principles: that concepts can vary and that we cannot count on language to identify them. As I explained, taking these principles seriously imposed significant theoretical and methodological challenges. Now that I have developed my theoretical position while considering the implications of those principles, it is time to face the methodological challenges.

Therefore, in this chapter, I first lay out the criteria for an experimental framework, given my research questions. I then justify the use of behavioural categorisation experiments for investigating them. Next I list some important considerations regarding the selection of stimulus domains. After that, I discuss existing categorisation paradigms, and argue for the adoption of a particular type of task for my experiments. I then discuss possible measures for comparing two outputs from this task, and settle on the ones I will adapt. Finally, I describe how the framework can be readily imposed with appropriate linguistic manipulations.

4.1 Criteria

Before developing my experimental framework, it's important to be clear what we want it to do and what the constraints are. The general purpose reflects my theoretical goals: to support a set of experiments which investigates how language might have an effect on conceptual coordination. In this section, I list specific methodological criteria concerning the experimental tasks, measure, stimuli and manipulations.

First, the experimental tasks should involve conceptualisation, and yield output that captures it. However, the task output cannot be identified by reference to words or externally-specified categories. Since conceptualisation is flexible (see Chapter 2), the tasks should also yield snapshots of participants' concepts, allowing for potential change over time.

Second, there must be a well-defined, objective, and meaningful measure for comparing participants' task outputs. In other words, the measure should quantitatively capture, as well as possible, the difference between two acts of conceptualisation. This measure is very important because it will produce the values for the primary dependent variable in all of my experiments.

Third, it should be straightforward to superimpose the framework with various manipulations, primarily, but not only, linguistic in nature. In particular, it needs to be possible and meaningful to control whether participants can speak to each other, whether they can exchange word and/or category information, and how and whether they interact. Moreover, the framework should be appropriate for use with different languages.

Fourth, the stimuli should be selected with care. They should be presented non-linguistically, so that we can manipulate language independently of them. They should come from a familiar domain, so that words already have pre-established mappings with them.[1] And the domain should be continuous and finely sampled, to maximise the potential for variation in conceptualisation.

Note that, to some extent, these criteria are interrelated. For instance, the kind of task we use obviously impacts what kind of outputs it produces, and thus what kind of measures for comparing them are possible. Thus, although I will deal with the criteria one at a time, it's important not to lose track of the bigger picture.

[1] Novel domains would also be interesting to explore, but that would involve somewhat different research questions, which my thesis does not address directly.

4.2 Categorisation tasks

4.2.1 Justification

What kinds of experimental tasks should we use to query people's concepts? In principle, there are different approaches we could take, ranging from high-level methods such as asking participants a set of survey-style questions (Vosniadou and Brewer 1992), to low-level methods like analysing brain activation patterns (Koenig et al. 2005). Like most cognitive psychologists (Murphy 2002), however, I will take the middle ground, and use categorisation tasks. As such, people's categorisation decisions will be the source of my comparison of their conceptualisations.

However, this brings the distinction that I made between concepts and categories (see Section 2.4.1) uncomfortably back to the surface, and harks back to the introduction of Section 2.5. Since I have insisted for a clear separation between concepts and categories, is it not hypocritical to then turn back and use categories as a measure of concepts? The way I have incorporated my theoretical model from Chapter 2 into the questions raised in Chapter 3 (see Figures 3.1, 3.2, and 3.3) specifically targets the level of concepts as the one we ought to be comparing, in contrast to the levels of words or categories. As such, the proposal to use categorisation tasks can be seen as rather blatantly at odds with that program. In particular, I am proposing to shift from the comparison depicted in Figure 4.1 to the one shown in Figure 4.2. Since I will be looking at how people align their concepts and manipulating word and category feedback (see especially Experiment 4), this is a significant compromise. How can this be justified?

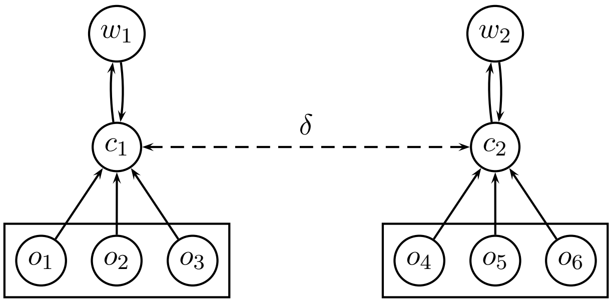


Figure 4.1: Comparison of two people's conceptualisations.

My answer is that yes, it is a significant compromise, but it's the best that we can do. We do not have direct access to people's concepts, so we have to access them indirectly.

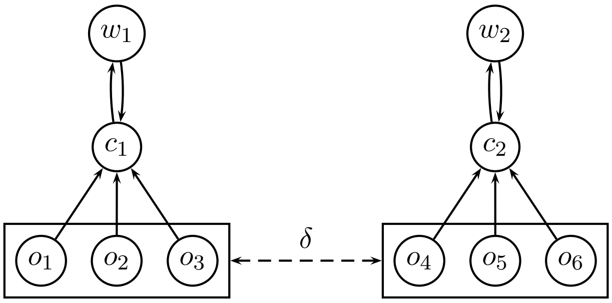


Figure 4.2: Comparison of two people's categorisations. This may the best we can do if we are actually trying to compare conceptualisations.

And although concepts and categories aren't the same thing, concepts determine categories, so their relationship is nevertheless very tight. There doesn't seem to be a better way, as is evident from the fact that Murphy's (2002) "Big book of concepts" is almost entirely devoted to discussing categorisation experiments. More importantly, my other conclusion from Section 2.4.1 was that categories are not autonomous: they exist only as conceptual extensions. They tell us about, and only about, the concepts inside someone's mind. They appear to be unique in this, and should therefore be exploited.

As a result, despite the imperfection of this solution, the practical constraints give us no clearly superior alternative. Therefore, I adopt the use of categorisation experiments, while acknowledging that this may bias results a little. However, this doesn't solve our problems yet, because there are many categorisation paradigms, and as we will see next, most of them are unsuitable for my purposes.

4.2.2 Categorisation paradigms

Categorisation tasks will occupy a very central position in my experiments. But what sort of tasks should we use? As we will see, although different options are available from the literature, most of them are incompatible with my criteria from Section 4.1. Moreover, the choice that I will actually converge on has some other issues that we will have to be prepared to swallow. Therefore, it is important to survey the different possibilities relatively carefully.

4.2.2.1 Picture naming task

Perhaps the most obvious categorisation task is the picture naming task. Participants are shown picture stimuli one at a time and asked to say what they are, perhaps with a single word or a longer phrase. The answer they give is taken as identifying their current conceptualisation. This paradigm can be used for various purposes, such as the investigation of semantic priming and domain differences (e.g., Lloyd-Jones and Humphreys 1997).

However, in terms of the framework criteria listed above, the picture naming task has an obvious shortcoming. It violates the criterion that concept identification must be separate from language. This is problematic on three fronts. First, since we are trying to assess the role of language in categorisation, we cannot identify participants' categories linguistically. Second, if words are the participants' only output, then there is no non-linguistic interaction between participants. This means that in manipulating language, the control condition would have no information exchanged. Third, we cannot assume identical word-concept associations, so the same word may relate to somewhat different concepts in different instances.

Does this mean that the picture naming paradigm is of no use to us? One potential workaround would be to borrow the same structure as is used in the picture naming task, but to replace linguistic naming with some other non-linguistic concept identifying elicitation. Perhaps participants could produce some other kinds of symbols (e.g., colours, numbers, geometric shapes) instead of words, while everything else is kept the same. However, although such an approach would make the identification of concepts language-independent, it would still suffer from the most important problem with using words: how could we compare the meanings associated with these symbols for different people? Indeed, symbol-concept associations of this sort are likely to be less conventionalised and shared than word-concept associations, so in fact the comparison would only become harder.

Moreover, the serial nature of the task also means that we are inducing participants' conceptual snapshots from temporally scattered responses. We learn what a participant's 'dog' concept looks like by considering which items they called "dog" and which they didn't across the series of tasks in the experiment. However, this neglects the possibility that conceptualisation may change over the course of the experiments.

The unsuitability of the picture naming task is unfortunate, because it would be readily extendable to cases involving pairs of participants. Indeed, as we have seen, dialogue studies have examined how speakers refer to items and how hearers interpret their expressions (e.g., Clark and Wilkes-Gibbs 1986), and such natural interactions provide a good natural context in which online conceptual coordination might occur. Moreover, it would not be too difficult to measure conceptual coordination. A crude binary measure could be given by simply seeing whether participants applied the same label to a particular referent. A more sophisticated method could be applied for cases with differing labels which takes into account the difference in meaning between words, by using statistical techniques such as latent semantic analysis (Deerwester, Dumais, Landauer, Furnas and Harshman 1990).

4.2.2.2 Category training

Another common categorisation paradigm is the use of category training. In such experiments, the researcher usually has a predefined set of categories that the participant is supposed to try to learn. Normally the stimuli are shown one at a time, and the participant responds with one of the categories. The category may be identified by a name, a button, or a location on the screen. Participants are often given feedback after their response, helping them to learn the categories over time. Such paradigms have been used to explore category learning, looking at how easily people learn different kinds of categories, under what situations they do best, etc (e.g., Smith and Minda 1998; Lupyan, Rakison and McClelland 2007).

At first glance, this paradigm seems promising, since it doesn't necessarily depend on linguistic responses. Participants can select from a set of categories by non-linguistic means, such as the left vs right sides of the screen. Moreover, it would be easy, though crude, to measure agreement between participants' categorisations, by simply seeing, on any particular task, whether they make the same judgement. Finally, it would be possible to manipulate linguistic conditions, by controlling whether participants could talk or exchange their own linguistic labels for the stimuli.

However, upon closer inspection, the paradigm actually suffers from the same fundamental problems as picture naming, and more. From a functional point of view, any means of identifying the concepts in such a task is symbolic, and therefore again faces the problem that the link between identifiers and concepts is not fixed. Indeed, these paradigms implicitly acknowledge this, since they are concerned with how the concept associated with an identifier changes, and perhaps does so differently for participants in different conditions.

In addition, category learning experiments involve an externally defined category, which participants are trying to acquire. As such, rather than emphasising what participants' own concepts are, they are focused on how a participant mentally represents categories set up by the experimenter. Therefore, while these tasks may tell us which of the experimenters' defined categories are associated with particular stimuli for a given participant, this restricts the participants to using these categories rather than ones that they would naturally choose themselves.

Moreover, as participants are only learning these categories in the course of the experiments, the conceptual snapshots are again scattered across time. As in the case of the picture naming tasks, this is problematic, since the concept may change over time; indeed, since these are online learning tasks, they depend on this kind of learning to take place.

Also, it is unclear how a category learning paradigm could be extended to study conceptual coordination. Confusion is likely to arise from participants simultaneously trying to learn the experimenter-defined categories and adapt to their partner's categories. Alternatively, learning can and perhaps should be envisioned as involving a teacher and a learner, and the participants could be assigned to these roles. However, it is not clear how compatible this asymmetric relationship is with an investigation of conceptual coordination. Of course, there are potential solutions: for example, perhaps coordination could be framed in terms of a series of bouts of learning, with the teacher and learner roles alternating within a participant pair. Nevertheless, the interactive solution doesn't flow cleanly from category learning as a starting point.

4.2.2.3 Familiarisation and novelty

Another place to look for help with categorisation tasks is in developmental and comparative psychology, since these disciplines are interested in categorisation, and must often research it with prelinguistic subjects. As such, they have the experience of necessarily gathering non-linguistic responses from participants. Various paradigms have been used, but they have in common a reliance on the coding of behavioural responses to stimuli. Here I discuss one such technique, which is based on the idea of familiarising participants to a series of items from one category and then showing them an item from a different category. While different behavioural measures can be used in conjunction with this technique, I discuss the looking time measure here because it makes it easier to understand how this technique works and is dominant in infant categorisation studies.

In this method, infants are typically exposed to a sequence of stimuli on a screen, and the experimenters measure how long the infants look at each stimulus. As infants become familiarised to a category by seeing repeated exemplars of it, their interest, as manifested in their looking times, tends to decrease. In contrast, if they exhibit relatively long looking times upon presentation of an item, then this is assumed to mean that the infants find the stimulus novel, and thus of a novel category. In this way, infants' discrimination of categories is studied by analysing their looking time patterns (e.g., Fulkerson and Waxman 2007). This method has also been recruited in experiments with non-human primates (e.g., Hauser and Carey 2003). Similar methods have been used based on other behavioural measures, such as the order in which infants touch objects (Mandler, Bauer and McDonough 1991).

In a way, these techniques are similar to the categorisation training tasks described previously. Usually, the experimenter has some pre-determined category distinction in mind, and wants to see if the subjects can make it. However, there is a little bit more freedom here, as there may be multiple different category boundaries that the subjects may distinguish, and the experimenter may be interested in which one the subjects tend to react to most (Mandler 2000). In this sense, to some extent, the experimenters are interested not in training the participants on certain categories, but in determining which categories the subjects use "spontaneously".

However, these paradigms have a number of important limitations with respect to my goals and criteria. First, they are very noisy and insensitive. This is important, because it is one thing to study whether a participant can distinguish airplanes from birds, but quite another to look at more subtle conceptualisations with fine differences between people, and how these may dynamically change through interaction. Second, these methods have not been used much with adults, and indeed, it is highly questionable whether they are appropriate for adults. There are likely to be far more complicated factors in explaining an adult's looking patterns than an infant's. Third, it would be difficult to come up with a reasonable measure which compared participants' conceptualisations in these tasks. The simple thing to do would be to quantitatively compare their looking times, but this could be interpreted in many different ways.

4.2.2.4 *Tasks involving two or three stimuli at a time*

Two other common kinds of categorisation tasks consist of judgements involving a minimal number of stimuli. In one type of task, a participant is simultaneously shown two items and asked to indicate whether they are the same kinds of things or different (e.g., Neiworth and Wright 1994). In another type, the participant is shown three items and asked which one doesn't belong with the other two (e.g., Lupyan 2009), or which of two items better matches a third (Rips 1989). The basis for the judgements could be unspecified in the task instructions, or specific criteria could be provided, such as the feature of interest.

For my purposes, these paradigms are improvements over the previously discussed methods for a couple of reasons. First, they do not involve any sort of identification using words, symbols, or other top-down associations. As such, there is no concern over how the token used to identify a concept may have different associations for different people or in different contexts. Second, they do not necessarily constrain participants with a predetermined set of experimenter-specified concepts to choose from. If the task instructions do not explicitly ask for them to use certain categories or criteria, then their output can be treated as reflecting their spontaneous personal choices. Third, these tasks involve judgements which explicitly put items together or split them apart. As such, these tasks do give partial conceptual snapshots. Fourth, it would be relatively simple to extrapolate these tasks for interactive purposes and with linguistic manipulations. Participants could carry out these tasks while communicating, name the items and exchange their labels, and see each others' judgements. Fifth, measuring conceptual coordination could be pretty easy, judged simply by whether or not the participants made the same judgement on a given task or not.

The main limitation with these tasks for my purposes, however, is that they only give snapshots of a tiny part of participants' categories, leaving the rest to the interpretation of the experimenter. In the dyad task, all we know directly is that the participant either places or doesn't place a category boundary between the items. Similarly, in the triad task, we only get a grouping of two items separated from a third. However, building up of the categories to span across multiple tasks is only possible with extra assumptions from the experimenter. For instance, if a participant judges an orange and a grapefruit to be the same kind of item, but a lemon and an apple to be different, then we might conclude that the grouping is based on citrus versus non-citrus fruit. However, having information of only two items at a time is rather crude, and combining results across tasks relies on speculative higher-level interpretations. This will not do for a study examining fine differences between people's conceptualisations. While we cannot completely remove this kind of uncertainty, we could do better if we had a larger number of items but fewer tasks to have to bridge across.

4.2.2.5 Free classification

A further categorisation paradigm involves sorting, or "free classification" (Imai and Garner 1965). Participants are given a relatively large set of stimuli at once, and asked to sort them into groups. Within practical constraints, the number, nature and sizes of the categories is up to the participant, and the end result of such a task is a partitioning of the stimulus set (e.g., Malt, Sloman, Gennari, Shi and Wang 1999; Clopper 2008).

This method has definite advantages for my purposes. First of all, it does not presuppose what concepts might be used. Participants can freely choose, consciously or not, on what basis they conceptualise items, and where they draw the boundaries between concepts. The relative lack of restrictions on the outputs means that participants can output categories which more faithfully reflect the concepts they correspond to. Moreover, the task does not depend on language. The groups of stimuli that participants come up with are sufficient as outputs, and are independently meaningful. Indeed, this method has been used in categorisation experiments with non-human primates with fruitful results (Spinozzi and Langer 1999). Also, if the sets of stimuli being grouped are not too small, there is no intrinsic need to try to relate the outputs across tasks, in contrast to the dyad and triad tasks. Indeed, the categories that are produced in a single task can be interpreted as referential snapshots of the concepts that the participant used to conceptualise the items. In other words, these experimental outputs are finite but meaningful subsets of categories.

Moreover, it is not difficult to extend these tasks to involve pairs of participants and to superimpose linguistic manipulations. Tasks involving participant pairs could be constructed in a couple of ways. First, participant pairs could perform a free classification task together, perhaps by taking turns or by deciding each step in unison. Alternatively, participants could carry out the tasks individually, but they might be allowed to communicate during the tasks. Notice that in contrast to the previous paradigms, where communicating one's categorisation decision on a given task would make coordination trivial, that would not be the case here, since there are many stimuli to deal with at once and the task output does not consist of a single decision. Finally, participants could carry out a sequence of tasks individually, but get feedback on each other's output between tasks. This feedback could consist of the category groupings, and, if participants were also asked to label their categories, of the category names. Although we might give a participant his partner's category name, this is not a problem (as it would be, for instance, in the picture naming tasks), since we can still define the main category output and analysis in terms of the groupings. Thus, this type of task can be readily used in interactive experiments, while simultaneously handling the linguistic manipulations we want.

However, this method involves a few complications, which may help explain why it is not particularly popular in categorisation experiments. First of all, there's the question of how we measure the degree of coordination between two task outputs. As we have seen, in some of the other paradigms, this is quite easy, often coming down to a binary value: agreement or disagreement. But in this case the situation is much more complex: how do we compare one set of sets of stimuli to another? Fortunately, as we will see in Section 4.3, there are well-defined and meaningful methods for this. These methods don't give binary outputs, but rather a value in some range indicating the degree of agreement between participants. For our purposes, this is actually an advantage rather than a weakness, since it may give us a more sensitive measure that can be used in analysis. Nevertheless, analysis and interpretation is more difficult in this case.

Furthermore, although, as I have outlined, free classification tasks can be adapted for an interactive coordination experiment, these tasks may not be as natural as some of their counterparts in previously discussed methods. It is normal for us to communicate linguistically about individual objects in the world, as would happen in an interactive extension of the picture naming task, for example. In contrast, in an interactive version of free classification, participants exchange information about several items at a time, and may do so through the category outputs alone, in experimental conditions where language is not available. As such, the ecological validity and relevance of such tasks could be questioned. However, there are real-world counterparts to these experimental situations as well. To give an intuitive example: when sharing a kitchen in a flat, flatmates coordinate how they sort dishes in the cupboards, possibly but not necessarily with the help of language. This can be seen when flat visitors come and wash the dishes, and may be unsure at first about where to put some of them afterwards. Therefore, these tasks might not map onto the most common situations involving conceptual coordination, but they are nevertheless grounded in normal behaviour and interaction.

4.2.2.6 Coordination games

So far, I have evaluated different kinds of tasks used to study the categorisation of individuals, and discussed how they might be adapted in coordination experiments. However, some researchers have also designed tasks specifically for studying coordination between individuals. Therefore, here I consider these kinds of tasks and whether they can be recruited for my purposes.

In typical coordination tasks, participants have to coordinate their behaviour or language to achieve particular joint goals. Among these, quite a few have looked at how people conceptualise things and how language is used in this process. Typically, these tasks are framed as some kind of game, to make them more meaningful for the participants, and to give them a clear goal. Often the tasks involve the description of items by one participant and their identification from among a set by the other (e.g., Clark and

Wilkes-Gibbs 1986). However, such studies also often emphasise that communication is not a one-way process with a distinct hearer and speaker, but that linguistic expressions and their referents are established through joint interaction. This is often reflected in the methods and data, by allowing participants to talk freely and converge on referring expressions (Garrod and Anderson 1987).

Can such paradigms help us here? On one hand, such interactive tasks have certain definite advantages. First of all, participants are explicitly motivated to coordinate how they conceptualise things, and are entertained in the process. Second, these tasks can help us see how dyads converge on particular expressions and contrast their solutions with those of other dyads. Divergences between different dyads can then be interpreted as demonstrating relative convergence within dyads. Third, since these experiments involve identification of a referent from among a set, we can study whether participants have the same referent in mind for a given expression, and under what conditions they match better or worse.

On the other hand, these tasks suffer from some of the same problems as previous paradigms. In particular, categorisation is again indicated through linguistic expressions, much as in the picture naming tasks. As a result, we are once again relying on words as top-down concept identifiers, we do not have referential snapshots of participants' categories, and we have no reasonable way of setting non-linguistic control conditions. Although some interactive coordination experiments have explicitly avoided reliance on language, they have still been dependent on some kind of response which relates to referents in a symbolic or iconic manner (e.g., Galantucci 2005).

Moreover, since the task output is now largely achieved together, it becomes difficult to identify individuals' categorisation. Rather than having separate results from individual participants, we must rely on more qualitative or indirect measures of how well they seem to understand each other. This therefore complicates the problem of defining a measure which compares two people's separate outputs. However, it is possible to partially get around this by following coordination tasks with individual ones, to see whether interaction has an effect on subsequent individual categorisation (Markman and Makin 1998).

Therefore, although I cannot adopt typical coordination game tasks directly here, it is still helpful to gather their insights more generally. In particular, it will be important to motivate participants with clear goals, to compare results not only within pairs but also across pairs, and to acknowledge that the richness of interaction permitted in the tasks may have a profound impact on the degree of conceptual convergence.

4.2.3 Conclusion

I have discussed a variety of different categorisation task paradigms with respect to my theoretical aims and methodological criteria. Most approaches suffer from a reliance on language or other symbols to identify categories, and constrain how participants can conceptualise items. Another common problem is how to make tasks which can be easily adapted to interactive sessions with pairs of participants, and how linguistic variables can be manipulated along with sensible control conditions. The best option overall appears to be the use and elaboration of free classification tasks, in which participants sort stimuli into groups. This method allows participants to conceptualise things relatively freely according to their own concepts, in a way which provides referential snapshots of their concepts in use. Moreover, it does so independently of language, although language can be easily superimposed on the task in different ways. However, this technique also involves certain challenges, especially with regards to comparing two people's conceptualisations. This is a crucial consideration that I turn to next.

Notice that in such a paradigm, each task outputs a subset of the category determined by a concept. This recalls the definition of categorisation that I arrived at from a theoretical point of view in Section 2.6.2, in which categorisation involves assigning (multiple) items to (multiple) categories. In effect, I have now independently argued for the appropriateness of that decision on methodological grounds.

4.3 Comparing two task outputs

4.3.1 The requirements

As I have mentioned, it is relatively difficult to define a suitable measure of agreement for the free classification paradigm. Multiple stimuli, multiple categories and a lack of predefined (by the experimenter) concepts result in complications. Figure 4.3 shows a schematic example involving six stimuli, where both participants produced two (almost identical) categories. And yet even here, it's not clear a priori how we would want the measure to work or what it should output. And things only get worse if there are more items and variable numbers of categories. So how should we compare in general?

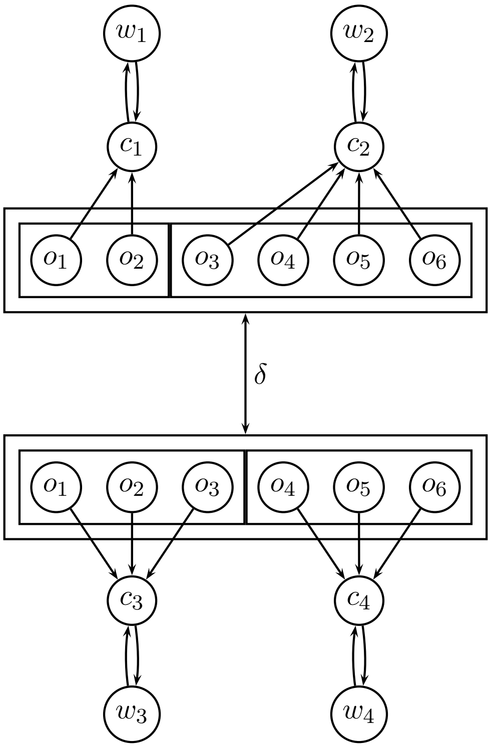


Figure 4.3: The comparison of two full categorisation outputs (top and bottom). The outputs are shown by the sets of sets of objects, and the measure is again represented by the δ between them. Note that the two outputs contain the same sets of objects.

Before looking into the problem more closely, let's be clear about what the measure is supposed to do, and in particular, what it should take as input and yield as output. First, the inputs. Free classification tasks are done individually by pairs of participants. Both participants are given the same set of n items and put them into between k_{min} and k_{max} categories of their own individual choosing (participants are also asked to name their categories, but that is irrelevant for the agreement measure). Mathematically speaking, each participant produces a "partitioning" (following Hubert and Arabie 1985), which is a set of subsets, where each of the subsets represents a category and each element of a subset represents a stimulus. The inputs to the agreement measure will be the partitionings of the two participants.

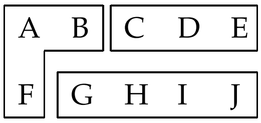
The output of the agreement measure needs to be an evaluation of the agreement between the two partitionings. For the sake of concreteness, we could require that the output be a number between 0 and 1 (inclusive), with 1 meaning exact match (i.e., the partitionings are identical). The ideal meaning of 0 would be complete disagreement, although as we will see, that is not so simple but not so important either. We also generally want lower values to indicate lower degrees of agreement. Indeed, as we will see, our measure would ideally have certain other properties as well.

It's worth noting that the simplest possible measure might be one which yields 1 if the two partitionings are identical, and 0 if they are not. However, such a measure would collapse a lot of information and would thus be much less sensitive, which makes it unattractive for my purposes.

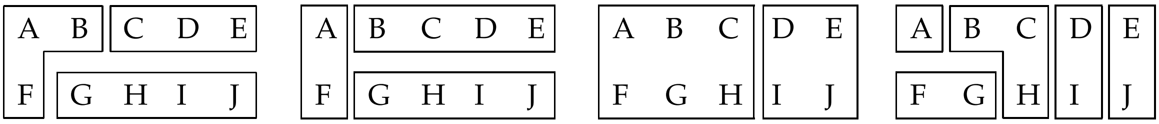
But if we do try to define a more granular measure, then the fact that different partitionings can contain various numbers of categories of potentially different sizes adds complexity to the problem, and makes it difficult to define a measure that matches intuitions. For instance, if two partitionings are identical except for one item, we may want to treat the measure differently depending on whether that item was put (by one or both of the participants) in a large category, a small category, or a category of its own, or if the total number of categories (for one or both participants) is large or small. We shall see an example of this complication later in this section.

4.3.2 An example

I start with a schematic example. Suppose that two participants are both given a set of 10 items, identified schematically here with letters A through J. They individually partition the set into groups. Then our agreement measure must assign a number comparing the two partitionings. Suppose that the first participant puts A, B, and F in one category, C, D, E in a second, and G, H, I, and J in a third. We can write this in set notation as $\{\{A, B, F\}, \{C, D, E\}, \{G, H, I, J\}\}$. Or, diagrammatically,



Now suppose the second participant partitions the set in one of four ways, shown below.



(a) (b) (c) (d)

What would we want the agreement measure to give us for these four cases? The partitioning for (a) is identical to that for the first participant, so the measure should yield a perfect score of 1. As for (b) and (c), it's not clear what we'd want the absolute scores to be, but intuitively the score should be higher in (b) than in (c). This is because in (b), the partitioning is nearly the same as for the first participant except for the categorisation of stimulus B, while (c) is quite different altogether. Finally, (d) should get the worst score, since the match between the partitionings is about as bad as possible: notice, for instance, that there is no pair of items that is put into the same category for both partitionings. Depending on how we define the measure, the score in this case may even be 0. In summary, for these four cases, we'd want the agreement measure M to satisfy $0 <= M(d) < M(c) < M(b) < M(a) = 1$.

4.3.3 Possible measures

Fortunately, the formal problem of comparing partitionings has been addressed in numerous ways in the literature. There are at least four families of approaches: based on stimulus pairs, category association, set matching, and variation of information.

Rand (1971) defined an early measure based on counting pairs. The idea is to look at each stimulus pair individually, and see if the partitionings are in "agreement". The partitionings agree if they either both put the pair in the same category, or both put the pair in different categories. The "Rand Index" is then the ratio of the agreements to the total number of pairs.

More formally, given two partitionings p_1 and p_2, we first define four values:

- $a =$ number of pairs put in the same category in p_1 and in the same category in p_2
- $b =$ number of pairs put in the same category in p_1 and in different categories in p_2
- $c =$ number of pairs put in different categories in p_1 and in the same category in p_2
- $d =$ number of pairs put in different categories in p_1 and in different categories in p_2

Then the Rand Index, RI, is given by:

$$RI = (a + d)/(a + b + c + d) \tag{4.1}$$

It can be easily verified that for our four examples above, RI yields (a) 1.00, (b) 0.89, (c) 0.53, and (d) 0.60. Notice that this doesn't quite satisfy the conditions specified earlier: the score in case (d) is too high relative to (c) and far above 0.

In fact, the Rand Index, although simple and reasonably intuitive, has other problems. For example, this measure (and other similar measures) is not corrected for what values are expected by chance. In particular, it does not take into account the varying numbers and sizes of categories that may be used by participants. Indeed, this is largely responsible for the counterintuitively high value on this measure for Example (d).

To address these problems, various modifications of the Rand Index have been proposed. For instance, Hubert and Arabie (1985) modify the Rand Index by correcting for its expected value. Under their modification, the measure yields a value of zero for the level of agreement which is expected by chance, taking into account the number of categories and their sizes. However, under such a definition, the measure can now sometimes take negative values (so it's no longer bounded by $[0, 1]$), along with other undesirable properties (Meila 2007).

Another kind of agreement measure, developed by Wills and McLaren (1998), is based on Cramer's phi statistic (Cramer 1946). The original statistic is designed to measure the association between two categorical variables, and is in turn based on Pearson's standard χ^2 statistic. The formula is given as:

$$\phi_c = \sqrt{\frac{\chi^2}{N(k-1)}} \tag{4.2}$$

where χ^2 is the χ^2 statistic based on the contingency table between the two sets of categories, N is the number of items being categorised, and k is the number of categories used in the smaller of the two sets.

However, as Wills and McLaren (1998) pointed out, this measure does not take into account the fact that participants can use different numbers of categories. To address this, they defined an adjusted measure, which corrected the value for what would be expected by chance. The value expected by chance, ϕ_{chance} (see their paper for mathematical justifications), is given by

$$\phi_{chance} = \sqrt{\frac{2(r-1)(c-1)-1}{2(N-1)(k-1)}} \tag{4.3}$$

where r and c are the number of categories in the two sets. The adjusted Cramer's phi, ϕ_{adj}, is then

$$\phi_{adj} = \frac{\phi_c - \phi_{chance}}{1 - \phi_{chance}} \tag{4.4}$$

With this adjustment, claim Wills and McLaren (1998), the measure yields a value of 0 for the degree of agreement expected by chance, independent of the number of categories in the two sets.

For the four examples above, ϕ_{adj} gives (a) 1.00, (b) 0.76, (c) 0.24, (d) -0.39. Thus, although the measure decreases monotonically for these examples, it does not satisfy the conditions laid out earlier, since it is not bounded on $[0, 1]$ (like Hubert and Arabie's 1985 measure).

A third basis for obtaining an agreement measure is set matching. In this approach, the sets in one partitioning are matched with the sets in the other partitioning in every possible way, and the total number of matching elements is counted for each alignment. The score is then based on the alignment for which this count is maximised. For instance, Meila and Heckerman (2001) defined the measure

$$H(p_1, p_2) = \frac{1}{n} \max_{\pi} \sum_{k=1}^{K} n_{k,\pi(k)} \tag{4.5}$$

where k indexes a category, K is the number of categories in the partitioning with less categories, and π is an injective mapping from $\{1, ..., K\}$ to $\{1, ..., K^{'}\}$.

Under this definition, for our examples, M gives: (a) 1.00, (b) 0.90, (c) 0.50, and (d) 0.30. These values do satisfy the criteria above.

Other measures based on set matching have also been proposed, with different ways of determining how to match the sets. None of these methods, however, considers the "unmatched" parts of the sets, and thereby fails to capture certain intuitions (Meila 2007). For example, if two partitionings differ with respect to a few stimuli, then we would expect the measure to yield higher numbers if all of these stimuli were together in a separate category than if they were scattered across the other categories.

A fourth agreement measure is derived from information theory, based on the notions of entropy and information. The basic idea is to define a measure in terms of how much information is lost or gained in shifting from one partitioning to another. The more difference in information there is between the two partitionings, the higher is the value along this measure.

The full mathematical derivation of the measure and its properties (see Meila 2007) is relatively complex. Here I present only the equations needed for calculations of the measure, and a brief summary of its most relevant properties.

First, given a particular partitioning C, the probability $P(k)$ that a randomly selected item is in a particular category C_k is given by:

$$P(k) = \frac{|C_k|}{n}. \tag{4.6}$$

By extension, the joint probability $P(k, k')$ can be defined as

$$P(k, k') = \frac{|\int C_k C_{k'}|}{n} \tag{4.7}$$

Then define the entropy of a partitioning C, $H(C)$, as

$$H(C) = -\sum_{k=1}^{K} P(k) \log P(k). \tag{4.8}$$

Next, define the mutual information of two partitionings to be:

$$I(C, C') = \sum_{k=1}^{K} \sum_{k'=1}^{K'} P(k, k') \log \frac{P(k, k')}{P(k)P'(k')} \tag{4.9}$$

Finally, the variation of information between two partitionings is then a function of the entropies and the mutual information:

$$VI(C, C') = 1 - \frac{H(C) + H(C') - I(C, C')}{\log n}. \tag{4.10}$$

Note that the $\log n$ term normalises the measure so that it is bounded on $[0, 1]$, and the subtraction from 1 reverses the direction of the scores so that a value of 1 indicates that the two partitionings are identical, as before.

For the example above, this measure yields: (a) 1.00, (b) 0.82, (c) 0.41, and (d) 0.15. Like the set-matching measure, this one satisfies the conditions laid out earlier. In addition, it also gives more divergent values between the cases.

Although this kind of measure is more complicated than the two discussed previously, it has several favourable properties. Most importantly, for any fixed number of stimuli n, it is, mathematically speaking, a metric. In particular, it is symmetric and satisfies the triangle inequality. This means that we can visualise different partitionings as occupying distinct points in a geometric Euclidean space. Moreover, this measure handles the problem mentioned earlier of the unmatched parts of the sets in an intuitive way, handling cases where categories are split or merged. For details, see Meila (2007).

For my purposes, I will generally use Meila's (2007) measure based on information theory, due to its useful properties. However, when participants are constrained to partition the items into exactly two categories, it seems better to use a set-matching approach: with only two categories, the drawbacks of this approach do not apply, making its simplicity preferable. In practice, this means that I will use the information theory method in Experiments 1, 2 and 3, and the set-matching method in Experiment 4.

4.4 Experimental manipulations

In Section 4.1, I described how the framework should support various kinds of manipulations. Therefore, before we commit to free classification tasks, it's important to consider whether they can support them effectively. Fortunately, they are quite amenable to the kinds of manipulations that I need.

The framework can be easily adapted to study coordination due to one's native language (see Section 3.3), or brought about by interaction (see Section 3.4). In the first case, participants with different native languages are tested individually, on either linguistic or non-linguistic categorisation tasks (which can be achieved by manipulating whether they are asked for category labels or not). Their free classification outputs are then compared both to those of other native speakers of the same language and to native speakers of other languages. The interactive case can also be addressed. Two participants can carry out a categorisation task at the same time, with the same stimuli and the same constraints. Alternatively, they could be asked to sort stimuli together, generating just one output between them. Depending on the experimental condition, they may be allowed to speak to each other during the tasks, and they could be given feedback on their partner's categories and/or labels. Participants could also be asked to carry out multiple tasks in succession, to give them time to learn and coordinate better. Many other manipulations are also possible, but these are the main types that I will use.

However, the success of the framework in handling these manipulations will no doubt be judged by whether or not they yield significant and interesting results. This is an empirical matter that cannot be decided a priori. Therefore, I do not discuss it further here: we will see from the experiments themselves the real answers about how well the framework supports these manipulations.

4.5 Stimulus considerations

In setting out my general framework criteria, I identified three subcriteria relating directly to stimuli. First, stimuli should be non-linguistic, so that we can separate language from conceptualisation. Second, they should come from a familiar domain, so that preexisting word-concept associations can be studied. Third, they should be prone to variability in categorisation, to maximise potential for variation and coordination. How can these criteria be satisfied?

The first two criteria are straight-forward, and have a lot of precedents. Although many categorisation experiments use verbal stimuli, whether auditory or written, visual pictures are probably even more common, and often involve familiar stimuli. Indeed, the categorisation literature has long been dominated by object concepts, which are commonly presented in pictorial form. Particularly frequently used stimulus domains which are easily presented visually include animals (e.g., Rousselet, Fabre-Thorpe and Thorpe 2002), foods (e.g., Ross and Murphy 1999), faces (e.g., Goldstone, Lippa and Shiffrin 2001), artifacts (e.g., Malt et al. 1999), shapes (e.g., Rothbart and Lewis 1988) and dot patterns (e.g., Zaki 2004). Most such domains are highly familiar to participants of various backgrounds.

It is less clear how to best satisfy the third criterion: what kinds of stimuli should be selected to maximise the potential variation in conceptualisation? First of all, we can use fluid stimulus domains, in which the boundaries between concepts are not obvious or even stable. Fluidity can be achieved through fine sampling of a domain which is already naturally fluid, or through the use of computer programs. An example of the former is Malt et al. (1999), in which the experimenters used photographs of real containers. While prototypical bottles, jars and tubes may be distinct, there are many intermediate items which might be ambiguous with respect to these categories. The use of computer software to achieve fluidity is exemplified in Goldstone et al. (2001), in which pairs of distinct human faces were fed into a morphing program, yielding a sequence of faces with interpolated features with images that still looked human. Computer programs can be used even more easily with artificial stimuli, by manipulating the parameters used to generate them along a continuum (Goldstone 1994).

Another simple way to produce variability in categorisation could be to use stimuli with multiple prominent dimensions. For instance, fruit vary visually in terms of at least colour, size, shape and texture. However, what is harder is getting the changes along different dimensions to be of roughly equal salience. How much of a shift in colour corresponds to a particular shift in size? This cannot be answered in purely objective, physical terms, but must be treated psychophysically (Shepard 1987), since we know that human perceptual sensitivity is not uniform even along a single dimension like colour hue (Wright and Pitt 1934). As a result, even for artificial geometric stimuli, for which we can often finely and precisely define values along particular dimensions, it is not clear a priori how these shifts correspond to shifts in psychological representations. Therefore, in selecting stimuli which vary along multiple dimensions, we cannot guarantee to sample the domain in a perfectly uniform manner.

Finally, a third way of maximising conceptual variation is by using domains for which cross-linguistic variability has already been demonstrated. If a stimulus domain exhibits cross-linguistic differences in its conceptual partitioning, this is evidence that its stimuli do not fall into uncontroversial natural kinds. As a result, such a domain is relatively likely to exhibit differences in how individuals conceptualise items, even for speakers of the same language. Researchers have argued that there are differences in how much different domains vary cross-linguistically. For instance, it has been claimed that there is more cross-linguistic variation in verbs than in nouns (Gentner 2006). On the other hand, there is also evidence for at least some cross-linguistic differences in concrete object domains like containers (Malt et al. 1999) and dishes (Ameel, Storms, Malt and Sloman 2005).

Putting these considerations together, I have settled on two kinds of domains for use in my experiments: dishes and triangles. Both of these domains are familiar and consist of non-linguistic stimuli. They can be sampled in such a way as to create a highly fluid set, augmented further with the help of morphing software or other computer programs. They vary along multiple dimensions such as shape and size, and the dishes domain has already been shown to exhibit cross-linguistic variation (Ameel et al. 2005). Moreover, using one natural and one artificial domain increases the generalisability of potential findings. Further details on how the particular stimulus sets were chosen and generated are provided in the individual experimental chapters.

4.6 Summary

In this chapter, I have laid out a general experimental framework aimed at addressing specific research questions. I started by laying out a few critical criteria for my framework, and then developing a framework which matched them. I began by arguing that categorisation tasks are the best way for us to assess conceptualisation, even though this is theoretically dissatisfying. I then reviewed different kinds of categorisation tasks, and argued that free classification, in which people sort things relatively freely into groups, was the most appropriate for my purposes. I then listed and assessed different measures for comparing two outputs of free classification tasks, and suggested how the framework could be manipulated suitably in my experiments. Finally, I considered some specific subcriteria for stimulus domain selection and sampling, and how they will be satisfied.

CHAPTER 5

Experiment 1: native language and prior conceptual alignment

5.1 Introduction

In Chapter 3, I distinguished two processes, operating at different timescales, in which language could plausibly play a causal convergent role: the developmental shaping of concepts, and the online application of concepts. In this chapter, I consider the first of these options empirically, albeit by examining the adult end state. As such, we are in the realm of the controversial linguistic relativity hypothesis, whereby our native language shapes our conceptual repertoire (see Section 3.3). To the extent that this view may be correct, it should follow that people's concepts would generally be more in line with those of other speakers of the same language. This is the main question behind Experiment 1: do speakers of the same native language conceptualise things more similarly to each other than speakers of different languages?

Recall that, according to Whorf's (1956) strong original formulations of linguistic relativity, our native language would determine our conceptual system. If that were the case, two native speakers of the same language would conceptualise things in the same way as each other, but in general differently than two native speakers of different languages. In other words, Whorf's position immediately predicts a positive answer to the question above.

However, as we have seen, the empirical evidence for linguistic relativity is mixed, which has resulted in weaker modern forms of the hypothesis. While current theories vary in their details, they share the basic principle that our native language influences how we see things in the world, without fully determining it. It has now been established that our language *can* have an effect on our conceptualisation, but this effect is limited and not pervasive. Therefore, further study in this area should stop asking whether language affects conceptualisation and focus instead on the scope and depth of this influence (Kay and Regier 2006).

This means, among other things, that linguistic relativity research should no longer confine itself to particular favourite domains, like colour. Indeed, one striking aspect of the literature is that the domain of object concepts, otherwise so pervasive in cognitive psychology experiments and cognitive science research in general, is relatively unrepresented in studies of linguistic relativity. As we have seen, although a fair number of studies have looked at how cross-linguistic grammatical differences may result in differences in object conceptualisation, very few have considered the potential effects of different languages *lexically* partitioning an object domain in different ways.

An important exception to this is Malt, Sloman, Gennari, Shi and Wang's (1999) cross-linguistic experiment, which suggested that native language has little or no effect on the non-linguistic categorisation of artifacts. In line with my methodological choices in Chapter 4, they used a free classification paradigm to probe individuals' conceptualisations. However, their analysis was not focused on the question of conceptual coordination between individuals, and those of their results that relate to it are equivocal.

As a result, in this chapter, I present an experiment which is based heavily on Malt et al.'s (1999) study but addresses slightly different target questions. I aim both to test the robustness of their findings by using different languages and a different set of stimuli, and to supplement their analysis by using different analytical measures and statistical tests.

In the rest of the chapter, I first discuss Malt et al.'s (1999) experiment in more detail, and how and why I propose to modify their design. Then I present my study, and finally draw conclusions concerning whether native languages induce conceptual coordination.

5.2 Malt et al. (1999)

5.2.1 *Purpose*

The main purpose of Malt et al.'s study was to explore the relationship between how people recognise objects and how they name them. They called these two types "recognition categories" and "linguistic categories", respectively. It is often assumed, they point out, that these two types of categorisation are virtually isomorphic, so that the name used for an object directly reflects the kind of thing that the object is recognised to be. And intuitively, this makes sense: after all, words refer to categories of things in the world (even though, as I argued in Section 2.4, categories are determined by concepts, and so can vary from person to person).

However, as Malt et al. point out, the correspondence may not be as tight as we might at first expect, because the two processes serve different functions. Object naming is a communicative act, while recognition is not. Consequently, there may be different pressures on how the two types of categories are formed, resulting in a potential dissociation between recognition categories and linguistic categories. This issue is of course related to the Sapir-Whorf hypothesis: discovering that recognition categories are not the same as linguistic categories would be evidence against the idea of language determining thought. It also relates to my model of conceptualisation, and Malt et al.'s (1999) linguistic and recognition categories can be loosely identified with lexicalised and non-lexicalised conceptualisation, respectively. This is shown in Figure 5.1.

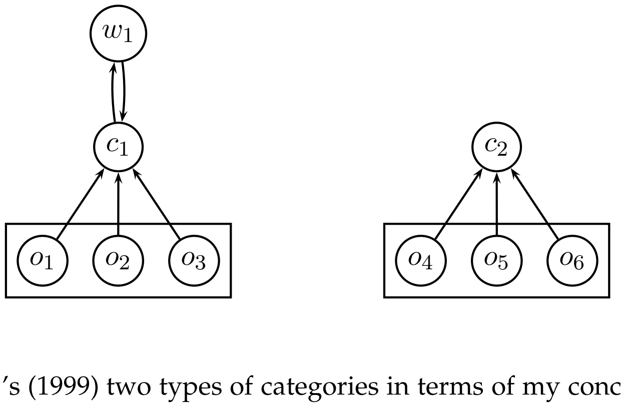


Figure 5.1: Malt et al.'s (1999) two types of categories in terms of my conceptualisation model: linguistic (left) and recognition (right).

Malt et al.'s experiment was designed to test whether there was a dissociation between the two kinds of categories. More specifically, they aimed to answer three questions (p. 237):

1. "Does the division of objects into linguistic categories differ across the three languages for this set of objects?"
2. "Does the perception of similarity among the objects differ across speakers of the three languages?"
3. "If at least some differences in linguistic categorization and perceived similarity are found, do these differences parallel one another?"

5.2.2 Design

The participants in Malt et al.'s study were native speakers of Chinese, English and Spanish, who were undergraduate students in China, the United States, and Argentina, respectively. They were mostly monolinguals, although some of the Chinese and Spanish speakers did know limited amounts of English. They were individually tested in their native language in their respective countries.

The stimuli consisted of photographs taken by the experimenters of sixty common containers. They were selected to represent as wide a range as possible of containers that were likely to be called "bottle" or "jar" in English, or to share features in common with these categories.

All participants carried out two kinds of tasks: sorting tasks and naming tasks. These were meant to capture participants' recognition and linguistic categories, respectively. In the sorting tasks, they were asked to sort the full set of sixty stimuli into categories based on similarity. The number, size, and nature of their categories were up to them. There were three kinds of sorting tasks, depending on what kinds of features participants were asked to base their sorts on: physical, functional or overall. All participants carried out two of these three sorts in succession. These were thus free classification tasks (see Section 4.2.2.5), with the constraint that the general basis for categorisation was provided by the experimenter. In the naming tasks, participants were asked to independently name each stimulus (in their native language). They were to name things the way they normally would to refer to the objects, and their names could consist of one or more words. All participants named each of the 60 stimuli. The resulting bodies of linguistic and non-linguistic categorisation data for each of the three languages groups provided the basis for testing Malt et al.'s research questions.

5.2.3 Analysis and results

The first step in Malt et al. (1999)'s analysis was to make the naming data comparable in form to the sorting data. Although participants named each object independently, linguistic categories could be induced by bringing together all items that were given the same label. However, since participants were given complete freedom in the basis and length of their expressions, they could, in principle, name each item (slightly) differently by using long and detailed descriptions. Malt et al. dealt with this by collapsing participants' free labels to the head nouns they used in their descriptions. For instance, "a small white bottle" would be treated as "bottle". By doing this for all the labels of a participant, they obtained a partitioning of the stimuli into a manageable number of basic noun categories for that participant, parallelling the sorting data.

Having done this, Malt et al. (1999) conducted different kinds of analysis. Since it is very relevant for the experiment I present in this chapter, I discuss a quick summary here (but see their paper for details). First, for each language, they looked at the the distribution of dominant names, which they defined as the most commonly used head nouns for particular objects. They showed that these distributions for the different languages diverged from each other. In a subsequent analysis of the same data, Malt et al. (2003) elaborated this analysis further and showed that the differences in the naming patterns could not be attributed to any single explanation, such as one language simply using more superordinate terms than another.

Also, for each language, they derived a measure of inter-object similarity of every pair of objects for both sorting and naming, based on the extent to which they were treated the same (i.e., were sorted into the same groups, or had similar naming distributions, respectively). This resulted in several sets of inter-object distance data, one for each combination of language and task, which were then used as input to further analyses. First, they computed correlations of the resulting values between different pairs of datasets to address their three research questions. The relative values of these correlations showed that the languages differed more in naming than in sorting, and that the naming of a given language did not correspond to its sorting any more than it did to other languages' sorting. Since these results capture their main findings succinctly, I reproduce the correlation values in Table 5.1. Second, they presented multidimensional scaling diagrams which plotted the objects in the similarity space derived from sorting, with each object labelled with its dominant name. This showed visually how the dominant names of a language did not correspond closely with the way objects mapped onto a similarity space based on their sorting data. Third, they used the inter-object similarities in both naming and sorting to test whether the discrepancies between naming and sorting tended to occur for the same pairs of objects in the different languages. Correlations suggested that there were correspondences, but that they were not strong.

Naming		
	American	Argentinean
Chinese	.35	.55
American		.54

Sorting		
	American	Argentinean
Chinese	.89	.82
American		.88

	Naming		
Sorting	American	Argentinean	Chinese
American	.70	.71	.43
Argentinean	.65	.78	.48
Chinese	.65	.68	.47

Table 5.1: Correlations comparing the naming patterns between languages (top left), the sorting patterns between languages (top right), and the sorting-naming correspondence both between and within languages (bottom). The sorting values come from the physical sorting tasks.

Finally, they derived a further measure, quite similar to the Rand Index (see Section 4.3.3), which was meant to capture the difference between the naming or sorting outputs of two individuals, irrespective of their language group. Principal components analyses were then conducted on the resulting sets of values to see if factors emerged that distinguished the groups. The analyses generally did find such factors, but more distinguishing factors emerged for naming than for sorting, and the differences in the factor loadings were also greater in the case of naming. This too suggests that the different languages named things differently from each other but sorted things relatively similarly.

5.2.4 Interpretation

Based on their results using these complementary forms of analysis, Malt et al. conclude that the different language groups name the objects differently from each other despite sorting them quite similarly, and that there is no clean direct correspondence between sorting and naming. In terms of linguistic relativity, this means that one's native language appears not to determine one's non-linguistic categorisation of the world. Similarly, in terms of my hypothesis of conceptual coordination, it suggests that different speakers of the same language do not generally conceptualise objects in line with their language and hence in line with each other.

These conclusions make a good case against a deterministic role of language on non-linguistic categorisation. However, it's important to recognise that they only reject a strong Whorfian position, which, as mentioned earlier, has been already abandoned and replaced by weaker versions in recent years. Malt et al. emphasise the fact that there is greater divergence in linguistic than in non-linguistic categorisation between language groups. But that merely shows that language affects the former *more* than the latter, not that it fails to affect the latter at all. In other words, just because non-linguistic categorisation fails to directly reflect linguistic categorisation does not imply that it is unaffected by language. Indeed, particularly weak formulations of linguistic relativity, such as Slobin's (2003) "thinking-for-speaking" (see Section 3.3.5), might even predict that linguistic effects on conceptualisation would only occur during explicit language use (and so not during non-linguistic sorting, for example). So what do Malt et al.'s findings tell us in terms of weaker forms of linguistic relativity?

To answer that, we need to focus on the sorting results, and whether they showed an effect of native language. In the correlational analysis based on object pairs, Malt et al. found high degrees of agreement for all three types of non-linguistic sorting (physical, functional and overall), consistently higher than the results for naming. But what is "high"? As Malt et al. acknowledge, their analysis is difficult to assess with significance tests, because looking at all object pairs means that the data entries are not independent of each other. This makes it difficult to put these numbers in perspective: is there no difference between the language groups, or are the differences merely small? Similarly, in the other relevant analysis, based on comparisons of results from participant pairs using principal components analysis, Malt et al. found that for two of the three sorting types, factors emerged which distinguished between the linguistic groups. In other words, there was evidence that speakers of different languages did sort things differently. Thus, although Malt et al. emphasised that their results were less divergent for sorting than for naming, they may still be consistent with a weak linguistic relativity position.

On the other hand, even if cross-linguistic differences in non-linguistic categorisation are found, it does not necessarily imply that language is responsible. This is why collecting data on linguistic categorisation is also important. However, the naming data from Malt et al.'s study needs to be treated with caution. Recall that the first step in their analysis was to collapse all the linguistic expressions down to their head nouns. This is an intuitively sensible and necessary simplification to make the rest of the analysis possible. However, it does also risk throwing away important information and making the naming differences between languages seem larger (or smaller) than they actually are. Indeed, in their discussion of naming patterns in the data, Malt, Sloman and Gennari (2003) noted that some of the cross-linguistic variation may be due to systematic morphosyntactic differences between the languages (e.g., diminutives in Spanish, classifiers in Chinese), which may result in an unequal burden on head nouns. This may be important, because all of the naming results were based on these head nouns. Therefore, it is possible that the analysis exaggerates the lack of correspondence between linguistic and non-linguistic categorisation.

It's also worth noting that Malt et al.'s (1999) analysis did not generally focus on the differences between individuals and thus directly test the possibility that sharing the same native language results in conceptual coordination. The one type of analysis that did focus on inter-individual comparisons yielded relatively clear inter-group differences for naming and weaker results for sorting. Moreover, the measure used for quantitative comparisons of two people's categorisations was closely related to the Rand Index. However, as I showed in Section 4.3.3, this measure has shortcomings and there are better alternatives available. Since comparing categorisations of individuals yielded mixed results in Malt et al.'s study and is central to my thesis, the choice of measure may be important. In addition, as I will show later in this chapter, it is possible to test the conceptual coordination hypothesis more directly and with a different type of analysis.

Finally, the robustness and scope of Malt et al.'s findings are unclear. As I surveyed in Section 3.3.3, linguistic relativity research has barely investigated how different languages might partition object domains differently, and whether or not the way they do so impacts how their native speakers perceive or categorise the domain non-linguistically. An important exception is Ameel, Storms, Malt and Sloman (2005), who replicated Malt et al.'s experiment as part of their bilingual study. Their replication tested French and Dutch monolinguals in Belgium, and tested two object domains: containers (as in Malt et al.'s study) and dishes. Both domains were sampled in a similar way to the original study, using as wide a range as possible of relevant stimuli. Their analysis yielded similar results to the original study: there were greater divergences between the languages in naming than in sorting (although the difference was smaller for the dishes domain), and sorting was less affected by one's native language (if at all) than naming.

In summary, Malt et al.'s results document divergence in the linguistic categorisation of different languages, but this does not appear to be reflected in non-linguistic categorisation. However, while the results refute a strong Whorfian position, this has already been generally dismissed in linguistic relativity research. It is not surprising that our native languages affect linguistic categorisation more than non-linguistic categorisation. In contrast, with respect to current weaker Whorfian positions, the results sit on the fence. In particular, Malt et al.'s results are equivocal with respect to the question of whether our native language affects the non-linguistic categorisation of objects at all. Malt et al.'s did not place emphasis on this, but their forms of analysis can be modified slightly and supplemented to address the hypothesis of conceptual convergence more directly. Subsequent work has yet to achieve this. I therefore address these issues in a modified replication of Malt et al.'s experiment, which I present next.

5.3 Experiment 1

5.3.1 *Overview*

Malt et al.'s (1999) study used novel methodological methods to provide evidence for a dissociation between linguistic and non-linguistic categorisation in the object domain. However, the object domain remains largely understudied in linguistic relativity research, so that the scope and depth of their findings are unclear.

More importantly for my purposes, their experiment yielded equivocal results concerning one of the central hypotheses of this thesis. In particular, it remains unclear whether native speakers of the same language exhibit a significant degree of conceptual alignment with each other when categorising non-linguistically. The basic comparison under consideration is shown generically in terms of my model of conceptualisation in Figure 5.2. Thus, my main focus in analysis will be on inter-lingual and intra-lingual comparisons in sorting tasks. However, I will also look at the naming data and the correspondences between them, in order to contextualise the results and make the best use of my data.

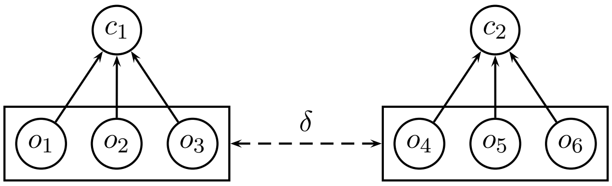


Figure 5.2: A comparison between the outputs on a non-linguistic categorisation task for two participants. The experimental hypothesis is that δ will be smaller when the two categorisers are native speakers of the same language, even in non-linguistic categorisation tasks.

Therefore, Experiment 1 has two main purposes:

1. **Do linguistic and non-linguistic categorisation dissociate in the way suggested by Malt et al.'s (1999)?**
2. **Are native speakers of the same language more conceptually aligned, by default, than speakers of different languages?**

Given the shortcomings I have identified of the original study with respect to my hypothesis, these two goals are partly at odds and wedding them requires a bit of compromise. Nevertheless, provided both purposes can be met, this yields obvious benefits, since it would allow me to directly address my thesis goals while also clearly relating my results to the limited amount of previous work that has been done in this area.

In terms of replication, I use largely the same experimental procedures as in the original study. Thus, as in the original experiments, each participant carries out both naming and sorting tasks. Despite the potential problems I identified with the naming tasks, I use the same procedure, and also analyse the naming data exclusively in terms of the head nouns used. I stick to the original procedure for three reasons. First, it allows for direct comparison of the results with the original study. Second, despite its shortcomings, it is still an intuitive basis for analysis and it is not clear what would be a superior alternative. Third, the primary emphasis of my study is on whether one's native language has an effect on sorting, not naming.

However, my study will differ from the original (as well as Ameel et al.'s 2005 replication) in the native languages selected, the language profiles and location of the participants, and the choice and sampling of the object domain. In particular, I will compare native speakers of English, Polish and Japanese. In contrast to the original study, the speakers are not necessarily monolingual (in fact, all the Polish and Japanese participants had at least a fairly good grasp of English), and rather than being tested in their home countries, they were all residing in Edinburgh, UK. This means that they could be expected to be less divergent both linguistically and culturally (even compared to the participants in Ameel et al.'s replication), and thus, if anything, are less likely to exhibit group differences in both linguistic and non-linguistic categorisation. The stimulus domain is dishes, which was not the case in Malt et al.'s (1999) original experiment, but was one of the two domains used in Ameel et al.'s replication (which found a smaller dissociation between naming and sorting in the dish domain than the container domain). This domain is appropriate because it exhibits all of the important properties identified in developing my framework in section 4.5: it is familiar, non-linguistic, fluid, multi-dimensional, and cross-linguistically variable. However, I sampled this domain differently than in the previous studies: rather than attempting to sample as large a variety of relevant stimuli as possible, a fairly narrow set was chosen, often with subtle differences between stimuli. This was done in hopes of increasing the amount of variation in categorisation among participants, by having more ambiguous items whose status with respect to certain category boundaries was debatable. Together, these modifications allow us to test the scope and robustness of the original results.

Moreover, I will introduce a couple of changes in analysis, which aim to improve and supplement the original methods used. First, I will use a slightly different method for calculating the distance between object pairs in naming, which is more directly analogous and comparable to the method used for sorting. Second, in order to compare the categorisations of two individuals, I will use Meila's (2007) information-theory method, which has advantages over methods based on the Rand Index (such as Malt et al.'s 1999, as argued in Section 4.3.3). Third, and most crucially, I will add a form of analysis which will allow for the comparison of conceptual agreement between individuals in different conditions. In particular, I will (among other tests) directly test whether pairs of participants who are native speakers of the same language categorised the stimuli more similarly than native speakers of different languages.

This last method of analysis will also have the advantage of providing concrete significance tests. Such tests are lacking in Malt et al.'s (1999) methods, due largely to the complexity of the data and the non-independence of their units of observation. This resulted in many of their analyses and conclusions (and inevitably also those in my replication analyses) being based on rather subjective comparisons of different values. Supplementing their methods with objective tests is beneficial, as it will serve to either increase confidence in their results or provide good reasons for putting them in doubt.

Finally, I will also counterbalance the order in which participants carry out the naming and sorting tasks. In the original experiment, participants always did sorting first. However, given the kinds of dynamic effects of words on conceptualisation that have been documented (see Section 2.6.5), it is possible that the order would matter. As a result, in my experiment, half of the participants (in each language) sorted the stimuli first, and the other half named them first. I then checked whether there is greater alignment between individuals in one condition or the other.

5.3.2 Method

5.3.2.1 Participants

Participants were mainly undergraduate students recruited through the University of Edinburgh student employment website. However, some postgraduate students and non-students also participated, especially in the Japanese group. There were 24 native speakers for each of three different languages: English (age: M=20.8, SD=2.8; gender: female=20, male=4), Polish (age: M=20.4, SD=2.5; gender: female=17, male=7) and Japanese (age: M=26.8, SD=6.1; gender: female=17, male=7). However, all of the Polish- and Japanese-speaking participants also had an intermediate or high level of English, and some participants also spoke other languages.

5.3.2.2 Stimuli

The stimuli consisted of sixty simple photographs of dish-like objects taken from online catalogues of IKEA and other stores, printed and laminated on 6cm × 6cm size white paper (Figure 5.3). The items were primarily chosen to be variants of what might be called "plates" and "bowls" in English, together with other items that shared many features in common with such items. The amount of variability among these stimuli was substantially smaller than in the original studies, so that effectively the domain was smaller but was sampled more finely.

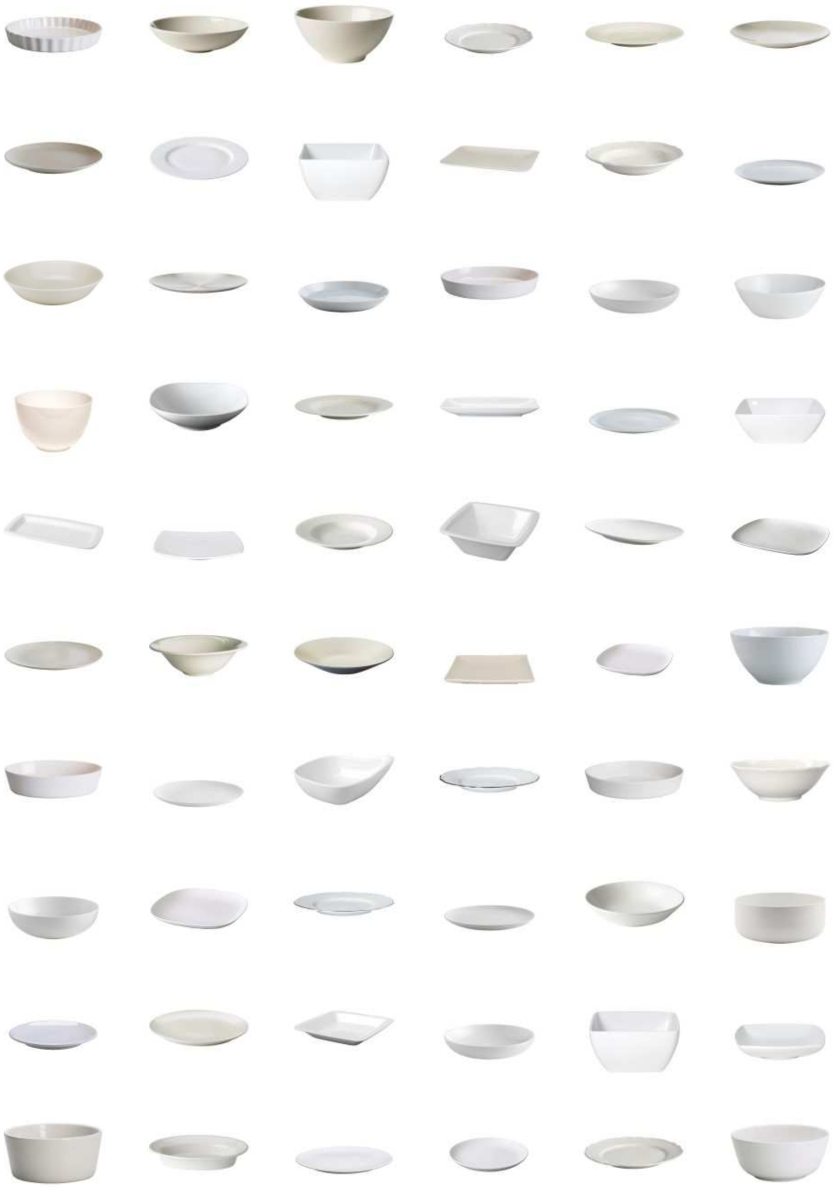


Figure 5.3: The stimuli used in Experiment 1.

5.3.2.3 Conditions

The main experimental condition was the participant's native language. To the extent possible, the experiment sessions took place entirely in the native language of the participants. The experimenter (who is native in English, fluent in Polish, and intermediate in Japanese) greeted and conversed with the participants in their native language. The written instructions (which were the main form provided) for the tasks were also in their native language, having been translated from English into Polish and Japanese together with the experimenter by native speakers of the respective languages who were also fluent in English. Naming data was collected in the native language of the participants (and the Japanese participants were asked to write the names in hiragana or katakana).

In addition, the order of the naming and sorting task was counterbalanced, so that half of the participants in each linguistic condition carried out the sorting task first (as in the original experiment), while the other half did the naming task first.

5.3.2.4 Procedure

Participants carried out one sorting task and one naming task, with the order of the tasks depending on the experimental condition they were randomly assigned to. After the tasks, they filled out a short questionnaire.

The tasks were nearly identical to the original experiment. In the naming task, participants were given the pile of stimulus cards (in random order) and asked to write down, for each stimulus, the number on the back of the card together with the name that they would naturally apply to the item in normal conversation. In the sorting task, the experimenter laid out all sixty cards (in random order) on a table, and the participants were asked to arrange them into groups based on similarity. Participants were asked to sort the stimuli based on physical features (recall that participants sorted based on either physical, functional or overall features in Malt et al. 1999, and based on overall features in Ameel et al. 2005). This was done because the categorisation criteria made little difference in the original experiment, and instructions for physical sorts seem the most concrete and therefore least susceptible to translation-induced differences. Also, the range of the allowed number of categories in the sorting task was more restricted, with a minimum of three categories and a maximum of eight (in contrast with Malt et al.'s original range of 2-15), because the stimulus domain was more restricted in this experiment.

5.3.3 Results

As this is largely a replication study, my analysis will, for the most part, follow the methods of Malt et al. (1999). In presenting my results, I will relate them to theirs (and, where more relevant, to those of Ameel et al. 2005). I first present results for the naming data, then move on to the sorting data, and finish by looking at the relations between them.

However, in order to address some of the issues identified in the introduction and test my own hypotheses more directly, my analysis will diverge from theirs in three main ways. First, I adapt a measure they used to analyse the naming patterns of the different languages to be analogous (and hence more comparable) to the analysis of sorting patterns. Second, I use a different measure for comparing the categorisation outputs of individuals, in line with my discussion of such measures in Section 4.3.3. Third, I add a new set of tests which compare the degrees of alignment between individuals in different conditions. These tests are based on randomly pairing up the outputs of different people under the constraints of particular conditions of interest. My analytic modifications will be explained in greater detail as they come up.

5.3.3.1 Example output

Before analysing the data, I first show an example of the experiment output from a randomly selected native English speaking participant. Figure 5.4 shows the sorting output, and Figures 5.5 and 5.6 show the naming output.

5.3.3.2 Comparisons of naming patterns

For each of the three languages, participants' full labels for each object were first collapsed to their head nouns. Lexical categories were then induced from the resulting labels. For instance, "large round bowl", "small bowl", "bowl" and "bowl for cereal" are all collapsed to the category "bowl". In addition, Polish diminutives were collapsed to the canonical form. For example, "miska" (bowl) and "miseczka" (little bowl) were both

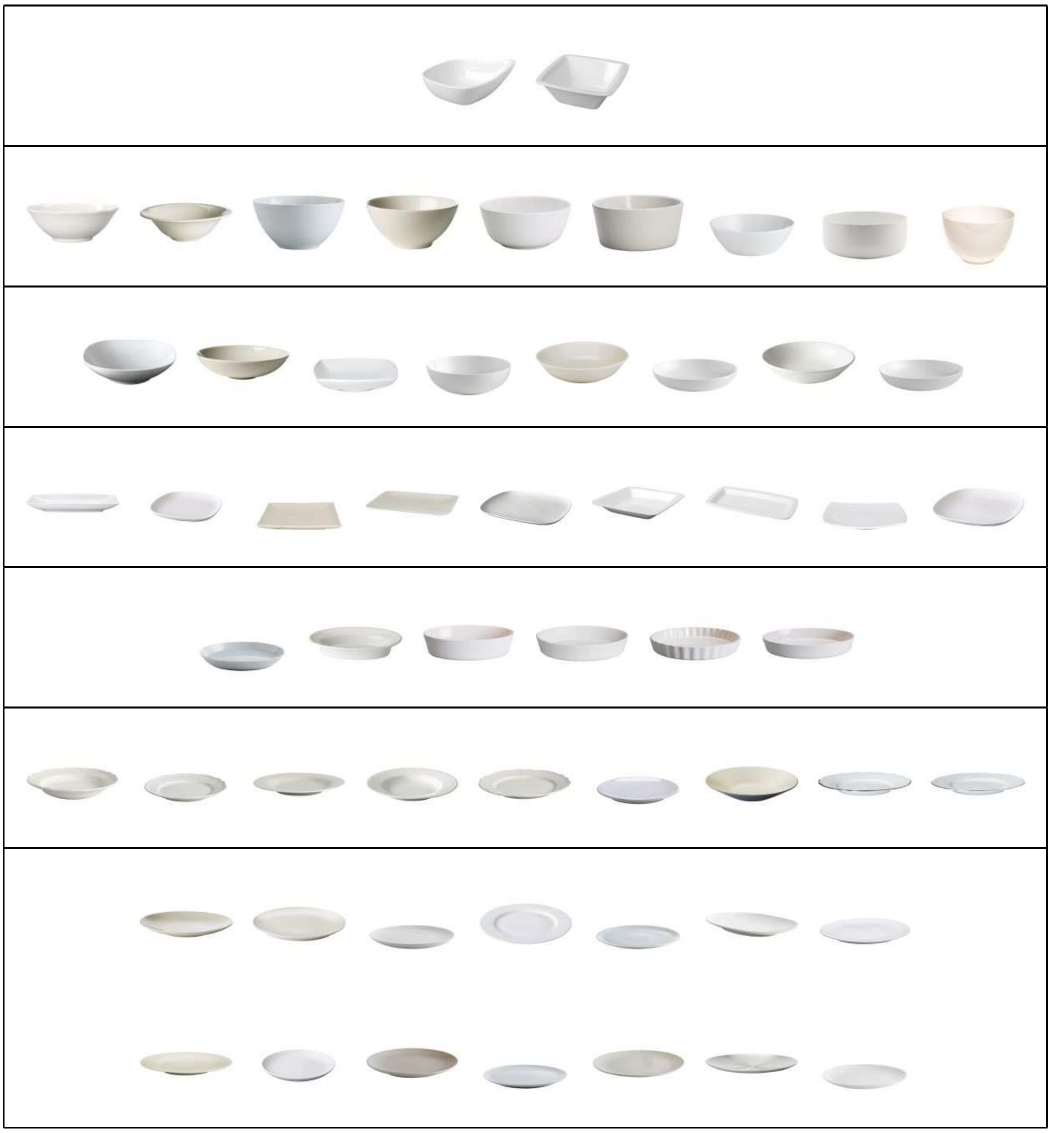
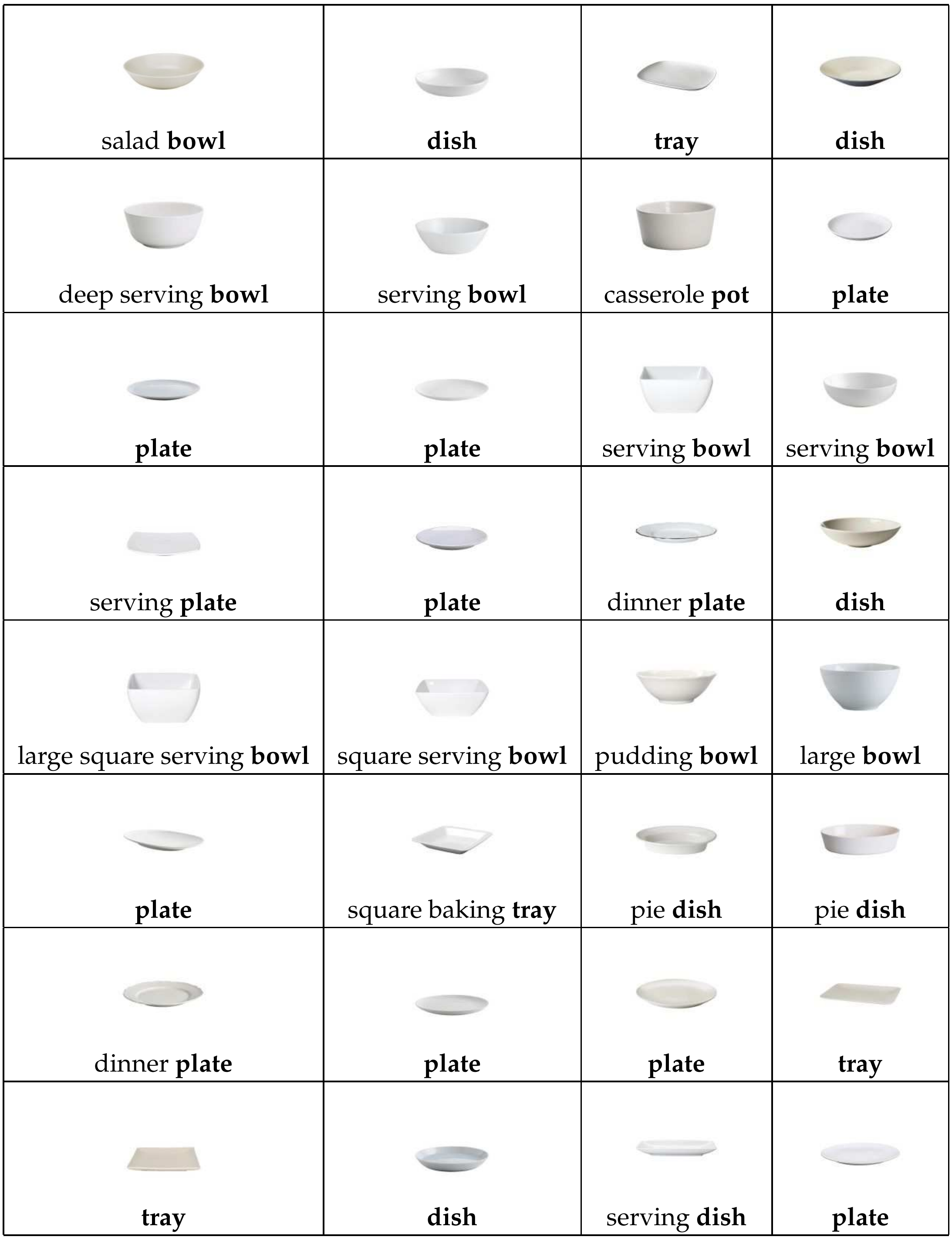


Figure 5.4: An example of the sorting task of an English speaker from Experiment 1. Each row represents one category.

treated as “miska”. Similarly, Japanese words which incorporated adjectival and honorific prefixes were collapsed analogously, so that “sara” (plate), “osara” (o = honorific prefix), and “fukazara” (fukai = ’deep’) were all treated as “sara”. This process dismisses a lot of information, but allows us to examine the lexical categories corresponding to object noun concepts, and to make the naming and sorting data fairly analogous. All of the remaining analysis involving object names operates on the resultant head noun data.

Now we can identify the “dominant names” for each language with respect to the stimuli of the experiment. A dominant name is defined as the most frequently used name for a given object. Of course, since the objects come from quite a restricted domain, many of



salad **bowl**	**dish**	**tray**	**dish**
deep serving **bowl**	serving **bowl**	casserole **pot**	**plate**
plate	**plate**	serving **bowl**	serving **bowl**
serving **plate**	**plate**	dinner **plate**	**dish**
large square serving **bowl**	square serving **bowl**	pudding **bowl**	large **bowl**
plate	square baking **tray**	pie **dish**	pie **dish**
dinner **plate**	**plate**	**plate**	**tray**
tray	**dish**	serving **dish**	**plate**

Figure 5.5: An example of the naming task of an English speaker from Experiment 1 (first half). The head nouns used in analysis are shown in bold.

the 60 objects will have the same dominant names. Table 5.2 shows the dominant names for English, Polish and Japanese, and their composition in terms of the dominant names for the other two languages.

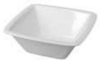
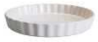
 plate	 **plate**	 **plate**	 **dish**
 square **plate**	 soup **bowl**	 **dish**	 pasta **dish**
 large **bowl**	 round serving **bowl**	 **plate**	 pie **dish**
 square **plate**	 dinner **plate**	 cake **tin**	 salad **bowl**
 serving **dish**	 **plate**	 **plate**	 **plate**
 tray	 **bowl**	 soup **bowl**	 dinner **plate**
 plate	 avocado **bowl**	 washing-up **bowl**	 quiche **dish**

Figure 5.6: An example of the naming task of an English speaker from Experiment 1 (second half). The head nouns used in analysis are shown in bold.

English, Japanese and Polish each had four dominant names, although, as we will see, they varied substantially in their scope. The English and Polish dominant names exhibited some correspondences. English speakers called over 80% of the items either “plate” or “bowl”, which corresponded roughly to Polish “talerz” and “miska”, respectively. However, “talerz” had a wider scope than “plate”, encompassing a few items that were called “bowl”, while the reverse was not the case (i.e., the dominant names of the same item being “plate” and “miska”). Both English and Polish also had two other dominant

English name	N	Polish composition	Japanese composition
plate	26	talerz(24), taca(2)	sara (26)
bowl	24	miska(21), talerz(3)	sara (10), hachi (8), wan (5), booru (1)
dish	8	miska(4), talerz(3), forma(1)	sara (8)
tray	2	talerz(1), taca(1)	sara (2)

Polish name	N	English composition	Japanese composition
talerz	31	plate(24), bowl(3), dish(3), tray(1)	sara (31)
miska	25	bowl(21), dish(4)	sara (11), hachi (8), wan (5), booru (1)
taca	3	plate(2), tray(1)	sara (3)
forma	1	dish(1)	sara (1)

Japanese name	N	English composition	Polish composition
sara	46	plate(26), bowl(10), dish(8), tray(2)	talerz (31), miska (11), taca (3), forma (1)
hachi	8	bowl(8)	miska (8)
wan	5	bowl(5)	miska (5)
booru	1	bowl(1)	miska (1)

Table 5.2: The dominant names for each language, and their composition in terms of the dominant names of the other two languages. The numbers (under the N column for the first language, and in brackets for each of the names for the other two) indicate the quantity of stimuli for which the names were dominant.

names each that covered a small range of items, but they did not correspond to each other in an obvious way.

The Japanese dominant names, on the other hand, partitioned the space of objects quite differently. The relatively generic term “sara” was used for over 75% of the items. Strikingly, these included items which corresponded to each of the dominant names for English and Polish. On the other hand, “sara” does not appear to be a generic superordinate term that covers the whole space. Indeed, each of the items with one of three remaining Japanese dominant names were called “bowl” in English and “miska” in Polish. In other words, Japanese seems to have one term which covers all of the non-bowls and some (in fact, almost half) of the bowls, but makes several fine distinctions among the remaining bowls.

Malt et al. (1999) do not show the compositions for all of the languages in terms of the dominant names for the others, so it is difficult to compare the results closely. However, from what they do show, it seems that the naming patterns in their experiment are harder to pinpoint. For instance, although Chinese also had one highly dominant term for containers, much like Japanese has for dishes in my experiment (i.e., “sara”), it is not the case that the remaining Chinese terms all map onto the same English category, or that they map cleanly onto separate ones. Similarly, in their results, there is no close correspondence between terms in two languages, like English "plate" and Polish "talerz". Much the same is true of Ameel et al.'s (2005) replication with dishes in French and Dutch: there too, there are no close correspondences between the scope of terms in the languages (with the exception of "pot").

As Malt et al. (1999) point out, the above form of analysis only considers the most frequently used names for each item, which ignores the fact that there was substantial intra-lingual variability in naming. Indeed, only 3 (two "bowl" and one "plate"), 3 (all "talerz") and 11 (all "sara") items were given the same name by all native English, Polish and Japanese speakers, respectively. The subsequent forms of analysis thus take the full head noun distributions into account.

The next analysis is based on pairs of objects. First, for each language, I made a list of all the head nouns that were used. Then, for each object, I counted the number of times that it was named with each of the head nouns, generating a name distribution for each object. Next, a correlation was calculated between the naming distributions of every pair of stimuli, giving a measure of naming similarity for each of the 1770 ($\frac{60\times59}{2}$) pairs. These values provide measures of how similar the two objects were in terms of naming for the language group as a whole. Finally, the resulting correlation vectors for one language were (second-order) correlated with each of the other languages to derive a measure of naming similarity between the languages. The resulting correlations are shown in Table 5.3.

	Naming	
	English	Japanese
Polish	.72	.57
English		.38

Table 5.3: Naming pattern correlations between languages, based on their naming distributions over object pairs.

As Malt et al. (1999) point out, the first-order correlations for each language are not independent of each other, since the same stimulus enters into 59 correlations. Consequently, it is not possible to do a simple significance test on the correlations. The focus is thus on how these correlation values compare to each other and to those for the sorting data to follow. Here, the correlations seem substantial, and similar to those in Malt et al.'s experiment. However, the relatively high correlation of English-Polish was not matched for any of the language pairs in the original experiment. On the other hand, naming agreement in Ameel et al.'s (2005) study suggested greater cross-linguistic variability in the container domain than in the dishes domain, and in fact the correlation between French and Dutch for dishes in their study was higher than any of the values found here (0.80).

The lack of significance tests and reliance on relative values makes it particularly important that the naming and sorting comparisons be as similar and fair as possible (if we are to compare them to each other). However, it is not possible to derive a directly analogous measure to the one above for sorting. Instead, since sorting does not offer category labels that can be used to bridge different participants' outputs together, a label-independent measure is needed which can be used for both naming and sorting. Fortunately, the derivation of Malt et al.'s (1999) object pair correlations for sorting (to be replicated in the next section) can be applied directly to the naming data as well. I conduct this analysis next.

In particular, for each pair of objects, we count the number of participants within each language who gave the two objects the same name. Note that this has a counterintuitive consequence, since the count can include different name-pair associations. For instance, a given pair of objects might both be called "dish" by some participants, "bowl" by others, and "plate" by still others: all of these count as the two items being put into the same category. Thus, unlike the previous analysis, here we are focusing on participants' category boundaries in their partitionings of the object space, without considering whether participants agree with each other on the names used. Again, this is necessary if we want to compare the naming and sorting correlations fairly. And, as before, the object pair values give a measure of how similarly the two objects were treated in naming space for each language group. Once these counts are obtained for each pair of objects, we can then carry out correlations between the languages, as in the previous analysis. The results are shown in Table 5.4.

Naming		
	English	Japanese
Polish	.80	.68
English		.50

Table 5.4: Naming pattern correlations between languages based on how often object pairs were given the same name.

Notice that each of the naming correlations is higher in this analysis than it was when we looked at naming distributions. In other words, the cross-linguistic differences in naming are smaller according to this measure than according to the one originally used by Malt et al. (1999) and Ameel et al. (2005). This may be important when we compare the naming and sorting results.

The next analysis compares the naming output of pairs of individuals, regardless of native language, and tests whether there are group differences underlying the results that correspond to native languages. To do this, first the degree of naming agreement between each pair of participants was calculated, yielding a symmetric distance matrix. However, in order to calculate the degree of inter-participant categorisation agreement, unlike Malt et al. (1999), I use Meila's (2007) measure based on information theory (as proposed in Section 4.3.3). The calculations resulted in a symmetric distance matrix which was then subjected to principal components analysis using the 'principal' function in R's psych library. Following Malt et al., if there are underlying group differences corresponding to the three languages, we expect at least three different principal components to emerge, and for them to distinguish between the languages. If not, we expect just one factor to emerge and for all languages to load equally on it. The analysis revealed five factors with eigenvalues above 1 (Malt et al.'s criteria for how many factors to keep). Nevertheless, while the first eigenvalue (48.24) was much higher than the second (2.38), there was a much smaller drop down to the third eigenvalue (1.39) and beyond. This suggests that actually only one factor should be extracted. Therefore, an analysis was first conducted to extract one factor. This factor accounted for 69% of the variance, which was much higher than in Malt et al.'s results. However, this factor did distinguish between the languages, with mean factor loadings significantly higher for Japanese than the other two languages (English: $M = 0.81$, $SD = 0.05$; Japanese: $M = 0.84$, $SD = 0.03$; Polish: $M = 0.80$, $SD = 0.07$; $F(2, 69) = 4.30$, $p < .05$). As a result, a second principal components analysis was run, this time extracting three factors, to see if the different factors would correspond to the different languages. The analysis suggests that they do, with the first factor loading most on Japanese (English: $M = 0.40$, $SD = 0.07$; Japanese: $M = 0.70$, $SD = 0.08$; Polish: $M = 0.46$, $SD = 0.12$; $F(2, 69) = 67.63$, $p < .001$), the second most on English (English: $M = 0.63$, $SD = 0.10$; Japanese: $M = 0.38$, $SD = 0.05$; Polish: $M = 0.42$, $SD = 0.09$; $F(2, 69) = 64.40$, $p < .001$), and the third most on Polish (English: $M = 0.37$, $SD = 0.09$; Japanese: $M = 0.35$, $SD = 0.06$; Polish: $M = 0.52$,

$SD = 0.09$; $F(2, 69) = 30.00$, $p < .001$). However, the three factors accounted for only slightly more variance (i.e., 72%) than the analysis with a single factor. These analyses, while being more equivocal than Malt et al.'s, are largely in agreement with them.

So far, the analyses presented have mainly considered group-level differences. However, my thesis is focused on conceptual coordination between individuals. Therefore, I also analyse the data by looking at the degrees of agreement between pairs of participants, in order to see whether conceptual coordination is relatively high in certain conditions. Since such analyses were not conducted by Malt et al. (1999), and I will present several of them both here and in subsequent subsections, I first explain the general method abstractly here, which will then be demonstrated through application.

In order to compare the degree of coordination between pairs of individuals in different conditions, I adopt the following technique. The main idea is to randomly assign participants to pairs under two conditions, subject to corresponding constraints. For instance, in order to see whether there is more coordination in naming between pairs of native speakers of the same language compared to native speakers of different languages, we can first randomly assign half of the participants to within-language pairs and the other half to across-language pairs. Then we again use Meila's (2007) measure to compare the categorisation outputs for each pair, and conduct a t-test to see whether there is a difference between the two conditions. However, since there is an enormous number of possible pairing permutations for most of the conditions I will test, we can sample many different pairing configurations and run such a test for each case. Then we can look at how often we get significant results in these tests.

However, there is a difficulty in interpreting this number, since it's not obvious what number of tests should yield significant results by chance. As several preliminary suites of tests revealed, the ratio of significant results was often well below 5%, which at first glance is surprising if we use a 5% threshold for significance. The reason for this seems to be that the inter-participant scores are largely dependent on which participants are involved (as shown by preliminary examinations of the scores), so that some participants' average coordination with others tended to be relatively high or relatively low. In other words, participants varied in terms of how idiosyncratic their categorisation tended to be. However, since the pairing configurations always include exactly one score involving each participant, then whatever samplings of scores are taken (under these constraints) can still show substantial variation, which means high standard deviations and thereby lower chances of significant differences.

As a result, the following strategy was used. For each test of this type, 10,000 random permutations were generated, and for each one, two one-tailed t-tests were conducted, one for each tail. If there is no difference between conditions, we would expect about the same number of significant results in both directions. Otherwise, we expect the number of significant one-tailed tests in one direction to be significantly higher than the number in the other direction. Chi-squared tests can be used on these totals (comparing the number of significant and non-significant results for both tails) to assess whether there is a difference between the conditions.

Here, I present the first two tests using this technique. First, I compared the degree of coordination in linguistic categorisation between languages versus within languages. In each of the 10,000 tests, the 72 experiment participants were randomly assigned to pairs in such a way that 18 pairs consisted of two native speakers of the same language (with 6 pairs for each language), and the remaining 18 pairs were two native speakers of different languages (with 6 pairs for each language combination). After calculating the degree of agreement for each pair, each test was then subjected to the two one-tailed t-tests. Out of 10,000 different pairing configurations, 5280 suggested that there was more naming agreement within languages, and only 1 supported the opposite claim. The chi-squared test confirmed that the difference between the two conditions is highly significant ($\chi^2(1) = 7167.61$, $p < .001$). Thus, speakers of the same languages named things more similarly than speakers of different languages.

A second test examined whether the degree of naming agreement within languages depended on the order in which the tasks were carried out. In particular, was naming agreement higher (or lower, or equal) if naming was done after sorting? Participants were randomly paired with other speakers of the same language who did the tasks in the same order. In this case, 153 tests suggested that naming agreement was higher if naming followed sorting, while none suggested the opposite, which was highly significant ($\chi^2(1) = 152.17$, $p < .001$). Thus naming agreement was increased by doing sorting first.

5.3.3.3 *Comparisons of sorting patterns*

The sorting data does not offer the possibility of any analysis that relies on category identifiers (such as labels). In particular, we cannot ask whether two participants put a particular stimulus into the "same" category as each other. However, many of the analyses conducted with the naming data in the previous section can be directly applied to the sorting data as well.

First, we can still consider how object pairs are treated by the different language groups and then compare the groups to each other. While there is no direct analogue for looking at naming distributions, we can look at how often items are grouped together. In particular, for every pair of objects, we can count the number of speakers of a given language that put the pair into the same category. This method is directly analogous to the modified method I proposed for dealing with object pairs in the naming analysis, and gives a measure of how similar the objects were for a language group as a whole. Again, this procedure results in a vector of 1770 values for each language, which can then be correlated with the corresponding vector for another language. The resulting correlations are shown in Table 5.5.

Sorting		
	English	Japanese
Polish	.96	.93
English		.92

Table 5.5: Sorting pattern correlations between languages based on how often object pairs were put into the same category.

Although again we cannot conduct significance tests since the vector values are not independent, the correlations are noticeably higher than the ones for naming. This is the case regardless of whether we are comparing to Malt et al.'s (1999) method based on naming distributions or the alternative I proposed based on how often pairs of objects are given the same names. Either way, the language groups seem to differ more in naming than in sorting. These results are consistent with Malt et al.'s and Ameel et al.'s (2005) results, although the correlations in my study are slightly higher: in both of those studies, the sorting correlations were consistently around .9.

Next, as for naming, principal components analysis was used to test whether there were group differences underlying patterns in inter-participant sorting agreement. If so, we again expect at least three different principal components to emerge, and for the languages to distinguish between them. Otherwise, we expect just one emergent factor which loads equally on the different languages. The analysis revealed that there were four components with eigenvalues greater than 1, but only the drop between the first (49.41) and second (2.13) was large, levelling off after the second (e.g., the third eigenvalue was 1.44). However, unlike for naming, this factor, which accounted for 69% of the variance, did not distinguish between the languages (English: $M = 0.83$, $SD = 0.04$; Japanese: $M = 0.82$, $SD = 0.06$; Polish: $M = 0.83$, $SD = 0.04$; $F(2, 69) = 0.17$, $p > .05$). Nevertheless, for comparison with the naming results, another analysis was run, extracting three principal components. In contrast to naming, and in agreement with Malt et al.'s (1999) results, none of the three principal components (which together accounted for 74% of the variance) distinguished between the languages (component 1: English: $M = 0.53$, $SD = 0.13$; Japanese: $M = 0.53$, $SD = 0.13$; Polish: $M = 0.53$, $SD = 0.16$; $F(2, 69) = 0.01$, $p > .05$; component 2: English: $M = 0.51$, $SD = 0.13$; Japanese: $M = 0.44$, $SD = 0.13$; Polish: $M = 0.48$, $SD = 0.14$; $F(2, 69) = 1.52$, $p > .05$; component 3: English: $M = 0.38$, $SD = 0.09$; Japanese: $M = 0.46$, $SD = 0.15$; Polish: $M = 0.42$, $SD = 0.11$; $F(2, 69) = 2.25$, $p > .05$). This suggests that speakers of the different languages sorted things in the same way, in line with Malt et al.'s results.

Finally, using the randomised pairing method introduced in the naming section, I conducted two analogous sets of tests involving randomised pairings of participants. First, I tested whether sorting agreement within languages was higher than sorting agreement across languages. 844 out of 10,000 tests suggested that it was, while 275 suggested the opposite. While the difference is not nearly as striking as for naming (consistent with the previous analyses and Malt et al. 1999), the chi-squared test confirmed that speakers of the same language did sort more similarly to each other than speakers of different languages ($\chi^2(1) = 305.40$, $p < .001$).

Second, I tested whether sorting agreement was higher within languages if the sorting task was done before the naming task rather than vice versa. 13 tests suggested that it was, while 28 suggested that sorting agreement was higher if naming was done first.

This result is also significant ($\chi^2(1) = 4.79$, $p < .05$). However, given the very low numbers for significant tests in both directions (much less than 1%), this result is suspicious and may be an artifact of the particular statistical technique I am using.

5.3.3.4 *Correspondences between naming and sorting patterns*

So far we have looked at sorting patterns and naming patterns independently. However, if words play a causal role in non-linguistic categorisation, then not only should sorting differ for the different languages, but also sorting should reflect naming.

First, multi-dimensional scaling (MDS) was used in order to give a basic visualisation of the degree of correspondence between naming and sorting for each language. MDS is a technique which takes a set of distances between every pair of points and lays them out in a lower-dimensional space in such a way as to preserve the distances as much as possible. In our case, the points are objects, and the distances come from the inter-object similarity measure defined in the preceding section (for sorting). By labelling the objects in the diagram with their dominant names, we can get a general impression of how well clusters in a language group's "similarity space" correspond to its words. Figures 5.7, 5.8 and 5.9 show the resultant two-dimensional scaling solutions for the three languages.

The solution for English suggests that there is quite some correspondence between naming and sorting, although it is far from perfect. For example, most of the "plates" are found tightly in a cluster of their own (top right), with no other items nearby. However, there are several "plates" further off that are together with a couple of "dishes" and near a "tray" (bottom). The "bowls" and most of the "dishes" (top left) are found together and form a more diffuse group, but still fairly distinct from the "plates".

The Polish solution is broadly similar to the English. There is again a tight cluster of "talerz" (top right), and a few "talerz" among "taca" (bottom). Here too the "miska" are together, and form a somewhat more diffuse group (top left); however, there are also a few "talerz" and one "forma" nearby.

As was the case for the dominant name data, the Japanese solution is clearly different than the other two. "Sara", by far the most prevalent Japanese dominant name, forms several separate clusters (top right, bottom right, bottom left). Among these, there are

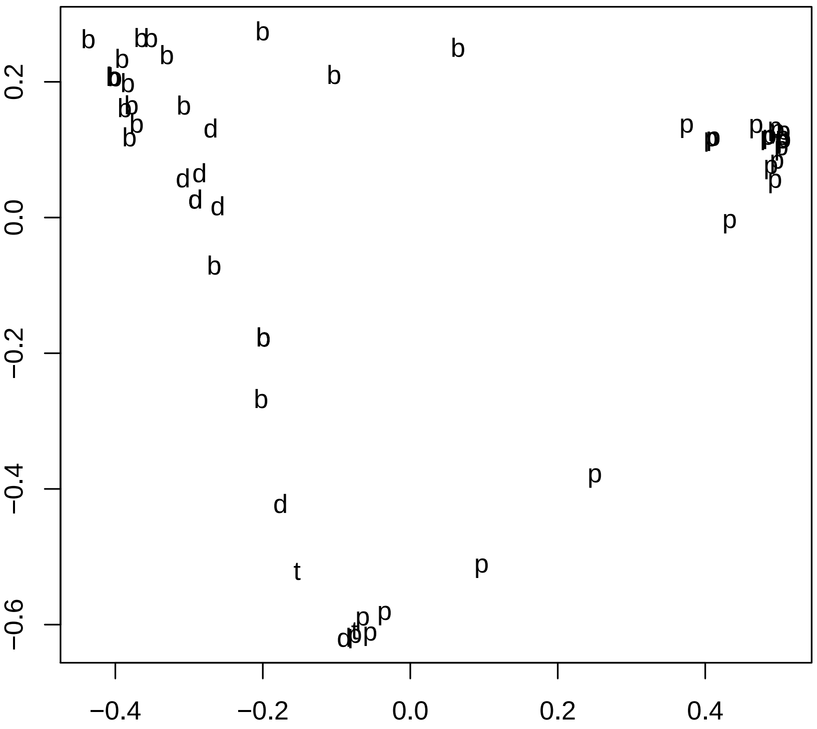


Figure 5.7: MDS solution for English: English sorting patterns are superimposed with English dominant names (b = bowl, d = dish, p = plate, t = tray).

no items whose dominant names are not “sara”, with the possible exception of the one “booru” which is on the edge of one of the clusters. However, the most diffuse group of objects has several “sara” and all the “hachi” and “wan” together. While the “wan” arguably form a subcluster, the rest are scattered around the cluster in no obvious pattern.

Comparing the multidimensional scaling diagrams of my study with those of Malt et al. (1999) is necessarily impressionistic, and I thus do not pursue this much here. I only make two observations. First, the results are in general agreement: in both cases, there was clearly some correspondence between lexical categories and object clusters in similarity space, but these correspondences were not clean and consistent. Second, the correspondences appear a little more chaotic in their data, with more distribution of certain lexical categories around the space and more cases of relatively adjacent items with different dominant names.

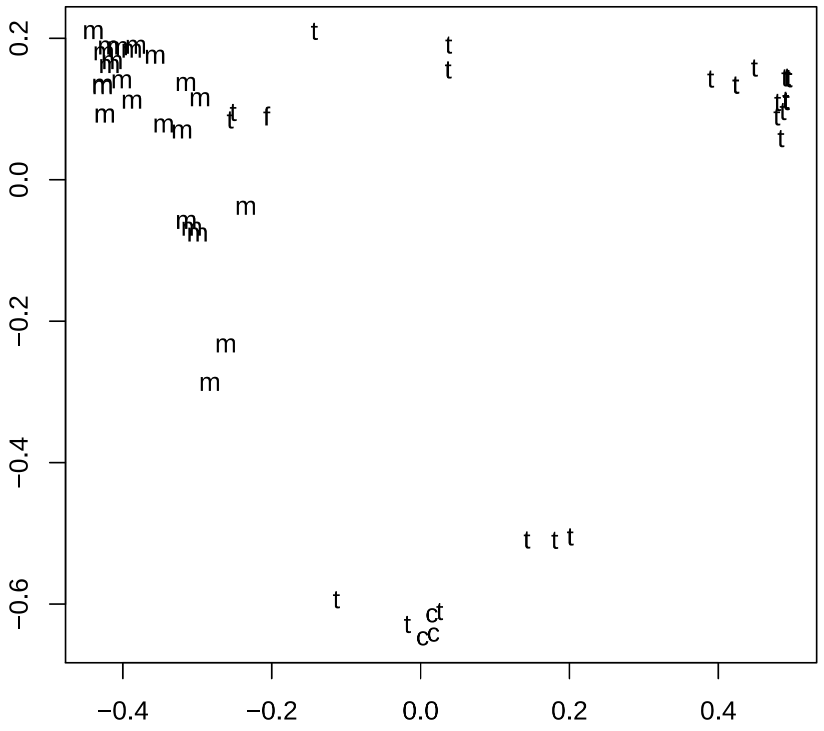


Figure 5.8: MDS solution for Polish: Polish sorting patterns are superimposed with Polish dominant names (m = miska, t = talerz, c = taca, f = forma).

Next, I return to the correlational methods involving object pairs. Previously I looked at how well the different languages agreed in their sorting or in their naming independently. In doing so, I obtained two vectors of inter-object distance data for each language, one for sorting and one for naming. Now we can see how well naming and sorting fit each other, and in particular whether the naming patterns of a given language fit that language's sorting patterns better than the sorting patterns of other languages. To that end, we can take correlations of each of the three naming vectors with each of the three sorting vectors. If naming and sorting correspond to each other within a language, we should expect higher correlations within a language (e.g., Polish naming and Polish sorting) than across languages (e.g., Polish naming and Japanese sorting). Moreover, since I carried out two kinds of analyses of this sort for naming (one according to Malt et al.'s 1999 naming distribution method and one that is more analogous to sorting), we can

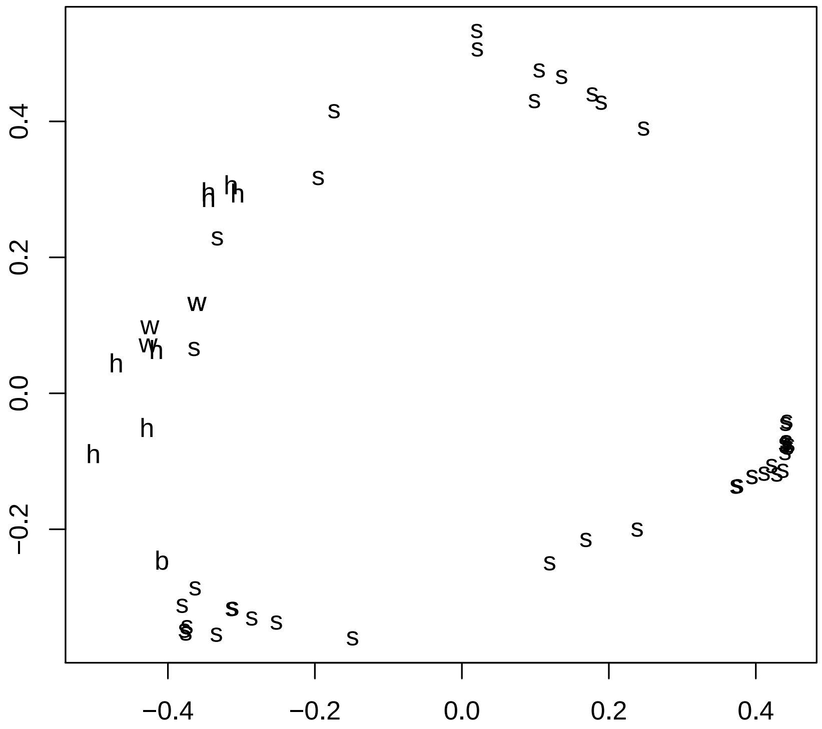


Figure 5.9: MDS solution for Japanese: Japanese sorting patterns are superimposed with Japanese dominant names (s = sara, b = booru, h = hachi, w = wan).

conduct these nine correlations twice, once for each method. The results are shown in Tables 5.6 and 5.7, respectively.

	Naming		
Sorting	English	Japanese	Polish
English	.69	.36	.57
Japanese	.67	.48	.65
Polish	.71	.37	.62

Table 5.6: Correlations of correspondences between naming (using the naming distribution measure) and sorting.

Since the two tables have virtually identical patterns of results, with the second table having consistently higher values across the board, I focus on the second table, as it is derived from the vectors obtained with analogous techniques for naming and sorting.

	Naming		
Sorting	English	Japanese	Polish
English	.80	.45	.71
Japanese	.76	.59	.78
Polish	.82	.47	.75

Table 5.7: Correlations of correspondences between naming (using the measure analogous to that for sorting) and sorting.

Although the non-independence of the data does not allow for simple significance testing, the relative values of the correlations do not fit cleanly with the prediction of a naming-sorting correspondence. The values along the diagonal are generally not higher than off the diagonal, showing that the naming of a language predicts that group's sorting no better than it predicts the sorting of other groups. Although English sorting is best predicted by English naming, Polish sorting also fits English naming better. Japanese sorting is most consistent with Polish naming, although English naming is almost as good, and in fact, according to the first table, fits Japanese sorting better. As such, for the manner of analysis borrowed from Malt et al. (1999), English naming actually predicts sorting best for all three languages.

Indeed, the predictability of naming appears to be independent of which language's sorting patterns are being compared to. As mentioned, English naming seems to generally fit the sorting data best. Polish naming is not far behind, while Japanese naming is by far the least predictive. Indeed, it is particularly striking that Japanese sorting is much more consistent with English or Polish naming than with Japanese naming.

These results are again consistent with the findings of Malt et al. (1999). They too found not only a lack of intralingual correspondence between naming and sorting, but also that the naming patterns of certain languages generally predicted sorting patterns better (or worse) irrespective of which language group did the sorting. In particular, they had found that Spanish naming tended to correlate with sorting the best, English was intermediate, and Chinese was the worst. Indeed, analogously to the finding in my data for Japanese, Chinese naming was the worst of the three at predicting Chinese sorting. This was the case for all three of their sorting data sets (recall that they had participants sort stimuli in three different ways).

Next, I look at whether the discrepancies between two languages in naming correspond to the discrepancies between them in sorting. The starting point is again the inter-object relation data derived in the previous sections for both naming and sorting. In this case, we can subtract two languages' resultant vectors for naming, getting a measure of how much the languages agreed in naming with regards to each pair of objects. The analogous calculation can be carried out for sorting. Then we can correlate the resulting two vectors. The resulting correlations reflect the extent to which sorting and naming discrepancies between pairs of languages match each other. The results (again for both kinds of inter-object naming measures) are shown in Tables 5.8 and 5.9.

Naming-sorting correspondence		
	English	Japanese
Polish	.10	.21
English		.30

Table 5.8: Correlations among groups in naming-sorting differences (using the naming distribution method for naming).

Naming-sorting correspondence		
	English	Japanese
Polish	.06	.34
English		.45

Table 5.9: Correlations among groups in naming-sorting differences (using the analogous method to sorting for naming).

These results are again generally consistent with Malt et al. (1999). The relatively low correlations here suggest that although speakers of different languages sort similarly, and to a lesser extent name similarly, the discrepancies between naming and sorting vary across the languages.

However, note that in Table 5.9, which reflect my alternative inter-object measure for naming, the results are a little different. Although the correlations are still relatively low, notice that they vary depending on whether Japanese is one of the languages involved. The correlation for English and Polish was quite low, but their comparisons with Japanese resulted in much higher values (note that the English-Polish correlation is also the lowest in Table 5.9, although the differences are less striking). This suggests that although Japanese sorting and naming do not fit each other well (as shown in the previous analysis), Japanese speakers may have a somewhat different (perhaps cultural) understanding of certain objects than their English and Polish counterparts, resulting in divergent patterns for both naming and sorting.

To finish analysing the correspondences between naming and sorting, I conducted more tests using the random partner permutations method introduced in the naming section. In this case, several more kinds of tests are possible than when we only looked at naming or sorting independently. First, I checked whether the degree of agreement between one participant's naming and another's sorting differed for native speakers of the same language rather than native speakers of different languages. Of the 10,000 tests, 407 suggested greater within-language agreement, and 424 suggested greater across-language agreement, which was not significantly different ($\chi^2(1) = 0.32$, $p > .05$). Next, I tested whether the amount of correspondence between naming and sorting within languages was greater if sorting was done before naming: 159 supported this claim while only 56 favoured the opposite. This difference was highly significant ($\chi^2(1) = 48.92$, $p < .001$). Third, I tested whether native speakers of the same language agreed with each other more in naming than in sorting. Thus, here the randomised pairing configurations consisted of 36 pairs of speakers of the same language, and each pair's degree of agreement was measured for both naming and sorting. 165 paired t-tests suggested greater agreement in naming than in sorting, while none suggested the opposite, and the difference was significant ($\chi^2(1) = 164.36$, $p < .001$). In contrast, conducting the same kind of test but with participants matched with native speakers of a different (rather than same) language revealed the opposite pattern, and even more strongly, with 5958 tests showing more agreement in sorting than in naming and none showing the opposite pattern ($\chi^2(1) = 8483.12$, $p < .001$). Next, I tested whether there was greater agreement between naming and sorting of the same individuals than there was between pairs of native speakers of the same language. In this case, each random configuration consisted of 24 individual participants (8 from each language) and 24 participant pairs (also 8 for each language). Of the 10,000 tests, 2116 favoured the claim that there was more intra-individual agreement, while only 62 supported the opposite position, which again constitutes a significant difference ($\chi^2(1) = 2171.67$, $p < .001$). Finally, I tested whether the agreement between sorting and naming within individuals depended on the order of the tasks. In this case, there is only one "pairing" configuration, with no randomisation possible: 36 participants did naming first and 36 did sorting first. A simple t-test revealed no significant difference ($t(70) = 0.03$, $p > .05$).

5.3.3.5 Questionnaire data

The responses to the questionnaire (see Appendix A.2) confirmed that, as expected, participants had a variety of linguistic and cultural backgrounds. Many participants spoke a few languages at different levels of proficiency, and/or had lived in a variety of countries. The Polish and Japanese native speakers in particular nearly all had an advanced level of English. This was thus not a controlled variable in this experiment.

5.3.4 Discussion

I first summarise my results while discussing how they replicate and extend Malt et al.'s (1999) findings. Then I discuss the implications for conceptual and lexical alignment, and linguistic relativity. Finally, I discuss some outstanding methodological issues.

5.3.4.1 Replication and extension of Malt et al. (1999)

Experiment 1 used the same general procedure as Malt et al.'s (1999) study, but modified a few other aspects of the design, and made a few changes in analysis. Participants carried out sorting tasks, in which they freely grouped a set of object stimuli into categories. They also did naming tasks, in which they freely named individual items. However, unlike the original experiment, my study tested a different set of languages, recruited participants who all lived in the same city and who were often bilingual or even multilingual, sampled the stimulus domain in a more restricted way, and modified a few of the original analytic methods. Despite these changes, my results were broadly similar to the original ones.

Most of my analyses followed the same general statistical techniques as Malt et al. (1999), including the use of correlations between languages based on the categorisation of object pairs, principal components analyses searching for factors which distinguished the different languages, and multi-dimensional scaling to visualise the extent of naming-sorting correspondences within languages. The results were quite similar to Malt et al.'s. The most commonly used names for the different languages varied in their scope, and Japanese naming was particularly distinct from Polish and English. More generally, although there was plenty of overlap, the languages differed substantially from each other in how they partitioned up the objects into linguistic categories. In contrast, sorting patterns were largely similar across languages. People tended to partition the objects non-linguistically in quite similar ways, regardless of their native language. However, like the original study, there was no evidence for a simple correspondence between naming and sorting, even though they are clearly related to some extent. The naming patterns of a particular language did not reflect the same language's sorting patterns any better than they did the sorting patterns of other languages. In brief, speakers of different languages seem to differ in how they name objects, but perceive things in largely the same way, and differences in perception do not generally correspond to naming differences.

Unlike Malt et al.'s (1999) original study, my experiment also looked more directly at the factors involved in determining the degree of categorisation agreement between pairs of individuals. This was done by assigning all of the participants in the experiment to random partners, subject to the conditions in question. These analyses fit with the previous ones and Malt et al.'s results, while also extending them. First, they confirmed that there were substantial cross-linguistic differences in naming objects. Also, in contrast to previous analyses, such cross-linguistic differences (albeit much smaller) were also found in sorting. This is a particularly relevant finding for my thesis, and I return to it later. Next, the order of the tasks was found to make a difference in the degree of categorisation agreement between people, especially in naming, and least clearly in sorting; however, the degree of naming-sorting correspondence for individuals did not depend on the order of the tasks. Moreover, people agreed more in naming than in sorting when compared with native speakers of the same language. In contrast, when compared with native speakers of other languages, agreement was higher in sorting than in naming. Finally, people's naming was found to be more consistent with their own sorting than with that of other speakers of the same language.

5.3.4.2 *Prior conceptual (and lexical) alignment*

In terms of my thesis, the main issue addressed by Experiment 1 is whether having the same native language makes people conceptualise things more similarly to each other. This issue can be broken down into two parts: (1) do people conceptualise things more similarly to other speakers of the same language, and (2), if so, is this due to language? I address the first part here, and the second in the next section.

One of the clearest findings of the experiment is that people's linguistic categories are in greater agreement with speakers of the same language than with speakers of other languages. This is not surprising, and is consistent with the results showing cross-linguistic differences in naming found in both this experiment and in Malt et al. (1999). Different languages divide up the object space in different ways, which is simply reflected in people's linguistic categorisation.

However, as I have argued in Chapter 2, the relationship between concepts and words is not trivial, so that there may be cross-linguistic differences in linguistic conceptualisation without corresponding differences in non-linguistic conceptualisation. If we take the categories people produce in the sorting tasks as snapshots of their concepts (see Chapter 4), then the main question concerns whether people sorted things more similarly to speakers of the same language than speakers of different languages. While the results of Malt et al. (1999) and my replicated analyses suggested that they probably do not, an analysis of the sorting data designed to test this question directly found that they did. Speakers of the same language agreed more in their non-linguistic categorisation than speakers of different languages.

It is worth emphasising that the degree of conceptual agreement was sandwiched between intralingual and interlingual lexical agreement. Although speakers of the same language agreed more in how they named things than in how they sorted, the opposite was true when comparing speakers of different languages. Indeed, chaining a few of the results together suggests a hierarchy of levels of agreement: intralingual naming, intralingual sorting, interlingual sorting, interlingual naming.

This hierarchy and the relative strengths of the effects suggest that languages polarise categorisation. Speakers of the same language may apply words to things in very similar ways, while exhibiting less underlying conceptual uniformity. At the same time, speakers of different languages may conceptualise things more similarly than their conflicting naming patterns may at first suggest. This is consistent with the idea that conceptual structure is largely universal, while linguistic structures differ in how exactly they map onto it (Chomsky 1986).

These results have implications for theories of online alignment during dialogue (Pickering and Garrod 2004; see Section 3.4.3). Pickering and Garrod argue that interlocutors align different levels of mental representations when engaged in dialogue. While my experiment does not investigate alignment processes in interaction between individuals, it does say something about how far they need to go to become fully aligned. In particular, my results can be interpreted as indicating relative levels of "prior alignment". Speakers of the same language seem to agree to a large extent on what words should be used for what objects, but exhibit a little more variation in how they conceptualise those objects. If they are to align during dialogue, then they have a larger road to travel at the conceptual level than at the lexical level. However, the conceptual distance will generally be smaller if both interlocutors are native speakers of the same language. If they are speakers of different languages, then their prior alignment in conceptualisation could be higher than their prior alignment in their lexical patterns. This reflects some of Pickering and Garrod's commentators' arguments that interlocutors may align relatively little at the situational level (Schober 2004; Branigan 2004). It also relates to the extent to which interlocutors already share common ground upon engaging in joint action (Clark 1996).

Importantly, the different degrees of agreement between people in linguistic and non-linguistic categorisation support the view that lexical and conceptual alignment cannot be equated a priori (Schober 2005). Although words certainly provide some indication of how people conceptualise things (Brennan and Clark 1996), they do not determine concepts (which is consistent with the view that I developed in Chapter 2). Studying co-ordination in conceptualisation must be done at least partly independently of language.

5.3.4.3 Is language responsible?

The experimental results showed that even non-linguistically, people conceptualised things more similarly with speakers of the same language than with speakers of different languages. But is this coordination due to language? Since participants cannot be "assigned" to native languages during the experiment, the study is only correlational. Therefore, differences found between the language groups could be due to language, but could equally well be due to other things that they have in common, including other cultural factors.

Indeed, the results do not give much reason to believe that the coordination effects are due to language. Although several different analyses showed that naming and sorting patterns were certainly related, they failed to show that the correspondences were very close or language-specific. Perhaps most strikingly, and as in Malt et al. (1999), the naming patterns of a given language were not correlated any more closely with the sorting patterns of that language than with the sorting patterns of other languages. Indeed, English seemed to be generally better at predicting the sorting of the other languages, while Japanese was the worst. This runs counter to the possibility that language could be causally related to sorting. Even the randomised pairing tests, which seemed to be relatively sensitive and gave significant results in many cases, did not show a significant difference here: while people named things much more similarly and sorted things a little more similarly with speakers of the same language, this was not at all the case for naming-sorting correspondences.

What about the result that there was more agreement between the naming and sorting within individuals than there was between native speakers of the same language? At first glance, this might suggest that naming is in fact responsible, but also that we have to look more closely, considering idiosyncratic differences between individuals. However, this result can also be explained differently, and consistently with the rest of the above findings: perhaps there are other characteristics that speakers of the same language have in common beyond the language itself which may be responsible for similarity in sorting. While it is beyond the scope of my work to determine what such factors may be, it seems likely that other cultural factors may be responsible.

Indeed, if we just revisit the main differences between the dominant names of the languages, we can see that these could be associated with cultural differences. The main difference between Polish and English was that although they both had two dominant names that covered roughly the same set of items (“plate”/“talerz”, “bowl”/“miska”), their boundaries were a little different, with “talerz” subsuming some “bowls”. This could be related with the fact that it is common in Poland to serve soup in relatively flat dishes, and to refer to them as “talerz”. Similarly, the greater degree of divergence of Japanese naming from English and Polish is consistent with the (subjective) observation that Japanese cuisine and tableware vary more radically from their western counterparts than the latter do among themselves. This explanation is admittedly speculative and vague, but it does fit the data better than a view involving a strong naming-sorting correspondence. Moreover, it is in line with experimentally motivated proposals of cross-cultural differences in categorisation reflecting cultural variation in domain expertise (e.g., Tanaka and Taylor 1991).

But if such non-linguistic cultural factors are to explain the differences between the languages, then why are differences in sorting smaller than those in naming? This may be because linguistic categories, via words, are more public and thus more susceptible to cultural influence, while non-linguistic categories are more necessarily private and sheltered from culture (Hurford 2007; Voiklis 2008). Indeed, the setup of the experimental tasks may have made this contrast particularly stark. In the naming task, participants were asked what they would normally call the items, which may conjure up situations in which they are interacting with other people with the same cultural background. In the sorting task, they were asked to sort things by physical similarity, which means a relatively low-level perceptual basis that has little or nothing to do with cultural practice. While cultural factors might still influence the latter, they are likely to be more directly implicated and hence stronger in the former.

This conclusion bears directly on the issue of linguistic relativity. Malt et al. (1999) found no cross-linguistic differences in non-linguistic categorisation, from which they argued against a (strong) Whorfian position. I argued, however, that the verdict was still open with respect to weaker Whorfian positions, whereby language merely influences rather than determines thought. And indeed, the new forms of analysis that I used in analysing the data of my experiment did discover some cross-linguistic differences in sorting, albeit small. Nevertheless, it appears that these differences might not be due to language after all, but to other cultural factors. As such, they do not provide support to even a weaker Whorfian position, since the causal role of language is crucial therein.

5.3.4.4 *Methodological considerations*

There is of course, as always, an important alternative explanation for the lack of correspondence between naming and sorting in the experiment: it may be due to methodological limitations. While this is always a possibility, there is a specific aspect of the results that lends itself to suspicion.

In particular, why would the naming-sorting correspondences in some cases be better across languages than within languages? The Sapir-Whorf hypothesis would predict that the naming patterns of a language should fit the sorting patterns of that language better than the sorting patterns of other languages. In contrast, an anti-Whorfian position would say that this should not be the case, with naming in one language being no more predictive of that language group's sorting than any other. However, noone has proposed that a "reverse" Whorfian pattern should occur, with some group's linguistic categorisation actually fitting worst with its own non-linguistic categorisation. And yet this is precisely what the correlations in Table 5.7 suggest. The most striking example of this is that of Japanese sorting being much better predicted by both English and Polish naming than by Japanese naming. Actually, as mentioned earlier, each of the language's sorting patterns fit best with English naming, and worst with Japanese naming. Why would this be?

One possible explanation is that languages differ in how "naturally" they carve up the conceptual space. For whatever reasons, perhaps even historical accident, certain languages, like English, may have formed lexical categories which are based more heavily on perceptual similarity and deviate less from universal conceptualisation tendencies. In contrast, other languages, such as Japanese, may have diverged more from non-linguistic categorisation, perhaps by more often undergoing the kinds of processes suggested by Malt et al. (2003) that could lead to a dissociation in the two kinds of categories.

However, a simpler and more likely explanation for the findings here would relate to natural tendencies in sorting. When you are asked to classify a set of items into groups, there is probably a natural bias towards putting roughly an equal number of items into different groups. However, when it comes to naming, the data showed much more domination of the object space by a single term in Japanese than for English or Polish. As a result, it is likely that there would be greater discrepancies between Japanese naming and sorting of any of the languages. Indeed, if a different sampling of the object space was taken, which focused on items that were not called "sara" in Japanese, perhaps the opposite pattern would be found, with Japanese names being more predictive than English or Polish (which might be tempted to call most of the items "bowl" or "miska", respectively). This explanation is reflected by the fact that Japanese participants assigned many more items to their largest linguistic category than English or Polish (Japanese: $M = 40.75$, $SD = 6.19$; English: $M = 24.42$, $SD = 3.78$; Polish: $M = 28.5$, $SD = 6.06$; $F(69, 2) = 58.19$, $p < .001$), but did not differ in this respect for non-linguistic categories

(Japanese: $M = 20.54$, $SD = 6.16$; English: $M = 18.12$, $SD = 3.67$; Polish: $M = 18.50$, $SD = 4.40$; $F(69, 2) = 1.72$, $p > .05$).

There are other reasons to be cautious about both the sorting and the naming tasks. I have already discussed the potential concern of truncating the naming data to the head nouns used (see Section 5.2.4). The naming task also may be problematic in that it is done serially, and thus does not give a snapshot of a concept in the same way as the sorting task does. Since precedence and priming have been shown to play an important role in the choice of referring expressions used for objects (Brennan and Clark 1996; Garrod and Anderson 1987), it is likely that participants prime themselves with names, meaning that the order in which they name things may have an effect. Moreover, assigning a name to an object in a decontextualised non-communicative setting may be artificial, and subtle unintended cross-linguistic differences in the task instructions may have had an impact. The sorting task is less problematic, but also raises concerns. Malt et al. (1999) designed the sorting task to probe similarity-based categorisation specifically, and thus this aspect was explicitly incorporated into their (and my) instructions. However, this places constraints on the kinds of categories that people produce, which are possibly at odds with the way they would naturally conceptualise, especially given the dissociation that has been demonstrated between categorisation and similarity (Rips 1989).

It's worth emphasising that these issues with naming and sorting are not necessarily so troublesome in themselves: any experimental task has limits that should be acknowledged. The problem is in expecting the data from the tasks to be analogous, and thus shedding light on the correspondence between linguistic and non-linguistic categorisation. It is possible that thought does reflect language, but that the different demands of the two types of experimental tasks introduce a methodological dissociation. Thus, while the present results seem to be best framed as showing cross-linguistic categorisation differences that are not due to language but to other cultural factors, there is certainly ample room and need for further exploration of this issue.

At this point, I should also mention the effects that were found concerning the order of the tasks. Recall that the results suggested that the degree of categorisation agreement between individuals was a little higher if participants first carried out the sorting task and then the naming task. This was the case for naming itself, for naming-sorting correspondences, and even, more debatably, for sorting. Although these effects were weak, they do draw attention to the fact that conceptualisation is a dynamic process, that may not always lead to the same results. The task context, goal and other factors may all play a role in how people conceptualise objects. And when it comes to coordination between individuals, this may especially be the case when people are engaged in interaction. These considerations will take a more central role when I shift my theoretical and experimental focus from the next chapter onwards.

5.3.5 Conclusion

In this chapter, I asked whether language brings about conceptual coordination on a relatively large timescale. In particular, the question was whether having the same native language results in people conceptualising things more similarly to each other. In this light, I first discussed a particular experiment that had previously been conducted which found a dissociation between linguistic and non-linguistic categorisation in the object domain. I pointed out that the study had different theoretical priorities from mine, and that a replication supplemented with new analytic methods could both test the robustness of their findings and zero in on my research questions. Experiment 1 was designed with this dual purpose in mind. The results confirmed the dissociation documented previously, while also qualifying it. Although the effects were much stronger in the linguistic than in the non-linguistic case, speakers of different languages did categorise things more similarly to each other than to speakers of other languages, for both kinds of categorisation. A close look at the data, however, suggests that these effects are more likely due to other cultural factors rather than language. In sum, native speakers do seem to be a little conceptually pre-aligned, but it does not appear that language is responsible.

CHAPTER 6

Experiment 2: conceptual convergence due to dialogue

6.1 Introduction

Experiment 1 found some evidence that even when categorising non-linguistically, native speakers of the same language do categorise objects a little more similarly than speakers of different languages. However, as I argued, this phenomenon does not seem to be attributable to language rather than other cultural factors. I therefore tentatively concluded that sharing the same native language does not seem to bring together people's conceptualisations in the object domain.

Nevertheless, this is only one side of the story. Malt, Sloman, Gennari, Shi and Wang's (1999) study and my replication implicitly assume, in a sense, a static conceptual system and a static lexicon. The two different tasks that participants carried out were meant to capture how they categorised objects, both linguistically and non-linguistically. For each experimental participant, a single answer was obtained for each of these tasks. However, in Chapters 2 and 3, I reviewed different kinds of evidence showing how both lexicalisation and conceptualisation are flexible, dynamic processes whose outcomes vary both between and within individuals. While it may be reasonable to assume that in the absence of an explicit context people have some kind of "default" conceptualisations, we need to treat this as a starting point and look beyond it to get a more complete understanding of the role of language in conceptual coordination. In particular, as discussed in

Section 3.4, another important possibility for how language may coordinate conceptualisation is online through interaction. However, unlike the question of linguistic relativity, this issue still largely remains to be investigated.

Therefore, I now take up the issue of the role of language in conceptual coordination between individuals on a much shorter timescale. Does interactive language use bring together people's conceptualisations? The next three chapters present three experiments which explore different aspects of this issue, and constitute the bulk of the empirical work of my thesis. All three experiments fully employ the general framework developed in Chapter 4.

Similarly, all three experiments will also take the model in Figure 6.1 as their theoretical starting point. The diagram is the same as in Figure 3.3, except that I have incorporated the unfortunate category-concept compromise acknowledged in Section 4.2.1.

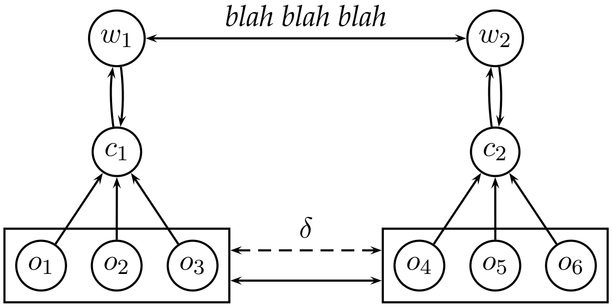


Figure 6.1: The general interactive model that will serve as the theoretical starting point for Experiments 2-4. Recall that δ is the numerical difference between the two people's categorisations, as discussed and defined in Section 4.3.

Recall that previous studies have already suggested that dialogue can bring together people's categorisation (see Section 3.4.3.3). In particular, Markman and Makin (1998) and Voiklis (2008) showed how communicating together during a shared task can result in relatively similar subsequent categorisation with other people who had communicated. However, these studies have lacked a crucial manipulation which prevents us from concluding that it is dialogue itself that is responsible for the coordination. In both studies, the main contrast in conditions involved one in which participants performed a task on their own, and another in which they worked together on the task while communicating linguistically. As such, the difference between the conditions cannot necessarily be attributed to language specifically, rather than interaction in general (or some other specific aspect of interaction). In order to determine whether dialogue is causing the subsequent convergence in categorisation, we need to have an experiment that manipulates whether or not partners can talk to each other *on top of* a condition which already has some form of interaction. In other words, we need a joint task that can be meaningfully performed with or without language, while still exhibiting other features that we want. This is addressed in Experiment 2.

6.2 Overview

The main purpose of Experiment 2 is to determine whether linguistic dialogue in a joint task temporarily brings together people's conceptualisations. As mentioned above, the main challenge is what to use as the joint task which dialogue gets added to. In particular, we need a joint task which can in principle be carried out without dialogue, prior to having participants carry out individual categorisation tasks.

Rather than constructing a special kind of interactive task, the solution I adopt is simply to adapt the free classification task so that participant pairs categorise a set of items together. That is, they are given a set of stimuli, and must decide together on how to put them into categories. Nevertheless, they still have their individual concepts, through which all decisions must be mediated. Figure 6.2 shows how this fits into the model I have developed. As the diagram highlights, the task can be thought of as a kind of intimate joint categorisation task.

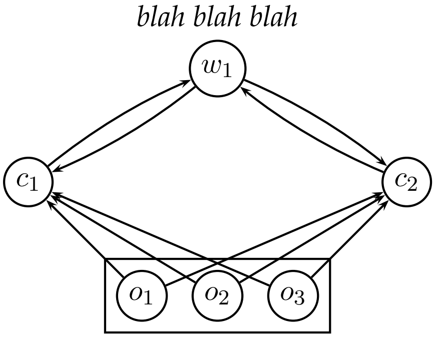


Figure 6.2: An interactive joint task between two people, in which they jointly make categorisation and lexicalisation decisions, although their concepts are still distinct. This was the form of the joint task in Experiment 2.

Indeed, as I argued in Section 4.4, while free classification tasks can probe people's individual conceptualisations, their properties are also suitable for a joint task. First, they can be done without resorting to dialogue, provided that partners can share the same perceptual input. Recall that reliance and manipulation of a shared visual space has already been fruitfully applied in dialogue studies (Clark and Krych 2004; Gergle, Kraut and Fussell 2004; Kraut, Fussell and Siegel 2003; see Section 3.4.2). Second, the task can be set up so that it is one large task in which participants can work together to arrive at a solution, rather than a sequence of individual tasks where the participants take turns but cannot affect each other′s decisions. This makes the task more interactive and cooperative, which, as we saw in Section 3.4.2, seems to be important for supporting linguistic effects. Third, by virtue of being directly analogous to the individual categorisation tasks of the experiment, we may be able to optimise the sensitivity of the experiment and thus the chance of finding differences between conditions.

The experiment has three conditions. In each condition, participants start and end the experiment by carrying out free classification tasks individually. However, the conditions differ in terms of what happens in between. In two of the conditions, they perform a joint task, as described above. In one of them, they carry out a joint free classification task, during which they are allowed to talk freely with each other. A second condition is identical except that no talking is allowed. Finally, in a third condition, no interaction or joint task takes place, and the participants simply carry out another free classification task on their own. Thus the structure of the experiment will follow a standard pre-task, treatment task, and post-task type of design, in which all participants carry out the same pre-tasks and post-tasks, but vary in the treatment.

A secondary but also important design decision is whether participant pairs also label their categories in the joint task. On one hand, if participants label their categories and these labels are visible to their partners, this allows for a limited mode of linguistic communication, which seems to run counter to the goal of isolating language in one condition. However, on the other hand, participants may implicitly label their categories anyway. By requiring them to do so explicitly in the task, we would control across all participants, so that they *all* label the categories, regardless of condition. Therefore, since in this experiment I want to see whether dialogue specifically brings about conceptual coordination, I have opted for the second possibility. But Experiment 4 in Chapter 8 addresses the problem that this leaves behind: whether merely exchanging category labels brings about conceptual coordination.

Using this design, the experiment is mainly meant to address two questions:

1. **Does engaging in a joint categorisation task bring about conceptual convergence after interaction?**
2. **If so, is this effect stronger if people can also communicate through dialogue?**

Thus the aim is to tease apart the role of dialogue from other aspects of interaction in conceptual coordination. While the studies of Markman and Makin (1998) and Voiklis (2008) suggest that there should be a difference between the two extreme conditions, it is unclear how each of them will fare relative to the middle one (i.e., interaction without dialogue).

A secondary but also important purpose of Experiment 2 concerns lexical alignment. As discussed in the previous section, we can separately compare the labels and categories between interacting pairs. Doing so will allow us to see whether the conceptual and lexical convergence results parallel each other, and thus how the two levels and processes might be related. However, it's worth noting the difference between how lexical and conceptual alignment were investigated in Experiment 1 and how they will be now. In Experiment 1, there were two kinds of conceptualisation tasks: sorting without labels, and labelling. Here there is one conceptualisation task, which merges those two tasks together: sorting with labels. While these are complementary investigations, the difference in approach can be theoretically confusing, and I return to it in Section 9.3.1.

Finally, Experiment 2 also serves as a general methodological test. Since this experiment constitutes the first attempt to apply the full interactive experimental framework developed in Chapter 4, there are various general problems that could potentially surface with the experimental stimuli, tasks, and program interface. Therefore, an informal but important overarching objective of this experiment was to evaluate the general framework. This was done mainly by obtaining feedback from participants via a questionnaire following the categorisation tasks.

Unfortunately, as we will see, this experiment fell short of addressing its primary purpose. After running 24 participant pairs, preliminary analysis did not show any signs of the predicted differences between conditions. Combined with unanticipated methodological shortcomings of the experimental design that revealed themselves in the meantime, this led me not to run any more sessions. Nevertheless, the experiment did provide other interesting data concerning lexical choices and alignment (via the category labels), the use of dialogue in coordination, and evaluation of the framework (via feedback from the questionnaire). Therefore, Experiment 2 is still an important stepping stone to Experiments 3 and 4.

6.3 Methods

6.3.1 Participants

Participants were 48 adult native English speakers (age: $M = 21.49$, $SD = 2.34$), mostly undergraduate students recruited through the University of Edinburgh student employment website. There were 31 female participants and 17 male participants. Participants were assigned randomly to pairings, and pairings were assigned randomly to conditions. Partners did not know each other before the experiment.

6.3.2 Conditions

Experimental sessions were conducted with pairs of participants, and there were three between-pair conditions: *control*, *silent*, and *talking*. In the *control* condition, participants independently carried out all three categorisation tasks individually in succession, without ever working or speaking with their partner. In the *silent* condition, participants conducted the pre-task and post-tasks independently, but in between silently carried out a joint categorisation task together. The *talking* condition was the same as the *silent* condition, except that participants were allowed to talk to each other during the joint categorisation task.

6.3.3 Stimuli

The dish domain exhibited inter-participant variation in categorisation in Experiment 1, and was therefore used again in Experiment 2. However, this time, we do not just need variation, but also room for potentially different degrees of alignment. Therefore, in order to maximise the stimulus set's fluidity and thereby potential variability in participants' categorisation, morphing software was used to generate many stimuli interpolated between prototypical dishes. This generated a large domain of stimuli from which different subsets were then sampled to be used in different categorisation tasks. Below I first explain how exactly the stimulus domain was generated, and then describe how the subsets were sampled and ordered.

6.3.3.1 Generation

The first step in generating the stimulus domain was selecting the starting points (which I call base stimuli) for the morphing sequences. For that purpose, ten pictures of dishes were selected from online catalogues like IKEA's. I selected the base stimuli to be likely members of what in English might be called "bowls", "plates", "trays", "cups" or "dishes". All of the items were square black-and-white photographs with a white background. The objects were simple in form (e.g., no handles), had no ornamentation, and were of a white or light grey colour. This was done mainly to facilitate the morphing process, since oddly shaped or coloured items tended to produce strange interpolations. It was also done in order to limit overly obvious features that distinguished the items, in the hope that this would increase variation in categorisation. The ten base stimuli are shown in Figure 6.3.

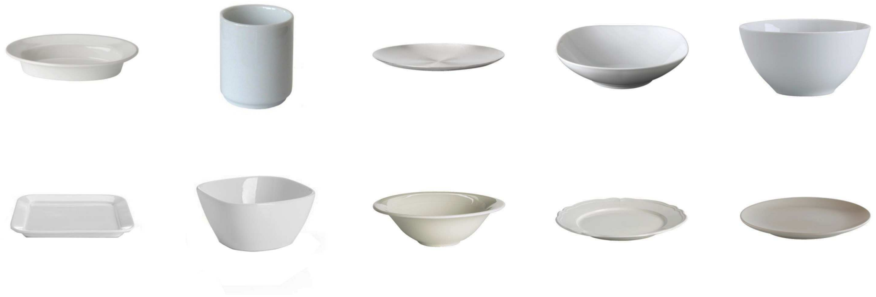


Figure 6.3: The base stimuli used in Experiment 2.

Morphing was done with the Morpheus Photo Morpher program (Version 3.10 Standard). Pairs of images were loaded into the program and many calibration points were aligned manually between corresponding positions of the images. Then the program was run and created a sequence of interpolated images. After that, the new images were visually inspected, and those that looked particularly unnatural were removed from the set. This yields many intermediate and therefore potentially ambiguous items.

In particular for each of the 45 possible pairs of stimuli, a sequence of thirteen intermediate stimuli was generated. An example sequence is shown in Figure 6.4. All of the resulting pictures were then visually examined to assess whether they looked natural, and on this basis, 25 stimuli were eliminated. Thus, at the end of this process there was a collection of 570 stimuli ($10 + 45 \times 13 - 25$).

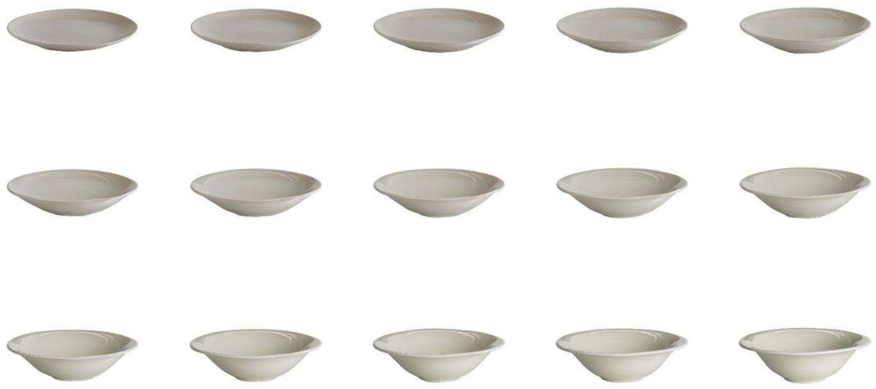


Figure 6.4: An example morphing sequence between two base stimuli used in Experiment 2.

The resulting stimulus domain, while highly continuous within the scope that it did cover, was relatively restricted in terms of how much of overall 'dish space' it covered. This is in contrast to the stimulus selection methods of Malt et al. (1999) (and even my replication), which specifically tried to collect as varied and diverse a set as possible. However, while Malt et al. were partly interested in how different languages actually divided up the dish domain, here the interest lay exclusively in how well people coordinated their concepts, and so emphasis was placed on fluidity rather than representativeness.

6.3.3.2 *Allotment*

From the superset of 570 stimuli, subsets were selected for the categorisation tasks. Ten sets of 120 stimuli were randomly and independently assembled. Selection was random rather than strategically selecting morphing sequences because all the stimuli for each task are visible at once, so such systematic sets would perhaps have been transparent to the participants. Each set was divided randomly into three subsets of 40 stimuli each, one for each of the three categorisation tasks. An example subset for one categorisation task from the experiment is shown in Figure 6.5.

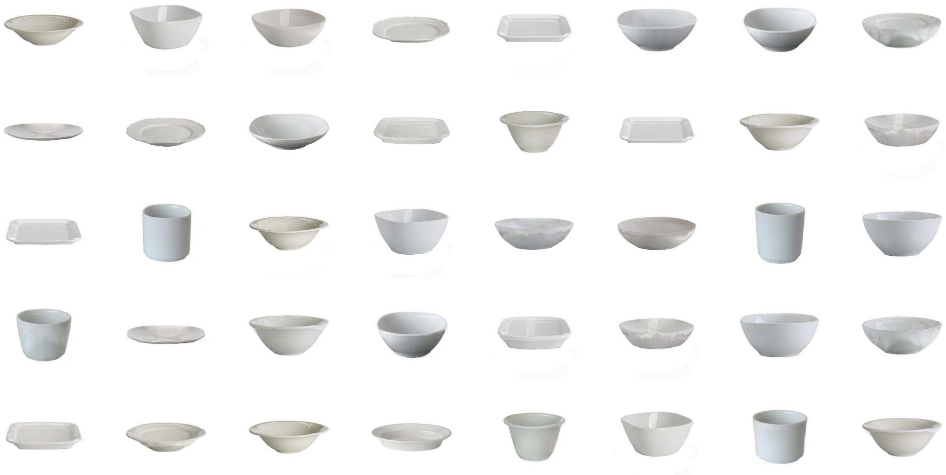


Figure 6.5: One of the stimulus subsets used in categorisation tasks in Experiment 2.

Each of the ten stimulus sets above identified a level in a secondary experimental condition. One participant pair from each main condition was assigned to each such level. Thus, different pairs in the same condition were in different stimulus groups, but they had a corresponding pair in each of the other two conditions which were in the same stimulus group. This was done in order to both prevent idiosyncratic properties of a particular set to affect the experiment too much, while still controlling for all conditions being as similar as possible.

As a result of this random procedure, the extent of usage of stimuli from the global set was not uniform across them: some stimuli were used multiple times, others just once, while others not at all. In addition, the tasks in which a stimulus was used were also random, so that, for example, the same stimulus could appear in the third task for one stimulus group but the first task in another group.

The visual order in which stimuli appeared on the screen for a given task was chosen randomly for every participant for all the pre-tasks and post-tasks (see Section 6.3.4 for descriptions of the task sequences). However, for the treatment task, the order of the stimuli was the same for both participants of a pair (but different from other pairs), regardless of whether or not they were carrying out the task together (in the *silent* and *talking* conditions) or separately (in the *control* condition). This was done to give the two participants in a pair the same experience (to the extent possible) in the treatment task. This could be a problem if we wanted to compare participants' outputs to other participants who weren't their partner. However, that is fortunately not the focus of this experiment.

6.3.4 Procedure

The experiment was conducted through a computer interface and consisted of a small practice categorisation task, three large categorisation tasks, and a short questionnaire. Pairs of participants were brought into the experimental lab and assigned randomly to two separate computer cubicles. After filling out a consent form, they began the experiment.

The practice task was a free classification task, but involved only 10 stimuli, consisting of pictures of pieces of furniture. Participants sorted the items into the groups on their own (much as in the sorting task in Experiment 1). In addition to sorting the items, the practice task also required participants to perform all the possible interface functions: naming a category, adding an item into a category, magnifying a category, changing the item in focus, renaming a category, and changing the category of an item. During this task, the doors to the experimental cubicles were left open, and participants were encouraged to ask the experimenter questions about the interface.

The subsequent three tasks were all relatively large free classification tasks, in which a set of 40 stimuli was to be sorted into 3 to 9 categories. Some of the tasks were carried out individually, while others, depending on the condition, were done together with another participant. In particular, the experiment design included a pre-task, treatment task, and post-task, with the conditions varying only in the treatment task. Both the pre-task and post-tasks consisted of individual categorisation tasks. In these tasks, there was no interaction between partners, no feedback concerning each other's categories or degree of categorisation agreement, and no reference to working with one's partner in the task instructions. Participants simply sorted the stimuli into categories and named the categories.

In the treatment, the task depended on the condition. In the *control* condition, participants carried out an individual categorisation task, just as in the pre- and post-tasks. In the *silent* and *talking* conditions, pairs worked together on a *joint categorisation* task (see the next section for mechanical details). Moreover, in the *talking* condition, pairs were allowed to talk freely during this task, while in the *silent* condition (and in the *control* condition), they were not. The task sequence is schematised in Figure 6.6.

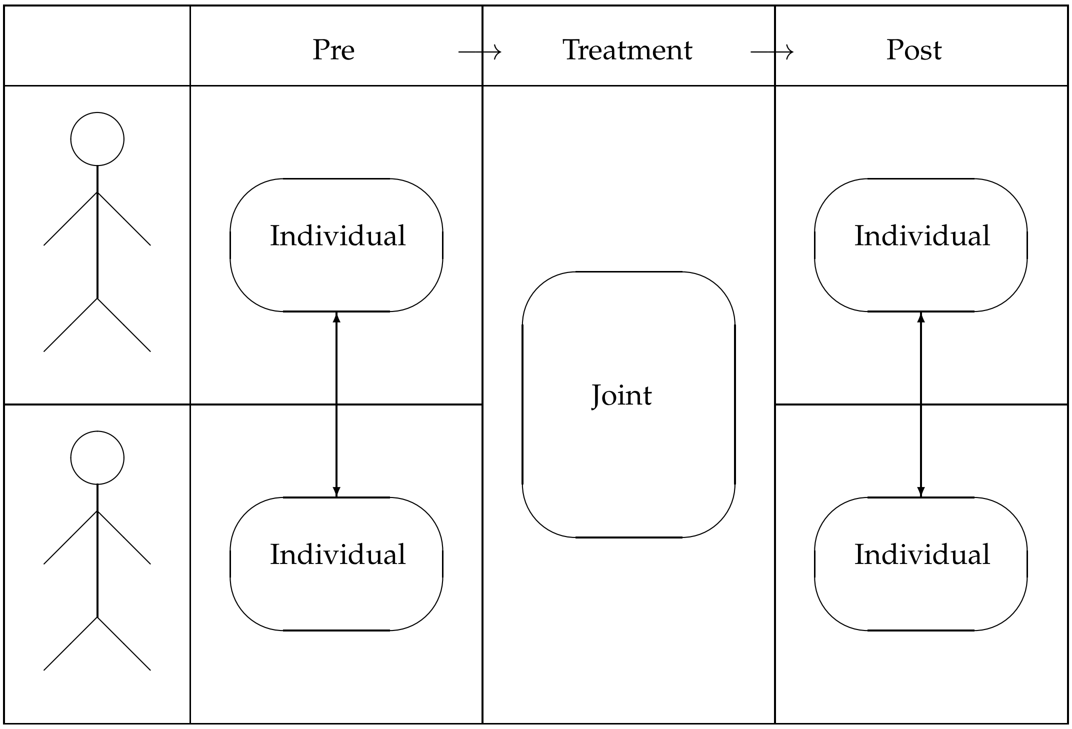


Figure 6.6: Task sequence (for the *silent* and *talking* conditions): Participants first engage in individual categorisation tasks, then carry out a joint task together (during which the *talking* condition participants can talk), and finish with another individual task. In the *control* condition, the treatment task consists of another individual categorisation task (so that control participants simply carry out three individual tasks in a row).

6.3.5 Program

The experiment was run through a computer program, which was specially written in Java, with the data stored in a MySQL database. Sessions were conducted in an experimental lab with dual computer cubicles, such that two participants sat side-by-side in front of individual computers, but were separated visually by a screen (so they could not see each other or the other screen). The computer monitors were all of the same model, and the different monitors used had the same display settings (e.g., size, contrast, brightness).

I first describe the program interface and operation for the individual free classification task. Snapshots of the interface are provided in Appendix B.3. The program window initially contains several colour-coded boxes: the nine blue category boxes on the left, the green focus box in the top right, and the black pool in the bottom right. The stimuli initially all show up in the pool, and the participants need to move the stimuli into the categories. Each of the category boxes also has an associated text field in which category labels are typed, and an "Add" button for putting a stimulus into it. Participants must name a category box before they can put any stimuli in it.

When participants click on an item, it is moved to the focus box, which allows them to inspect a larger version of it. Thus participants always see such a large version before they categorise any item. This was done because under the size constraints of the monitor, items must otherwise be shown quite small so that they all appear on the screen at once. Moreover, each category box (and the pool as well) also has a "Big" button, which allows participants to see all the stimuli in that category in the larger version together (alongside whatever object is currently in the focus box). Thus the focus box and "Big" buttons effectively serve as a virtual magnifying glass.

After all the stimuli have been moved into categories, a 'Done' button appears underneath the pool. Once this button is clicked on, the task is over. However, before pressing the button (and in fact at any stage prior to that as well), the participants can make as many changes as they want to the categories. This includes recategorising stimuli (i.e., putting them into different category boxes) and relabelling the categories. Only the participants' final partitioning of items into categories is submitted to analysis.

The interface and operation of the joint categorisation task is nearly the same as for the individual task, with a few minor changes. First and most obviously, the task is carried out by pairs of participants. In particular, participants see the same program window on their two computer monitors, and take turns in performing operations, while their partner's window is updated. During their turn, participants can also make changes to either their own or their partner's previous contributions, including recategorisation of stimuli and relabelling of categories. A participant's turn includes any magnifying and naming (or renaming) actions they perform, and ends when they perform a categorisation (or recategorisation) action. The changes that a participant makes during his turn are marked with a temporary red border around category labels (for category name changes) or stimuli (for categorisation changes). Whose turn is also indicated by the background colour of the focus box: during one's turn it is green (like it is during an entire individual categorisation task), and when the turn ends, it becomes red (until it is one's turn again). The interface for a participant is passive when it is not his turn, so that any typing or mouse-clicking in the interface window at this time has no effect. Once all the stimuli are categorised and both partners do not want to make any further changes, the task is finished (as indicated by both partners clicking on the "Done" button in succession). Thus, participants negotiate the categories during the task through their categorisation decisions and changes (and possibly through dialogue as well, depending on the experimental condition). The final result is interpreted as the joint categorisation of the two participants. Note, however, that since the output is joint, there is no clear way to obtain an agreement score between the participants for the task itself.

6.4 Results

I first present the results addressing conceptual alignment, as this is the main purpose of the experiment. I then show analogous analyses for lexical alignment, followed by tabulation of the most common category labels used by participants. Next I provide some examples of how dialogue was actually used by pairs in the *talking* condition. Finally, I present questionnaire data relevant to evaluating the experimental design and the general framework.

But first, in order to give a concrete example of the data, Figure 6.7 shows the output from a joint categorisation task. This is the output from a random pair in the *talking* condition, which means that the pair worked together and were able to talk during the task. In this case, the pair chose to use eight (of nine possible) categories.

6.4.1 *Conceptual convergence*

The main empirical issues concerned whether participant pairs categorised things more similarly as a result of having done a joint categorisation task, and whether dialogue made this effect even stronger. First, I calculated the degree of agreement in every partnership's scores for both the pre-tasks and post-tasks (using Meila's 2007 information theory measure). The resulting means and standard deviations for each condition are shown in Table 6.1, and boxplots are shown in Figure 6.8.

I conducted three kinds of statistical tests to evaluate whether pairs were more aligned in the post-task than in the pre-task, and whether the degree to which this was the case

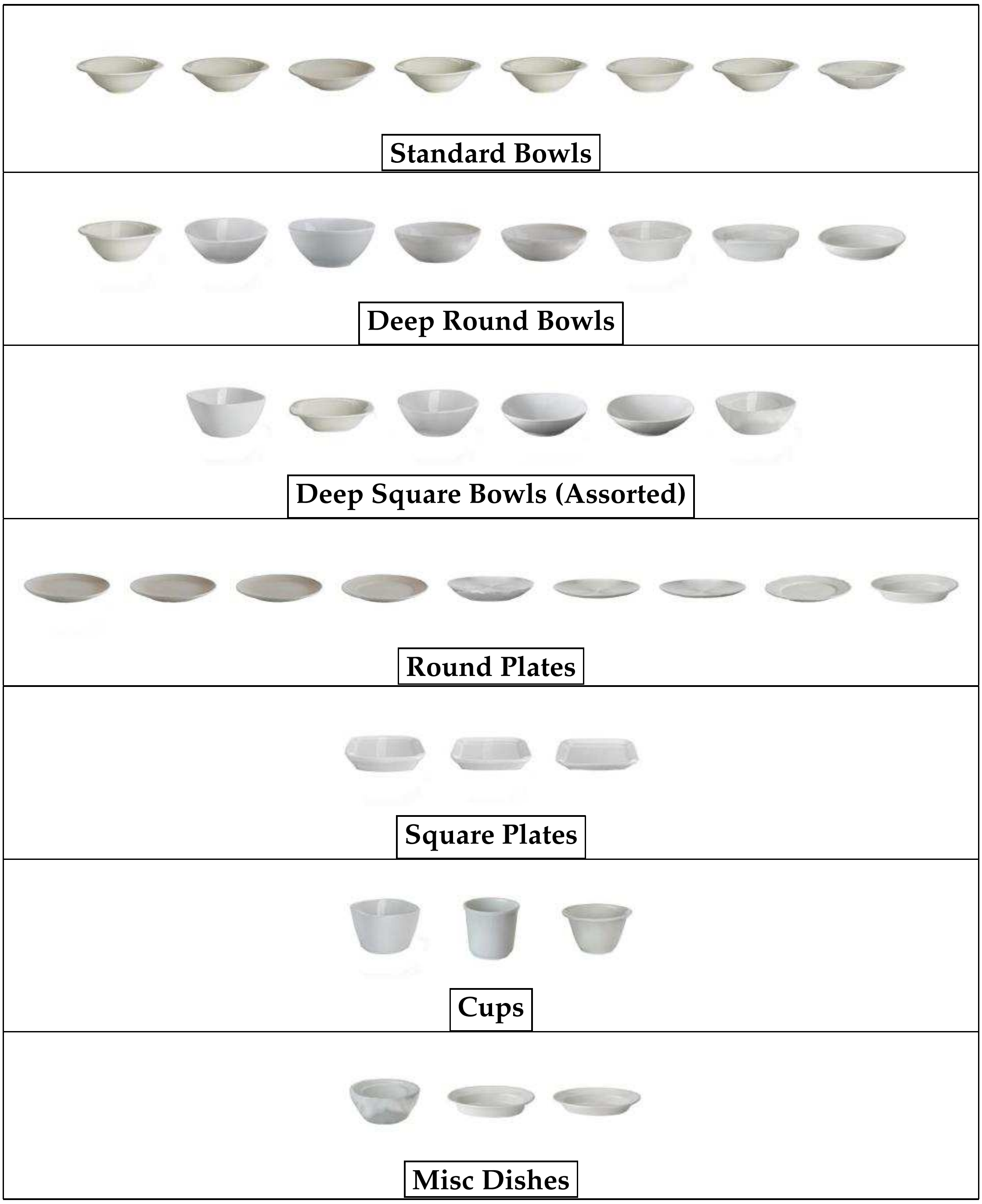


Figure 6.7: An example of a joint task from Experiment 2.

	control	*silent*	*talking*
Pre	0.68 (0.09)	0.65 (0.08)	0.70 (0.09)
Post	0.70 (0.14)	0.74 (0.16)	0.66 (0.09)

Table 6.1: The means and standard deviations for conceptual alignment scores for participant pairs in each condition and for both pre-tasks and post-tasks.

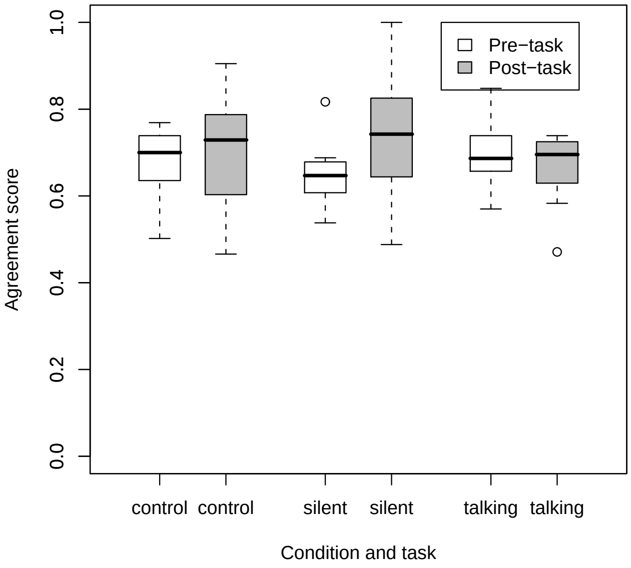


Figure 6.8: Distributions of conceptual alignment scores in both pre-tasks and post-tasks for each condition.

depended on the condition. The first test consisted of a mixed-effects ANOVA, with condition as the between-pairs variable and task (pre-task, post-task) as the within-pairs variable. There was no significant main effect of condition ($F(2, 21) = 0.05$, $p > .05$), task ($F(1, 21) = 1.58, p > .05$) or the interaction ($F(2, 21) = 2.97, p > .05$).

A second test, following the recommendation of Twisk and Proper (2004), consisted of an ANCOVA, with pairs' post-task scores as the dependent variable, pre-task scores as the covariate, and condition as the independent variable. Although pre-task scores were found to make a significant contribution ($F(1, 20) = 12.18$, $p < .01$), the analysis also showed no significant effect of condition ($F(2, 20) = 2.71, p > .05$).

Finally, a third test was run using Jacobson and Truax (1991)'s reliable change index. This method considers the amount of variation in all the participant pairs' pre-task scores, and then uses it independently for each pair to evaluate whether the pair's change from pre-task to post-task scores is a reliable change (i.e., less than 5% chance of it occurring by chance); see their paper for calculation details. Support for my hypotheses would be found if the number of pairs for which a reliable change was found was low in the *control* condition, higher in the *silent* condition, and highest in the *talking* condition. However, the results revealed that only 1 of the 24 pairs exhibited such a reliable change. This pair was in the *silent* condition and was the only pair with a post-task score of 1.00 (their pre-task score was 0.688).

Thus, all three tests failed to find support for my hypotheses. Conceptual alignment did not increase between pre-task and post-task, and no difference was found between the conditions.

6.4.2 Lexical convergence

Although I am mainly concerned with convergence in categorisation, it is also insightful to consider category label convergence, especially given the lack of the former. To that end, I define here a simple measure of lexical agreement, and make analogous tests to the three used above for categorisation. In addition, I will be able to conduct a few extra tests whose analogues were not possible for categorisation.

Defining a measure for comparing two sets of labels is tricky because participants can vary in their number of categories (and thus labels), and because they were free to use whatever labels they liked, including labels composed of multiple words. To deal with this, I define a relatively simple measure. In particular, given two sets of labels, I count the number of labels that occur in both, and divide it by the average number of labels in the two sets. I count only labels that match exactly, except for discrepancies in irrelevant aspects such as spelling, capitalisation, and pluralisation. For example, given one label set of "plates", "big bowls", "small bowls", and "trays", and a second set of "Plaet", "Big round bowl", "Big square bowl", "Smal bowl" and "Large tray", there are two matches (i.e., "plate" and "small bowl") and an average of 4.5 labels between the two sets, which yields an agreement of 0.444 ($2/4.5$). Although this measure is quite strict, it makes objective coding easier (which was done blind to the condition and task). The measure was used to calculate a degree of label agreement for every participant pair in both the pre-task and the post-task, which was then fed into the following analyses.

Using this measure, I can now compute the degree of agreement between partners in both pre-tasks and post-tasks for every pair, just as I did for the categorisation data. The resulting means and standard deviations for each condition are shown in Table 6.2, and boxplots are shown in Figure 6.9.

	control	*silent*	*talking*
Pre	0.31 (0.38)	0.18 (0.20)	0.18 (0.15)
Post	0.40 (0.38)	0.35 (0.29)	0.49 (0.33)

Table 6.2: The means and standard deviation for lexical alignment scores for participant pairs in each condition and for both pre-tasks and post-tasks.

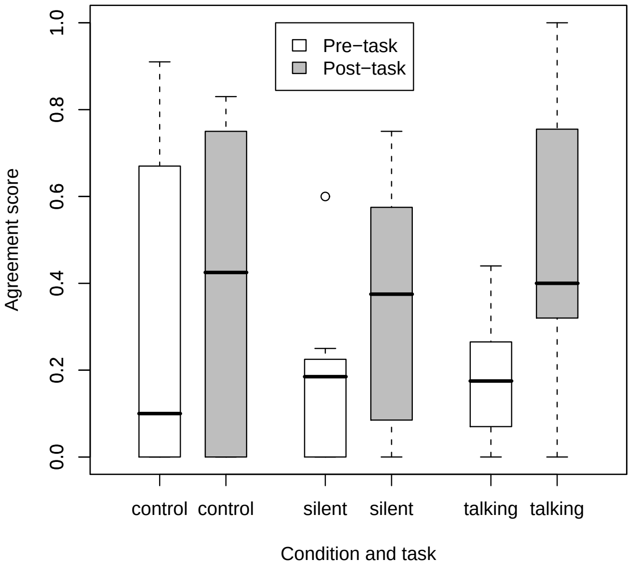


Figure 6.9: Distributions of lexical alignment scores in both pre-tasks and post-tasks for each condition.

I begin again with a mixed ANOVA (between: condition; within: task). Unlike the finding for categorisation, the main effect of task was highly significant ($F(1, 21) = 13.68$, $p < .01$). However, there was no significant effect of condition ($F(2, 21) = 0.23$, $p > .05$), nor of the interaction ($F(2, 21) = 1.52$, $p > .05$).

Next, I ran an ANCOVA again (dependent variable: post-task scores; covariate: pre-task scores; independent variable: condition). As for conceptual convergence, the covariate's contribution was again highly significant ($F(1, 20) = 14.35$, $p < .01$), while there was no significant effect of condition ($F(2, 20) = 1.23$, $p > .05$).

The third analysis was based, as before, on the reliable change index, measuring how many participants in each condition exhibited reliable changes between the pre-task and post-task, based on the amount of variability exhibited overall in the pre-task. Remarkably, despite the effects of task found in the first analysis, reliable changes were not found for any pair in any of the conditions: no pair showed higher levels of label agreement in the post-task than would be expected if they had not undergone interaction.

Together, these three tests do provide some evidence that lexical alignment occurred between the pre-task and post-task, although it is not unanimous. However, no difference was found between conditions.

The labelling data also allows for an extra kind of analysis that was not possible for categorisation. In particular, we can also look at the labels that partnerships used in the joint categorisation task and see if participants preferred to stick with those labels or to revert to their pre-task labels in the post-task. I use the same measure of label agreement as before, but this time compare the label set of each participant in the *silent* or *talking* condition (the *control* condition is excluded because it did not include a joint task) with both their own label set in the pre-task and the label set they used with their partner in the joint categorisation task. The resulting means and standard deviations for both interactive conditions are shown in Table 6.3, and boxplots are shown in Figure 6.10.

	silent	*talking*
Pre	0.37 (0.23)	0.51 (0.22)
Joint	0.58 (0.32)	0.63 (0.34)

Table 6.3: The means and standard deviation for label agreement scores in both interactive conditions between participants' own post-task labels and both their pre-task labels and their partnership joint-task labels.

A mixed ANOVA (between: condition; within: task) showed a main effect of task ($F(1, 30) = 10.59$, $p < .01$), but no main effect of condition ($F(1, 30) = 1.19$, $p > .05$) and no significant task-condition interaction ($F(1, 30) = 0.66$, $p > .05$). As for the previous analyses, it appears that participant pairs do converge in their lexical choices, but that this does not depend on the experimental condition.

For the sake of comparison, and to help verify that the naming agreement measure is sensible, I also performed a loosely analogous test with the *control* condition. Recall that in this condition, participants carried out another (their second) individual categorisation task instead of a joint task with their partner. As such, they never actually interacted with their partner, and were fully naive to both their categories and labels. In this case, I compared each *control* participant's post-task labels with both their pre-task labels (as for the other conditions above) and their partner's mid-task labels. Since again they would

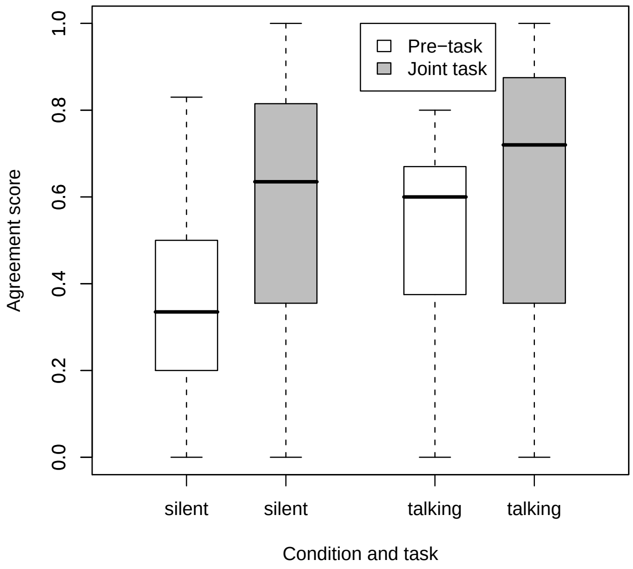


Figure 6.10: Distributions of lexical alignment scores in each condition between partners' own post-task labels with both their own pre-task labels and their partnership joint task labels.

have had no exposure to the latter, we should expect higher agreement with their pre-task labels. A paired t-test revealed that this was indeed the case: *control* participants' post-task labels agreed more with their own pre-task labels ($M = 0.61$, $SD = 0.40$) than with their partner's mid-task labels ($M = 0.37$, $SD = 0.34$), and the difference was significant ($t(15) = 2.91$, $p < .05$). Thus, *control* participants remained relatively consistent with themselves across tasks.

6.4.3 Lexical choices

Before finishing with the label data, it is informative to see what kinds of labels participants actually came up with. Recall that participants were free to use whatever labels they liked, and that their labels could consist of multiple words. And indeed, as we will see, they often made use of these possibilities.

In order to give an impression of the range of labels that participants used, I conducted two kinds of tallies, one based on the full labels and one on individual words used. The first tally counts the number of tasks in which each label was used (after having pruned the label data of differences in spelling, capitalisation, pluralisation, and punctuation). To put the resulting numbers in perspective, note that the total number of tasks in the experiment was 128 (3 pairs of individual tasks in each of eight sessions in the *control* condition, plus 2 pairs of individual tasks and 1 joint task in each of the other sixteen sessions: $6 \times 8 + 5 \times 16 = 128$). The most frequent label, "plate", was one of the category labels in 60 of those 128 tasks. The second word-based tally counts the number of categories across all tasks whose names included each word that occurred in labels. The total number of categories was 707 (although the sum across all word frequencies will be substantially higher, since many labels included multiple words). The most frequently occurring word in category labels was "bowl", which was included in the label of 232 of the 707 categories. Table 6.4 shows the fifteen most frequent labels and words, according to these two respective tallies.

Full label	Frequency	Word	Frequency
plate	60	bowl	232
cup	51	plate	178
bowl	45	round	98
square plate	32	square	88
round plate	31	dish	78
dish	30	cup	64
square bowl	24	deep	59
tray	23	shallow	57
deep bowl	22	rim	38
round bowl	21	tray	24
shallow bowl	19	saucer	16
serving dish	12	mug	16
saucer	12	circular	15
mug	11	lip	14
pot	10	serving	14

Table 6.4: The 15 most common labels (left) and words occurring in labels (right), with their respective frequencies. The label tallies are based on all the tasks from all the sessions, while the word tallies are based on all the categories from those tasks.

Other less common types of labels framed the objects in alternative ways. For instance, one participant divided up the items in terms of different kinds of "crockery", including "breakfast service crockery" and "lunch service crockery". Another participant used geometric terms, such as "tall cylinders" and "sliced spheres". Some participants used prepositional phrases or adjectives in brackets to describe the objects in more detail (e.g., "bowls with a rim", "cup (shallow)"). Other categories were explicitly given hybrid names, such as "cups and mugs" or "saucers/side plates". And a few labels presumably identified the kinds of things they may be used for (e.g., "noodles", "ice cream and strawberries") or things they resembled (e.g., "flying saucer"). Finally, a few labels from one participant seemed completely senseless (e.g., "X marks the spot", "Patagonia").

6.4.4 Use of dialogue

In this section, I focus on the recorded audio data from the joint categorisation tasks in the *talking* condition. Although dialogue strategies are not the focus of my experiment, it is useful to take a brief look at how participants actually used dialogue in the tasks. As a result, here I present several excerpts from the *talking* condition pairs which exemplify a few different ways in which dialogue was used.

The first general observation is that all of the participant pairs in the *talking* condition did in fact exploit the opportunity to talk. However, they varied in the extent to which they did so. Some pairs discussed their general categorisation schemes in some detail, and asked each other about uncertainties concerning which category to place an item (that was currently in the focus box) into or what to call a category. Other pairs were less communicative, with one pair in particular, after an initial exchange, only seeking confirmation a few times during an otherwise silent interaction.

Pairs generally began their linguistic interaction by asking each other and explaining what kind of categorisation schemes they had used in the previous task. In some cases, they began by settling on one category label (e.g., "Alright, what's the first category?") and populating it extensively before moving on to others. More often, they asked each other about the categorisation schemes they had used in the previous, individual task (e.g., "What kind of categories - what did you name your categories before?"), and/or discussed what scheme to use together (e.g., "How do we wanna categorise these?"). In some cases, one participant's scheme would be immediately accepted, without discussing the other's, as in this example:

A: How did you categorise them before?

B: Ummm, I categorised them, in, uh, plates, dishes and bowls. And categorised each by round or square, but I don't know whether that's very accurate, because I can't tell the difference between some of them, between -

A: Yeah.

B: - the plates and the dishes.

A: Yeah. We'll go for that.

B: Yeah, sure.

A: That'll be fine.

Participants often asked their partners for their opinion about categorisation decisions, such as the features of an object (e.g., "Would you say that's quite deep?") or what category to put it into (e.g., "Would you say that's a plate?", "What about this one?"). Occasionally, participants also asked each other for explicit explanations of how they understood different categories (e.g., "Wait. What's the difference, well, what would you say is the difference between a saucer and a plate?") Such questions were particularly prevalent when dealing with boundary cases, often indicated by two specific options (e.g, "Do you reckon that one's more a dish or a bowl?").

Here's an example of an uncertain boundary case and how it was readily resolved:

A: What'd you think, that's a bowl?

B: Yeah.

A: Or a plate.

B: It could be um, I reckon it could be either a plate or a bowl.

A: I don't know. I'll go for bowl.

B: Cool.

Occasionally, discussion of boundary cases triggered a partial restructuring of the category system, as in this example:

A: Is that a dish or a plate?

B: I'd say that's probably a plate.

A: Okay.

B: We could scrap our dishes section, and -

A: What, just have bowls and uh,

B: - have bowls or plates and settle on.

A: Yeah, I like that idea.

B: Cause that one is probably a round plate, isn't it?

A: Yeah.

Sometimes pairs would explicitly disagree, but normally, one participant would readily conform with the other, as shown below. This example also illustrates how decisions were often made by reference to similarity with other items already placed into categories:

A: Shallow bowl?

B: Uh, it's looking quite deep compared to the other shallow bowls.

A: Yeah.

B: It's another serving dish I think.

In a few cases, one participant suggested a category which changed how their partner categorised things. One way in which this happened was when a category label was introduced which was relatively unfamiliar to one participant. For example, one participant wanted to call an item a "trough", which was initially rejected by their partner for the more familiar "bowl". However, the first participant later explained what troughs were and created a trough category, and his partner later used the trough category as well. Another type of case was when one participant's categorisation made their partner see things in a way they had not considered before. This is demonstrated in the example below, which also shows how participants often appealed to what they had done in the previous task, and were even sometimes under the impression that the stimuli were the same across tasks.

A: What would you say these things are?

B: I, well ...

A: Oh, that's. I can't even remember what I put. I think I put mug.

B: Oh, right. I put, for one of the tall ones I put kitchen utensils -

A: Yeah.

B: - utensil holder, but that one, I think you could be right, yeah, I didn't think of it as a mug.

A: Cool, I'll put mug.

Finally, it's worth noting that pairings usually explicitly checked with each other before finalising their categorisations by clicking on the "Done" button. Here's the final exchange of a particularly enthusiastic pair:

A: Right, so are we done?

B: Yeah.

A: Are you happy with everything?

B: Couldn't be more happy.

A: Perfect. Done.

6.4.5 *Framework and experiment feedback*

Given that Experiment 2 is the first of my experiments to use the full-fledged general experimental framework developed in Chapter 4, as well as its lack of evidence for conceptual convergence, it is especially important to evaluate different aspects of both the specific experiment and the general framework. To that end, here I present the relevant questionnaire data, which concerns the stimuli, categorisation strategies and tendencies, and the experiment and program interface more generally. The feedback question numbers are marked with '#' in brackets and refer to Appendix B.2.

I begin with general feedback concerning the stimuli. Participants were asked to rate the set of experimental stimuli on a scale of 1 (very low) to 5 (very high) in terms of two criteria: how familiar the items were (#6), and how natural they looked (#7). The stimuli were judged to be both familiar ($M = 4.27$, $SD = 0.79$) and natural ($M = 3.81$, $SD = 0.89$). When asked to indicate whether they noticed differences between the stimulus sets from the three categorisation tasks (#8), participants varied in their responses, with 13 participants responding that they did not notice differences, while the rest claimed that they did to varying extents (e.g., "Yes", "There were sometimes more of one type of item than another", "Slight variances, but not particularly"). Finally, the questionnaire asked participants whether they mentally labelled stimuli linguistically (#5). A vast majority of participants (40) claimed that they did mentally label items with English words, with only four saying that they did not. The remaining four participants gave more subtle responses (e.g., "Not every item", "Generally, unless it was too ambiguous").

Participants were also asked about their basis for categorisation, whether this changed over the course of the experiment, and how they coordinated categorisation with their partner. For the basis (#2), most participants mentioned one or more criteria which they followed (so that many participants in the tallies below belong to two or more tallies). 37 participants mentioned some kind of physical basis (especially shape (26) and size (10)); 22 participants listed object function, use or purpose, sometimes indicating more specific criteria (e.g., "what kind of food/drink I think I would put in them"); 7 participants simply referred to the kind of object (e.g., "type") as a criterion; 4 participants mentioned that they categorised by evaluating similarity between objects (either different objects from the experiment, or other objects they were familiar with); and 1 participant referred to organisation ("how I would organise them in a cupboard"). In terms of changing their basis for categorisation across tasks (#3), 11 participants said they did not, 25 indicated that they did, while 11 claimed that they did but only slightly. Of those who said they did change their basis, a few offered details, such as "used size also later on" and "developed more complex categories as the tasks progressed". One participant mentioned methodological constraints, saying how they needed to be more flexible due to limited space allowed per category. And three participants mentioned that working with a partner affected the joint categorisation or their own subsequent individual categorisations. When the participants in the *silent* and *talking* conditions were asked about how they agreed with their partner on categories (#4) , there was a variety of responses, partly because they seemed to interpret the question differently. Participants had different opinions on how well they agreed (e.g., "agreed very well", "a little", "I disagreed with her categories"), the procedure which they followed (e.g., *silent* condition: "I went with her categories", "one person would enter a category name, generally the other person agreed with them"; *talking*: "by discussing what I had done in the first round and then discussing each picture on screen", "we picked an item and decided together on a label"), and the basis that they used together (e.g., "function, shape and depth", "both went for shape of opening"). Thus, the questionnaires offered independent evidence that participants conceptualised the items differently, that they sometimes underwent changes, and that they varied in their strategies.

A further question asked participants what they thought was the purpose of the experiment (#1). Only 9 participants gave no answer, claimed they had no idea, or made guesses that were well off (e.g., "to code different objects"). 26 participants gave responses that were correct but quite general or that were slightly off (e.g., "to see how people categorise every day things", "examine how we categorise visual objects and how we refine our thinking", "to see how your mind can be changed by another person"). The remaining 13 participants, all of whom had engaged in a joint categorisation task (6 in the *silent* condition, 7 in the *talking* condition) guessed the purpose remarkably well (e.g., "To see if you categorised things differently after working with a partner", "To find out the effects that the other persons categorizations had on our own"). From these responses, it seems that the experiment's purpose was pretty transparent to most of the participants, especially those who had worked together on a joint categorisation task.

The final two questions on the questionnaire were quite general: the first asked participants if they had any comments about the experimental program interface (#9), and the second asked for any more general comments about the experiment (#10). In both cases, the vast majority of participants either gave no comment or gave positive feedback (e.g., "Very straight forward to use"). However, 6 participants did provide some constructive criticism about the interface (e.g., “Sometimes there was a delay when choosing a picture”, “A bit sticky sometimes”, “the way items swap when you have one and then click another can be quite confusing”). Most of these comments concerned technical issues with mouse clicking in Java that were difficult to avoid, but fortunately, as suggested by the relative lack of negative comments, these things were only minor annoyances. Overall, then, the participants felt that both the program interface and the experiment in general were well-designed and easy to use and follow.

6.5 Discussion

6.5.1 *Summary of results*

The main hypothesis of this experiment was that engaging in a joint task would bring together people’s categorisation, and would do so more if they were able to talk to each other. The results do not provide support for this position. Pairs did not categorise any more similarly before interaction than they did after it.

However, the experiment also gathered other data of theoretical and methodological value. First, unlike categorisation, label agreement between participants did increase from pre-tasks to post-tasks. However, the extent to which this occurred did not depend on the experimental condition. Moreover, inspection of category labels suggested that a range of different kinds of words and phrases were used, with nouns referring to types of dishes (e.g., “plate”, “bowl”, “cup”) and adjectives identifying dimensions of variation (e.g., “square”, “round”, “deep”) being particularly common. Despite the lack of effect on conceptual or lexical convergence, dialogue was exploited by all the participant pairs in the *talking* condition, in order to agree on general categorisation schemes, category labels and the categorisation of particular (especially ambiguous) items. Finally, questionnaire feedback on different aspects of the experiment showed that participants found the stimuli suitable, varied in their bases and strategies for categorisation, guessed the purpose of the experiment with surprising accuracy, and were comfortable in using the program interface.

6.5.2 Conceptual alignment, lexical alignment and dialogue

The results of this experiment do not provide support for my hypotheses. In contrast with the past findings of Markman and Makin (1998), conceptual convergence was not increased by engaging in a joint task.

How can this be explained? Previous work has shown that engaging in a joint task with linguistic interaction does bring together people's subsequent categorisation (Markman and Makin 1998; Voiklis 2008; see Section 3.4.3.3). While these experiments did not tease apart the effects of interaction from those of dialogue specifically, the results here show no effect of either. This suggests that methodological factors may be responsible for the results, which I discuss in the next section.

Given the lack of expected conceptual convergence results, the lexical results take on more weight and offer indirect insight. The weak lexical convergence results are also surprising, although they did yield some significant effects. Pairs did show more lexical agreement in the post-task than in the pre-task, but the differences did not depend on the condition. This seems to conflict with previous work which has shown how interlocutors converge on shared referential expressions for objects (e.g., Clark and Wilkes-Gibbs 1986; Brennan and Clark 1996; see Section 3.4.3.2). In contrast to those studies, there is no evidence in Experiment 2 that either a joint task in general or dialogue specifically resulted in greater amounts of lexical convergence after interaction. The lack of difference between conditions suggests that even pairs in the *control* condition were converging more across the tasks, so that there may be an effect of familiarity with the stimuli and tasks on label convergence. Perhaps as people continue to work with the same objects they begin to give them less idiosyncratic names, leading to more agreement with each other. If so, the lexical agreement may be due to this kind of standardisation rather than partner coordination (Voiklis 2008).

However, this cannot be the full story, because there was also some evidence for convergence. Participants who engaged in a joint task tended to stick with their joint labels after interaction, rather than reverting to their own individual labels from before. This result cannot just be explained as all participants becoming less idiosyncratic across the tasks, because an analogous comparison with the *control* participants showed that people who did not interact were significantly more consistent with their own pre-task labels than their partner's mid-task labels. This shows that interaction does in fact have an impact on lexical convergence. To reconcile this with the previous finding, it seems that the general lexical convergence evident from pre-tasks to post-tasks may still be primarily due to the *silent* and *talking* conditions, but that the difference was statistically too small to manifest itself as a significant effect between the conditions in the other analyses. This would also be a little more consistent with previous lexical alignment work.

To the extent that the results do show lexical alignment, they suggest that interaction between individuals may result in lexical convergence without necessarily implying conceptual convergence. Interlocutors' agreements on "conceptual pacts" (Brennan and Clark 1996) may not reflect a corresponding underlying agreement in conceptualisation. Although it may seem surprising that this would not be the case at all, we should at least be cautious about assuming the latter from the former. Indeed, such a conclusion would be broadly compatible with the findings of Malt et al. (1999) (and my replication in Experiment 1) of a dissociation between naming and sorting. Echoing Schober (2005), labels do not necessarily reflect conceptualisation to the extent that we might imagine.

Nevertheless, overall, there was no support found for dialogue-induced conceptual alignment, and the lexical alignment results were relatively weak. The relative lack of dialogue effects for both conceptual and lexical convergence are also surprising given how pairs in the *talking* condition exploited the opportunity to use the speech channel. As shown in Section 6.4.4, participants used dialogue extensively and for a variety of purposes. They negotiated general categorisation systems, the nature and names of particular categories, and the categorisation of particular items. Why wouldn't such extensive and categorisation-relevant interaction result in some degree of conceptual convergence? While this may be a real effect, I suggested above that the results may perhaps be primarily due to methodological factors. Therefore, I discuss methodological issues concerning both the current experiment and general framework in the next section.

6.5.3 Methodological issues

I have already raised some methodological concerns with this experiment, and hinted at several others. Indeed, the experiment was stopped early due to these concerns. In this section, I discuss the methodological issues in more detail, suggest that they may be largely responsible for the experimental results (and lack thereof), and point to potential solutions.

One source of concern lies in the lack of a clear goal and potentially motivation for participants. Although the task procedure, mechanics and endpoint were well-defined, and participants seemed to have no trouble understanding and following them, there was no clear objective for them to reach. No particular categorisation was treated as "better" than any other, in that neither the individual nor joint tasks rewarded participants for how they categorised the items. As a result, in principle, nothing prevented them from creating arbitrary categories and assigning things to them randomly. A particularly striking example of this was the participant who seemed to just give random names to his categories, such as "X marks the spot" and "Patagonia", although fortunately, that was by far the exception rather than the rule. Nevertheless, lack of incentive for producing coherent categories and meaningful labels are a potential factor. This contrasts my study with those of Markman and Makin (1998) and Voiklis (2008), in which participant pairs were given a common goal that was intended to be interesting and challenging. Recall that both of those studies demonstrated how working on a communicative joint task could result in converging categorisation. It is possible that the lack of such convergence in my experiment is due to a lack of goal and motivation rather than more theoretical reasons. Given the fact that many participants guessed the purpose of the experiment fairly accurately, it would potentially cost little and gain much to modify the design so that participants are given the explicit goal of trying to converge their categories.

Another unfortunate aspect of the current experiment is a lack of convergence data available for analysis from the joint task itself. For any pair, we can compare the two participants' categorisations in both the pre-tasks and post-tasks, because they produce independent outputs. However, there is no clear way to do this for the current joint task (as is highlighted by Figure 6.2). On one hand, in that task, participants make decisions jointly, suggesting that we could treat them as converged by definition. On the other hand, they are still obviously conceptualising the items with their own minds, and there is no good evidence to indicate that they were in perfect agreement on all of their decisions. The lack of comparison available for participant pairs in the joint task means that we are not able to say, on the basis of the data, how much participants conceptually converge *while* interacting. Yet this is an important open question in the literature: while on-line alignment between interlocutors has been demonstrated at different linguistic levels, it is much more controversial at the non-linguistic conceptual level (see Section 3.4.3).

A shortage of analysable data is actually a more general problem with the design of Experiment 2. Although we do get convergence data for the pre-tasks and post-tasks, there are just these two values for each participant pair. In other words, from the 48 participants in the experiment, only 48 datapoints were obtained, 24 for pre-task comparisons and 24 for post-task comparisons. And since the main analysis is based on the difference between the two (or just the post-task), and there are three different conditions, that effectively means just 8 datapoints per condition. Finally, assigning random subsets of the stimulus space to participant pairs makes it difficult to run an alternative, item-based analysis. Of course, part of the reason for the shortage of data is that the experiment was stopped early, but still, the design is inefficient.

The previous point also raises the question of whether a pre-task is really needed. If we had no between-subject conditions, and were comparing only the effect of interaction on one group of participants, then it would be essential. But since the main comparison of interest is between conditions, and the *control* condition is meant to provide a baseline, it could be argued that the pre-task is redundant, with the main analyses being possible with just the post-task data. Indeed, there are two further reasons why the pre-task should perhaps be eliminated. First, the time that it takes up in a standard experimental session could be exploited instead to gather other kinds of more useful data, so that we get more datapoints per condition. Second, it may be that engaging in the pre-task "self-primes" participants to stick more to their own categorisations, making it more difficult for them to be influenced by their partner.

Also, it appears that trying to maximise the amount of variation in categorisation in the experiment may have backfired statistically. Recall that for both kinds of alignment, the residual change index analysis suggested that in fact no convergence was occurring of either sort. These results may be due to the particular emphasis that this method puts on the amount of variation in pre-task scores. However, pre-task scores varied widely for both lexical and conceptual alignment, since participants' output varied, and only chance determined whether they had been paired with someone whose output was very similar or different to their own. As a result, score changes from pre-task to post-task needed to be very high to achieve significance thresholds, so that only one pair in one type of alignment actually met these criteria.

Finally, it may also be the case that learning about another person's concepts is more easily done through a series of short tasks rather than a single large one. Indeed, most category learning experiments in psychology present only one stimulus at a time. Although free classification tasks inherently require multiple stimuli and categories at once, the number could be reduced, which would also allow for a greater number of tasks, and thus a greater number of datapoints.

Nevertheless, despite the methodological shortcomings of Experiment 2, the questionnaire feedback gives plenty of room for optimism regarding the general experimental framework developed in Chapter 4. Participants varied in how they categorised and labelled the items, and generally found the stimuli familiar and natural, despite the fact that most of them were morphed. Moreover, aside from a couple of minor annoyances, the questionnaire feedback indicated that participants found the program interface easy to use and the task instructions clear. This is particularly encouraging with regards to the paired aspect of the experiment, especially in the *silent* condition, since it demonstrates that it is possible to conduct meaningful interactive experiments using free classification tasks either with or without dialogue.

In summary, Experiment 2 had a few serious methodological shortcomings, including the lack of a clear goal, no simple way of assessing conceptual alignment during interaction, and limited data obtained per experimental session. However, the general experimental framework seems to have resolved certain methodological challenges, and holds promise for future experiments.

6.6 Conclusion

Experiment 2 was intended to test the hypothesis that dialogue could result in conceptual convergence between individuals, over and above that which occurs as a result of working together on a joint task (without dialogue). However, the results not only showed no support for this hypothesis, but even failed to show conceptual convergence due to interaction in general. This runs counter to previous findings, suggesting that the experiment may have methodological issues. Indeed, the design did feature several flaws that were not identified earlier, which I address in subsequent experiments. Nevertheless, the experiment did provide some other useful and interesting data. First, it showed how participants used a variety of different category labels, and how there was some evidence for lexical alignment, despite the lack of conceptual alignment. Second, it showed that even though dialogue wasn't necessary to complete the joint tasks, it was actively used when available to make various kinds of joint decisions. Finally, feedback from the participants suggested that it was the particular design of Experiment 2 that was methodologically at fault, and not the general framework. Therefore, more experiments can and should be conducted within this framework to explore unresolved issues.

Chapter 7

Experiment 3: conceptual alignment during and after dialogue

7.1 Introduction

Experiment 2 fell short of its ambitions. It was intended to investigate whether dialogue played an important role in conceptual coordination. To that end, it examined whether the use of dialogue in a joint categorisation task brought together participants' individual categorisations, but found that a joint task does not bring about conceptual convergence. Moreover, there were also surprisingly few differences found between the conditions in terms of lexical convergence. However, I then pointed out various methodological shortcomings of the experiment and argued that the results may have been methodological artifacts rather than real theoretical findings. As a result, I concluded that further experiments are needed to evaluate these issues.

An obvious strategy in designing follow-up experiments to address these concerns is to make methodological modifications which lead to a more sensitive design and are more likely to reveal differences between conditions. If doing so yields the same pattern of results, then this would add credibility to the findings of Experiment 2. Otherwise, if expected differences do come up, then we could reevaluate that conclusion and my experimental hypotheses.

However, the lack of expected results in Experiment 2 also suggests that we shift the theoretical focus a little. In particular, we should not restrict ourselves to studying conceptual coordination following interaction. If conceptual convergence between pairs of participants is not evident after their interaction, then what about even during the interaction itself? In effect, Experiment 2 presupposed that participants conceptualise things the same way while they interact, even though their actual concepts are still their own (as could be seen in Figure 6.2). This was similar to Brennan and Clark's (1996) conceptual pacts, whereby interlocutors are assumed to agree on a joint conceptualisation when they converge on a particular referential expression to refer to an object. However, as has been emphasised throughout this thesis, lexical alignment does not necessarily imply conceptual alignment. Moreover, while Pickering and Garrod (2004) have argued that interlocutors align cognitive representations during dialogue, little is known about whether such alignment occurs in joint tasks not involving dialogue, and how much, if at all, dialogue augments it.

Experiment 3 is also meant to address the methodological concerns raised by Experiment 2 (see Section 6.5.3). First of all, the participants should be given a clear and challenging goal, in hopes of increasing both their motivation and the sensitivity of the experiment. Secondly, and more generally, the task structure should be designed to produce more readily analysable and relevant data. At the same time, there would ideally be less variation in the pre-task, or the pre-task would be dropped altogether. Finally, the joint part of the experiment should be modified to accommodate the change in theoretical scope.

Therefore, in this chapter, I present an experiment which attempts to address the methodological shortcomings of Experiment 2, while also expanding its theoretical scope of inquiry to include conceptual coordination both during and after interaction. However, since the issues with Experiment 2 did not appear to be due to problems with my general experimental framework, Experiment 3 continues to adopt this framework.

7.2 Overview

The main theoretical purpose of Experiment 3 is similar to that of Experiment 2: to investigate whether dialogue brings about conceptual alignment between interacting individuals. However, an important difference is that rather than exclusively inquiring whether joint action and dialogue bring about conceptual coordination *after* interaction, this time I also investigate conceptual coordination *during* interaction. Also since Experiment 2 showed evidence of lexical but not of conceptual convergence, this raises the possibility of a dissociation between the two. I therefore investigate lexical alignment as well, asking the same questions as for conceptual alignment.

The main concrete modification relative to Experiment 2 concerns changes to the interactive part of the experiment. Firstly, the two participants no longer categorised a joint set of stimuli together through a synchronised interface. Instead, they categorised the items independently (except that they could talk in the *talking* condition), but were explicitly asked to categorise the objects as similarly as possible as each other. Therefore, the focus now was shifted to comparing their conceptual alignment during interaction under the *control*, *silent* and *talking* conditions. The three cases are visualised in Figures 7.1, 7.2 and 7.3, respectively.

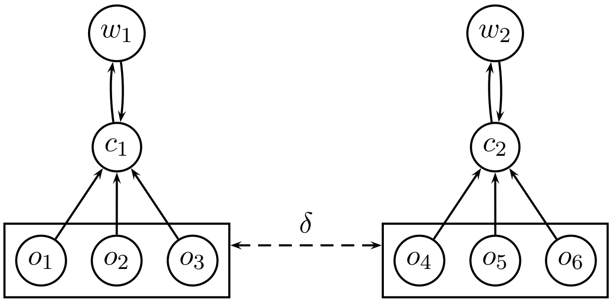


Figure 7.1: Interaction and feedback in the *control* condition in Experiment 3.

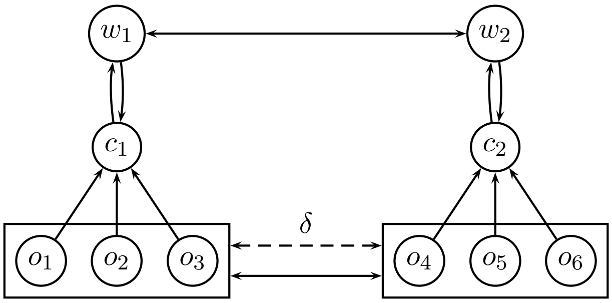


Figure 7.2: Interaction and feedback in the *silent* condition in Experiment 3.

Moreover, the single large joint task was replaced with a series of ten smaller tasks. After each task, they were also given feedback by their partner's categories and labels appearing on their screen next to their own (potentially helping them to learn across tasks), along with a score indicating how well they had coordinated. These modifications

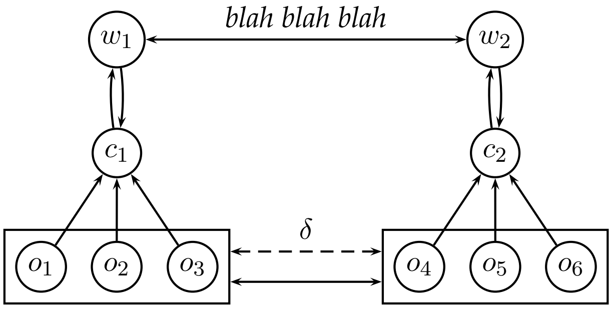


Figure 7.3: Interaction and feedback in the *talking* condition in Experiment 3.

addressed several methodological concerns: they gave the participants a particular goal, and they produced data (ten datapoints per pair) for analysing the degree of conceptual alignment during interaction. However, the modified task instructions also change the theoretical question asked by the joint phase of the experiment. Rather than looking at whether dialogue *spontaneously* improves conceptual coordination during interaction, the experiment now asks whether it *potentially* can be used to improve it. Thus, this experiment does not focus on automatic alignment (Pickering and Garrod 2004), but also encourages the conscious use of language as a coordination tool.

Next, the post-interaction part of the experiment was also expanded. However, rather than increasing the number of categorisation tasks, a large set of similarity judgements was added. Recall that acquiring concepts can lead to categorical perception, a phenomenon by which the mind exaggerates the similarity of things that fall within the same category and the dissimilarity of things that are in different categories (Goldstone 1994; see Section 2.6.4). But if categorisation can affect similarity systematically in this way, then conceptual alignment could also result in alignment of similarity spaces. As a result, the potential convergent effects of engaging in a joint task could also manifest themselves in similarity judgement tasks. Indeed, previous work has documented such results (Malt and Sloman 2004).

Finally, the pre-task of the experiment was entirely eliminated. This not only addressed the fact that this part was unnecessary, if not problematic, but also made more room for the expanded joint phase and individual phase of the experiment. Also, it addresses the potential concern of self-priming in Experiment 2 (see Section 6.5.3).

In summary, Experiment 3 makes several methodological changes relative to Experiment 2, which not only seek to make the experiment more sensitive and interesting, but also result in slightly shifted theoretical questions. In particular, Experiment 3 asks:

1. **Can engaging in joint categorisation tasks bring people's categorisation, lexicalisation, and similarity spaces closer together, both during and after interaction?**
2. **If so, are these effects stronger if people can also communicate through dialogue?**

7.3 Methods

The experiment had two phases: a joint phase and an individual phase. The joint phase consisted of a sequence of ten joint categorisation tasks. The availability of feedback and dialogue was manipulated, as in Experiment 2. In contrast, the individual phase, like the pre-tasks and post-tasks in Experiment 2, was individual and the same for all conditions. It consisted of a sequence of 60 similarity judgements and one categorisation task.

7.3.1 Participants

Participants were 60 adult native English speakers (age: $M = 22.0$, $SD = 3.8$), mostly undergraduate students recruited through the University of Edinburgh student employment website. There were 36 female and 24 male participants.

Participants were assigned to pairings randomly, and pairings were assigned to conditions randomly, except that the relative numbers of different gender pairings was controlled and kept the same in each condition. In particular, there were four female-female pairs, four female-male pairs, and two male-male pairs assigned to each of the three conditions (described in Section 7.3.2). These numbers were chosen to balance the composition of gender pairings across conditions, even though gender was not expected to make a difference in the experiment. Again, partners did not know each other before the experiment.

7.3.2 Conditions

Experiment 3 had the same three conditions as Experiment 2: *control*, *silent* and *talking*. In the *control* condition, participants were not allowed to talk during the experiment, and they also did not receive feedback on their partner's category groupings or labels. In the *silent* condition, they also could not talk, but they did receive feedback on their partner's category groupings and labels. In the *talking* condition, participants received feedback on their partner's category groupings and labels, and they were allowed to talk freely during the joint categorisation tasks.

7.3.3 Stimuli

7.3.3.1 Generation

As in the previous experiment, the stimuli used in Experiment 3 consisted of pictures of dish-like objects. Again, morphing software was used to generate a more fluid space, but in this case attempts were made to make the space even more fluid and to increase the level of ambiguity of the items.

In particular, the first step was the selection of four pictures out of the ten base stimuli from Experiment 2. I chose items which I considered to be prototypical members of what in English would be called 'bowl', 'plate', 'tray' and 'cup', respectively. This contrasted with Experiment 2, where, according to the distribution of participants' labels (see Section 6.4.3), there were unequal ratios of different kinds of dishes. For each of the six possible pairs of these four images, I then took the morphing sequences from Experiment 2, consisting of 13 intermediate stimuli for each pair.

Then, for each of these six morphing sequences, I used the central interpolant image (which I call a derived base stimulus) for a further set of "second-generation" morphing sequences. All fifteen possible pairs of the six derived base stimuli were then used to morph a further thirteen interpolants. All together, this resulted in a domain of 277 pictures of dish-like objects ($277 = 4 + 6 \times 13 + 15 \times 13$). Figure 7.4 shows the four original base stimuli, Figure 7.5 shows the six combinations of these and the resulting derived base stimuli, and Figure 7.6 shows an example morphing sequence taken between two derived base stimuli.

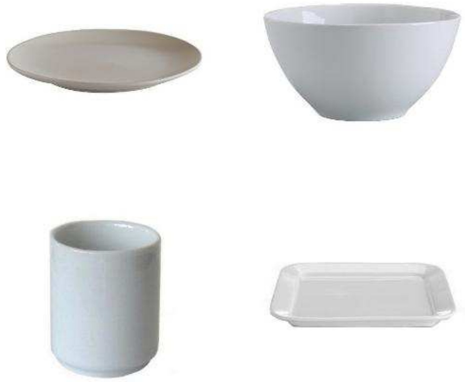


Figure 7.4: The original base stimuli used in Experiment 3.

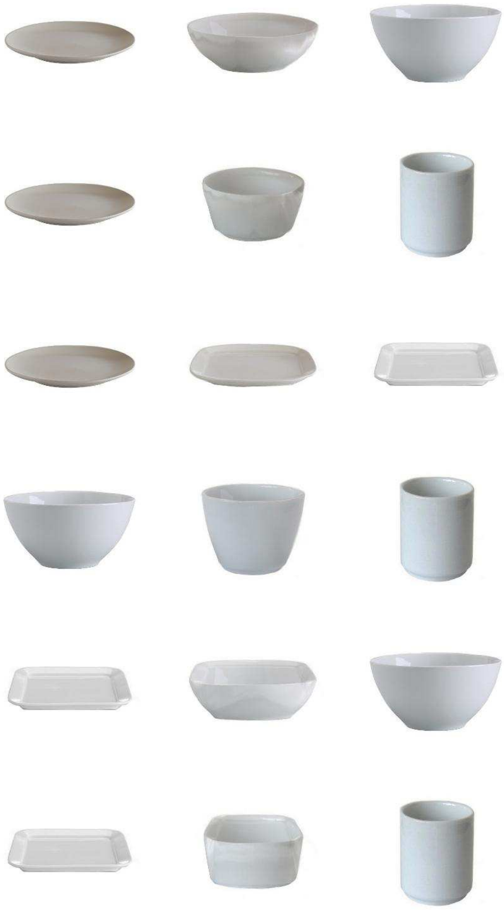


Figure 7.5: The derived base stimuli of Experiment 3 (in the middle column), derived from the midpoint interpolations of each pair of original base stimuli (left and right columns).

7.3.3.2 *Allotment*

From the superset of 277 generated stimuli, 100 were randomly selected to be used in the joint categorisation tasks. These were then randomly divided into ten sets of ten

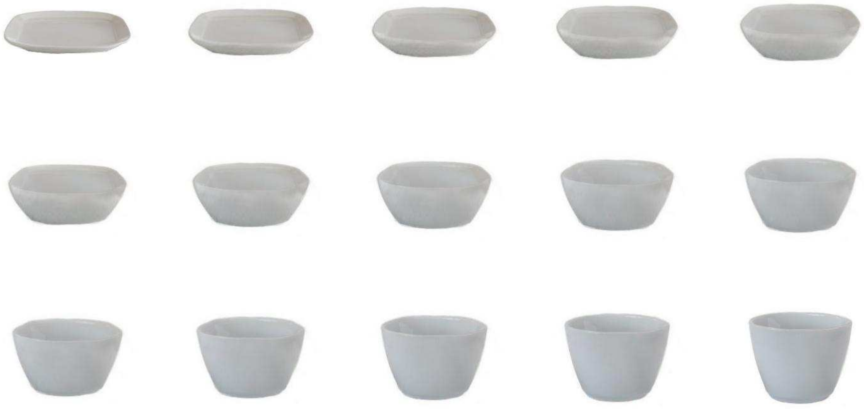


Figure 7.6: An example morphing sequence from two derived base stimuli used in Experiment 3.

stimuli, with each set being assigned to one task. Ten different random orderings of these tasks were then defined. Each of these task orderings was assigned to one participant pair of the same gender combination (female-female, female-male, or male-male) in each condition.

The similarity judgement tasks used 60 novel and 60 familiar stimuli. The novel stimuli were randomly chosen from the remaining 177 in the global set. The familiar stimuli were chosen from the 100 stimuli previously chosen, with three random pairs chosen from each joint categorisation task. The order of these tasks was randomised, but kept the same for all participants.

The individual categorisation task used 35 familiar stimuli (and no novel stimuli), randomly selected from the remaining 40 stimuli which were used in the joint categorisation tasks (100) but not reused in the similarity judgement tasks (60). Familiar stimuli were used in order to increase the chances of the joint phase tasks affecting these individual tasks.

The visual order in which stimuli were displayed within tasks varied (within pairs) in the joint phase, so that participants could not rely on stimulus order to achieve coordination. In contrast, it was kept constant in both parts of the individual phase for all participants in all pairs and conditions. This was done so that the individual phase could be treated like a post-treatment test, with all participants undertaking exactly the same test.

7.3.4 *Procedure*

The experiment consisted of two experimental phases, the *joint* and the *individual* phase, and a short questionnaire. As in Experiment 2, pairs of participants were brought into the experimental lab and assigned randomly to two separate computer cubicles. After filling out a consent form, they began the experiment.

7.3.4.1 *Joint phase*

7.3.4.1.1 Joint categorisation tasks After a brief practice individual categorisation task (as in Experiment 2), participant pairs carried out a sequence of ten joint categorisation tasks, in each of which they sorted 10 items into 2-4 categories (with a maximum of 5 items per category). However, these tasks were fundamentally different from Experiment 2. This time, both participants first categorised the items individually (just as in the individual categorisation tasks), although the *talking* condition pairs were allowed to talk during these tasks. At the end of each task, in the *silent* and *talking* conditions, they saw their partner's categories and labels, and had a chance to compare them to their own. They also saw their task score (in all conditions), which was computed using Meila's (2007) measure, but rounded to two decimal places and converted to percents before being shown to participants. Their task was to try to group things as similarly as possible to their partner (and thus to maximise their score), and they were told that the pair with the highest overall scores would receive an extra financial reward (£5 each). The instructions also emphasised that the labels did not matter to the score: only the category groupings affected it. As before, participants could change their decisions as many times as they liked before committing their choices, including both the category labels and their members.

7.3.4.2 *Individual phase*

Once participants had completed the ten joint categorisation tasks of the joint phase, the individual phase began. For the rest of the experiment, participants worked on their own, and they were told that at this point. This phase of the experiment consisted of two parts: a sequence of similarity judgement tasks, followed by a single individual categorisation task.

7.3.4.2.1 Similarity judgement tasks In the similarity judgement tasks, participants saw two items at a time, side-by-side, and were asked to assess their degree of similarity. Participants rated items on a scale of 1 to 7, with 1 meaning that the items were very dissimilar and 7 meaning that they were very similar. The first two similarity tasks were for practice, and involved pictures of furniture. After that, participants carried out a sequence of sixty similarity judgement tasks with the dish stimuli. Half of the tasks involved previously seen stimuli, while the other half involved novel ones (but from the same dish domain).

7.3.4.2.2 Individual categorisation task After the similarity tasks, participants carried out a single individual categorisation task. This task was the same as the individual categorisation tasks in Experiment 2, except that participants sorted 35 items into 2 to 4 categories (with a maximum of 20 items per category). The stimuli were a subset of those used from the joint tasks.

7.3.4.3 Questionnaire

Once participants had finished with both the joint phase and the individual phase of the experiment, they filled out a questionnaire. The questionnaire was very similar to that used in Experiment 2. It asked them for demographic information such as age, gender, and language background, as well as feedback questions concerning the experiment, such as what they thought the experiment was about or what strategies they used to align their categories.

7.3.5 Program

The experiment was run through a revised version of the computer program developed for Experiment 2. The revisions reflected the methodological changes, as described here. Snapshots of the program interface for the different tasks can be found in Appendix C.3.

The interface for the joint categorisation tasks worked much the same way as for individual tasks in Experiment 2, with participants individually sorting the items into categories. However, the program window was arranged a little differently, mainly reflecting the reduction in the number of categories available and the fact that the window now had to visually accommodate both people's categories (during feedback). The part of the window with category boxes was split vertically in half, with the participant's four category boxes appearing in the left half. The right half was reserved for the participant's partner's categories, and was left empty (and black) during the task. After both partners had finished each task, the program calculated the agreement score between the two sets of categories, and the score was shown underneath the focus box, highlighted by a surrounding red border. In addition, in the *silent* and *talking* conditions, the participants were also then shown their partner's category groupings and labels to the right of their own (they were warned this would happen in the task instructions). The partner's categories were visually ordered to be in the sequence which gave the best categorisation match (and hence also which reflects the score). The participants then had 45 seconds to see the score and study the category feedback (if any), and then the next task automatically began.

The interface for the similarity tasks was very simple. Pairs of stimuli were shown side-by-side, one pair at time. Underneath the pair there were seven radio buttons, identifying degrees of similarity, from lowest to highest. Participants clicked on the button of their choice, which immediately triggered the next similarity judgement task.

The individual categorisation task interface was the same as in Experiment 2, except that there were now only four category boxes (in a 2x2 arrangement), rather than nine, and the category boxes occupied the right half of the window.

7.4 Results

I present the data in three general parts, corresponding to three types of alignment between participants: conceptual, lexical, and similarity. However, unlike Experiment 2, I do not go on to present the dialogue data, and I only skim over the questionnaire data. This is because the audio files, unfortunately, were lost, while the main purpose of the questionnaire (i.e., assessing the general framework), was already addressed in the previous experiment.

But first, in order to give a concrete example of the data, I first show the output from a joint categorisation task. Figure 7.7 is output from a randomly selected task. It is the fifth joint task of one participant pair in the *silent* condition. This means that they have already had feedback on each other's groupings and labels four times each. In this case, both participants used all four categories.

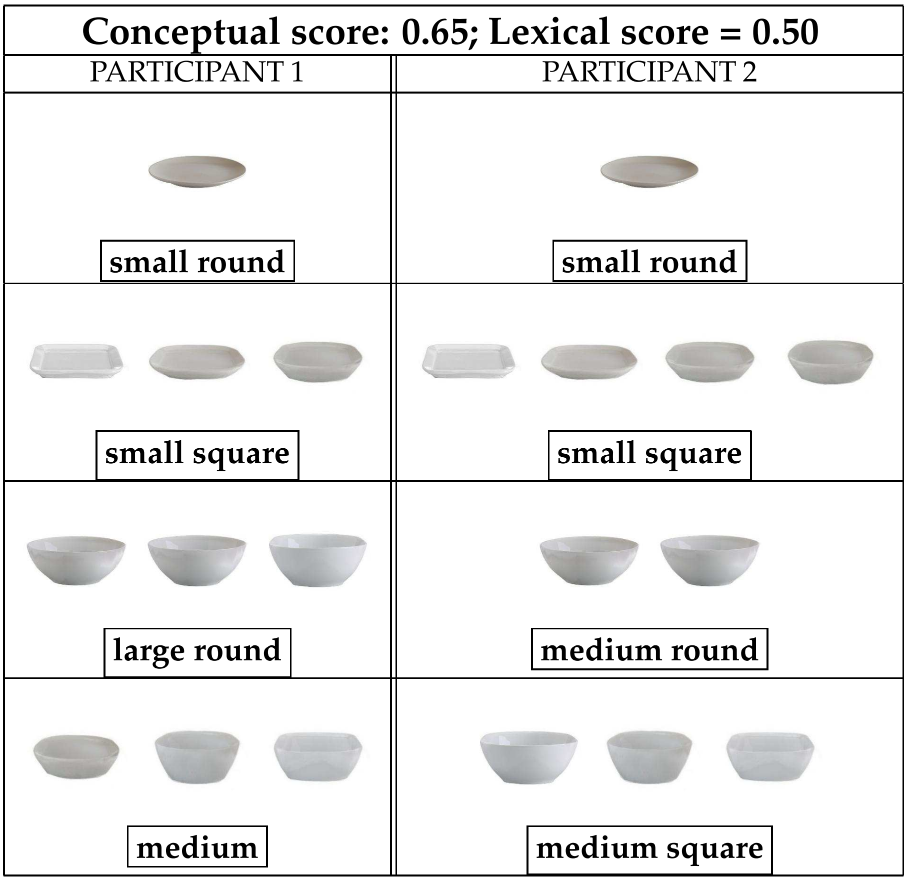


Figure 7.7: An example of a joint task from Experiment 3.

7.4.1 Conceptual alignment

7.4.1.1 Joint categorisation tasks

Recall that the main hypothesis regarding the joint tasks was that interacting pairs would align their categorisation over time, and that this would occur more if they were allowed to talk. Before the main analyses, however, I show, in Figure 7.8, the sequence of scores for every pair, as well as the mean scores, organised by condition. From these plots, which provide a quick overview of the raw data, I note three informal observations. First, there was a lot of variability in scores, even within pairs. Second, there were quite a few perfect scores of 1, especially in the *talking* condition. Third, the scores look highest in general in the *talking* condition, but, surprisingly, the difference between *control* and *silent* conditions is less clear.

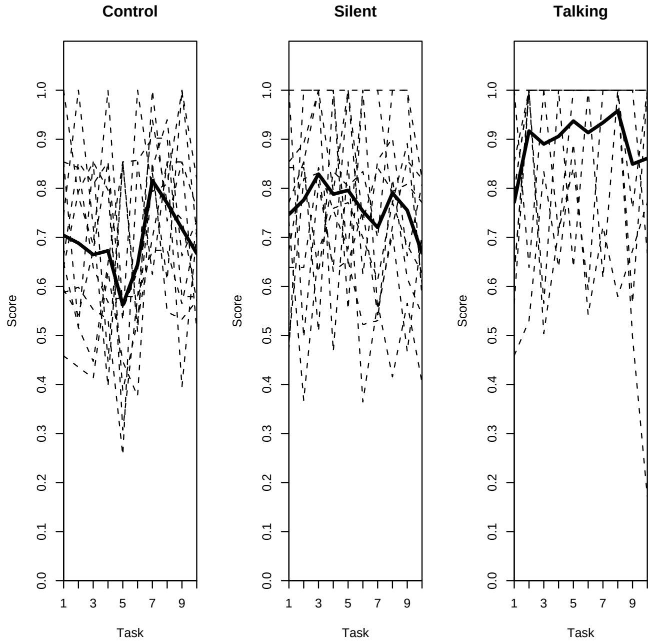


Figure 7.8: Conceptual alignment for each task in all sessions, by condition.

These preliminary impressions point to caution in conducting analysis. In particular, a high number of ceiling scores in the *talking* condition suggests that a standard parametric test is both inappropriate and relatively unnecessary. Indeed, the score distributions by condition, shown in Figure 7.9, highlight the obvious ceiling effect in the *talking* condition. As a result, I conduct standard parametric tests with only the *control* and *silent* conditions, but include all conditions for the tests which should be less statistically sensitive.

First, I look at whether pairs in the *control* and *silent* conditions increased conceptual coordination over time across the tasks, and whether the extent to which they did so depended on the condition. Table 7.1 shows the mean scores for each joint task in each condition, although I exclude the *talking* condition pairs from the analysis because of the high number of ceiling scores. As before, the means are clearly highest in the *talking* condition, but here we can also see that the mean scores tend to be higher in the *silent* condition than in the *control* condition. However, there is no obvious learning effect, with mean scores staying about the same across tasks.

A linear mixed effects analysis (using R's 'lmer' and 'anova' functions) supports this impression. In order to assess whether there was general improvement in scores over time,

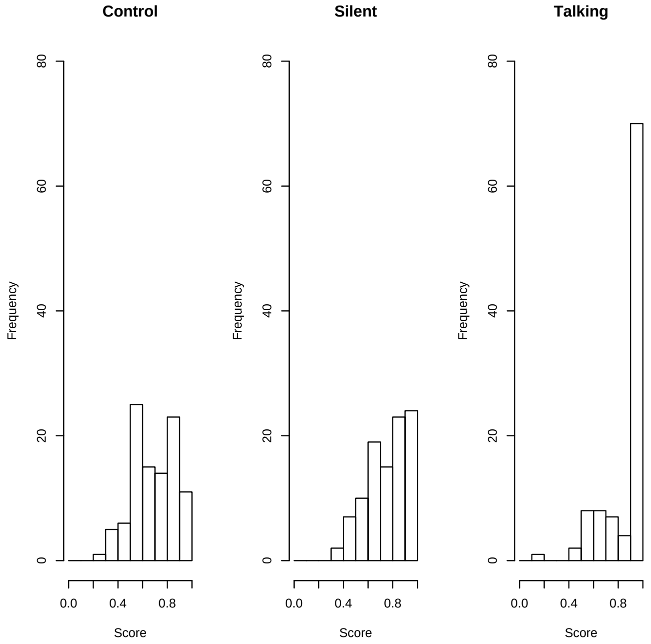


Figure 7.9: Conceptual alignment score distributions, by condition.

Task	*control*	*silent*	*talking*
1	0.70 (0.16)	0.75 (0.19)	0.77 (0.20)
2	0.69 (0.19)	0.78 (0.21)	0.92 (0.18)
3	0.66 (0.16)	0.83 (0.18)	0.89 (0.20)
4	0.67 (0.20)	0.79 (0.18)	0.91 (0.15)
5	0.56 (0.23)	0.80 (0.16)	0.94 (0.12)
6	0.64 (0.19)	0.75 (0.21)	0.91 (0.18)
7	0.82 (0.12)	0.72 (0.19)	0.93 (0.14)
8	0.77 (0.13)	0.79 (0.17)	0.96 (0.13)
9	0.72 (0.21)	0.76 (0.18)	0.85 (0.21)
10	0.67 (0.10)	0.67 (0.14)	0.86 (0.27)

Table 7.1: The means and standard deviations for conceptual alignment scores for each joint task in each condition.

I applied a simple model[1] with task as a fixed effect and pair-specific score intercepts as the only random effects. Although such linear mixed models do not provide clear p-values, I used the heuristic, as recommended by Baayen (2008), of looking at the t-value for the fixed effect of task: an absolute value of this greater than about 2 indicates a significant effect. However, in this case, the t-value fell far short of this (-0.29), confirming that pairs did not manage to improve over time. As a result, for the rest of the analysis, I ignore the longitudinal variable of task order.

[1] The R lmer call: lmer(score ˜task + (1 |session)).

Next, since there were so many ceiling scores in the *talking* condition, I compare that condition with the other two just by looking at the number of ceiling scores. There were 68 ceiling scores in the *talking* condition, and a total of 28 (*control*: 7; *silent*: 21) for the other two. A χ^2 test confirmed that the difference was highly significant ($\chi^2(1) = 86.87$, $p < .001$). Thus, the conditions did differ in how often pairs obtained ceiling scores.

I also compare the different conditions based on the mean scores across all tasks for each participant pair. This analysis has the statistical advantage of reducing the number of ceiling effects (there were only two pairs for which the mean scores were 1, both in the *talking* condition). The relative values of the mean scores, as shown in Table 7.2, were consistent with my hypotheses. A one-way ANOVA showed a highly significant main effect of condition ($F(2, 27) = 10.23$, $p < .001$). Pre-planned comparisons revealed that the *talking* condition did significantly better than the *silent* condition ($t(27) = 2.89$, $p < .01$), while there was no difference between the *control* and *silent* conditions ($t(27) = -1.57$, $p > .05$). This shows that, surprisingly, feedback on partner groupings and labels did not increase conceptual coordination, although being able to talk was beneficial.

control	*silent*	*talking*
0.69 (0.18)	0.76 (0.18)	0.89 (0.18)

Table 7.2: The means and standard deviations for conceptual alignment scores across the joint tasks in each condition.

7.4.1.2 *Individual categorisation tasks*

So far the analysis has focused on conceptual alignment during interaction. Now I analyse the results from the individual part of the experiment to see whether interaction and dialogue resulted in conceptual alignment that persisted beyond the interaction. I turn first to whether there was conceptual agreement between partners in the subsequent individual categorisation task. In between, of course, partners also underwent the sequence of similarity judgment tasks, but I analyse that separately in Section 7.4.3.

First, I computed the degree of agreement between partners on the individual categorisation task, using Meila's (2007) measure. The resulting means for each condition are shown in Table 7.3, and their spread is plotted in Figure 7.10.

control	*silent*	*talking*
0.72 (0.09)	0.72 (0.08)	0.74 (0.11)

Table 7.3: The means and standard deviations for categorisation agreement scores in the individual categorisation tasks in each condition.

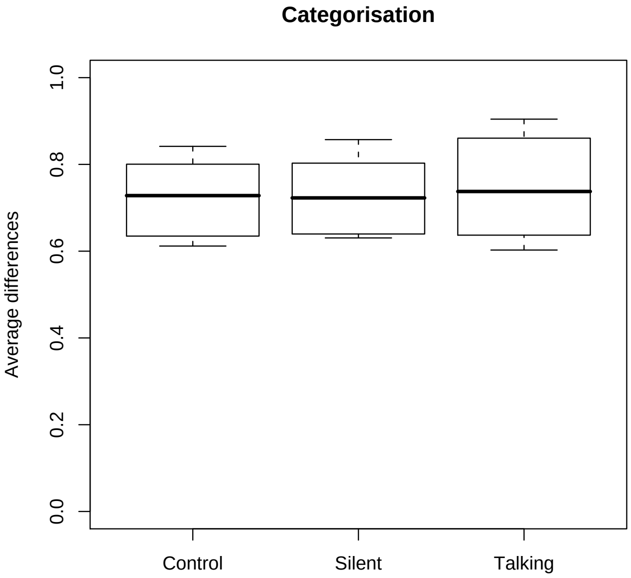


Figure 7.10: Conceptual alignment in the individual categorisation tasks.

A one-way ANOVA was conducted to test for the effects of condition on subsequent categorisation agreement. The analysis revealed no significant effect of condition ($F(2, 27) = 0.22, p > .05$). Thus, interaction and dialogue did not cause conceptual alignment during the individual categorisation task.

I also checked for a relationship between agreement in the joint and individual categorisation tasks, regardless of condition. A linear regression analysis between pairs' mean scores on the joint tasks and their scores on the individual categorisation task revealed no significant relationship ($\beta = 0.16$, $p > .05$, $R^2 = 0.05$).

7.4.2 Lexical alignment

7.4.2.1 Joint categorisation tasks

Although my hypotheses are focused on conceptual convergence, analogous analyses to those above can be performed on the category label data as well. In order to calculate the degree of label agreement on a given task for a participant pair, I use the same strict measure as was defined in Experiment 2, based on the proportion of full labels shared among the pair (see Section 6.4.2 for details).

Figure 7.11 shows the resulting lexical agreement scores for every pair, organised by condition. There appears to be even more variation and fluctuation in agreement scores than there was for conceptual agreement (see Figure 7.8), with a fair number of both ceiling and floor scores. However, it also appears clear that scores are highest in the *talking* condition and lowest in the *control* condition, as expected.

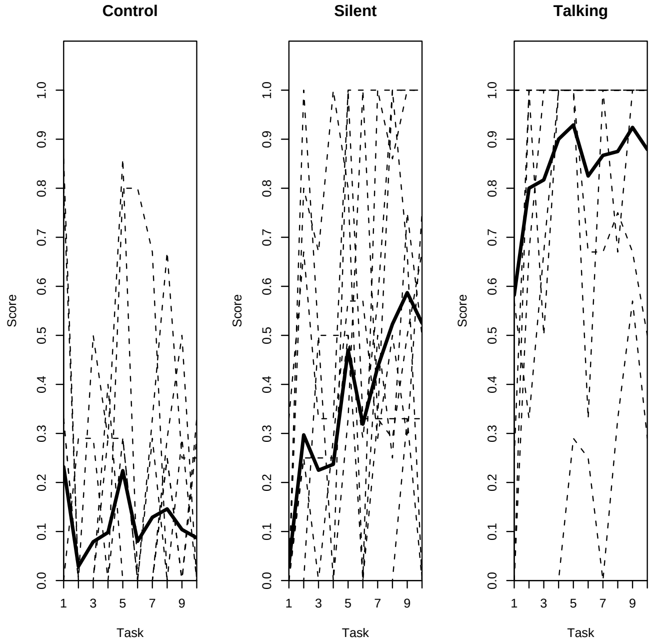


Figure 7.11: Lexical agreement on all runs for each condition.

I consider first whether pairs' lexical agreement scores increased over time, and how that might depend on the condition. Table 7.4 shows the means for each task and each condition. The values in the table give the impression that scores in the *silent* condition go up over time, but that this is less clear for the *control* condition (which has consistently low scores) or the *talking* condition (which has consistently high scores). This is intimately related with the distribution across conditions of floor and ceiling scores, and makes a linear mixed model analysis (as was used for conceptual convergence) problematic. In particular, such an analysis would inappropriately penalise pairs in the *talking* condition in terms of improvement over time, since they do so well from early on that they have far less potential room for improvement. Moreover, treating the scores in the early task(s) as covariates (analogously to Experiment 2) would also be inappropriate, because the pairs in the *talking* condition had an explicit advantage from the very first task.

Task	*control*	*silent*	*talking*
1	0.23 (0.34)	0.03 (0.10)	0.58 (0.47)
2	0.03 (0.09)	0.30 (0.38)	0.80 (0.36)
3	0.08 (0.17)	0.22 (0.26)	0.82 (0.34)
4	0.10 (0.16)	0.24 (0.32)	0.90 (0.32)
5	0.22 (0.34)	0.47 (0.39)	0.93 (0.22)
6	0.08 (0.25)	0.32 (0.41)	0.82 (0.30)
7	0.13 (0.23)	0.44 (0.35)	0.87 (0.32)
8	0.15 (0.22)	0.52 (0.41)	0.88 (0.23)
9	0.10 (0.18)	0.59 (0.35)	0.92 (0.16)
10	0.09 (0.14)	0.52 (0.42)	0.88 (0.26)

Table 7.4: The means and standard deviations for lexical alignment scores in the joint tasks in each condition.

Given these difficulties and the fact that this issue is not central to my main hypotheses, I resorted to a simplified analysis. In particular, I considered each condition independently, looking at the mean scores for each task. These means were shown graphically with the bold lines in Figure 7.11. Using the means also has the advantage of smoothing out the data and eliminating many floor and ceiling values. I performed a simple linear regression for each condition separately, testing specifically for the effect of task. The results showed that the effect of task was not significant for the *control* condition ($\beta = 0.00$, $p > .05$, $R^2 = 0.02$), highly significant for the *silent* condition ($\beta = 0.05$, $p < .001$, $R^2 = 0.79$), but also significant for the *talking* condition ($\beta = 0.02$, $p < .05$, $R^2 = 0.44$). Thus, pairs converged in their choices of category labels if they interacted, even if they were not engaged in full-fledged dialogue.

Next, I compare the three conditions, collapsing across task order. First, I consider the ceiling and floor scores more closely. There were a large number of ceiling scores (i.e., pairs used identical sets of labels), especially in the *talking* condition, but also many floor scores (i.e., pairs shared no full labels in common), especially in the *control* condition. Table 7.5 shows the totals of both for each condition. The relative values of these counts are in line with the experimental hypotheses, and the differences between conditions were significant for both ceiling scores ($\chi^2(2) = 145.45$, $p < .001$) and floor scores ($\chi^2(2) = 88.28$, $p < .001$). This confirms that the conditions did differ in how often partners used identical label sets and how often they used completely different labels.

	control	*silent*	*talking*
Ceiling	0	17	75
Floor	72	40	7

Table 7.5: The total number of ceiling and floor scores in each condition.

I turn next to mean scores, again collapsing across tasks. The means, shown in Table 7.6, were lowest in the *control* condition and highest in the *talking* condition, as expected. Due to the high number of ceiling and floor scores, and generally non-normal score distributions, I applied a non-parametric test to compare scores across conditions. In particular, a Kruskal-Wallis test was applied and confirmed that the differences between the conditions were highly significant ($\chi^2(2) = 142.84$, $p < .001$). Moreover, preplanned comparisons were significant for both the *control* ($U = 14$, $n_1 = 10$, $n_2 = 10$, $p < .01$) and *talking* ($U = 91$, $n_1 = 10$, $n_2 = 10$, $p < .01$) conditions relative to the *silent* condition. Thus, *control* pairs were less lexically aligned than *silent* pairs, who in turn were less aligned than *talking* pairs.

control	*silent*	*talking*
0.12 (0.23)	0.37 (0.37)	0.84 (0.31)

Table 7.6: The means and standard deviations for lexical alignment scores across the joint tasks in each condition.

7.4.2.2 *Individual categorisation tasks*

We can next examine whether lexical alignment was preserved in the individual categorisation task. I used the same measure of labeling agreement as above, and computed it for each pair. The means for each condition are shown in Table 7.7, and their spread is plotted in Figure 7.12.

Due to large fluctuation and non-normal distribution of the data, I again used a Kruskal-Wallis test to compare the conditions. The result was significant ($\chi^2(2) = 6.42$, $p < .05$).

control	*silent*	*talking*
0.09 (0.22)	0.33 (0.31)	0.47 (0.40)

Table 7.7: The means and standard deviations for label agreement scores in the individual categorisation tasks in each condition.

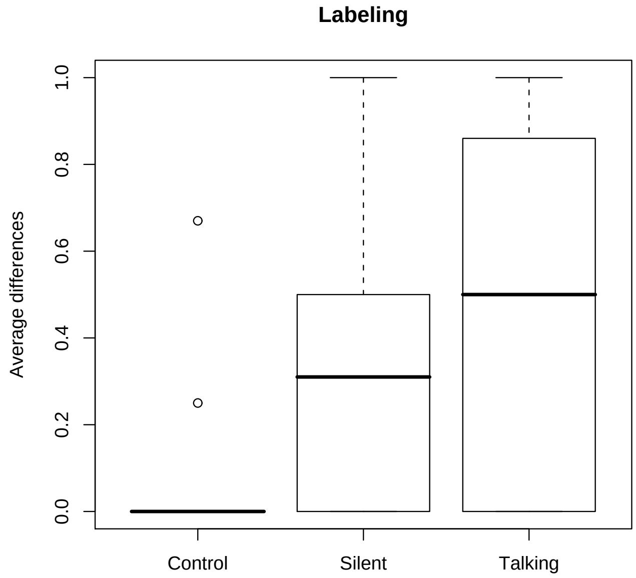


Figure 7.12: Lexical alignment in the individual categorisation tasks.

However, preplanned comparisons showed that while the difference between *control* and *silent* conditions was also significant ($U = 24.5$, $n_1 = 10$, $n_2 = 10$, $p < .05$), that between the *talking* and *silent* conditions was not ($U = 60$, $n_1 = 10$, $n_2 = 10$, $p > .05$). In other words, pairs that interacted ended up with more similar sets of labels in the individual categorisation task, regardless of whether they could talk.

Finally, I checked for a relationship between label agreement in the joint and individual categorisation tasks, regardless of condition. In contrast to the case for conceptual convergence, a linear regression showed a significant relationship ($\beta = 0.52$, $p < .01$, $R^2 = 0.29$). In other words, the degree to which partners aligned lexically in the joint tasks was reflected in the degree to which they remained lexically aligned afterwards.

7.4.3 Similarity convergence

Recall that all participants carried out the same 60 similarity judgments, for each of which they gave a number between 1 and 7 indicating how similar they found the two objects. I now compare the judgments of participants and their partners, and test whether how close they are depends on the condition.

Quantifying the degree of agreement between two people's similarity judgments is not trivial, however. The meaning of a similarity assignment of 2 or 5 (for example) is not well-defined and thus may depend on the participant. Therefore, I first normalised each participant's judgments, by subtracting each judgment from that participant's mean judgments, and dividing it by the standard deviation of that participant's judgments. Then, for each participant pair and each similarity task, I defined the similarity agreement between the pair as the absolute value of the difference between their standardised judgments. The means of the resulting judgments for each condition are shown in Table 7.8, and their spread is plotted in Figure 7.13.

control	*silent*	*talking*
0.79 (0.71)	0.76 (0.70)	0.70 (0.64)

Table 7.8: The means and standard deviations for agreement scores in the similarity tasks in each condition.

A one-way ANOVA was conducted to test for the effects of condition on similarity agreement. The analysis revealed no significant effect of condition ($F(2, 27) = 0.16$, $p > .05$). Thus, interaction and dialogue in the joint tasks had no effect on alignment in the similarity tasks.

I also checked for relationships between similarity agreements with both conceptual and lexical agreement in the joint categorisation tasks. Linear regressions for the mean degrees of agreement did not show significant effects in either case (conceptual: $\beta = 0.30$, $p > .05$, $R^2 = 0.01$; lexical: $\beta = 0.12$, $p > .05$, $R^2 = 0.02$). Thus, convergence in similarity judgments did not appear to be related to either conceptual or lexical convergence.

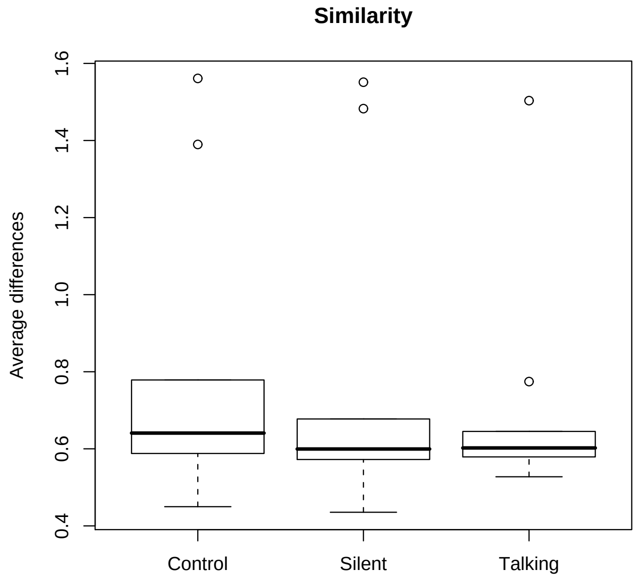


Figure 7.13: Similarity judgment agreement in the individual similarity tasks.

7.4.4 Questionnaire data

The questionnaire asked mostly the same questions as were asked in Experiment 2, aimed largely at assessing the framework. However, I do not go into detail here as I did there, for three reasons. First, although Experiment 2 had several methodological shortcomings, its questionnaire data already suggested that the general framework was not to blame (see Section 6.5.3). Second, the questionnaire responses were generally quite similar to those in Experiment 2. Third, and most important, the experiment yielded several significant effects, all of which were in line with my hypotheses. This provides perhaps the best and most direct evidence that the framework is fruitful, and makes close scrutiny of the questionnaire data largely redundant. However, given the change to the experiment and joint task structure, it is worth taking a brief look at the questions concerning the basis of categorisation. The feedback question numbers are marked with '#' in brackets and refer to Appendix C.2.

The most dominant bases for categorisation (#2) were different physical properties, which were mentioned by 56 different participants. In contrast, only 12 participants appealed to the kind of object, 7 resorted to object type or function and 2 mentioned the judgment of similarity relative to other objects. Finally, two participants mentioned the use of task coordination strategy as a basis for categorisation (e.g., "tried to standardize my own for her while also adopting to her style"). When asked about whether they changed their basis of categorisation (#3), 13 indicated clearly that they did, and 18 that they did not; 3 more mentioned their partner and how they tried to coordinate together; 2 said that they made changes when the constraints of the experiment (e.g., the maximum number or sizes of categories) required this of them, and the remaining 24 gave other kinds of responses that did not fall clearly into one of the previous criteria. Participant responses considering coordination strategies (#4) depended a lot on the experimental condition. In the *control* condition, participants generally either listed their basis for categorisation, indicated that they did nothing specific to coordinate, that they tried to use simple or small category systems, or that they studied the score feedback and tried to deduce from it how their partner categorised. In the *silent* condition, participants also indicated a reliance on simple, small or consistent systems, but also mentioned trying to identify their common basis, or switching to their partners' basis, or, conversely, sticking to their own system in hopes that their partner would switch. In the *talking* condition, some participants mentioned similar strategies to those in the other conditions, but others mentioned how they communicated together to arrive at decisions, sometimes in quite specific ways (e.g., "Talked about properties of objects, worked out obviously distinct objects first").

7.5 Discussion

7.5.1 *Summary of results*

The main hypothesis of Experiment 3 was that interaction on a joint task would increase conceptual alignment, and that dialogue would make the effect even stronger. Alignment was tested for both during and beyond interaction. Moreover, although conceptual alignment was the main focus, I also looked at lexical and similarity alignment.

The results depended substantially on whether we were looking at alignment during or beyond interaction, and on the type of alignment being investigated. Conceptual alignment was very significant in the *talking* condition during interaction, but did not occur in the other conditions, and was not evident in any condition after interaction (including the similarity tasks). In other words, feedback on category labels and groupings did not enhance conceptual convergence, but being able to resort to dialogue did. Lexical alignment, on the other hand, did occur during interaction with or without dialogue, but was stronger in the *talking* condition. Moreover, in this case, there was more convergence evident beyond the interaction in the *silent* condition than in the *control* condition, but no difference between the *talking* and *silent* conditions. For lexical convergence, then, just getting feedback on category groupings and labels was enough to get people to converge on their labels, although dialogue did enhance this further during interaction. Finally, there was no effect of condition on the alignment of similarity judgements, nor any evidence that similarity alignment correlated with lexical or conceptual alignment.

7.5.2 Conceptual and lexical alignment during interaction

Recall that Experiment 2 had failed to find lasting effects of a joint task in general, and dialogue specifically, on conceptual convergence between individuals following interaction. This surprising result begged the question of whether conceptual convergence would even occur during interaction. The mechanics and structure of the joint categorisation tasks were modified for Experiment 3 so that they would provide data that would bear directly on this question.

The results showed clear effects of dialogue on both conceptual and lexical alignment during interaction. Participants who could talk to each other during the joint categorisation tasks categorised things much more similarly to each other, and also used more similar category labels. Indeed, unlike in the other two conditions, *talking* pairs often managed to adopt identical sets of categories and labels. Although the loss of audio data prevents us from being able to analyse how participants used dialogue to achieve these results, dialogue was clearly very helpful to this end.

However, without recourse to dialogue, the results were strikingly different. Participants who received full feedback on their partners' category groupings and labels did not manage to conceptually converge across the ten joint tasks, and did not align any more with each other than pairs who did not interact at all. On the other hand, this did not apply to lexical convergence: feedback was sufficient for pairs to lexically align with each other, although not to the same extent as pairs who could talk together.

These results thus show a dissociation between lexical and conceptual convergence, but only in the absence of full-fledged language use between interacting participants. When participant pairs could talk, they aligned very well both lexically and conceptually, and these presumably corresponded to each other. But when their linguistic interaction was impoverished, so that they got feedback without being able to talk, they still aligned lexically, but failed to do so conceptually. This is particularly striking since, unlike in Experiment 2, participant pairs were specifically instructed to try to coordinate their categories rather than their labels.

The results could serve to reconcile perspectives in the literature (discussed mainly in Section 3.4) that may at first seem to be in conflict with each other. On one hand, some authors have argued that there is a close correspondence between conceptual and lexical alignment. Brennan and Clark (1996) equate the two a priori, beginning their study with the claim that "labels reflect conceptualisations" (p. 1482). Similarly, Pickering and Garrod (2004) argue that interlocutors align their mental representations at different levels, including linguistic and non-linguistic ones. On the other hand, other authors have shown a dissociation between lexical and conceptual processes, including alignment. Both Malt et al. (1999) and my replication in Experiment 1 have shown that the way speakers of a language divide up certain conceptual domains linguistically does not neatly correspond to how they do so non-linguistically. And Schober et al. (2004) have shown that respondents in telephone surveys, although faced with questions wrapped in the same linguistic forms, diverged from the intended meanings of them, especially if the interviewers did not interactively clarify key terms.

The current results suggest that both perspectives are right. When full dialogue is available, interlocutors can align both lexically and conceptually, and then these two levels presumably correspond to each other. But when it is not, people can still align lexically or accept each other's lexicalisations, while failing to conceptualise things in the same way. Indeed, this opposition has already been captured in the theoretical debate between the commentary and responses in Pickering and Garrod (2004)'s BBS article. Schober (2004) argued that the authors' interactive alignment account went too far in claiming that interlocutors also aligned conceptually. Pickering and Garrod's response was that Schober's arguments were based on data that did not involve full-fledged dialogue, and that this form of language was central to their claims. The results here provide an empirical basis for this debate, showing that lexical and conceptual alignment do correspond, but only when full dialogue is available.

It remains to be explained why feedback should be ineffective, while dialogue improves performance so much. After all, participants in the *silent* condition were given a full view after each task of their partner's categories along with their labels. In terms of conceptual structure, this information provided them with an indication of both the senses and exemplars of their partner's concepts (see Section 2.5). Why shouldn't they be able to use it effectively to align with those concepts? The most obvious explanation is that the task was still too hard, and that participants did not have enough opportunity to learn and coordinate. This may be because the subtle perceptual differences among the stimuli that were intended to maximise ambiguity and potential variation in categorisation may have also had the effect of making learning from feedback a very subtle and challenging task, which required not only conceptual flexibility but also fine perceptual attention. And as we already know, perceptual learning is slow (Goldstone 1998). Moreover, the fact that participants could not interact during the tasks meant that the joint phase was effectively a type of coordination game, in which two people have a shared goal and their actions depend on each other, but they cannot check with each other (Schelling 1960). They must independently decide, after each task, whether to stick to their own concepts, switch to their partner's, aim for a middle ground, or try something completely different. This adds an extra dimension to the already difficult challenge of conceptual coordination.

Dialogue, on the other hand, had a very significant effect on conceptual convergence. I suggest that this is primarily due to the process of grounding which participants can achieve through dialogue. Clark and Brennan (1991) argued that the development of common ground between people depends significantly on the medium of communication. The more that a medium lacks the properties of natural, face-to-face conversation, the harder it gets to establish common ground. By being able to talk freely with each other, people can interactively build up a shared way of both seeing and describing objects (Brennan and Clark 1996). Recall that there were also signs of this in the dialogue data from Experiment 2, suggesting that this process also applies to my categorisation tasks. In addition, this grounding process also addresses the coordination problem that participants face. Participants can decide together, explicitly or implicitly, whose or what categorisation approach to adopt in each successive task, and they can even revisit their strategy during the tasks. As such, language helps solve the coordination problem (Lewis 1969).

7.5.3 Conceptual, lexical and similarity alignment beyond interaction

The data and analysis of Experiment 2 had been designed to target conceptual and lexical convergence after engaging in a joint task. However, the results had been counterintuitive. The results were thus questioned on methodological grounds, and Experiment 3 was meant to address the shortcomings.

The first issue was whether conceptual alignment was even to be found **during** interaction, a possibility that had been assumed rather than tested in Experiment 2. As we have seen in the previous section, Experiment 3 found strong evidence for conceptual alignment in the *talking* condition, but none in the *silent* condition. This suggests that if conceptual convergence was to be found beyond the interaction, this should occur in the *talking* condition, but not in the other two.

However, this was not the case. In fact, no differences were found among the conditions in post-interaction conceptual alignment. The differences among pairs' categorisations were no smaller in the *talking* condition than the other two. In other words, although dialogue brought together people's categorisations, it only seemed to do so for the period of interaction, not beyond. The same was true for the similarity judgement results, with no differences turning up between conditions.

These results are inconsistent with previous studies demonstrating how a linguistically mediated joint task can result in categorisation alignment lasting beyond the interaction (see Section 3.4.3.3). The experiments of Markman and Makin (1998) and Voiklis (2008) had involved individual categorisation tasks, but found that pairs that had worked together with the help of language categorised more similarly to each other afterwards. Moreover, Malt and Sloman (2004) had found that using the same terms for things (from two reasonable alternatives) also brought together people's similarity judgements, which again clashes with my results.

How can we explain these discrepancies between my results and past work? I appeal again to the stimulus domain, and the perceptual challenges it produces. My stimuli varied subtly along perceptual dimensions such as depth and roundedness, so that for participants to converge in a persistent way on similar categorisations would partly implicate a significant degree of perceptual learning. In contrast, the stimuli in previous work were not specifically designed to be maximally ambiguous and featured more discrete differences between them. For instance, Voiklis's (2008) alien stimuli varied along six observable dimensions with binary values. Therefore, participants might align by settling on the same relative saliences of the dimensions, rather than subtly shifting category boundaries along perceptually continuous dimensions. Thus, the contrast in results may be due to conceptual alignment in my experiment relying more on perceptual change than previous work. This is consistent with Goldstone's (1998) argument that "perceptual processing is slower to change than higher-level conceptual processing" (p. 242).

Unlike the conceptual and similarity alignment results, the analysis based on the lexical data revealed that lexical alignment was evident in the two interactive conditions. In other words, participants tended to use similar category labels to their partners in the individual categorisation task if they at least had feedback about each other's category groupings and labels during the joint tasks. This result is consistent with the large body of literature showing how referential expressions that are used in dialogue can get entrenched for people even after they stop interacting with a particular partner (e.g., Brennan and Clark 1996; Malt and Sloman 2004; Metzing and Brennan 2003; see Section 3.4.2).

However, the lack of conceptual alignment here despite lexical alignment again warns us against the presumption that they are equivalent or always go hand in hand. Indeed, these results again serve to highlight the special role that dialogue can play online in conceptual coordination. It appears that dialogue can bring together people's conceptualisations, but only for the duration of the interaction (at least of the kind of stimuli used in my experiment). Together with my previous findings, Experiment 3 suggests that lexical alignment can induce and correspond to conceptual alignment, but only if supported by full and current dialogue.

7.5.4 Methodological evaluation

One of the main goals of Experiment 3 was to address the methodological shortcomings of Experiment 2. This included establishing a clear goal and motivation for participants, redesigning the joint task so it provides alignment data, and setting up smaller but more numerous tasks. Most importantly, the experiment was meant to be more sensitive and to yield more interesting results with respect to my hypotheses. So how did it do?

The modifications to the joint task were generally successful. There were more joint tasks, and each of these tasks now readily yielded both conceptual and lexical alignment scores for each participant pair. Analysis of this data provided some significant results, each of which supported prior intuitions (unlike Experiment 2).

On the other hand, it remained surprising that no significant differences were found in conceptual convergence between the *control* and *silent* conditions, which suggests that feedback from another's category groupings and labels is useless for coordinating one's categories with them. Such a conclusion seems counterintuitive, and raises the question whether further methodological adjustments are not needed, such as a still greater number of joint tasks. This is taken up in Experiment 4.

Given the large differences in convergence between the *talking* condition and the other two, it is important to point to a substantial mechanical difference between them. In particular, in the *talking* condition, participant pairs were not just restricted to feedback between tasks, but could actively interact during the tasks as well. Therefore, they could discuss, negotiate and confirm individual categorisation decisions with each other before committing to them. Moreover, participants could "meta-negotiate" how to go about the task in general, such as what basis of categorisation to use. This could help them avoid the convention problem facing the *silent* (and *control*) condition participants of whether to continue categorising their own way (in hopes that their partner will adapt to them) or adapting to their partner (in hopes that their partner will not change their basis). It is possible that if simple labels could be exchanged during the tasks, or if, conversely, dialogue was only allowed between tasks, the large differences between the *silent* and *talking* conditions would diminish. However, dynamic interaction could also be considered a crucial feature of dialogue (Clark 1996), so that it is not clear whether this is a theoretical or methodological point.

Overall, the design of the joint phase of the experiment was undoubtedly a substantial improvement over Experiment 2. On the other hand, the individual phase of Experiment 3 yielded mostly null results. There were no differences between conditions in either the individual similarity or categorisation tasks. In particular, even pairs in the *talking* condition, who had aligned so well in the joint tasks, failed to reveal any alignment in the individual phase tasks. While this could simply be a theoretically real result, its scope must be limited, given that it is inconsistent with several other findings in the literature (as discussed in the previous section).

It is possible that the divergence between the results in the joint and individual phases of the experiment could be due to the differences in instructions and goals given to the participants. In the joint tasks, participants were told to try to coordinate as well as they could with their partner, were given an extra financial incentive, and were given score feedback at the end of each task. In the individual tasks, participants lacked these goals, motivation, and information. This difference could be responsible for the relatively rich results in the former tasks and the mostly null effects in the latter. However, even if this were the case, it could not explain the interesting differences between the conceptual and lexical alignment results, which cut across the two phases of the experiment. In both phases, there were greater differences and more significant effects in lexical alignment than in conceptual alignment, suggesting a dissociation between the two. Moreover, in the relevant past experiments also, participants had explicitly worked together on joint tasks and had not received any coordination instructions on the individual tasks (Markman and Makin 1998; Voiklis 2008).

Nevertheless, the experimental framework appears to be better suited to studying conceptual alignment during rather than beyond interaction. Fortunately, this is also where there is a greater gap in the literature. Several studies have addressed how conceptual convergence could result from linguistic interaction, by having participant pairs work together on a joint task and then carry out individual non-linguistic tasks such as sorting or similarity judgements (Markman and Makin 1998; Malt and Sloman 2004; Voiklis 2008). Thus, although my framework was designed to help separate the effects of dialogue from other interactional factors, the study of conceptual convergence at least potentially arising from linguistic interaction is not new. On the other hand, my framework offers a way to study online conceptual alignment during interaction, while simultaneously allowing for an independent examination of lexical alignment as well. While this does constrain the amount and type of interaction (compared to the range of joint tasks developed in other experiments), it does thus provide a solution to the difficult problem of getting at people's conceptualisations while they interact, independently of language. As such, these joint tasks are the strength of the experimental framework, and I will focus exclusively on them in Experiment 4, where I explore the potential role of category labels in particular in conceptual convergence.

7.6 Conclusion

Experiment 3 was designed to address the methodological shortcomings of Experiment 2, while also extending the scope of inquiry. As such, it asked not only whether interaction and dialogue result in conceptual convergence after interacting, but also whether there was any alignment during the interaction itself to begin with. The results were more extensive and satisfying than in Experiment 2. Conceptual alignment did occur, but was heavily restricted in scope: it only occurred when pairs could engage in free dialogue, and it did not persist beyond the interaction. Post-interaction similarity judgements also did not exhibit any convergence. On the other hand, lexical alignment occurred to varying extents depending on whether participants only received feedback on their partner's category groupings and labels after each task, and outlasted the interaction. The results suggest a potential reconciliation between two theoretical positions: lexical and conceptual alignment may be generally dissociated, but do come together during full, normal dialogue.

CHAPTER 8

Experiment 4: Category versus word feedback in conceptual and lexical alignment

8.1 Introduction

Experiment 3 demonstrated that dialogue is a very useful tool for conceptual coordination. However, it is important to realise that in Experiment 3, the linguistic resources available to participant pairs in the *talking* condition were enormous. Since there was no restriction on when and how much they could talk, people were free to negotiate not just the basis of categorisation, the category boundaries, and the category labels, but even "meta-negotiate" such aspects as the number of items per category and strategies they used after obtaining low scores.

On the other hand, Experiment 3 found no effect of feedback on conceptual convergence. Partners who had access to each other's words and categories did not manage to conceptually align significantly more than those who did not. However, the finding that dialogue substantially helps in this regard shows that it is certainly possible for people to converge under the right conditions, and that language can play an important role in this process. Therefore, the potential usefulness of words as concept labels warrants further investigation.

The importance of words can also be framed from a different perspective. Recall that, as I discussed in Chapter 2, words are associated with concepts, but not as tightly as is often assumed. Concepts, on the other hand, determine categories and are thus in a much tighter relationship. Is this reflected in how easy or difficult it is for people to get at each other's concepts, depending on the kind of information they have access to? Notice that this question can be readily studied using my experimental framework. In fact, Experiments 2 and 3 manipulated both label and grouping feedback, but they were confounded (compare the *control* and *silent* conditions).

Notice that this question is related to the internal structure of concepts. According to the model I developed in Chapter 2, words seem to be more associated with senses, while categories are more closely linked to exemplars (see Figure 2.9 in particular). We can now incorporate that structure into the model I have been working with for the interactive experiments, as shown in Figure 8.1.

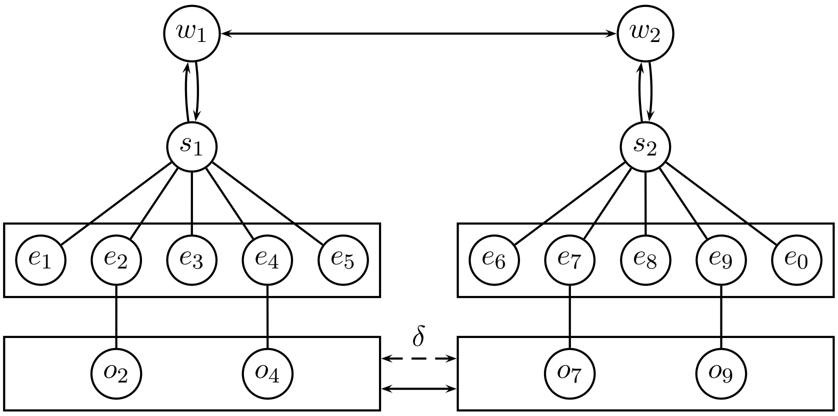


Figure 8.1: Interactive model which incorporates the internal conceptual structure, and how it is related to words and objects in the world.

Therefore, the question posed here can also shed new light on conceptual structure. In particular, if people rely more on words as concept identifiers, that would provide a new kind of support for prototype and theory theories, while if they do better with sets of categorised items, this would be evidence in favour of exemplar theory.

Furthermore, manipulating word and category feedback independently may be particularly fruitful for continuing the investigation of the relationship between conceptual and lexical alignment. If people rely primarily on one kind of information for the corresponding kind of alignment (and vice versa), that would provide further evidence for a dissociation between them when they are not talking. Otherwise, we would have novel support for an intimate connection between them.

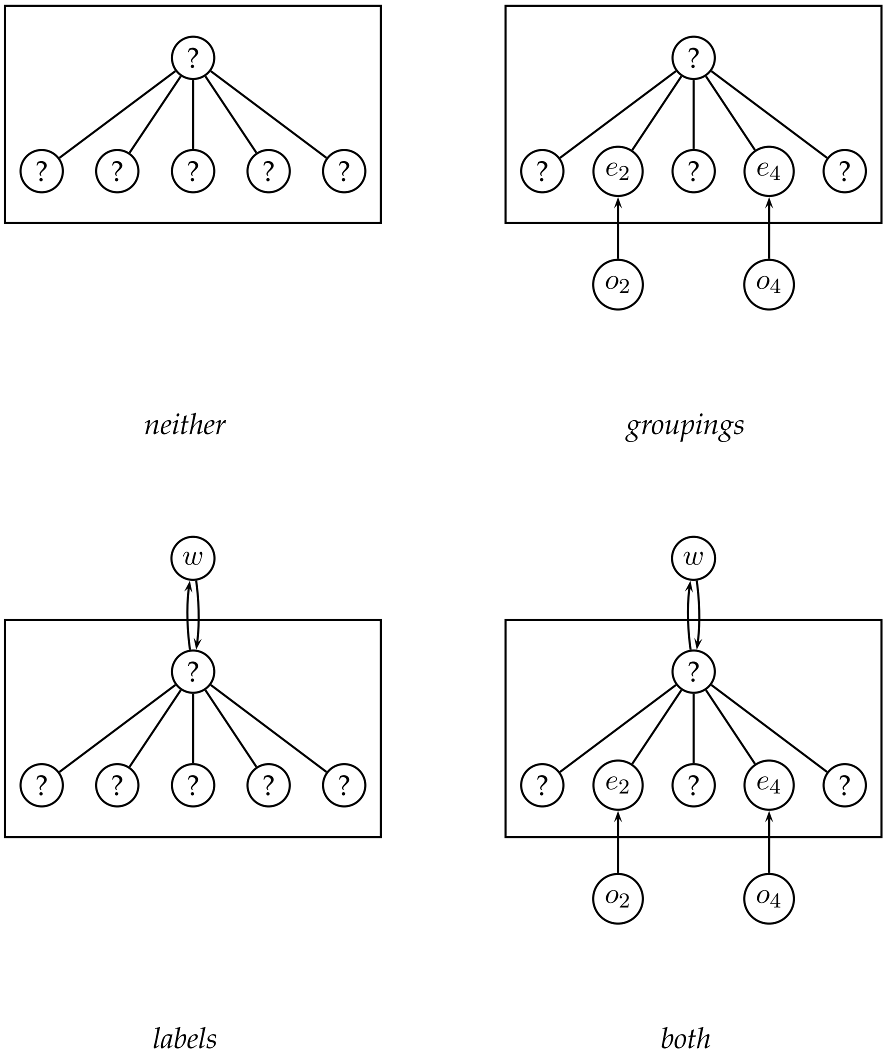
Therefore, in this chapter, I present an experiment which independently manipulates label and grouping feedback. The experiment should yield data which sheds light on the role of language in conceptual coordination, the relationship between lexical and conceptual alignment, and the internal structure of concepts.

8.2 Overview

Experiment 4 does not continue the investigation of dialogue, but instead looks more closely at the usefulness of feedback in alignment during interaction. As such, it effectively drops the *talking* condition from Experiment 3, while teasing apart confounded aspects of the *control* and *silent* conditions. To that end, the two types of feedback were separated, and the design was adjusted in a few ways to make it more sensitive.

The two kinds of feedback consisted of **label** and **grouping** feedback, which identify the kind of information that participants get about their partner's concepts at the end of each joint categorisation task. If there is label feedback, participants see their partners' category labels at the end of each task. If there is grouping feedback, then they see the groups of stimuli that their partner produced. I use these terms rather than "word" and "category" feedback to avoid confusion: the words are acting specifically as conceptual labels, and the groupings are only small subsets of the theoretically infinite categories. The experiment has a 2X2 between-subjects design, so that all four combinations of the two variables are investigated. Figure 8.2 shows the kind of feedback given in each condition and how they give participants different information regarding their partner's conceptualisations. As we can see, this design allows us to minimally pit the usefulness of lexical and conceptual information against each other, and to study how they interact. Moreover, it will also help shed light on the issues of internal conceptual structure, as raised in the previous section.

The data analysis will also place a greater emphasis on lexical alignment than was done in the previous experiments. In particular, since the two kinds of feedback correspond to the two kinds of alignment under investigation, the experiment offers an opportunity to look at the relationship between lexical and conceptual alignment more closely. Therefore, a good portion of the analysis will also consider how the lexical and conceptual alignment data match up to each other.



neither

groupings

labels

both

Figure 8.2: Feedback provided to participants on their partner's concept (represented by the box) in Experiment 4.

Recall that Experiment 3 did not find any evidence for conceptual convergence in the individual tasks following interaction, even in the *talking* condition. Moreover, Experiment 3 showed no impact of label feedback on conceptual alignment, even when accompanied with grouping feedback as well (i.e., there was no significant difference between the *control* and *silent* conditions). A larger set of joint tasks may be needed to give participants greater opportunity to learn from each other and conceptually converge. Therefore, Experiment 4 eliminated the post-interaction tasks, and increased the number of joint tasks.

Unlike the previous experiments, Experiment 4 adopted an artificial domain, consisting of pictures of triangle-like shapes. Such shapes offer several advantages. First, it is straightforward to get a fluid set of stimuli with many dimensions of variation. Rather than using a morphing program, I wrote a simple script which creates random triangles with a fluid range of different sizes, shapes, colours, angles, corner types, etc. Second, as these are artificial stimuli without obvious functional properties, there are probably less culturally entrenched conceptualisations of them than for plates, bowls, etc. This potentially allows more room for variation and negotiation. Third, they are perceptually simple, so that it is sufficient to see a relatively small version of the images on the screen. Fourth, using quite a different domain sheds light on the general scope of the empirical findings of this thesis.

A couple of further methodological changes were also introduced. First, the number of categories used by participants was fixed at two. Although this reduces some of the freedom in the classification tasks, it may be easier for participants to compare their categorisations with those of their partners. As we will see in Section 8.3.5, this also renders possible a more efficient program user interface. Moreover, it justifies the use of a simplified measure for comparing people's categorisations (Meila and Heckerman 2001; see Section 4.3.3). Second, two constraints were placed on possible category labels: they had to consist of a single word, and they had to be real English words (checked against an electronic dictionary). The restriction to a single word was incorporated since the focus of the experiment was on the effects of single words (rather than entire phrases). Words were restricted to existing English ones mainly to prevent participants from "cheating" (e.g., "bigtriangle", "followme") and from creating neologisms (e.g., "trianglish"). This also made it easier to focus on existing concepts and to compare different participants' labels.

Within this design, the main empirical hypotheses are:

1. **Do grouping and/or label feedback affect conceptual alignment?**
2. **Do grouping and/or label feedback affect lexical alignment?**
3. **Is there a correspondence between lexical and conceptual alignment?**

8.3 Methods

8.3.1 Participants

Participants were 80 adult native English speakers (age: $M = 21.9$, $SD = 2.9$), mostly undergraduate students recruited through the University of Edinburgh student employment website. There were 48 female participants, and 32 male participants.

Participants were assigned to pairings randomly, and pairings were assigned to conditions randomly, except that the numbers of different gender combinations was kept the same in each condition. In particular, there were four female-female pairs, four female-male pairs, and two male-male pairs assigned to each of the four conditions (described in Section 8.3.3). As in Experiment 3, these asymmetric numbers were chosen in accordance with the relative frequency of male and female participants available, because gender was not expected to make a difference in the experiment. Again, partners did not know each other before the experiment.

8.3.2 Stimuli

The stimuli were triangle-like shapes which varied along eight dimensions: colour (RGB), size, shape, orientation, corner size, and corner pointedness (to be defined below). As I will describe, the set of stimuli was chosen by first generating and piloting several different sets, and then divided up into particular tasks.

8.3.2.1 Generation

A perl script was written (included in Appendix E) which produced individual Postscript graphics files, one for each stimulus. Each stimulus was a dark geometric shape on a white background, and was obtained by randomly selecting values within certain criteria for the different parameters.

Size and shape were randomly generated according to the following algorithm. First, six random values were generated between 0 and 160 (since the images were 160 pixels in width). These defined the initial x and y coordinates of the triangle. If the area of the resulting triangle was less than 5/9 the area of the image (25,600 square pixels), then the triangle was discarded, and the algorithm started anew. Such a criterion was set so that excessively small stimuli would not be produced, although the particular threshold was arbitrary. If the triangle was big enough, then it was shifted so that its centroid (its middle) was at the centre of the image. In case this pushed one or more corners off the image, the triangle was again discarded. Notice that this can result in the rejection of triangles which are overly thin and long. If the triangle passed both of these criteria, it was kept, and then the other parameters were manipulated. Notice that this procedure automatically generates randomness in size, orientation and shape (within the constraints mentioned).

The other parameters were manipulated along different ranges of continuous values (i.e., subranges of $[0, 1]$), and different sets of ranges were defined to produce different sets of stimuli (to be tested in pilots). Colour was straightforward, being controlled through separate ranges for each RGB value. For instance, ranges of $[0, 1]$ for red, $[0, 0]$ for green, and $[0, 0.2]$ for blue would produce triangles with fully varying amounts of red, no green, and varying tinges of blue.

The corners of the triangles were also manipulated in such a way as to increase ambiguity concerning whether or not the shapes were triangles. The details for these manipulations were technical, and involved the determination of points used to smooth out each triangle corner with a Bézier curve using Postscript's 'curveto' operator. Intuitively, one parameter determines the "size" of the corners, which determines how far along a triangle's sides the corner should start rounding (a value of 0 would mean that the triangle was not rounded at all, while a value of 1 would mean that the sides of the triangle would be replaced by three Bézier curves between the midpoints of the sides). The other parameter could be thought of as determining the pointedness of the corners, as it determined the control points for the curves (a value of 0 would mean that the corners are chopped off with straight lines, while a value of 1 makes the curve go through the original corner).

Thus, all the sets varied freely (within the general constraints mentioned) in terms of the shapes, sizes, and orientations of the triangles, but they varied in terms of their manipulations of the colours and corners of the triangles. These then were tested in pilot experiments as well as informally in discussions with colleagues in order to get an idea of the relative psychological salience of different dimensions. The goal was to choose a stimulus set which exhibited a high degree of fluidity, as well as variation of roughly equal salience on several dimensions.

Based on the outcomes of these pilots, a final set of 330 stimuli was chosen for the actual experiment, from one particular parameter configuration. Since colour was an overly dominant dimension in the pilots, the RGB values for the shapes were set to vary only slightly along all three colour dimensions from solid black (along $[0, 0.2]$ for all three values). The corner size parameter range was set to $[0.1, 0.3]$, meaning that the corner rounding would begin 10-30% along the triangle's sides from the corner. The corner pointedness parameter range was set to $[0.7, 1]$, so that the corners either went through the original corners or nearly there. Figure 8.2 shows a random sample of 60 of the final stimuli.

8.3.2.2 Allotment

The stimuli were randomly split up into thirty sets of eleven shapes each, one set for each free classification task. An odd number of shapes was used to prevent participants from following a potential bias towards necessarily making their categories be the same size. Then five different random orderings of the thirty sets were defined. Each of these task orderings was assigned to two participant pairs of the same genders (female-female, female-male, or male-male) in each condition.

The order of the stimuli within a set was also randomised, but again under certain constraints. For each of the task orderings above, two stimulus orderings were defined for each task, and assigned randomly to the two participants in each pair. This was done so that (1) participants could not use the order of the items within a task to coordinate their categories, and (2) each participant would have a counterpart participant in a different pair in the same condition who would not only carry out the tasks in the same order, but would also see the items within tasks in the same order.

8.3.3 Conditions

The experiment has a 2x2 between-subject (or, more accurately, between-pair) design, as shown in Table 8.1. The two binary variables identify what kind of feedback participants

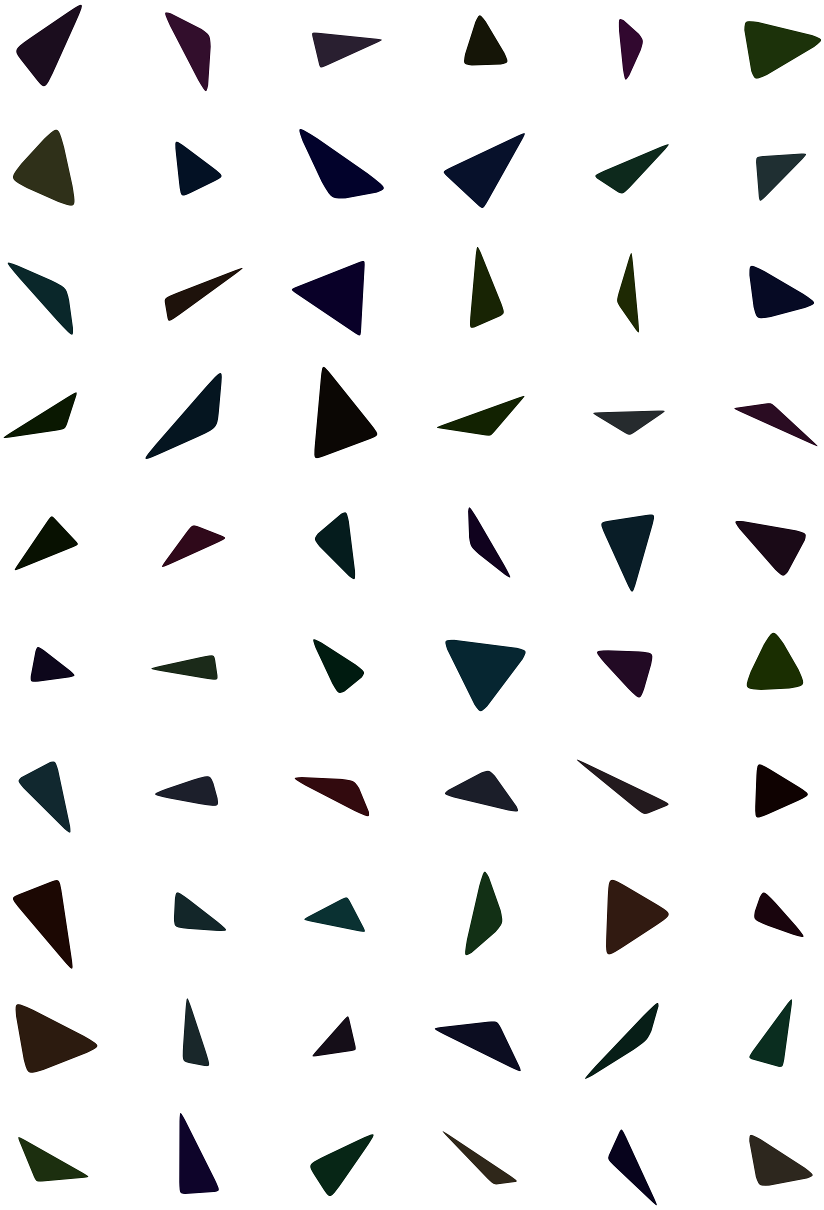


Figure 8.3: A random subset of the 330 stimuli used in Experiment 4.

receive of their partner's categories after each task. One variable refers to label feedback, and the other refers to grouping feedback. Crossing these variables gives four conditions: *labels* (only label feedback), *groupings* (only grouping feedback), *both* (both kinds of feedback), and *neither* (neither kind of feedback). Ten participant pairs were assigned to each of these conditions.

		Label feedback	
		NO	YES
Grouping feedback	NO	*neither*	*labels*
	YES	*groupings*	*both*

Table 8.1: Conditions (between-subjects) for Experiment 4: grouping feedback and labelling feedback.

8.3.4 Procedure

The experiment consisted of a series of joint categorisation tasks followed by a short questionnaire. Pairs of participants were brought into the experimental lab and assigned randomly to two separate computer cubicles.

After filling out a consent form, participants carried out one practice categorisation task, as in Experiments 2 and 3. After that they did thirty joint categorisation tasks, which followed the same procedure as in Experiment 3, with the following minor modifications. First, participants categorised the shapes into only two categories (which allowed for a simplified user interface, as explained in the next section). Second, the maximum number of items that participants could place in a category was increased from seven to eight (which reflects the slight increase in the number of stimuli per task). Third, category labels were required to consist of a single real English word, checked by the program against a dictionary file.

8.3.5 Program

Since the tasks involved only two categories and eleven stimuli each in this experiment, it was not a problem to have all of the stimuli appear relatively large on the screen. Therefore, the categorisation interface mechanism was simplified (for both the practice task and the main tasks). The focus box was removed, and participants categorised into either a left side category or a right side category. This was achieved by clicking on a stimulus with the appropriate mouse button: a left mouse button click categorised (or recategorised) an item as a member of the left category, and a right mouse button click placed it in the right category. Snapshots of the program at various stages during the tasks are shown in Appendix D.3.

After both partners were done the task, the program calculated the agreement score between the two sets of categories, and was shown in fraction form (e.g., “3/5”). In addition, depending on the experimental condition, the participants were also shown their partner's category groupings and/or labels underneath their own. The partner's categories were visually aligned to be in the order which gives the better category match (and hence also which reflects the score). The participants then had twenty seconds to see the score and study the category feedback (if any), and then the next task automatically began.

Once participants had carried out the thirty joint categorisation tasks, the final step in the program was to fill out a questionnaire. The questions were very similar to those used in Experiments 2 and 3. Participants were asked for demographic information like age, gender, and language background, as well as feedback questions concerning the experiment, such as what they thought the experiment was about or what strategies they used to align their categories.

8.4 Results

The results will be presented in the following order. First, I show a few examples of actual categorisation outputs from the experiment. After that, I analyse the data in three parts: conceptual convergence, lexical convergence, and the correspondence between them.

8.4.1 Example task output

Before analysing the data, I show three examples of the categorisation outputs of participant pairs. These are real examples from the experiment, and serve to illustrate how lexical and conceptual alignment can dissociate. Note that, as explained in the next section, the particular conceptual alignment measure used here is not the same as the one used in my first three experiments. To preview, scores take on fraction values out of 5 (i.e., 0, 0.2, 0.4, 0.6, 0.8, 1), with 1 meaning perfect agreement and 0 being the worst possible degree of agreement.

Figure 8.4 shows a case where a participant pair achieved a perfect score of 1. Both participants partitioned the set of items in exactly the same way, with four smaller stimuli in one category and the remaining seven larger ones in the other. Notice, however, that the participants used entirely different category labels for their categories. Participant 1 used “unequal” and “equal” (which seems to refer to whether the sides of the shapes are of roughly equal lengths). In contrast, Participant 2 used “small” and “large”. While these two pairs of labels concern different criteria, for this set of items and this pair of participants they resulted in the same groupings. This example thus demonstrates how participants can produce exactly the same categories despite using different labels.

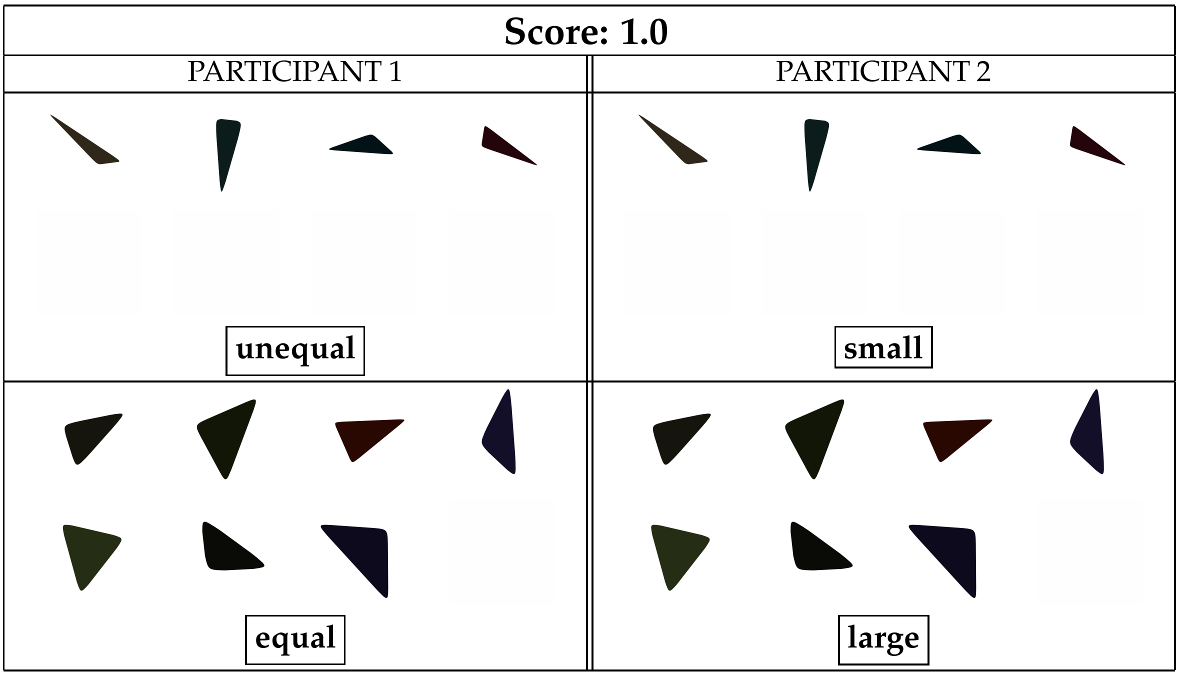


Figure 8.4: An example of a task for which the participant pair achieved a perfect score (1.0), since they have partitioned the stimuli into exactly the same two groups (even though they have used different labels).

Figure 8.5 shows a task for which participants achieved a score of 0.6. The partitionings are fairly similar, except that one of the stimuli that Participant 1 treated as ‘round’ was treated as ‘sharp’ by Participant 2, and vice versa for a second stimulus. Notice that, in contrast to the previous example, here participants failed to achieve a perfect score despite using the same labels.

Figure 8.6 shows an example of a task in which participants received a score of 0.0, which is the worst possible score. The partitionings are as different as possible, so that five of one participant’s items would have to be moved to their other category to make the partitionings identical. The labels are also different from each other.

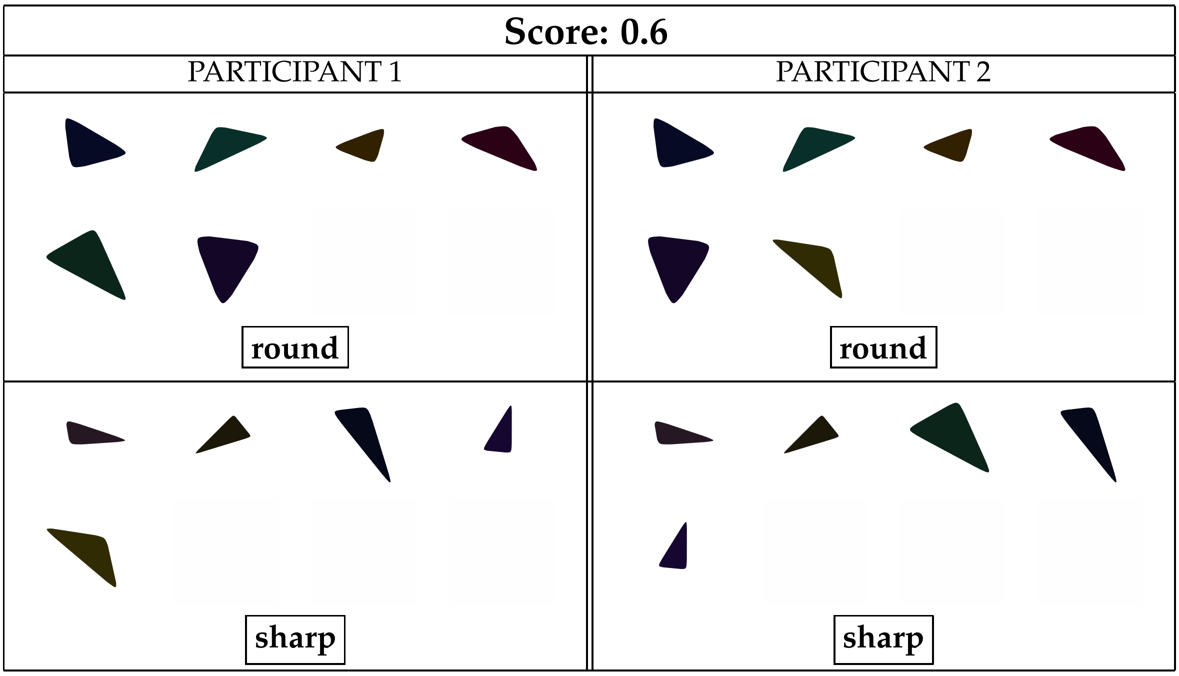


Figure 8.5: An example of a task for which the participant pair achieved a middle score (0.6). Notice that the participants' categories differed even though they used the same labels.

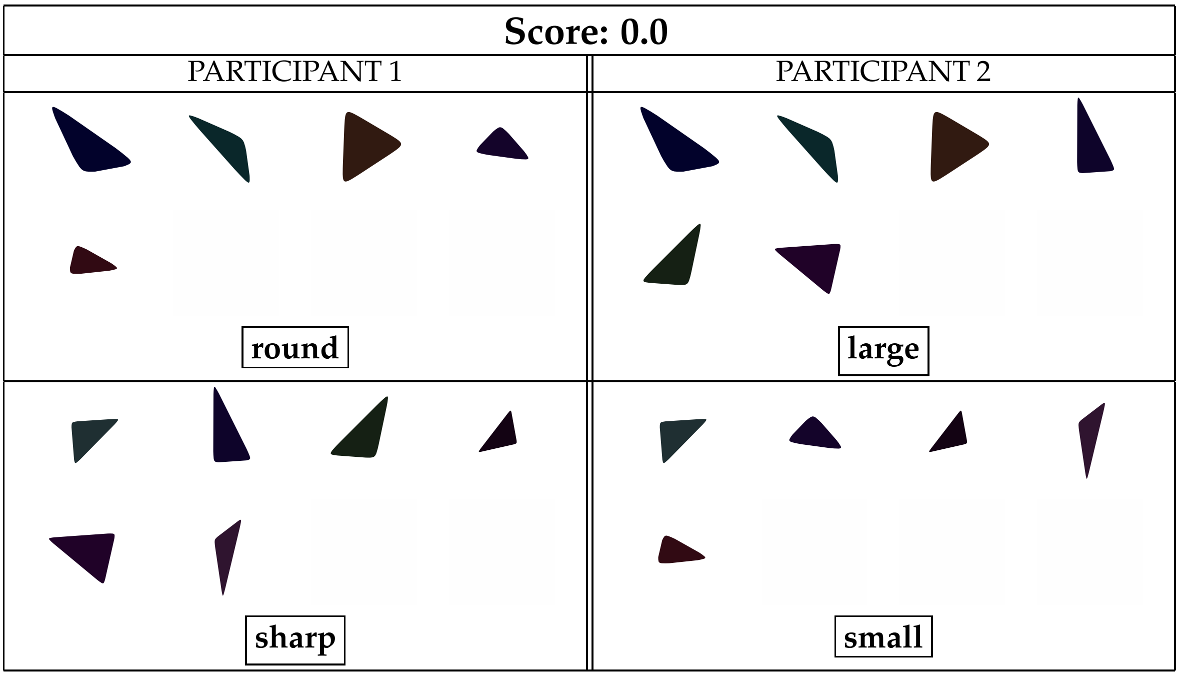


Figure 8.6: An example of a task for which the participant pair achieved the worst possible score (0.0). Five stimuli would have to be moved to make the partitionings the same.

8.4.2 Conceptual alignment

The main hypotheses, as in previous experiments, concern conceptual convergence. However, on this occasion, I use a different measure of conceptual alignment between participants. Since participants in this experiment sorted stimuli into only two categories, I adopt the score measure of Meila and Heckerman (2001) based on set-matching (see Section 4.3.3 for a description of this measure and why I adopt it for two-category experiments).

It can be easily verified that in the specific case of a set of eleven items being sorted into two categories, this measure reduces to:

$$score = 1 - \frac{m}{5} \tag{8.1}$$

where m is the number of stimuli that would need to be moved for one participant from one category to the other to make the two partitionings identical. Thus, identical partitionings yield a score of 1, whereas partitionings for which five stimuli would have to be moved (which is the worst possible case) give a score of 0. The other possible scores lie in between at 0.2, 0.4, 0.6, and 0.8, where 1, 2, 3 and 4 stimuli would need to be moved, respectively.

Figure 8.7 shows the raw convergence scores: for each participant pair, the thirty scores of all ten participant pairs are shown, along with the average over them for each task. Three features are noticeable from these plots. First, there is again quite a lot of variation within pairs, with scores fluctuating substantially. It is not uncommon, for example, for a pair to receive a high score on one task, and then a low score on the next one, or vice versa. Second, despite the fluctuations in individual pairs' scores, there does seem to be an increase in scores across tasks in most of the conditions, except perhaps the *groupings* condition. A longitudinal effect here would contrast with Experiment 3, where no such effect was found. Third, scores appear to start highest in the *groupings* condition, but to end highest in the *both* condition. However, these observations are obviously impressionistic, and need to be evaluated further.

Table 8.2 gives a compressed view of the data, showing the mean conceptual alignment scores for each task, by condition. A glance at these values too suggests an increase over time in all but the *groupings* condition, but highest final scores in the *both* condition.

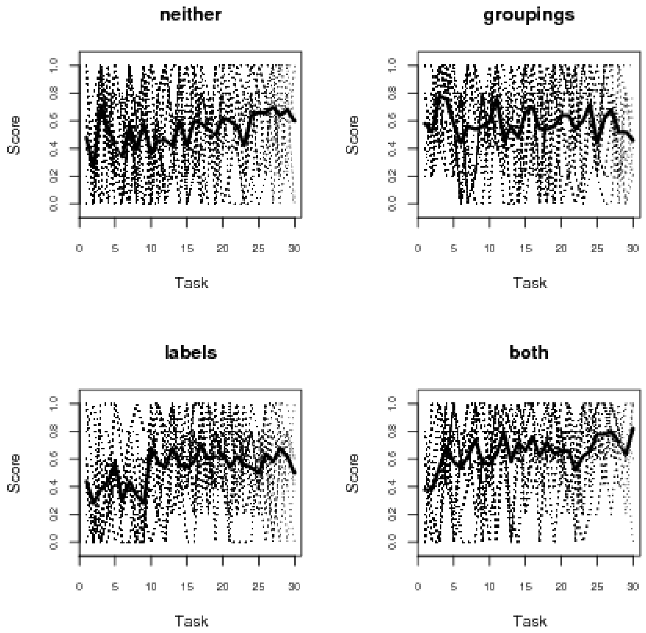


Figure 8.7: Scores over time per pair by condition. The dotted lines represent the scores of individual pairs. The bold lines represent averages across participant pairs.

In order to analyse the potential effects of grouping and label feedback across tasks, I used linear mixed effects models. First, as for Experiment 3, I applied a simple model with task as a fixed effect and score intercepts of participant pairs as random effects. The t-value for the fixed effect was 5.14, which easily surpasses Baayen's (2008) heuristic threshold of around 2. This confirms that participant pairs did in general coordinate with each other over time (in contrast to Experiment 3).

As a result, I continue with the mixed effects analysis. First, I compare a sequence of nested models, which all have task order as a fixed effect, but vary in their random effects term. The key resulting output of the model comparison (from R's 'anova' function) is shown in Table 8.3. I choose the best model based on the three criteria of AIC's, BIC's, and p-values. (AIC and BIC are measures of the goodness of fit of statistical models, which take into account the number of free parameters, and can be used to compare nested linear mixed models; see Pinheiro and Bates 2000). Based on these three criteria, the best model appears to be the one which allows for pair-specific variation in score intercepts and task slope effects, but without correlation between these two. This suggests that there is substantial variation between participant pairs, in both their starting scores and their improvement across tasks.

Tasks	*neither*	*labels*	*groupings*	*both*
1	0.48 (0.34)	0.44 (0.35)	0.58 (0.29)	0.38 (0.24)
2	0.26 (0.28)	0.28 (0.40)	0.52 (0.29)	0.38 (0.35)
3	0.70 (0.40)	0.38 (0.39)	0.78 (0.27)	0.54 (0.39)
4	0.54 (0.30)	0.44 (0.31)	0.76 (0.32)	0.68 (0.21)
5	0.44 (0.32)	0.56 (0.31)	0.58 (0.32)	0.58 (0.43)
6	0.34 (0.31)	0.30 (0.25)	0.44 (0.36)	0.54 (0.25)
7	0.56 (0.37)	0.44 (0.31)	0.56 (0.42)	0.62 (0.30)
8	0.40 (0.28)	0.34 (0.37)	0.54 (0.37)	0.74 (0.37)
9	0.58 (0.39)	0.28 (0.30)	0.56 (0.30)	0.58 (0.33)
10	0.36 (0.40)	0.68 (0.32)	0.60 (0.28)	0.56 (0.32)
11	0.46 (0.31)	0.58 (0.35)	0.76 (0.28)	0.64 (0.37)
12	0.46 (0.41)	0.54 (0.31)	0.46 (0.30)	0.80 (0.13)
13	0.42 (0.26)	0.68 (0.30)	0.56 (0.36)	0.58 (0.37)
14	0.60 (0.37)	0.56 (0.23)	0.48 (0.37)	0.72 (0.29)
15	0.42 (0.38)	0.54 (0.28)	0.66 (0.35)	0.66 (0.30)
16	0.60 (0.37)	0.58 (0.36)	0.70 (0.29)	0.76 (0.25)
17	0.58 (0.37)	0.70 (0.27)	0.54 (0.40)	0.62 (0.27)
18	0.54 (0.27)	0.60 (0.31)	0.54 (0.37)	0.72 (0.25)
19	0.50 (0.36)	0.62 (0.30)	0.56 (0.43)	0.64 (0.36)
20	0.62 (0.32)	0.64 (0.23)	0.64 (0.34)	0.66 (0.23)
21	0.60 (0.34)	0.54 (0.40)	0.64 (0.35)	0.66 (0.30)
22	0.56 (0.32)	0.62 (0.38)	0.54 (0.35)	0.52 (0.27)
23	0.42 (0.38)	0.56 (0.30)	0.60 (0.44)	0.62 (0.42)
24	0.64 (0.36)	0.54 (0.31)	0.72 (0.23)	0.66 (0.30)
25	0.66 (0.38)	0.50 (0.19)	0.44 (0.28)	0.78 (0.26)
26	0.66 (0.28)	0.64 (0.34)	0.62 (0.37)	0.78 (0.22)
27	0.70 (0.34)	0.58 (0.39)	0.68 (0.34)	0.80 (0.25)
28	0.64 (0.39)	0.68 (0.29)	0.52 (0.38)	0.72 (0.19)
29	0.68 (0.32)	0.62 (0.29)	0.52 (0.37)	0.64 (0.25)
30	0.60 (0.44)	0.50 (0.25)	0.46 (0.27)	0.82 (0.32)

Table 8.2: Mean scores across all participants within a condition for each task. Standard deviations are given in brackets.

Model	AIC	BIC	p
Intercept only	700.08	720.44	
Slope only	694.87	715.23	$p < .001$
Intercept + Slope	691.18	716.63	$p < .05$
Intercept * Slope	690.04	720.58	$p > .05$

Table 8.3: Comparisons between several nested linear mixed models, with a fixed effect term of task order, used to determine the random effects term. The difference between the last two models (marked by '+' and '*') is that the former does not model correlation between the pair-specific intercept and slope, while the latter does. The combination of AIC, BIC and p-value criteria suggests that we should adopt the Intercept + Slope model (shown in bold) at this stage.

Next, I consider the possible effects of grouping and label feedback, one at a time. First, I check whether there seems to be a fixed effect of grouping feedback, by fitting a simple model with it as the only fixed term, but keeping the random effects just identified. The t-value for the fixed effect was 3.81, which again is above Baayen's (2008) threshold. On the other hand, an analogous model for label feedback did not satisfy the criterion, with a t-value of only -0.13.

Stepwise model comparisons support these conclusions. I first added a fixed effect of groupings to the previous model with tasks as a fixed effect, and another model adding also a task-groupings interaction. The comparison, shown in Table 8.4, suggests we keep the groupings factor. However, although the AIC decreases for the model with the interaction, I do not keep it, because the comparison does not yield a significant p-value and the BIC increases. Table 8.5 shows the subsequent model comparison involving label feedback. In this case, it is clear that label feedback is not improving the model, as the more complex model fails on all three criteria. Therefore, a fixed effect for label feedback is not kept.

Model	AIC	BIC	p
Task only	691.18	716.63	
Task + groupings	681.29	711.84	$p < .001$
Task * groupings	679.91	715.54	$p > .05$

Table 8.4: Stepwise comparison between mixed models testing for the effects of grouping feedback. The sequence incrementally adds a grouping feedback term and then a task-groupings interaction term to the 'Intercept + Slope' model adopted in Table 8.3. Together, the AIC, BIC and p-values suggest that we keep the fixed effect term for groupings, but not the interaction.

Model	AIC	BIC	p
Task + groupings	681.29	711.84	
Task + groupings + labels	683.36	718.99	$p > .05$

Table 8.5: Comparison between the linear mixed model adopted in Table 8.4 and one which also adds a fixed effect term for label feedback. As the output shows, label feedback is clearly not making a significant contribution, and therefore should not be kept (as a result, I did not even test models with interactions involving label feedback).

The final mixed effect model[1], then, has fixed effects for task and grouping feedback, but no label feedback and no interaction terms (although the task-grouping interaction was noteworthy). This suggests that pairs who received groupings feedback were more

[1] The R lmer call: lmer(score ˜ task + groupings + ((0 + task) |session) + (1 |session))

conceptually aligned, and that, regardless of condition, pairs converged conceptually over time.

Note that this analysis is consistent in some but not all ways with my impressionistic observations earlier. The adopted model is consistent with the claim that pairs increased their scores over time, and that pairs in the *grouping* and *both* conditions (i.e., the conditions with grouping feedback) achieved relatively high scores. However, the analysis did not yield any evidence for an effect of labels, or for differential improvement over time between the conditions. This seems surprising, since as we saw in Table 8.2, the *both* condition appears to be achieving higher scores than the other conditions by the final tasks.

In order to look into this further, I turn to a simpler analysis, in which I collapse across tasks and consider mean scores. This has the drawback that it obviously cuts away the sequential information and dependency of the scores. However, performing the analyses based on the means also has clear statistical advantages, in that we can obtain more normally distributed data, and we can use simpler and more powerful analytic methods.

I begin with an analysis based on the mean scores across all tasks for each participant pair, by condition. The means of those means (across participant pairs) are shown in Table 8.6. A 2x2 ANOVA (grouping feedback and label feedback) was performed on the mean scores. The results were consistent with the linear mixed analysis earlier: there was a significant main effect of grouping feedback ($F(1,36) = 6.47$, $p < .05$), but no effects of label feedback ($F(1,36) = 0.83$, $p > .05$) or of the grouping-label interaction ($F(1,36) = 0.94$, $p > .05$).

		Label feedback	
		NO	YES
Grouping feedback	NO	0.53 (0.35)	0.53 (0.32)
	YES	0.59 (0.34)	0.65 (0.31)

Table 8.6: Mean scores across all participants within a condition across all the tasks. Standard deviations are given in brackets.

However, given the significant main effect of task reported earlier, averaging over the whole set of tasks is not the most relevant analysis. In comparing the conditions in terms of conceptual convergence, we should inquire rather into how well pairs are aligned in the latter stages of their interaction. Ideally, we would just look at their scores on the very final task. However, given the large fluctuations in scores, I use instead the means across each pair's final five tasks. Table 8.7 shows the resulting means by condition on the final five tasks. Note that although the choice of five is rather arbitrary, the same significance patterns reported below were found for each of the analogous analyses based on the means over the last seven tasks or less (including just the final one).

		Label feedback	
Grouping feedback		NO	YES
	NO	0.66 (0.35)	0.60 (0.31)
	YES	0.56 (0.34)	0.75 (0.25)

Table 8.7: Mean scores across all participants within a condition across the final five tasks. Standard deviations are given in brackets.

Another 2x2 ANOVA was then conducted, and yielded different results from previous analyses. This time, there were no significant main effects of either groupings ($F(1,36) = 0.20$, $p > .05$) or labels ($F(1,36) = 1.47$, $p > .05$), but the interaction ($F(1,36) = 4.47$, $p < .05$) was significant. The interaction appears to be of a cross-over type, as can be seen from the plot in Figure 8.8. Therefore, I followed this up with simple effects tests, assessing the effect of each factor for both levels of the other factor. And indeed, the effects of groupings were not significant when there was no label feedback ($F(1,36) = 1.38$, $p > .05$), but were nearly significant when there was label feedback ($F(1,36) = 3.29$, $p = .08$); and, similarly, the effects of labels were not significant in the absence of grouping feedback, ($F(1,36) = 0.41$, $p > .05$), but were significant when there was grouping feedback ($F(1,36) = 5.54$, $p < .05$). This is consistent with the fact that the mean of the *both* condition is relatively distinct from the other three. In order to conduct one final check of this, a t-test was conducted between the mean scores in the *both* condition ($M = 0.75$, $SD = 0.11$) and the means in the other three conditions combined ($M = 0.61$, $SD = 0.20$). The difference was highly significant ($t(28) = 2.88$, $p < .01$), confirming that the *both* condition does indeed have higher scores by the end of the experiment.

In order to help interpret these differences between conditions in the final tasks, it's useful to also make a simple comparison between the conditions on the first few tasks. To that end, I conducted an ANOVA based on the mean scores across the first five tasks of the experiment (analogous to the one above based on the last five tasks).

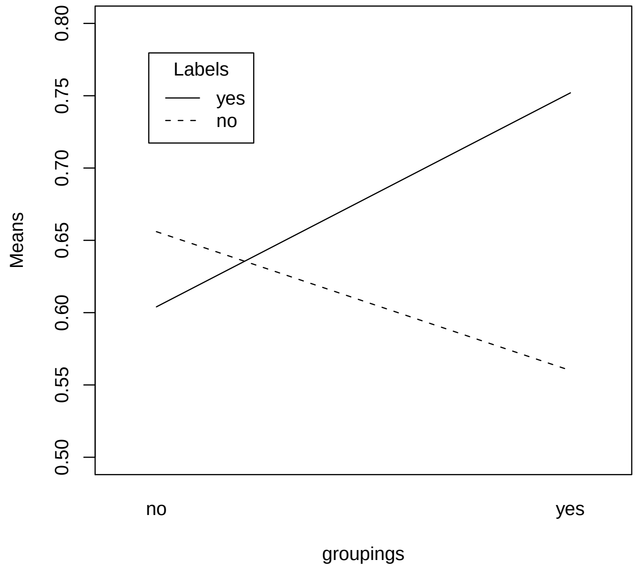


Figure 8.8: Interaction plot for mean scores over the final five tasks in the different conditions.

Table 8.8 shows the means by condition on the first five tasks. A 2x2 ANOVA (analogous to the one carried out on the last five tasks) was then conducted, with interesting results. There was a significant effect of groupings ($F(1,36) = 6.47, p < .05$), a nearly significant (but negative) effect of labels ($F(1,36) = 0.47$, $p = .06$), and no effect of the interaction ($F(1,36) = 0.47$, $p > .05$). In other words, at the outset of the experiment, getting grouping feedback was beneficial, while getting label feedback seems to actually have been counterproductive.

Tasks	*neither*	*labels*	*groupings*	*both*
1-5	0.48 (0.35)	0.42 (0.35)	0.64 (0.31)	0.51 (0.34)

Table 8.8: Mean scores across all participants within a condition across the first five tasks. Standard deviations are given in brackets.

Finally, we can get a simple idea of whether scores increased in the individual conditions between the beginning and end of the experiment by using t-tests comparing initial and final scores. In particular, for each condition, I conducted a paired t-test between the mean scores across the first five and last five tasks, respectively (therefore, these tests were based on the same means as the two ANOVAs above). The results were highly significant for the *both* condition ($t(9) = 3.57, p < .01$), significant for the *neither* condition ($t(9) = 2.91, p < .05$), and nearly significant for the *labels* condition ($t(9) = 2.18, p = .06$), but not significant for the *groupings* condition ($t(9) = 1.16$, $p > .05$). In other words, broadly consistent with the impressionistic observations earlier, pairs in the *groupings* condition did not improve over the course of the experiment, while they did improve in the other conditions, especially if they received both kinds of feedback (i.e., in the *both* condition).

8.4.3 Lexical alignment

Just as we examined how conceptual convergence is dependent on the different kinds of feedback, we can now turn to the same questions with regards to lexical convergence. However, I again modify the measure used for calculating lexical alignment relative to the previous experiments. In particular, this time I treat agreement as binary: participants are said to agree on category labels for a particular task if they use exactly the same two labels for their two categories; otherwise, they are said to disagree. Numerically, I assign agreement a score of 1, and disagreement a score of 0, which makes it possible to take meaningful averages across a number of tasks. Given such a binary score, in what follows I will sometimes use the terms **agreement** and **disagreement** when referring to lexical alignment scores of 1 and 0, respectively.

Note that it could be argued that treating all possible discrepancies, large or small, between two sets of labels equivalently is misleading and unnecessary. However, there are good reasons for adopting this simplification. With only two categories labelled with single words checked against a dictionary, it should be relatively easy for participants to adopt identical label sets (compared to the previous experiments). Although it would be possible to give an intermediate score if one label matched or to use tools like latent semantic analysis to numerically compare non-identical terms, such methods may not be meaningful in the context of a sorting task with a very restricted stimulus domain. Moreover, a binary measure greatly simplifies analysis.

Although it's not visually clarifying to show individual binary agreement scores graphically (as I did for conceptual alignment), we can look at the number of participant pairs (out of 10) in each condition which showed lexical agreement. This is shown for all four conditions in Figure 8.9. Notice that the two conditions with label feedback have many more agreeing pairs than those without label feedback, while there seems to be little or no difference made by grouping feedback. Thus, not surprisingly, label feedback seems to enhance label agreement, while grouping feedback does not.

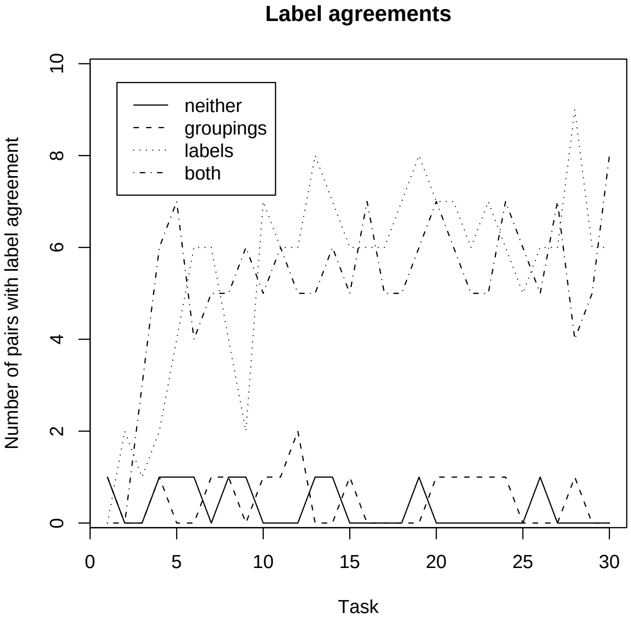


Figure 8.9: Label agreement over time, for each condition. The plots show the number of participant pairs with label agreement for each task.

In order to confirm these observations, I applied logistic linear mixed models. This is analogous to the analysis conducted in the previous section for conceptual convergence, and was also done using R's 'lmer' and 'anova' functions, but this time the output variable is binary rather than numerical. I begin again by checking for a main fixed effect of task, with intercepts for participant pairs as random effects. The t-value for the fixed effect was 5.63, which is much beyond Baayen's (2008) heuristic criterion of 2. Thus, participant pairs did generally coordinate their category labels over time.

Consequently, continuing with the mixed model analysis, I compare several nested models with task order as a fixed effect, in order to determine the random effects term. The key output from the model comparison can be seen in Table 8.9. The highly significant p-values and the decreasing AICs and BICs suggest that we should keep the most complex random effects, with correlated pair-specific intercept and slope terms.

Now I consider the effects of label and grouping feedback, one at a time. First, I fit a simple model with label feedback as the only fixed term, keeping the random effects I have just identified. The t-value for the fixed effect was 5.24, easily surpassing Baayen's

Model	AIC	BIC	p
Intercept only	720.80	741.16	
Slope only	708.21	728.57	$p < .001$
Intercept + Slope	627.76	653.21	$p < .001$
Intercept * Slope	619.61	650.15	$p < .01$

Table 8.9: Comparisons between several nested linear mixed models, with a fixed effect term of task order, used to determine the random effects term. The difference between the last two models (marked by '+' and '*') is that the former does not model correlation between the pair-specific intercept and slope, while the latter does. The AIC, BIC and p-value criteria all suggest that we should adopt the most complex model (shown in bold) at this stage.

(2008) threshold. In contrast, the analogous model for grouping feedback did not, with a t-value of only 0.53. This suggests the opposite pattern of results from those for conceptual convergence.

I next conducted further stepwise model comparisons to verify these conclusions. I first took the last model with the fixed effect of task, and added first a fixed effect of labels, and then a task-label interaction. The comparison criteria, based on the values in Table 8.10, strongly suggest to keep both. On the other hand, if we then try to add a fixed effect of groupings, as shown in Table 8.11, there is clearly no improvement.

Model	AIC	BIC	p
Task only	621.61	657.24	
Task + labels	605.73	641.37	$p < .001$
Task * labels	592.16	632.88	$p < .001$

Table 8.10: Stepwise comparison between mixed models testing for the effects of label feedback. The sequence incrementally adds a label feedback term and then a task-label interaction term to the 'Intercept * Slope' model adopted in Table 8.9. The AIC, BIC and p-values all suggest that we keep both the main effect and the interaction.

Model	AIC	BIC	p
Task * labels	592.16	632.88	
Task * labels + groupings	594.04	639.85	$p > .05$

Table 8.11: Comparison between the linear mixed model adopted in Table 8.10 and one which also adds a fixed effect term for grouping feedback. As the output shows, grouping feedback is clearly not making a significant contribution, and therefore should not be kept (as a result, I did not even test models with interactions involving grouping feedback).

Thus, the final mixed effect model[2] has fixed effects for task, labels, and their interaction, but no effect of groupings. This is, in a sense, nearly a mirror image of the model

[2] The R lmer call: lmer(sameLabels ˜ task * labels + (task |session))

adopted earlier for conceptual convergence, where groupings had an effect but not labels. However, in this case, the effects were stronger and the decisions of which models to keep easier, as the different criteria were in agreement. Overall, this model suggests that pairs who received label feedback aligned more lexically, and that pairs converged over time, although the extent to which they did so depended on whether they got label feedback.

I turn next to an analysis that collapses across the task variable. First, I calculate the mean lexical alignment scores for each pair across all tasks. The means of these means (across participant pairs) are shown in Table 8.12. A 2x2 ANOVA (grouping feedback and label feedback) yielded results in line with the earlier linear mixed model analysis: there was a significant main effect of label feedback ($F(1, 36) = 43.20$, $p < .001$), but no effects of grouping feedback ($F(1, 36) = 0.01$, $p > .05$) or of the grouping-label interaction ($F(1, 36) = 0.10$, $p > .05$).

		Label feedback	
		NO	YES
Grouping feedback	NO	0.03 (0.18)	0.55 (0.50)
	YES	0.05 (0.22)	0.52 (0.50)

Table 8.12: Mean lexical alignment scores across all participants within a condition across all the tasks. Standard deviations are given in brackets.

Recall that in the conceptual convergence section, I found that an ANOVA gives quite a different result if focused on just the final five tasks, and that these are the most relevant for a convergence hypothesis. Therefore, I also conduct an analogous second ANOVA here, focusing again on the final five tasks (Table 8.13 shows the means by condition). The results were just the same as when considering the whole range of all 30 tasks: a significant effect of label feedback ($F(1, 36) = 42.19$, $p < .001$), but none for either grouping feedback ($F(1, 36) = 0.19$, $p > .05$) or the interaction $F(1, 36) = 0.19$, $p > .05$). Thus, in contrast to the counterpart pattern for conceptual convergence, grouping feedback appears to be irrelevant for lexical convergence.

8.4.4 Correspondence between conceptual and lexical alignment

In the previous two sections, I have looked at the relationship between two kinds of alignment in the categorisation tasks with two kinds of feedback. For the most part,

		Label feedback	
Grouping feedback		NO	YES
	NO	0.02 (0.14)	0.66 (0.48)
	YES	0.02 (0.14)	0.58 (0.50)

Table 8.13: Mean lexical alignment scores across all participants within a condition across the final five tasks. Standard deviations are given in brackets.

the results showed a dissociation, so that lexical convergence depended on lexical but not conceptual feedback, and vice versa (although the effects were much stronger in the lexical case). Does this mean, however, that the two kinds of alignment bear no relationship to each other? In this section, I address this question, by evaluating potential correspondences and relationships between the two.

The analyses below are restricted to the participant pairs who received label feedback (i.e., those in the *label* or *both* conditions). I proceed in this way because, as we saw in Figure 8.9, participants almost never exhibited lexical agreement when they didn't get label feedback, which skews the overall distribution of lexical alignment scores enormously. On the other hand, when there was label feedback, lexical agreement happened roughly half of the time, which keeps things relatively balanced for statistical analysis. Moreover, as I noted, participants also seemed to treat labelling more seriously when they got label feedback. Finally, focusing on label feedback is more relevant to the general question of language's role in conceptual coordination.

I begin by asking whether there was a general correspondence between lexical agreement and high levels of conceptual alignment. Figure 8.10 shows the number of tasks with label agreement and disagreement corresponding to each conceptual alignment score. The ratio of lexical agreements clearly increases with increasing conceptual alignment scores, and this was confirmed by a linear regression analysis ($\beta = 1.74$, $p < .001$, $R^2 = 0.99$).

I look next at two different analyses based on participant pairs. The first analysis asks whether pairs obtain higher scores when they achieve label agreement. To that end, for each participant pair, I calculate the average conceptual alignment score for tasks where they agreed on their labels, and a second average for tasks where they did not. The averages were higher for label agreement ($M = 0.69$, $SD = 0.13$) than for label

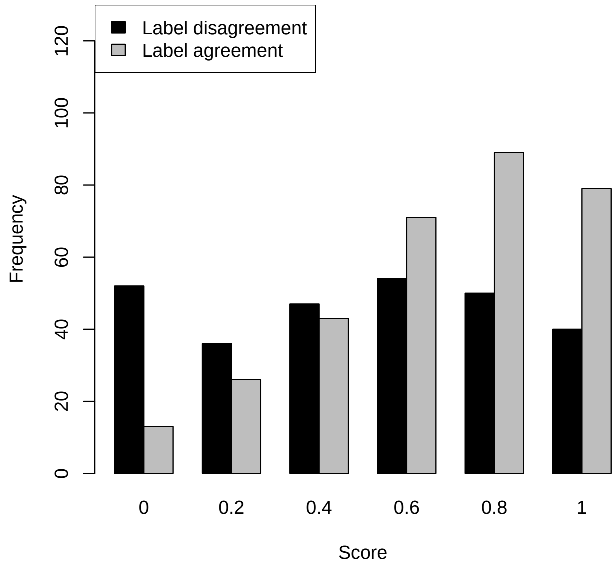


Figure 8.10: Numbers of label agreements and disagreements for each conceptual alignment score.

disagreement ($M = 0.46$, $SD = 0.16$), and a paired t-test showed that the differences were highly significant ($t(19) = 8.22, p < .001$).

We can also ask whether participant pairs who achieved lexical agreement more often also tended to have higher conceptual alignment scores. In order to test this, I took the mean lexical alignment scores and mean conceptual alignment scores for each pair, and conducted a linear regression between them (modelling the conceptual in terms of the lexical). The resulting data are plotted in Figure 8.11, along with the regression line. First note how, interestingly, there seem to be two distinct clusters of mean lexical scores, but there is no obvious difference between their conceptual alignment scores. In any case, the analysis revealed no significant relationship between them ($\beta = 0.09$, $p > .05$, $R^2 = 0.06$).

Finally, I consider the relationship between stability in a pair's set of labels and their average degree of conceptual alignment. On one hand, it's possible that pairs who settle on labels early and stick to them may achieve higher scores over the tasks. On the other hand, it may be that greater flexibility is more effective, so that participant pairs who adapt their labels and categories to the current task stimuli are rewarded. Figure 8.12 investigates this possibility, plotting participant pairs' mean scores by the total number

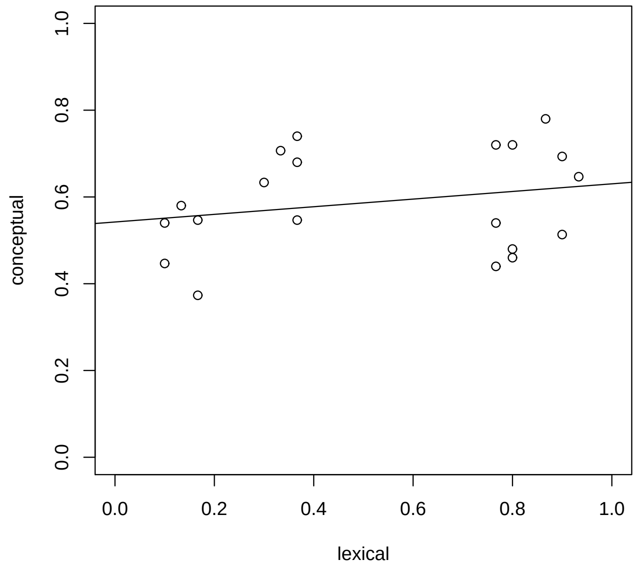


Figure 8.11: Mean lexical alignment scores plotted against mean conceptual alignment scores for each participant pair, along with a regression line.

of distinct labels they used between them across all of their tasks. Although there does appear to be a slight decrease in scores with increasing variability in labels, a regression analysis reveals no significant effect ($t(18) = -0.37$, $p > .05$).

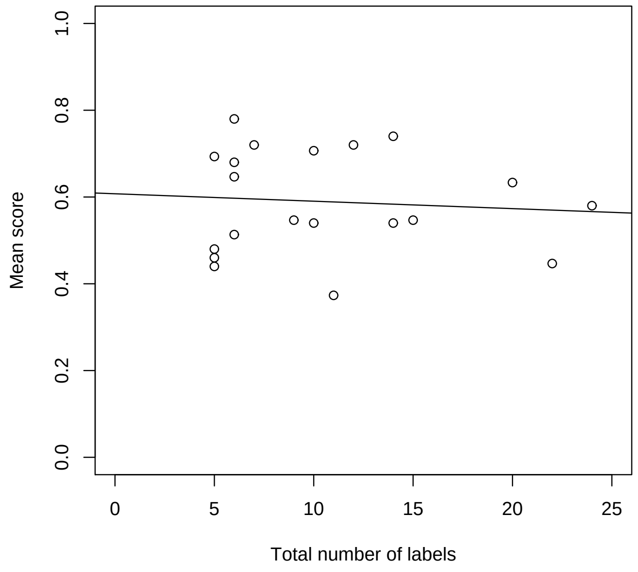


Figure 8.12: Mean scores of participant pairs who received label feedback against the total number of distinct labels they used between them over the course of all 30 tasks.

8.4.5 Questionnaire data

Participants were asked the same feedback questions as in Experiment 3, except that there was no question about how they judged similarity (since there was no similarity judgement task), and they were not asked to rate the familiarity and naturalness of the stimuli (since I was now using artificial stimuli). Since Experiment 3 had already run smoothly, and Experiment 4 did as well and had the same kind of joint categorisation tasks, I do not present the questionnaire data here in any detail.

However, given the different stimulus domain and the fact that there was evidence of learning across the tasks, it is worth mentioning the kinds of answers that people gave concerning the basis of categorisation (#2) and whether they changed it (#3) (the feedback question numbers marked with '#' refer to Appendix D.2). Not surprisingly, unlike Experiments 2 and 3, participants never mentioned functional properties of the stimuli. Most (73 out of 80) participants referred to a variety of physical dimensions, including shape, size, boundedness, pointedness, direction, angles, width, boundedness, corner types, and thickness. Three participants referred to the requirement of having to split the items into two categories. Two participants mentioned trying to agree with their partner, and one mentioned the task scores as having an effect. When asked whether they changed their basis for categorisation, 50 said that they did, 12 that they did not, 9 that they did originally but then stabilised, and the remaining 9 gave more ambivalent responses (e.g., "sometimes").

8.5 Discussion

8.5.1 Summary of results

Experiment 4 asked what role words play in conceptual alignment. The design pitted label feedback versus grouping feedback, and was meant to assess the relative importance of each. I also asked how these two forms of feedback affected lexical convergence, and how and to what extent the two types of convergence corresponded to each other.

The analyses yielded a variety of significant findings. First of all, unlike the previous experiments, participant pairs did generally converge over time, not only lexically, but also conceptually. Secondly, there was evidence for a dissociation between lexical and conceptual alignment. In general, conceptual alignment depended on conceptual feedback, and lexical alignment depended on lexical feedback. However, there was also a difference between the two cases. While for lexical alignment, the effect of lexical feedback was enormous and absent for grouping feedback, the exclusivity was weaker for conceptual alignment. Indeed, there was some evidence that by the end of the experiment, lexical feedback also significantly improved conceptual alignment, as long as it was provided in conjunction with grouping feedback. Thirdly, despite the general dissociation between the two levels of alignment, they did also exhibit some correspondence to each other. Tasks with higher conceptual alignment scores were also more likely to show lexical agreement, and participant pairs tended to have higher conceptual alignment scores on tasks where they agreed lexically. On the other hand, the correspondence also had limits: pairs who aligned lexically more often did not score higher conceptually, and relative stability in lexical choices also did not correspond to higher conceptual alignment scores.

8.5.2 *Conceptual and lexical alignment*

The results of this experiment shed light on the role of words in conceptual coordination, as well as the relationship between conceptual and lexical alignment more generally. I address these issues in this section. I first discuss how the results for lexical and conceptual alignment mostly mirror each other and suggest a dissociation between them. I then discuss the results that do not fit this pattern, and how they change the picture. Finally, I relate the findings here to previous work. I cover this separately and last because the first parts of the discussion are somewhat technical, and relating the finer details of my experiment to previous studies is less fruitful.

8.5.2.1 *A dissociation*

The main conceptual and lexical alignment results seem to be straightforward mirror images of each other. Both the linear mixed analyses and most of the ANOVAs collapsing across tasks suggest that conceptual alignment depends on grouping feedback, and lexical alignment depends on label feedback. When people were able to see each other's groupings after each task, they were better able to align them over time. And similarly, when they saw their partner's labels each time, they tended to converge with each other's labels. Each type of feedback thus seemed to have a direct influence only on the corresponding type of alignment.

Moreover, some of the analyses which more directly related participants' conceptual and lexical output also supported this separation. In particular, the wide variation found between pairs in how often they agreed on labels was not reflected in their average conceptual scores. In other words, pairs who rarely used the same labels coordinated their categorisation just as well, on average, as those who usually agreed on their labels. Moreover, stabilising on a set of labels, which might be interpreted as stabilising on conceptual pacts (Brennan and Clark 1996), did not correspond to better conceptual alignment scores either. Together, these results suggest a dissociation between lexical and conceptual alignment.

8.5.2.2 *Not entirely dissociated*

However, as I hinted at before, there were also other results which do not fit a simple explanation of a clean dissociation between lexical and conceptual alignment. In particular, the conceptual alignment results captured by the condition-specific paired t-tests, and the ANOVAs based on the first or last few tasks, are at odds with this explanation. These results are important, and I discuss them now in more detail, and assess how they fit with the other results.

The t-tests revealed differences between the conditions regarding whether or not pairs increased their conceptual alignment over time. The results showed that scores increased significantly (or nearly significantly) in the *both*, *neither* and *labels* conditions, while in the *groupings* condition there was no sign of improvement at all. But if conceptual alignment is based primarily on grouping feedback, then why would such feedback not lead to increasing scores?

The simplest explanation is that this is due to the fact that pairs in the *groupings* condition obtained initially higher scores than in other conditions (as was determined by the ANOVA based on the first few tasks). Combined with the fact that the range of possible conceptual alignment scores is bounded above, obtaining high initial scores limits the scope for possible improvement. Indeed, particularly striking is that despite the different and even seemingly opposing effects of grouping and label feedback in the early phase of the experiment (since the same ANOVA suggested that label feedback was actually counterproductive at the beginning), by the end the differences between the *neither*, *groupings* and *labels* conditions were not significant (as confirmed by the ANOVA conducted on averages over the last five tasks). It appears that pairs in these conditions plateaued at a certain level, beyond which they could not break through (at least not after thirty tasks). As a result, pairs who did not receive grouping feedback eventually caught up to those in the *groupings* condition, perhaps due to increasing adeptness in the task and feedback on the task score. Indeed, the fact that score feedback was available to participants in all conditions could explain why even pairs in the *neither* condition improved over time. It's interesting, however, that label feedback turned out not to contribute anything extra when provided on its own.

On the other hand, participants in the *both* condition not only caught up to those in the *groupings* condition, but also significantly surpassed them. In other words, in terms of conceptual alignment by the end of the experiment, only a combination of conceptual and lexical feedback produced any difference relative to the *control* condition. This was not captured by the ANOVA based on the means over all the tasks, probably due to the initial disadvantage that labels seemed to cause. But the ANOVA which relates to pairs' convergence after they have had some time to interact, learn and converge (e.g., based on the final five tasks) confirms this, as shown by the groupings-labels interaction and simple effects analyses. As this is a key result, I stress that although the choice of basing the ANOVA on the final five tasks is rather arbitrary, the same significance patterns were found for each analysis based on the last seven tasks or less.

An interesting nuance in the results is that the high performance of pairs in the *both* condition (and hence including label feedback) at the end of the experiment contrasts with an apparent negative effect of label feedback at the outset of the experiment (as shown by the ANOVA based on the first five tasks). In other words, labels appear to have even been detrimental at first, even though by the end they were useful, at least in conjunction with grouping feedback. Why would this be?

I speculatively appeal to the following practical explanation. At the outset of the experiment, although participants are already aware that their task is to coordinate their groupings and not their labels, they may be preoccupied by and unduly focused on the labels. As such they pay less attention to not only their partner's groupings, but even their own and their agreement score. In other words, the labels may be initially distracting, which is not surprising since in the real world, we normally use language to communicate our conceptualisations.

What about lexical convergence? In contrast to the conceptual convergence results, there is no evidence for any direct impact of conceptual feedback on lexical alignment. All of the analyses relating to lexical convergence suggest that it is influenced exclusively by lexical feedback. Participants tended to adopt each other's labels more when they saw them, but they did not induce those labels from their partner's groupings. It should be emphasised, however, that the goal of the task may make a difference here and bias the results (but see Section 4.2.1). Perhaps if participants were asked to coordinate their words rather than their categories, they would be able to use their partner's groupings to figure out their partner's labels. However, at this stage, this is a matter of speculation.

Overall, then, my results do suggest a greater dissociation between lexical and conceptual alignment than is often assumed. To a large extent, conceptual coordination relies on conceptual feedback, and, even more, lexical coordination on lexical feedback. However, the results also highlight how words can play a role in conceptual coordination, as long as they are available in conjunction with samplings of the categories they are associated with.

8.5.2.3 Relation to previous alignment results

Until now I have been discussing the results of Experiment 4 in isolation from other work. In this section, I relate my experiment to previous dialogue and alignment research, before looking further out in the next section.

As reviewed in Section 3.4.3.2, many dialogue studies have looked into lexical and conceptual alignment in tasks where participants could communicate with language and shared perceptual information (e.g., Garrod and Anderson 1987; Brennan and Clark 1996). These studies have generally shown that interlocutors collaborate and converge on shared expressions for particular referents. Since different pairs converge on different terms, and these terms conceptually frame the referents in different ways, this has generally been taken as evidence that partners align both lexically and conceptually.

Experiment 4 sheds further light on these issues and qualifies past conclusions. First of all, consistent with the past studies, lexical feedback was shown to enormously increase lexical alignment. This was the case even though participants were not engaging in full dialogue but could only exchange category labels, a difference which was shown to be very significant in Experiment 3. Also, while in most dialogue experiments, participants needed to use language to figure out what they were referring to, here the instructions explicitly emphasised that label coordination was irrelevant (to their task scores). Thus, lexical alignment is not confined to full-fledged dialogue interaction, and it can occur even when it is theoretically redundant. This finding is consistent with Pickering and Garrod's (2004) theory of dialogue, according to which alignment is usually automatic.

On the other hand, the results here conflict with an assumption of a clean correspondence between lexical and conceptual alignment in particular. Even though participants may converge on lexical choices which seem to reflect distinct ways of categorising items, it does not automatically imply that they are conceptualising them in the same way. We may still try to defend the claim that "labels reflect conceptualisations" (Brennan and Clark 1996, p. 1482), by sticking to the individual's mental lexicon, but we cannot trivially then make the logical jump to maintain that lexical convergence entails conceptual pacts between individuals. The same word can correspond to different concepts for different people, even among interacting native speakers of the same language (see Section 2.4.3).

A few dialogue studies have shown that sharing visual information improved participants' performance in tasks with one participant directing another (Gergle et al. 2004; Clark and Krych 2004; see Section 3.4.2). The visual feedback could be loosely related to the grouping feedback in my experiment, and the results could be explained by one participant having some kind of access to their partner's conceptualisations. However, this explanation is speculative, and those results could also be explained in a variety of alternative ways.

As we saw in Section 3.4.3.3, a couple of dialogue studies have also looked more directly at conceptual alignment, and whether it is increased after dialogue-mediated interaction on a joint task (Markman and Makin 1998; Voiklis 2008). The main findings of relevance here were that people who engaged in interaction categorised things more similarly with each other than those who did not. However, these experiments did not separate linguistic interaction from other kinds of interaction. On the other hand, Experiment 4 manipulated the type of interaction at a finer level: although all pairs carried out the tasks in silence and received a coordination score after each task, they also received, depending on the condition, linguistic and/or non-linguistic feedback. Recall that the results revealed that conceptual coordination was primarily improved by conceptual feedback, and only secondarily (and at later stages) by lexical feedback. This suggests that although language can certainly enhance conceptual coordination, it may actually play a secondary role in doing so when other kinds of conceptual information are also available.

Overall, my results are consistent with previous research, although they caution against unfounded assumptions regarding the relationship between lexical and conceptual alignment. With regards to the role of words in conceptual alignment, they suggest that words do play a causal role, but a relatively small one.

8.5.3 Conceptual structure

The results of Experiment 4 also shed light on the relationships between words, concepts and conceptualisation. Concepts and words do not appear to be locked tightly together: their relationships are not uniform across speakers of the same language, nor are they stable over time. I will return to these issues in the next chapter (see Section 9.5), since to some extent, these are stronger versions of conclusions that came out of the previous experiments. In this section, however, I consider an issue which relates specifically to the design of Experiment 4: the internal structure of concepts.

Although, strictly speaking, participants in Experiment 4 categorised items on their own, the goal was of course to align conceptually with others. The fact that alignment generally increased across the tasks shows that, to some extent, participants were successful at aligning with their partners. Since the same stimuli did not recur in subsequent tasks, people could not just rely on remembering individual items and reproducing categorisations they had previously produced or seen (in the case of grouping feedback). Nor could they simply reproduce the concept labels their partner had used (if they got label feedback). In order to converge conceptually, they needed to at least partially infer and adopt their partner's concepts (unless their partner did so with theirs), from which the subsequent categorisations would then be determined. The different conditions manipulated what information they had available after each task for carrying out these inferences (as shown in Figure 8.2). [3]

Therefore, we can now relate the results of Experiment 4 with the theories of conceptual representations. This will be illuminating because the debate between different concept theories or how they should be integrated has not been resolved, and my experiment provides a novel way to shed light on it. In fact, it provides a potentially good battleground between prototype and theory theories on one hand and exemplar theory on the other, since word meaning has generally been handled better by the former and category learning by the latter (Murphy 2002). Experiment 4 brings these two themes intimately together.

So what's the verdict? In brief, the main results of Experiment 4 could be interpreted as suggesting that access to someone's exemplars is more useful for conceptual alignment than access to their senses. According to this interpretation, we could argue that exemplars form the primary basis of concepts (Medin and Schaffer 1978), at least when it comes to conceptual coordination, since they are more intimately involved. Intuitively, this brings to mind commonplace statements like "I don't understand what you mean: give me an example".

However, this interpretation has a potentially significant weakness. In particular, it does not acknowledge any asymmetry between words and categories as conceptual identifiers. Indeed, I have brought together arguments and evidence (see Section 2.4) that concepts and categories are locked in an intimate causal bond (Frege 1892/1948), but that the relationship between words and concepts is more flexible (Geeraerts 1993). Therefore, the stronger effect of grouping feedback could equally be a result of a category subset giving a more accurate indication of someone's concept (via the strong concept-category link) than a label (via the weak concept-word link).

My experiment does not allow us to distinguish between these possibilities. However, my main results suggest that at least *one* of these two claims should hold. Either words

[3] "Inference" is an unintentionally loaded term here: people might align without explicitly modelling their partners' mental states (Pickering and Garrod 2004); I use it here primarily for convenience.

are not as reliable for getting at concepts as category subsets, or exemplars constitute the more dominant component of conceptual structure, at least for the purposes of interactive conceptual alignment.

We can also try to relate the time course of category learning to that of conceptual alignment. Smith and Minda (1998) showed in a series of experiments that a prototype model generally fit categorisation results early in learning, but that an exemplar model took over later on. This result is inconsistent with my findings, since in Experiment 4, exemplars dominated early on, but at later stages both exemplars and sense were important. However, Smith and Minda's study has been criticised by Nosofsky and Johansen (2000), who then showed that exemplar models could handle both stages.

More generally, on the basis of his extensive review, Murphy (2002) noted that exemplar theory has done better in category learning experiments, but that this has particularly been the case when they have dealt with small, unnatural sets of categories. My experiment also uses an unnatural stimulus domain, but it is much larger than those of most previous studies. While there were 330 stimuli in my experiment, and each was only categorised once by every participant, many influential category learning experiments supporting exemplar theory have only had a handful of items that are repeatedly categorised over and over (e.g., Medin and Schaffer 1978).

Therefore, my perspective on the learning trajectory is like this: exemplar theory does well when there are a limited number of items to deal with, as then our memory can manage them and we can exploit their individual properties. However, if the set of stimuli just keeps expanding (as in my experiment but unlike most category learning studies), then at some point, memory starts to overload and a greater reliance on an abstraction (like a prototype) begins to be more useful. If this is correct, then we might expect that if my study had gone on longer, label feedback would not only "join" grouping feedback as it did by the end, but would eventually take over.

As such, we should guard against assuming that Experiment 4 is the end of the story. Although access to others' concept exemplars seems to be more useful than access to their corresponding concept senses, there is also some evidence that access to senses becomes more important over time, and much more research is required to resolve this.

In any case, these results show how, following Markman and Makin (1998), interactive studies of communication can be fruitfully employed to study categorisation.

8.5.4 Methodological comments

Experiment 4 built off the methodological successes of Experiment 3, while also tweaking the design. As I explained in Section 7.5.4, the framework seems both more suited and more needed for studying conceptual alignment during rather than after interaction. As a result, Experiment 4 consisted exclusively of a series of joint categorisation tasks, in which participants categorised sets of stimuli independently, but with condition-specific feedback after each task.

However, since Experiment 3 did not show any effects of increased convergence across tasks, or any significant difference in conceptual convergence between the *control* and *silent* conditions, I made a few specific modifications, including reducing the number of categories to two, increasing the number of tasks from ten to thirty, and switching to a potentially more variably conceptualised stimulus domain. These changes were successful, in that Experiment 4 showed a significant effect of task. In addition to clearly showing that the framework could be used to study changes in lexical and conceptual alignment over the course of short experimental sessions, this also made it possible to study the time course of alignment, which yielded interesting results. The differences between conditions in conceptual convergence were wider in the first few tasks of the sessions than they were by the end. However, since the patterns of scores did not clearly plateau, it would be interesting to see what would happen in such an experiment with a much larger number of tasks (e.g., 100). It is possible that the rather arbitrary imposed (due to practical constraints) cutoff of thirty only reveals a snapshot at a particular moment of generally complex and changing relationships.

A remaining issue with the framework is that, when dialogue is not permitted (as in Experiment 4), there is actually no interaction between participants during the individual tasks. Participants do get feedback after each task, but they cannot coordinate how to act on it. In particular, this means that in cases where they do not get perfect scores, they have to independently decide what to do on the subsequent task: stick to the way they did things before, switch to what their partner was doing, or try something else entirely.

While this is not an intrinsic problem and can be considered part of the challenge, it can slow down convergence. Fortunately, this issue was apparently not large enough to eliminate interesting effects.

In general, the framework proved effective in Experiment 4, continuing off of the improvements seen in Experiment 3. However, I return to a general evaluation of it in Section 9.2.

8.6 Conclusion

Although Experiment 3 showed that dialogue was hugely beneficial for conceptual convergence, it left unanswered questions regarding the usefulness of category labels on their own. As a result, Experiment 4 was designed to look into this in more detail, and pitted conceptual feedback versus lexical feedback over a larger number of tasks using a different domain of stimuli. The results primarily pointed to a kind of dissociation between lexical and conceptual alignment: lexical alignment depended primarily on lexical feedback, while conceptual alignment depended mainly on conceptual feedback. However, a closer look at the data, especially when considering different stages of the interaction, revealed that conceptual alignment was also sensitive to lexical feedback, and in a striking way. At first, lexical feedback seemed to be a distraction and actually impaired conceptual alignment. But by the end of the tasks, pairs were using it to increase convergence, as long as it was provided in conjunction with conceptual convergence. Therefore, it appears that lexical and conceptual alignment, although not in direct correspondence, are related, and that exchanging category labels (as opposed to full dialogue) does (eventually) increase conceptual coordination.

CHAPTER 9

General discussion

In this thesis, I have set out to investigate a novel research question, proposed an empirically motivated theoretical model to ground the investigation, developed a corresponding experimental framework, and conducted four experiments which examined different aspects of the question. It is now time to evaluate the results of my experiments as a whole in terms of my general hypotheses, and to consider their broader implications.

This discussion is organised as follows. I first review my four experiments, summarising the designs, results, and interpretations. I then evaluate the experimental framework, assessing its strengths, weaknesses and potential. I then integrate my results and past literature and discuss them in terms of language's role in conceptual alignment and closely related issues. After that, I consider the implications of my findings for other issues, including those that my experiments have tackled directly, that lay the foundations for my theoretical model, and that are intimately related with my work but have been hitherto neglected. Finally, I draw my overall conclusions regarding the question I began with.

9.1 Summary of empirical findings

The experiments in this thesis have investigated whether language brings together people's categorisation. They have looked at different timescales of conceptual alignment and manipulated the extent of language availability.

Experiment 1 focused on the possible effects of our native language on our categorisation, and in particular, on whether speakers of the same language categorised things more similarly to each other than speakers of different languages. The experiment was meant to partially replicate and extend the previous study of Malt, Sloman, Gennari, Shi and Wang (1999) but also address my research questions more specifically. Native speakers of English, Japanese and Polish carried out both naming and free classification tasks with the same set of 60 dishes. Like the original study, my replication results showed that there was divergence between the linguistic groups in naming, but much less in sorting, and that naming and sorting patterns did not generally correspond to each other. However, I also supplemented the original forms of analysis with methods that were based on the degree of agreement between pairs of individuals, and these analyses qualified the original findings. The most important result for my thesis was that speakers of the same language did categorise things more similarly to each other than speakers of different languages, although the difference wasn't nearly as large as it was for the analogous analysis based on naming. However, I argued from the body of results that the cross-linguistic differences in sorting were not in fact due to language, but were more likely better explained by other cultural factors.

Experiment 2 shifted the focus to the role of linguistic interaction on conceptual alignment between individuals. It focused on whether engaging in a joint categorisation task brought people's conceptualisations together, and whether being able to talk freely during the task increased the effect. To that end, pairs of participants carried out a sequence of three free classification tasks. The first and third tasks were always carried out individually, but the second depended on the experimental condition. In the control condition, the second task was an individual one, just as the first and third. In the other two conditions, participants worked together to categorise the stimuli, taking turns through a synchronised interface. In one of these two conditions, pairs were also allowed to engage freely in dialogue during this task. Surprisingly, the results revealed no effects of either feedback or dialogue on conceptual convergence. They did converge lexically, but even then, there was no difference between the experimental conditions. However, the experiment had several methodological shortcomings, which cast doubt on the results and the shortage of differences between conditions.

Experiment 3 sought to address the methodological issues with Experiment 2, while also expanding the scope of inquiry. The most important methodological modification lay in the nature of the joint categorisation tasks, which was incorporated to permit the study of alignment during interaction as well. In particular, rather than sorting a set of items together, participants now did so independently of their partners, but with the explicit goal of aligning. In the control condition, they saw an alignment score after each task. In the other two conditions, they were also given feedback on their partners' categories. And as in Experiment 2, the difference between the latter two conditions was that in one of them they were allowed to talk freely during the joint tasks. The experiment consisted of a sequence of ten joint categorisation tasks of this kind, followed by an individual phase in which participants individually carried out similarity judgments and categorisation tasks. This modified design revealed a very strong effect of dialogue on conceptual convergence during interaction, but no effect of feedback, and no effect on either similarity or categorisation following interaction. On the other hand, lexical convergence did occur over the course of the tasks for interacting pairs, and the effect was stronger if participants could talk to each other. Moreover, lexical convergence persisted beyond the interaction, and was evident still in the individual categorisation tasks.

Experiment 4 shifted focus entirely to alignment during interaction, and eliminated the availability of full-fledged dialogue. The experiment manipulated the availability of category grouping and label feedback, inquiring whether they affected conceptual and lexical convergence. In order to give pairs more time to learn and adapt to each other, the individual post-interaction tasks were eliminated, and the number of joint categorisation tasks was expanded. Also, as the labels were now more in focus, they were restricted to single English words. The results showed that lexical alignment was primarily enhanced by lexical feedback, and conceptual alignment by conceptual feedback, suggesting a dissociation between the two processes. However, lexical feedback did also eventually increase conceptual convergence, as long as it occurred in conjunction with conceptual feedback. Moreover, other analyses showed that there was some correspondence between lexical and conceptual alignment, so the lexical-conceptual dissociation should not be overstated.

9.2 Evaluation of the empirical framework

One of the main goals of this thesis has been to design an experimental framework suitable for empirically exploring my hypotheses. As I pointed out in Chapter 1, this was particularly challenging because it required the explicit rejection of two related, convenient and frequently adopted idealisations: that concepts are the same for different people, and that there is a simple one-to-one relationship between words and concepts (at least within a language). The serious rejection of these idealisations imposed substantial constraints on methodological decisions, which largely determined the features of the experimental framework that I developed in Chapter 4 and subsequently used in my experiments.

The resulting framework had several key features. First, the main tasks were free classification tasks, in which participants partitioned a set of stimuli into categories. Within a few practical constraints, participants were generally free to decide on the number, sizes, names and bases for their categories. Second, the experimental stimuli were sampled from continuous perceptual domains, namely dishes (in Experiments 1-3) and triangles (in Experiment 4). This allowed for multiple stimuli to be presented non-linguistically and simultaneously, and increased the potential for variation between participants' categorisation. Third, measures were carefully chosen for comparing the outputs from two different participants in a task. The most important measures for my purposes concerned comparisons of two category partitionings, but measures were also defined for comparing category labels and similarity judgments. Fourth, the free classification tasks involved varying degrees of interaction between participants. Sometimes, participants simply sorted the items individually, with no interaction or shared goal; in other cases, they sorted individually, but linguistic interaction was allowed or feedback was provided to them; and in still other cases, they sorted things together, making joint decisions on a single stimulus set. Fifth, the role of language was evaluated with specific manipulations. In particular, I controlled whether participants were allowed to speak to each other during the experiments, and whether they received feedback on their partners' category labels.

But how did the framework perform in practice? Some aspects of the framework cannot be judged a priori, but only through actual application in experiments. In particular, in order to be useful, the experimental tasks must be meaningful and straight-forward for participants, they must have an appropriate level of sensitivity (being neither too challenging nor too trivial), they must yield variation in output between participants, and they need to show at least some significant differences between experimental conditions. After four different experiments and a few bumps on the way, it is now clear that these criteria were generally satisfied. Experimental sessions generally ran smoothly and efficiently, and questionnaire feedback consistently confirmed that participants understood the tasks and were comfortable with the program interface. There was plenty of variation between the conceptual, lexical, and similarity outputs of different participants, allowing for an examination of the conditions under which they could or would align. And there were numerous interesting differences found between experimental conditions in different experiments, shedding light on my hypotheses and raising new questions.

It is important to point out, however, that the framework has proven to be better suited to some theoretical questions than others. The main weak point is that it does not appear to be very powerful at capturing alignment phenomena that persist beyond interaction. In Experiments 2 and 3, most of the analyses which compared participant pairs' outputs on individual tasks following interaction yielded null results, especially for conceptual and similarity convergence. However, previous studies have demonstrated such effects of interaction (Markman and Makin 1998; Voiklis 2008), suggesting that this might be a methodological weakness, rather than a theoretically valid result. In contrast, the use of full-fledged dialogue during interactive tasks seems to make lexical and conceptual convergence almost trivial. In particular, Experiment 3 yielded a very large proportion of ceiling effects in the talking condition for the joint convergence tasks. While this was very useful for the purpose of demonstrating the powerful role of dialogue, it may be a potential worry when designing future experiments to explore the relationship between dialogue and conceptual alignment in more detail.

The framework also has a couple more general concerns. First of all, while I describe some of my experimental conditions as "interactive", the actual interaction which is supported and permitted is quite limited. Indeed, when dialogue is not allowed and participants carry out tasks individually (as in Experiment 4, for example), interaction is restricted to certain forms of feedback between tasks. In such conditions, participants cannot speak to each other, cannot follow each other's body movements and gestures, and are not operating continuously within the same problem environment. A side-effect of this is that participants cannot then negotiate anything while actually carrying out a task, which can lead to "cross-over" effects while attempting to coordinate categories, as discussed in Section 7.5.4. In particular, both participants of a pair may successfully adopt each other's categories from the previous task, and yet see no increase in their agreement score. However, these issues are not intrinsic limitations of the framework, but could be addressed in different ways in future experiments (e.g., by embedding the tasks in more natural interaction, where participants can see each other). Eliminating these effects should help decrease the amount of tasks (and thus time from participants) needed to study convergence under different conditions.

Also, the framework faces questions regarding the theoretical interpretation of concepts. The free classification tasks yield categories of stimuli, which have been interpreted as snapshots of participants' concepts. However, as I argued myself in Section 2.4.1, there is an important difference between concepts and categories, with categories being referential extensions of concepts, and multiple concepts potentially corresponding to a single category. Moreover, seeing people's category groupings tells us nothing absolute about how they are thinking of the stimuli. My exclusive reliance on referential information for identifying concepts also implies that the framework is not trivially applicable to abstract or perceptually complex conceptual domains. However, note that these theoretical issues are not unique to my experiments but apply to categorisation studies in general, especially when there is a need to separate lexical from conceptual processes.

These concerns notwithstanding, the experimental framework offers a concrete solution to an important methodological challenge which had not previously been met. Recall (see Section 3.4) that past experiments have examined conceptual alignment either during interaction but indirectly via lexical alignment (e.g., Garrod and Anderson 1987), or directly and independently of language but after interaction (e.g., Markman and Makin 1998). The fundamental problem lay in how to query conceptual alignment during interaction without relying directly on language. But this was precisely the strength of the current framework. While the framework did not reveal much of interest when examining conceptual alignment in tasks that followed interaction, it did yield novel and insightful results concerning online conceptual alignment during interaction. The key feature that made this possible lay in embedding free classification tasks in interaction between participants. Within this basic setup, we can control how participants interact, if and how they can use language, what sorts of things they are categorising, and what the task goals are, while still getting conceptual snapshots independently of language. Having participants also label their categories (as in Experiments 2-4) offers the additional possibility of independently measuring lexical alignment and comparing the two phenomena or testing for correspondences between them. However, if we want to minimise the interference of language in the free classification task, we may choose instead not to have participants label their categories (as in Experiment 1). In general, the framework offers a lot of flexibility while allowing for the separation both between lexical and conceptual processes and between individuals. As such, it provides a partial solution to the fundamental problem emphasised by Nuyts and Pederson (1997): in order to study the relationship between language and conceptualisation, we (researchers) need to be able to separate them.

Indeed, the generality of the framework and its success in examining online conceptual alignment suggest that it could be fruitfully applied to many research questions beyond the specific ones investigated in this thesis. The modifications and applications could vary widely from minor extensions to major adaptations. Participants could be allowed to talk, but only between rather than during the joint tasks (making the comparison between the roles of label feedback and dialogue more "fair"). Different stimulus domains could be used, and sampled in different ways. If the requirement of using non-linguistic stimuli was relaxed (although this would have some other implications), abstract domains could be studied as well by using words for stimulus items. Note that, given the apparent domain-dependence of findings in linguistic relativity research (see Section 3.3), the choice of stimulus domain might make a big difference. The kind of interaction could be further manipulated, such as allowing participants to see each other or where they were looking (perhaps with the use of eye-tracking technology). Issues of partner-specificity and community conventions could be investigated by manipulating participants' partnerships in particular ways (as was effectively done by Garrod and Doherty 1994). Since the framework has been implemented as a server-client computer program using a separate database for persistent data storage, experiments could easily be done through the Internet, potentially helping to amass much larger amounts of data from more participants and larger sequences of tasks. The specificity of effects to language could be tested by offering participants other ways to communicate or label their categories (Galantucci 2005). The importance of language proficiency or other bilingual issues could be tested by running the experiments with non-native speakers. Sociolinguistic questions could be investigated by focusing on how participants are paired up (e.g., by gender, nationality, profession, etc.). And since the free classification paradigm has been successfully used to study categorisation in animals and young children, the paradigm could even be used to explore the debatable relationship between conceptual development and language acquisition, or the equally debatable issue of what constitute evolutionary cognitive preadaptations to language.

In summary, although methodological and theoretical challenges remain for the framework to ultimately face, and despite not being equally successful on all fronts, the experimental framework I have developed marks one of the most important contributions of this thesis. Not only has it allowed me to address my hypotheses, it also provides a general and novel way to investigate conceptual alignment between people during interaction, which does not rely on language. As such, it could be used as a tool for many other important research problems, some of which I will mention more specifically in the subsequent sections of this discussion.

9.3 Conceptual (and lexical) alignment

9.3.1 The unit of alignment

Before relating the alignment results of my four experiments to each other and to previous work, it is important to discuss an ambiguity concerning the unit of analysis in alignment. The issue derives from the fact that the output of a single task in my experiments did not consist of a single category and/or label being assigned to a single item, but rather multiple categories and multiple labels which partitioned a relatively large set of items. Although, as I argued in Chapter 4, this was necessary and appropriate for investigating my research questions, it introduces ambiguity in what is meant by alignment, especially lexical alignment.

In particular, there are at least three possible interpretations and corresponding bases for measuring lexical alignment. Perhaps the most natural interpretation is where we characterise the output as assigning each item to a particular label, and then assessing the extent to which people agreed on particular labels for the particular items. This way we incorporate both the labels and the items in the output. However, we could also focus exclusively on the items, ignoring the actual labels used, and measure the extent to which people's labels partition the stimulus space into the same groups (the **item**-based approach). Alternatively, we could ignore the items and focus instead on the labels, so that our measure only compares the sets of labels used by different people (the **identifier**-based approach).

While the first method may seem the richest, there were specific reasons why I adopted the item-based approach in Experiment 1 and the identifier-based in Experiments 2-4. The main goal of Experiment 1 was to investigate the correspondence (or lack thereof) between linguistic and non-linguistic categorisation. As I argued in Section 5.3.3.2, it was important to make the measure of lexical alignment as analogous to the the measure of conceptual alignment as possible. Since the non-linguistic tasks and conceptual alignment measure did not, by definition, involve any linguistic labels, they were necessarily item-based. As a result, contra Malt et al.'s (1999) original analysis, my own lexical alignment measure in Experiment 1 focused exclusively on the item groupings induced by people's labels.

In contrast, in Experiments 2-4, the research focus was quite different. In this case, I wanted to study how conceptual alignment was influenced by various linguistic manipulations, and how it related to lexical alignment, both during and after interaction. As a result, I needed to have a single task which provided both conceptual and lexical output, and my solution to this problem was to adapt the free classification tasks so that participants also labelled their categories. However, this wedding also meant that, by definition, the labels and the categories partitioned the space up in identical ways. As such, using an item-based interpretation of lexical alignment in these experiments would make the comparison with conceptual alignment meaningless. Instead, the point was to separate the labels from the groupings. Therefore, the tasks were interpreted as yielding two distinct sets of output: a set of labels and a set of groupings. Comparing the former was interpreted as measuring lexical alignment, and comparing the latter as conceptual alignment. These two kinds of sets are fundamentally different things, and so direct comparisons between the values, as in Experiment 1, would be inappropriate. Fortunately, that was not required for the purposes of Experiments 2-4.

Note that the ambiguity between different interpretations of lexical alignment could also apply, in principle, to conceptual alignment. There too we could theoretically ask whether we are concerned with the categories into which individual items are placed, or whether we just care about the set of categories that is used to partition a set of items (or both). However, in practice, the distinction collapses because, in the absence of external category identifiers (like labels), I have chosen to actually identify categories by the set of items that constitute them. Consequently, asking what categories a person used to partition a set of items intrinsically contained all the information about which items were put into which categories. In other words, for conceptual alignment, there was effectively no difference between the item-based and the identifier-based interpretations. In a sense, in this case, we have the luxury of emphasising either interpretation, depending on our purpose.

It is important next to consider how my interpretations of alignment relate to previous work. The answer is again closely related to my adoption of free classification tasks for my experiments. Notice that the three-way distinction I made between possible interpretations of alignment also collapses if task output consists of only one item at a time. In that case, the "partitioning" reduces trivially (in terms of analysis) to a single category containing a single item, and the category label is inevitably associated with both that category and that item. Of course, many previous experiments have also involved multiple stimuli that participants are dealing with at the same time. However, in most cases, the tasks involve the identification (in production, comprehension and communication) of particular items from larger sets, so that single items still constitute the focus and unit of analysis (e.g., Horton and Gerrig 2002; Metzing and Brennan 2003; Brennan and Clark 1996). Yet despite the potential interpretative ambiguity concerning lexical alignment in past work, it makes sense to interpret it as relating to both the items involved **and** how they are lexicalised. The theoretical issue of lexical alignment is not normally whether people merely use the same words **or** attend to the same items, but rather whether they align on the same terms for the same items. As a result, although for the purposes of my research questions it was appropriate to take different views of lexical alignment, the interpretation and measure should be interpreted as less strict than in previous work, and this must be borne mind when discussing my findings more generally (which I undertake in the following section).

This perspective also sheds light on the two different kinds of conceptual alignment results in the literature. As discussed in Sections 3.4.3.2 and 3.4.3.3, past work has used different methods to inquire into conceptual alignment during or after interaction, respectively. In the case of alignment during interaction, concepts have been identified with words, so that lexical and conceptual alignment were essentially equivalent, by assumption (e.g., Brennan and Clark 1996). As a result, the unit of analysis was again usually a single item, and the most natural interpretation of alignment concerns both an item and its identifying label. In contrast, studies that have examined conceptual alignment following interaction have followed the same general free classification approach as I have (Markman and Makin 1998; Voiklis 2008). Therefore, as in my experiments, alignment in these studies could in principle be interpreted as either item-based, identifier-based, or both. However, such studies are rare, and they have not looked at alignment during interaction.

More generally, the usage of conceptualisation tasks in which the output consists of multiple categories (and labels) involves a certain idealisation that is not required in the individual case. In particular, the underlying assumption is that task output constitutes a referential snapshot reflecting a person's mental conceptualisations. However, this abstracts away from the fact that each such task does involve the individual conceptualisations of different items, and that these may change slightly over the course of the task. Indeed, giving participants the possibility of changing their categorisation decisions before committing them explicitly acknowledges this. While such an assumption is a necessary and justifiable compromise when our tasks depend on the categorisation of multiple items at once, it may become important when we consider different task conditions, which differ in terms of whether participants can interact within individual tasks. I return to this point in the next section.

9.3.2 *Prior, online and preserved alignment*

My experiments have investigated conceptual and lexical alignment at several different points in time relative to interaction. In order to avoid confusion, and to help structure the discussion of my findings and how they relate both to each other and to other literature, I make a three-way terminological distinction concerning alignment along the temporal dimension: **prior**, **online**, and **persistent**. Prior alignment is that which exists between speakers of the same native language who have not interacted directly with each other (this is analogous to the role that background knowledge plays in common ground; Clark 1996). Online alignment captures that which there is between people while they are interacting with each other in some way. Persistent alignment is that which remains between participants shortly after interaction. While these distinctions are not precise when applied to the world at large, they are sufficient to clearly divide up the relevant experimental results. In relation to my studies, Experiment 1 and the pre-task in Experiment 2 measured prior alignment, the joint tasks in Experiments 2-4 addressed online alignment, and the post-task in Experiment 2 and the individual phase in Experiment 3 attempted to capture persistent alignment. I now discuss each of the three types in turn.

Prior alignment was measured in both Experiments 1 and 2. In Experiment 2, prior alignment was measured strictly to provide a baseline for assessing preserved alignment afterwards. As such, these tasks were not granted any theoretical value of their own, and I do not discuss them further here. In contrast, in Experiment 1, prior alignment was a major focus. The key results were that speakers of different languages diverge from each other in their prior alignments, but that the divergence is greater for lexical than for conceptual alignment. This finding is consistent with the growing consensus in the linguistic relativity literature that language does not determine our thought, but that it does affect it (Gumperz and Levinson 1996; see Section 3.3.5). As such, it is expected that linguistic differences should have a stronger impact on linguistic than on non-linguistic representations, but that the latter are still partly affected and aligned. However, such an effect in an object domain but not based on formal grammatical distinctions such as gender (Lucy 1992) is relatively novel. Nevertheless, as I argued in Section 5.3.4.3, given the other results of my experiment, the relatively high degree of conceptual alignment among speakers of the same language is probably best explained here by other, non-linguistic cultural factors. I return to this issue in Section 9.4.

The investigation of online alignment took place in Experiments 2-4, and did so in quite different ways. In Experiment 2, this involved a joint categorisation task where participants worked together on the same stimulus set to produce a joint set of categories. While this made the task more interactive, it did not allow for a separation between the output of the participants, making it impossible to directly apply a measure which compares them. However, an examination of the audio data informally suggested that participants (in the talking condition) used dialogue in various ways to align their labels and conceptualisations, as has been demonstrated previously (e.g., Brennan and Clark 1996; Clark and Wilkes-Gibbs 1986).

In Experiments 3 and 4, participants carried out a sequence of joint categorisation tasks in which they actually worked independently, but received feedback after each task. However, the results were different in the two experiments. I discuss first the results for lexical alignment, and then for conceptual alignment. Experiment 3 showed that dialogue could enormously increase conceptual alignment, but failed to find any effect of feedback between the tasks or any improvement across tasks. On the other hand, Experiment 4 did find a general learning effect, as well as an effect of feedback, especially conceptual feedback. The appearance of a learning effect in Experiment 4 may have been due to a large increase in the number of tasks (from 10 to 30), combined with a reduction of the number of categories per task (from a maximum of 4 to 2). This combination gave pairs more time to learn to adapt to each other, while probably also making it easier to process their partners' categories and compare them to their own. The superior advantage provided by conceptual feedback over lexical feedback points out that while language can play a role in conceptual alignment, it is by no means the only cause, and probably not even the strongest one. However, the effect of dialogue shown in Experiment 3, on the other hand, was striking, showing a massive advantage over feedback, even when that feedback included both groupings and labels. The strong online effects of dialogue in Experiment 3 were among the clearest results of my experiments, and I return to them in Section 9.3.4.

The results for online lexical alignment in Experiments 3 and 4 were similar to but sharper than those for conceptual alignment. First, in Experiment 3, dialogue promoted lexical alignment, but so did feedback, although to a lesser extent. However, Experiment 3 confounded lexical and conceptual feedback (compare the control and silent conditions in Figures 7.1 and 7.2, respectively). Experiment 4 thus separated the two, and found that lexical alignment showed a strong dependence on lexical feedback, but none on conceptual feedback. So, people seem to have an easier time aligning their labels than their concepts. While participants' convergence on lexical choices is consistent with previous work, the difference between the lexical and conceptual results warns us against assuming an equivalence between lexical and conceptual alignment, as is often done in the literature (e.g., Brennan and Clark 1996). I discuss the lexical-conceptual dissociation further in the next section.

Finally, Experiments 2 and 3 also looked at preserved alignment. Participants carried out joint categorisation tasks (and, in the case of Experiment 3, similarity judgement tasks) after interacting together, to see whether their interaction had a lasting effect on their alignment. The results for both experiments found no such effects for conceptual alignment. This is particularly striking given the sharp difference in online alignment between the talking condition and the other conditions in Experiment 3: despite having coordinated so well conceptually during interaction, the convergence disappeared entirely afterwards. These results are in conflict with previous experiments that have investigated preserved conceptual alignment, which have found that people who interacted linguistically in a joint task were more aligned afterwards than those that did not (Markman and Makin 1998; Voiklis 2008). As discussed in Section 7.5.3, this may have been due to the properties of the stimulus domain in my experiment, since in contrast to previous experiments, I specifically aimed to adopt stimulus sets that were as challenging to categorise as possible.

Preserved lexical alignment showed different results. Experiment 2 found that some lexical convergence did occur, but did so regardless of condition. However, this experiment was generally not very sensitive, and the lack of difference between conditions here could also be an artifact of that. Indeed, Experiment 3 found that interaction did help convergence, although dialogue specifically did not enhance it further. Together, these results suggest again that it is relatively easy for pairs to converge on the same labels, which is consistent with lexical alignment results showing how participants sometimes stick to the same referential expressions even after switching communication partners (e.g., Brennan and Clark 1996; Malt and Sloman 2004). However, in conjunction with the conceptual results, there seems to again be evidence for a lexical-conceptual dissociation, although this might be partly explained by the unit of alignment issue I discussed in the previous section.

In the next two sections, I take up two central themes surrounding these results. First, I discuss the dissociation that has been repeatedly cropping up between lexical and conceptual alignment. Second, I return to the logically distinct but related question with which I began this thesis: the role that language still plays in conceptual alignment.

9.3.3 Dissociation between lexical and conceptual alignment

Before continuing, it is important to clearly distinguish two quite different ways in which language has been involved in my experiments. On one hand, the use of language has been manipulated as an independent variable. In particular, the experimental conditions in Experiments 2-4 varied in terms of whether people could use dialogue and whether they received label feedback. On the other hand, linguistic output was measured as a dependent variable. Participants' category labels were compared to calculate lexical alignment. The experiments have explored how the linguistic manipulations affect both lexical alignment and conceptual alignment. As a result, the issue of the relationship between lexical and conceptual alignment should not be confused with the question of how language affects conceptual alignment. Consequently, I treat the two issues separately.

I begin, in this section, with the relationship between lexical and conceptual alignment. I have emphasised from the beginning the importance of not assuming an a priori equivalence between words and concepts, and between lexical and conceptual alignment. My experiments have demonstrated that this was not just a minor theoretical quibble, but that once we manage to separate the two phenomena methodologically, we find a real separation between the two.

Indeed, all four experiments in this thesis have found evidence that lexical and conceptual alignment are not equivalent, as argued by Schober (2005). Experiment 1 revealed more lexical than conceptual alignment among speakers of the same language. In Experiment 2, participants converged lexically, while showing no conceptual convergence. Experiment 3 found that lexical alignment came more easily to participants than conceptual alignment, both during and after interaction. And Experiment 4 discovered that conceptual alignment was primarily dependent on conceptual feedback, while lexical alignment was only affected by lexical feedback.

Of course, this does not imply that words and concepts have nothing to do with each other, or that lexical and conceptual alignment never correspond. Clearly, that would be overstating the case, and there are also results in my experiments which qualify such an untenably strong stance. Moreover, it would be unfair to summarily dismiss previous work which has assumed equivalences between lexical and conceptual alignment. To a large extent, it is probably a matter of precision. In my experiments, the differences between stimuli were very small and the concepts that people used varied very subtly, but in other experiments the stimuli tended to be more distinct, and the conceptual candidates more clear. For instance, Brennan and Clark's (1996) experiment involved lexicalisations of footwear (among other things) at various degrees of specificity (e.g., "shoe", "loafer"). There can be little doubt that these terms do suggest different conceptualisations, at different levels of abstraction, and that these are at least partially shared by speakers of the same language. Similarly, there is no doubt that the different description schemes (e.g., "B3", "top right") that participants used in Garrod and Anderson's (1987) maze experiment reflected different conceptual bases of solving the referential challenge. Words clearly tell us something useful.

Nevertheless, what we must be on guard against is the logical jump from vague reflection to solid equivalence. Lexical alignment seems to be more superficial and straightforward than conceptual alignment. In other words, people agree readily on expressions for things, but that does not imply that they conceptualise them in entirely the same way. I pursue the implications of this in Section 9.5.3 for the relationship between words and concepts.

9.3.4 *The role of language in conceptual alignment*

We are now in a position to address the general research question that I posed way back in Section 1.2: does language bring about conceptual coordination between individuals? The results of my experiments, together with previous literature, suggest that it does. When interacting linguistically, people's conceptualisations of things in the world come closer together, at least for the duration of the interaction. However, the extent to which alignment occurs depends on the form of language that is available. If people can communicate freely using dialogue, then they can align quite precisely. But if they only have access to the labels that their interactants assign to their categories, then alignment is much weaker and is actually also (and more) dependent on perceptual support.

Why should there be such a big difference between the effects of dialogue and label feedback? After all, we might at first imagine that the main benefits of dialogue would come from being able to share lexicalisations of the stimuli, which would trigger specific conceptual associations. However, it does not appear that this was the case, since lexical and conceptual alignment have revealed a dissociation, and since the combination of lexical and conceptual feedback was nowhere near as beneficial as dialogue. Therefore, it seems more likely that it was other rich features of dialogue that were responsible. In particular, the possibility to describe items in detail, to focus on one at a time, and to check decisions with each other may have been more important factors. In effect, dialogue may have turned the problem from one of a single conceptual snapshot into a gradual cumulative negotiation, during which participants converge and align their conceptualisations. This is in line with the action view of language use developed by Clark and colleagues (e.g., Clark and Wilkes-Gibbs 1986; Schober and Clark 1989; Clark and Brennan 1991; Brennan and Clark 1996; Clark 1996).

This explanation of the extra benefits provided by dialogue raises a couple of important issues. First, one of the goals of my experiment designs has been to manipulate the availability of language while maintaining interaction and a joint goal. Consequently, the joint tasks in Experiments 3 and 4 asked participants to try to coordinate their categories, and gave them feedback, even when dialogue was not allowed. In some cases, the feedback included both groupings and labels, so that after each task participants effectively had a complete conceptual and lexical snapshot of their partner's categories. However, this form of interaction was rather impoverished: participants could not actually interact with each other at all during the individual tasks.

As mentioned in Section 7.5.4, there are different ways in which this point could be interpreted. On one hand, it could be argued that this places participants in the dialogue condition (in Experiment 3) at an unfair advantage: while the difference between the silent and talking conditions was only supposed to be the addition of dialogue while keeping the interaction constant, the change also fundamentally enhanced some of the features of the interaction itself. Under this interpretation, the manipulation failed to control for interaction, and thus did not isolate the effect of dialogue.

On the other hand, we could emphasise that the only mechanical difference between the silent and talking conditions was that partners were allowed to speak, which entails that we did isolate the role of dialogue. If dialogue has properties which result in a richer form of interaction, then that is an important part of what dialogue brings to the table. If we interpret in this way, then the manipulation, far from being a failure, is particularly illuminating because it highlights some of the special mechanical features of dialogue.

I do not try to resolve this issue categorically, because the points are actually complementary. It would be fruitful to push the framework in future experiments in a way which somehow maintains a high level of interaction in non-linguistic conditions. But it would also be useful to consider how the mechanical differences in such future experiments may themselves be due to language.

These considerations are reminiscent of Clark and Brennan's (1991) analysis of how features of communication media can affect the development of common ground. Clark and Brennan listed eight different dimensions along which communication media can differ, and claimed that the further you get away from normal face-to-face conversation along these dimensions, the more grounding is affected. Although it is not obvious how the features they identify apply to experimental conditions in my framework, the general idea is compelling and can be recruited for the current problem.

These perspectives point to a specific overarching view of language's role in conceptual alignment. In particular, I propose that the extent to which language affects conceptual alignment depends substantially on the richness of the form of language use available. The more that this form resembles full, natural, face-to-face conversation, the more people are able to exploit it to align their conceptualisations. This suggestion is reminiscent of Wilkes-Gibbs and Clark's (1992) proposal that there are different levels of shared information between speaker and hearer (or overhearer), depending on how intimately they are interacting.

Analogous predictions can be incorporated with respect to other relevant dimensions, such as time course, language proficiency, and non-linguistic information. In particular, conceptual alignment would be predicted to be strongest during the course of interaction (dissipating afterwards), for adult native speakers of a language, and with rich non-linguistic information (e.g., category grouping feedback, seeing each other's faces, etc.).

The results that I have obtained concerning the temporal and non-linguistic dimensions are broadly consistent with this proposal. While I did not test language proficiency, note that similar predictions have been made by Costa, Pickering and Sorace (2008) regarding alignment in non-native speakers.

The proposal has the advantage of making falsifiable predictions. As long as we could reasonably specify in advance how different conditions rank along the key dimensions above, we could test the model accordingly. And since my experimental framework could be readily applied to many different problems which explore this parameter space (see Section 9.2), the tools now largely exist to explore this further.

We can now also try to reincorporate lexical alignment, and what we have learned about its relationship to conceptual alignment. It is useful to recall a few key findings in this respect. First, we have seen that, at least for prior alignment, lexical alignment is much higher than conceptual alignment (for native speakers of the same language). Second, we have also seen that lexical alignment (at least of the identifier-based variety) seems to happen more easily and lasts longer than conceptual alignment. Third, despite the dissociation between lexical and conceptual alignment, conceptual alignment scores correlate strongly with lexical alignment scores.

In sum, putting these results together with my current proposal suggests the following further predictions for future work. Lexical alignment should normally be higher than conceptual alignment (if it were measured fairly based on both identifiers and items), and the closer they are to full alignment, the closer they will tend to be to each other, both in value (along the measure) and in correspondence (how they carve up the conceptual space).

As such, this proposal extends the discussion in Section 7.5.2. There I argued that the debate about the fit between lexical and conceptual alignment can be reconciled by considering the form and context of linguistic interaction. When people can resort to full dialogue, conceptual and lexical alignment will coincide to a high degree. But to the extent that the form of linguistic interaction at people's disposal is impoverished and lacking non-linguistic support, the two types of alignment dissociate from each other.

9.4 Linguistic relativity

My work also sheds light and suggests a certain perspective on the issue of linguistic relativity. Experiment 1 used cross-linguistic data to show how non-linguistic categorisation did not reliably reflect linguistic categorisation. However, it also revealed that speakers of the same language were more conceptually aligned than speakers of different languages. Nevertheless, I argued in Section 5.3.4.3 that this difference was probably not due to language, but rather other cultural factors.

Moreover, my other experiments also have complementary implications for this issue. Three clear conclusions that emerged from my findings were that there is substantial variation in how native speakers of the same language conceptualise things, that conceptualisation can change over a short period of time even within individuals (albeit perhaps temporarily), and that the associations between words and concepts are not fixed. To the extent that these claims are valid, they also cast doubt on a strong view of the effect of language on thought.

However, as discussed in Section 3.3.5, such a strong static view of linguistic relativity has already been generally abandoned, and replaced by weaker and more dynamic versions. In particular, Slobin (1996) suggested that perhaps our language only affects cognition when we are actively using it, and Kay and Kempton (1984) proposed that language is exploited as a strategy when solving otherwise difficult problems.

Notice that Experiments 3 and 4 could also be interpreted in these terms. In both experiments, participants were explicitly given a difficult problem (i.e., trying to coordinate their categories), for which language could be strategically used. Although only native English speakers were involved, the stimulus domain was so finely sampled that lexical patterns of individuals varied considerably. Therefore, we could think of participants as speakers of different dynamic idiolects, rather than a single monolithic language, and the experiments examine how they converge through interaction, both lexically and conceptually (Steels 1997). As Malt and Sloman (2004) pointed out, to the extent that linguistic relativity effects do exist, they must ultimately originate from some form of interaction with members of a language community. Experiments 3 and 4 could thus be interpreted as investigating how linguistic relativity effects could emerge through interaction. Since both experiments provided evidence for a linguistic manipulation impacting the degree of conceptual alignment between people, they show indirect support for a weak, dynamic view of linguistic relativity. In particular, our language does affect how we conceptualise things, but it does so online through linguistic use and interaction.

It is important to acknowledge that this perspective implicitly reconsiders how we think of the source of words in linguistic relativity investigations. A priori, it is important to distinguish whether we are examining the effects of hearing word input from others, or of the words internally stored in our minds. However, if we conclude that linguistic relativity effects are specific to instances of language use, and recognise that the fundamental arena of language use is social and interactive (Clark 1996), then the distinction about the source of words begins to break down. After all, even when a change in how we conceptualise something is brought about through the verbal input of someone else, it must still be processed through our own minds. For words to affect conceptualisation in explicit language use, they must still be triggered in the individual. How the triggering actually takes place, on the other hand, could be considered a separate issue, though also a potentially illuminating one.

Therefore, while I am suggesting that my experimental results do provide support for linguistic relativity, it is in the context of a particularly weak interpretation of it. While this may be consistent with previous anti-Whorfian findings in the object domain (e.g., Malt et al. 1999), it is substantially weaker than some of the conclusions that have come out of work in other areas (e.g, Levinson 2003; Roberson, Davies and Davidoff 2000). How do we reconcile these differences? I suggest that these differences may come from the nature of the domains and the cognitive constraints imposed by a language. The colour domain, for example, is inherently continuous and can in principle be divided up in any number of ways, which places language in a particularly fluid space. And different linguistically encoded spatial frames of reference require different information to be encoded, without which speakers could not express basic spatial propositions and hearers could not understand them. Objects, on the other hand, are relatively concrete, tending to cluster in natural kinds (Gentner and Boroditsky 2001), and do not need to be placed in a special cognitive framework to support communication. Therefore, it is not surprising that our native language does affect conceptualisation of objects too, but just not as deeply or pervasively as in some other areas.

9.5 Revisiting conceptual issues

In Chapter 2, I developed a simple model for the structures and processes involving concepts, and then had it ground the experiments of my thesis. Now that the experiments are behind us, it is time to reassess some of the major theoretical issues that underlay that development.

9.5.1 Concepts and categories

One important theme has been the distinction between concepts and categories. I argued in Section 2.4.1 that although the distinction between these is often blurred, that wasn't to happen here. Concepts determine categories, but they exist in the mind, while categories are constructs of the psychologist trying to make sense of them. However, despite being adamant about the distinction, I also followed the psychological route in using categorisation experiments to access people's concepts (see Section 4.2). Since we cannot access people's concepts directly, I argued, and concepts determine categories, then categories are still our best shot for accessing concepts. As a result, when I obtained categorisation outputs from pairs of participants and quantitatively compared them, I called the result a measure of "conceptual alignment". Should I have been calling it "category alignment"?

I think not, and stand behind my original decision. As I discussed in Section 8.5.3, for people to align their categories, they have to go through concepts. Each task involved a different subset of stimuli, so memorising the specific stimuli wouldn't do. Instead, participants had to try, at some level, to infer their partner's concepts, which in some cases was based on groupings, but in others (depending on their condition) would have involved labels or just a task score. Their subsequent categorisation output remained our best means of accessing the concept.

Notice that, in effect, Experiment 4 asked whether this reliance on the category as the primary indicator of a concept was also to be found in participants. When people have access to both categories and words, which do they manage better with for aligning their conceptualisations? The experiment found that participants primarily relied on the categories as well.

On the other hand, it's important to remember Frege's (1892/1948) lesson: multiple concepts can determine the same category, and so a category underdetermines its concept. In other words, just because two people produce the same category groupings, doesn't mean that they actually have identical concepts. In fact, perhaps if we actually discovered how to compare concepts without resorting exclusively to category information, we might find that some other factor outweighs it, and that we have been getting misleading comparisons all along.

Until then, however, we should continue using categorisation to access concepts. Both the literature and experiment participants suggest that it is the best way. Indeed, we should not shy away from interpreting such tasks conceptually, lest we revert to limiting forms of behaviourism (Malt 2006).

9.5.2 *Concepts, conceptualisation and variation*

One of the first tasks of this thesis was to dismiss the assumption of conceptual universals. Since my thesis is in psychology, that was fairly easy. In fact, I did not even have to prove that variation exists, but only argue that we must not assume universals a priori. And conceptual variation in my experiments, though sometimes very subtle, was pervasive, strengthening this position further. Indeed, the experiments suggest that concepts are very dynamic and flexible, although perhaps not as flexible as words.

This has the potential to push us towards a radical theoretical position, in which there are no fixed, static concepts (Smith and Samuelson 1997). Instead, we may conclude that concepts are "created at the moment of use" (Croft and Cruse 2004, p. 75). This, however, would be unfortunate, as it would lead us into a host of new problems, not the least of which concerns what concepts would actually be created from (Hurford, personal communication). Although there are proposals of how to deal with such dynamic concepts, such as the recruitment of rich perceptual representations (Barsalou 1999; Prinz 2002), they are still relatively new and radical.

A potential alternative solution comes from studies of how word meanings can be modulated by context (Murphy 2002). [1] For instance, Roth and Shoben (1983) got participants

[1] Note that since I merged word meanings with concepts (see Section 2.4.3.1), I can readily recruit such findings for a view of concepts as well.

to give typicality ratings on the interpretation of "beverage" in different contexts. They found that the word tended to be interpreted quite differently in the contexts of secretaries or truck drivers, for example. More generally, Cruse (1986) claims that "A single sense can be modified in an unlimited number of ways by different contexts, each context emphasising certain semantic traits, and obscuring or suppressing others" (p. 52). In other words, here we are having the best of both worlds: many subtly different meanings but subsumed under one unifying sense.

However, we could take this further. In fact, we could even try to take it as far as Fodor (1998)'s conceptual universals. Recall that Fodor (1998) had argued that conceptual variation was unacceptable, because then we would never be able to say that we are talking about the same thing. If we were feeling really adventurous, we could try to re-adopt universal concepts seriously and just stipulate that they undergo massive contextual modulation all the time. This could then account for all kinds of apparent conceptual variation, including the variation in my experiments.

Trying to resolve this issue would be a mess, however, without more theoretical tools. Fortunately, we can largely bypass this issue by focusing on instances of conceptualisation rather than concepts themselves. By doing so, we take a snapshot of a concept of a particular person at a particular time, and are not bound to make any theoretical claims about the integrity of the concept across people or time. Moreover, as we can see by comparing Figures 2.5 and 2.8 (reproduced here as Figure 9.1), a particular act of conceptualisation (in contrast to a concept in general) also has the advantage of having just one specific word associated with it (if any), and normally identifies real objects in the world. As such, it is easy to compare two people's conceptualisations, provided they are dealing with the same objects.

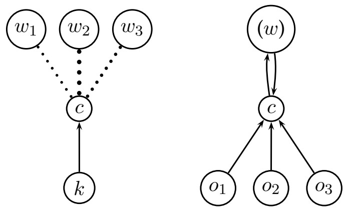


Figure 9.1: A concept's relationship with words and its category (left), versus an act of conceptualisation labeled by a particular word and applied to particular objects (right).

This is what my experiments and framework have sought to do. Conceptual snapshots were taken of people at different times, and compared with their experimental partners. By doing so, I bypassed the problem of conceptual identity or development altogether. While this problem should ultimately be resolved, we can go a long way without doing so, and thereby avoid all kinds of philosophical turmoil.

9.5.3 Concepts and words

The results of my experiments have shown that concepts and words do not seem to be as intimately related as is commonly assumed. In Experiment 1, people's non-linguistic sorting was shown to differ from the way they named things, consistent with previous findings (Malt et al. 1999; Ameel, Storms, Malt and Sloman 2005). Experiments 3 and 4 complemented these findings by showing that even when people aligned their lexical choices, this did not necessarily correspond to aligning conceptually as well. This corroborates results which have shown that people's interpretations of words can be surprisingly divergent (Schober, Conrad and Fricker 2004). Experiment 4 in particular assessed the correspondence between lexical and conceptual alignment, and revealed that it was not straightforward. Sometimes people used the same labels but categorised completely differently, and conversely, sometimes they would categorise identically despite using different labels.

Nevertheless, this dissociation should not be overstated. As was seen in Figure 8.10, despite all of the divergent examples found in Experiment 4, these were still a minority. Indeed, there was still a highly significant correspondence between words and concepts. Lexical alignment depended only on label feedback, and conceptual alignment depended mostly on grouping feedback, but they were still closely related. As such, despite all the variation, the connection is still constrained (Geeraerts 1993). We certainly don't just say anything to refer to a particular item, and the more we diverge from the prototypical meaning of a word the harder it is for listeners to process what we say and understand us (Garrod and Sanford 1977).

However, the relationships between words and concepts are arbitrary and must be learned, so it should not be surprising that they are less reliable than the deterministic connections between concepts and categories. Words may appear to be our best cue of others' concepts in our everyday lives, because we do not always have access to snapshots of their categories (as in my experiments), especially for abstract concepts. Nevertheless, the cases when we do are illuminating. Consider the example of going to the supermarket and looking for lemon juice. Do you look for the aisle with a certain label (e.g., "Juice"), or do you look for the kinds of things that are normally found near lemon juice? Strategies no doubt differ and their relative effectiveness will depend on the particular situation (e.g., going to a supermarket in China without knowing any Chinese) but in any case, we are certainly not at the mercy of words all the time. These considerations have implications for practical applications in the real world, from the labelling of recycling bins to the marketing of new car models.

Nevertheless, my results also add to our understanding of how hearing the words of others can affect our conceptualisations. Although this was particularly true when dialogue was available (Experiment 3), it also applied to a lesser extent when only category labels were exchanged (Experiment 4). These results are corroborated by previous findings (see Section 3.2). Some experiments have shown how conceptualisation or memory of visual stimuli can be affected by the linguistic label (if any) they are presented with (e.g., Carmichael, Hogan and Walters 1932; Billman and Krych 1998; Feist and Gentner 2007). More recently, labels were shown to enhance category learning even when completely redundant (Lupyan, Rakison and McClelland 2007), and there is ample evidence that such effects originate very early in development (e.g., Waxman and Markow 1995; Plunkett, Hu and Cohen 2008). At the same time, it should be acknowledged that these effects are generally subtle. As we have seen, language can help, but it seems to be too "sketchy" to determine thought (Gleitman and Papafragou 2005, p. 636).

Consider now the opposite direction. Recall that my experiments, and Experiment 4 in particular, did not show any evidence that access to other's category groupings leads to lexical alignment. Even when participants could see how their partner categorised the items and were able to conceptually align with them, they did not align lexically as well without lexical feedback. Therefore, they did not seem to be converging conceptually by determining what lexical concept their partner had in mind, but did so independently of language. This exemplifies how "conceptual preparation" in lexical production is not automatically determined by the referent, but depends on a speaker's intention and conceptualisation (Levelt, Roelofs and Meyer 1999, p. 3), and how there are often many reasonable ways of referring to things (Brennan and Clark 1996; Malt and Sloman 2004). The human mind is very flexible, of which our use of language is a particularly good example.

So how should we think of the dynamic relationship between words and concepts? Geeraerts (1993) offered the metaphor of a floodlight which lights up slightly different areas with different applications (see Section 2.4.3.2). However, this view does not naturally mesh with the fact that words enter into all kinds of semantic relationships with each other, while at the same time moving dynamically on top of conceptual space. I therefore propose the metaphor of a fishing net floating on the surface of the sea, while anchored to the bottom, with each knot in the net representing a word. As the waves move up and down, the net gets slightly displaced, and different weather conditions affect how it moves, stretches and makes contact with the sea. Similarly, words relate to concepts in a dynamic way, and the relationships depend on various factors. It does not mean that variation is rampant or random. As long as your sea is similar enough to mine, and you anchor and build your net in a similar way, we should still be able to communicate effectively (Churchland 1998). Indeed, the plausibility of such variation is supported by psychologically motivated multi-agent computer simulations which have yielded high communicative success despite substantial divergence between individuals' conceptual systems (Smith 2005).

Acknowledging the dynamic and variable nature of the relationship between words and concepts has implications for theories of meaning. First of all, although it can be tempting to ground meaning in universal objective categories (Putnam 1975), ultimately this cannot be maintained. Putnam (1981) was right when he abandoned his original view, because meanings "just ain't in the world". Similarly, theories which place meaning in the mind but rely on fixed, universal concepts are also problematic (Fodor 1975; Jackendoff 1983). Indeed, while some linguists have stressed semantic diversity across languages (Evans and Levinson 2009), we need to additionally consider such diversity within languages. To that end, linguists should focus on semantic theories which embrace variation and situate meaning where it belongs, in the mediating and dynamic human mind. Such views are emphasised and developed in cognitive linguistics (Lakoff 1987; Langacker 1987; Croft and Cruse 2004).

9.5.4 Internal structure

Earlier on in my thesis (see Section 2.5), I argued for and adopted a hybrid view of internal conceptual structure, based on a fusion of prototype, theory and exemplar theories. And in discussing Experiment 4, I interpreted some of my results within this view (see Section 8.5.3). The conclusion, although tentative at this point, was primarily in favour of exemplar theory. Here, I do not reopen that analysis, but only briefly discuss a more general point.

In particular, this thesis demonstrates how we can fruitfully investigate the interface between the internal and external relationships of concepts, bringing together insights from different disciplines. Philosophers, linguists and semioticians have long studied how words seem to relate to both categories of things in the world, and concepts in the mind. But in fact, analogous relationships exist in the human mind as well. Concepts seem to be composed of relatively abstract senses together with stores of concrete exemplars. With some important modifications, senses correspond to what concepts or meanings have traditionally been considered to be (Frege 1892/1948). Exemplars, on the other hand, are active in our mental models of the actual things in the world being (or having been) conceptualised (Johnson-Laird 1983). Both are stored in the mind, and both play important roles in conceptual processes. Moreover, while word forms are public, they too are stored in our minds, and are associated with concepts. Therefore, while semiotic triangles (see Section 2.4.4), for example, have traditionally been used to visualise the relationships between words, concepts and categories externally to the mind, they are also instantiated in a very real way inside each person's mind. However, they take on slightly different forms for each individual, and change over time. By integrating these mind-internal and mind-external perspectives from different disciplines, we can make important new steps in our understanding of language, mind and meaning.

9.6 Conceptual development, language acquisition, and language evolution

Although my experiments have used familiar stimulus domains, they clearly have an important element of learning. In Experiments 3 and 4 in particular, participants were asked to try to coordinate their categorisation as closely as possible with that of their partner. While we can argue about whether this involves conceptual acquisition, permanent change, or temporary modulation (see Section 9.5.2), some kind of learning is required, whether implicitly or explicitly. Indeed, in Section 2.3.1, I reviewed how concepts seem to vary at several different timescales. My experiments have dealt with intercultural and online levels of variation, but a full account of language's role in conceptual variation would also encompass the developmental and evolutionary timescales. As such, we should consider how the methods, findings and conclusions I have presented here relate to developmental and evolutionary issues. Unfortunately, I do not have the space to do justice to these massive themes, but I must at least accommodate a couple of brief remarks.

An important question in children's development concerns the relationship between conceptual development and language acquisition. Which comes first, the concept or the word? A lot of research has been devoted to the topic, and the emerging answer is not simple. Concepts do not require language (Bermúdez 2003), and even pre-linguistic infants have substantial conceptual abilities (Hespos and Spelke 2004). However, words do help shape and augment conceptual development, although there are different perspectives on the mechanisms and extent of that influence (Bowerman and Levinson 2001; Carey 2009).

This general consensus is in line with my position that conceptualisation does not rely on language, and yet can be affected by it. Moreover, my experimental framework offers new ways in which these issues can be investigated. Because the framework is explicitly designed not to rely on language and does not presuppose what specific concepts children have, we could readily test whether children align, both lexically and conceptually, and whether feedback or dialogue would have a different effect on them. Based on the view that I have developed, I would predict that language should have a relatively small effect on conceptual alignment in children, and that conceptual and lexical alignment should be more dissociated than in adults, since language is not yet as deeply entrenched in their minds. In fact, there is evidence that children align readily at the lexical level, but are relatively unconcerned when they discover an underlying conceptual misalignment (Garrod and Clark 1993).

Similar issues exist concerning the emergence of language on an evolutionary timescale. Although simple signaling systems do exist in some animal species, such as vervet monkeys (Seyfarth and Cheney 1990), human language appears to be unique in having an enormous number of arbitrary, symbolic and learned associations between signals and concepts. So how did words first emerge, and what effect did they have on the evolution of concepts? Again, researchers have made different claims concerning whether our hominid ancestors already had "proto-concepts" (Hurford 2007), or whether conceptual abilities were radically transformed with the emergence of a biological capacity for symbols (Deacon 1997). However, these positions agree that language helped transform the conceptual landscape.

While the above perspectives have offered biological explanations for language evolution, an alternative approach has suggested that many features of human language and cognition could have cultural origins (Tomasello 2000). Although this view is still dependent on a base of biological prerequisites for language to be in place (Hurford 1999), it argues that the cognitive abilities which emerge in human ontogeny, together with the social nature of human interaction, have themselves radically transformed our minds. As a result, some researchers have begun to study the cultural emergence of linguistic structure in the laboratory using chains of human participants (Kirby, Cornish and Smith 2008). While such experiments have generally assumed a uniform conceptual space, continuous meaning spaces (as in my experiments) have also been adopted recently with success (Matthews, Kirby and Cornish 2010).

My experimental framework could be applied to both biological and cultural explorations of the role of language in conceptual evolution. Since free classification tasks again do not rely on language or presuppose particular concepts, they are suitable for comparative studies, and have been used with monkeys and apes (Spinozzi 1996). Experiments could be conducted to see whether they categorised differently when hearing or uttering communicative signals, whether they be those of their own species, human words, or symbols acquired in previous training. It seems likely, though, that the relative lack of cognitive and signal flexibility in non-human primates would lead to weaker alignment in apes than in humans. On the cultural side, human cultural transmission experiments could investigate whether conceptual alignment could carry over and grow from one generation to the next. In effect, this has parallels with dialogue experiments where participants change partners (Garrod and Doherty 1994; Brennan and Clark 1996; Malt and Sloman 2004). Based on those studies and my own, I would predict that language should affect conceptual alignment across such transmission chains, but that it would depend on the richness of the linguistic interaction and shared perceptual information between participants.

9.7 Conclusion

This thesis began with a seemingly simple question: does language bring about conceptual coordination? In a way, the answer seems obvious. We use language to communicate with others, coordinate our behaviour in joint projects, and express our perspectives on all kinds of things: so how could language not coordinate conceptualisation? On the other hand, there are also good reasons for being suspicious of such a conclusion. In particular, given the private nature of concepts, we cannot assume a priori that words and concepts fall into neat relationships that are uniform within language communities. But if that uniformity turns out to be lacking, then perhaps language's apparent effectiveness at coordinating conceptualisation may be an illusion, and this warrants empirical study.

The experimental results confirm these suspicions. People with the same native language divide up the world into very similar lexical categories, and when they communicate, they can adopt each other's words with ease. But if we look closely enough, that does not seem to be generally reflected in how they conceptualise things: coordination of words does not imply coordination of concepts. In fact, conceptual coordination seems to benefit much more from seeing how someone sorts things into categories rather than how they label them. Indeed, in this sense, people's priorities seem to parallel those of cognitive psychologists, who assume a close correspondence between concepts and categories, while shying away from trying to relate words and concepts (Murphy 2002).

Yet that is not the whole story. Under some circumstances, language certainly does help conceptual coordination. But it is not trivial, and depends on the form of language use and the degree of shared information. In fact, if language is used in full natural conversation, it can greatly increase conceptual coordination. And to a smaller extent, coordination can also improve even with more impoverished forms of language, as long as it's supported by alternative forms of feedback. But some form of grounding is essential.

As I have suggested throughout this chapter, a lot of work remains to be done to test the emerging conclusions and explore the issues further. Moreover, the theoretical and methodological challenges that needed to be faced in order to get this project off the ground have resulted in other forms of contribution as well. In particular, a theoretical model of conceptualisation and conceptual coordination was developed, along with a corresponding experimental framework, both of which could be fruitfully recruited in addressing many different research questions. In addition, although I have explored various implications of my findings, limited space has not allowed me to do them justice, so that I have barely touched on relevant themes like bilingualism, artificial intelligence, semiotics and sociolinguistics.

Nevertheless, the main result is clear. Our language does not determine the concepts that we have, and language use does not automatically bring people's concepts together. Merely throwing words around is generally useless. However, if language is set free in its full forms of interactive usage, or is supported by rich alternative sources of information exchange, then it can bridge the gap in the way we see the world.

APPENDIX A

Experiment 1 materials

A.1 Instructions

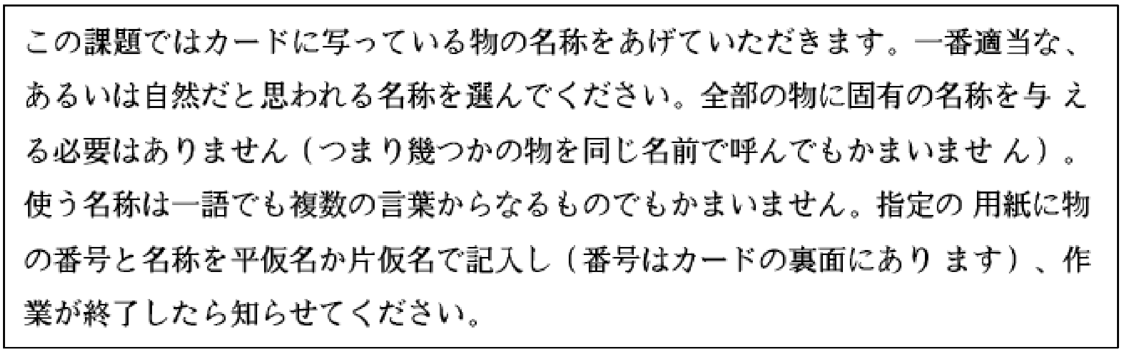
Naming task

Below are the instructions for the naming task in the three languages.

English

> In this part of the experiment, your task is to individually label each of the object pictures. Please choose a name that seems best or most natural to you. You do not need to use a unique name for each object (i.e., it's fine to call multiple items the same thing). The name can consist of one or more words. Please write the number and name on the form provided (the object's number is written on the back of its card). When you're finished, please let me know.

Japanese



この課題ではカードに写っている物の名称をあげていただきます。一番適当な、あるいは自然だと思われる名称を選んでください。全部の物に固有の名称を与 える必要はありません（つまり幾つかの物を同じ名前で呼んでもかまいませ ん）。使う名称は一語でも複数の言葉からなるものでもかまいません。指定の 用紙に物の番号と名称を平仮名か片仮名で記入し（番号はカードの裏面にあり ます）、作業が終了したら知らせてください。

Polish

W tej części eksperymentu zadanie polega na nazwaniu każdego z obiektów, które widzisz na obrazkach. Wybierz nazwę, która wydaje Ci się najlepsza lub najbardziej naturalna. Nazwy mogą sie powtarzać i mogą składać się z jednego lub kilka słów. Na formularzu zamieszczonym poniżej zapisz, wybraną przez Ciebie nazwę obiektu i jego numer (znajdziesz go z drugiej strony obrazka). Poinformuj mnie, proszę, kiedy skończysz.

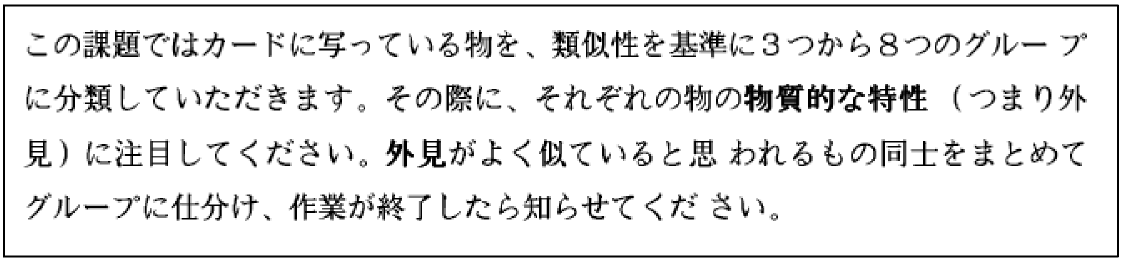
Sorting task

Below are the instructions for the sorting task in the three languages.

English

In this part of the experiment, your task is to arrange the object pictures into 3 to 8 groups based on similarity. Please focus on the **physical qualities** of each object (i.e., what it looks like). Put together into piles all the objects that you think are very similar to each other **physically**. When you're finished, please let me know.

Japanese



この課題ではカードに写っている物を、類似性を基準に３つから８つのグルー プに分類していただきます。その際に、それぞれの物の**物質的な特性**（つまり外見）に注目してください。**外見**がよく似ていると思 われるもの同士をまとめてグループに仕分け、作業が終了したら知らせてくだ さい。

Polish

W tej części eksperymentu zadanie polega na uporządkowaniu obiektów w grupy (powinno grup być nie mniej niż 3 i nie więcej niż 8). Skup sie na **cechach fizycznych** (wyglądzie) tych przedmiotów - przyporządkuj do tej samej grupy obiekty, ktore wydają Ci sie bardzo podobne **wyglądem**. Poinformuj mnie, proszę, kiedy skończysz.

A.2 Questionnaire

Participants were asked for their name, age, gender, study program (i.e.,. degree, subject, year), their length of stay in the UK, the languages they spoke (with self-rated proficiencies), the countries they had lived in (with the number of years), and their familiarity with the stimuli ("How familiar were the objects to you?") on a scale from 1 (very unfamiliar) to 7 (very familiar).

APPENDIX B

Experiment 2 materials

B.1 Instructions

Below are the instructions provided to participants, first for the interface in general, and then for each task. They are divided up by condition where appropriate. Note that interface instructions also came with a snapshot of the categorisation interface as it appears at the beginning of the training task.

Interface description

The tasks in this experiment involve sorting pictures of everyday objects into categories. In each task, you will sort the items into categories using the interface shown at right.

Your job in each task will be to put each item from the black pool of items in the bottom right into one of the nine blue category boxes on the left, and to label the categories. You do not need to use all of the boxes, but please use at least three, with a maximum of sixteen items per category (the program will warn you if you try to put more). There are 6 basic actions you can perform, described below.

A paper copy of these instructions has also been provided to you. Please feel free to refer to it at any time.

Actions

Naming a category Before putting items into a category, you must give the category a name. This is done by typing a name in the place provided beneath each category.

Adding an item to a category First click on the item, which will move it into the green focus box in the top right and magnify it for convenience. Then click on the 'Add' button underneath the category of your choice.

Magnifying a category You can magnify an entire category (or the pool) by clicking on the 'Big' button beneath it. This will show you all the items in that category in a larger size.

Changing the focus item If you have put an item in the focus box but change your mind and want to categorise a different item, simply click on another item. The original item will return to where it came from, and the new item will appear in the focus box.

Renaming a category You can rename a category at any time by simply editing the name you gave in the place below the category.

Changing the category of an item First click on the item, which will move it into the green focus box, and then click on the 'Add' button underneath the category you want to switch to.

Training task

The first task is a short practice task. Please put all the items you will see in the black pool of items into the blue category boxes, and name the categories.

Since the main purpose of this practice task is for you to become fully acquainted with the interface, please try out each of the six operations at least once (i.e., naming a category, adding an item to a category, magnifying a category, changing the focus item, renaming a category, and change the category of an item). The program will remind you at the end if you forget any of them.

First task

It’s time to start the first task. Please put the items into categories, and name the categories. Remember that you CANNOT speak with the other participant.

Second task

Control condition

It’s time to start the second task. Please put the items into categories, and name the categories. Remember that you CANNOT speak with the other participant.

Silent condition

It’s time to start the second task. In this task, you will again be putting items into categories and naming the categories, but this time you will be working **with a partner**. Remember that you CANNOT speak with your partner.

You and your partner will take turns at putting the items into categories. It is your turn when the focus window background is green, and your partner’s turn when it is red. A turn ends when you put an item in a category (i.e., when you click on an ‘Add’ button). You cannot carry out any actions when it is not your turn (other than observing your partner’s changes). As before, you can recategorise items or rename categories, including those categorised or named by your partner. The task ends when the pool of items is empty, neither partner wants to make any more changes, and both partners click on ’Done’ in succession.

Talking condition

> It’s time to start the second task. In this task, you will again be putting items into categories and naming the categories, but this time you will be working **with a partner**. This time, you MAY speak to your partner: please do so whenever you feel it would help.
>
> You and your partner will take turns at putting the items into categories. It is your turn when the focus window background is green, and your partner’s turn when it is red. A turn ends when you put an item in a category (i.e., when you click on an ‘Add’ button). You cannot carry out any actions when it is not your turn (other than observing your partner’s changes). As before, you can recategorise items or rename categories, including those categorised or named by your partner. The task ends when the pool of items is empty, neither partner wants to make any more changes, and both partners click on ’Done’ in succession.

Third task

Control condition

> It’s time to start the third task. Please put the items into categories, and name the categories. Remember that you CANNOT speak with the other participant.

Silent and talking conditions

> It’s time to start the third task. In this task, you will again be working on your own. Please put the items into categories, and name the categories. Remember that you CANNOT speak with the other participant.

B.2 Questionnaire

Participants were asked for their name, age, gender, nationality, study program (i.e.,. degree, subject, year), whether they knew their partner, and what languages they spoke (with self-rated proficiencies). The feedback questions on the questionnaire were as follows:

1. What do you think was the purpose of the experiment?
2. On what basis did you categorise the pictures?

3. Did this basis change between the different tasks?
4. How did you agree with your partner on categories for the pictures? pictures?
5. When you looked at an item, did you mentally label it with an English word?
6. How familiar were the pictures to you? (Options: very familiar, familiar, neutral, unfamiliar, very unfamiliar)
7. How natural-looking were the pictures? (Options: very natural, natural, neutral, unnatural, very unnatural)
8. Did you notice differences between the sets of pictures from the different tasks?
9. Do you have any comments about the program interface?
10. Do you have any other comments about the experiment?

B.3 Program interface

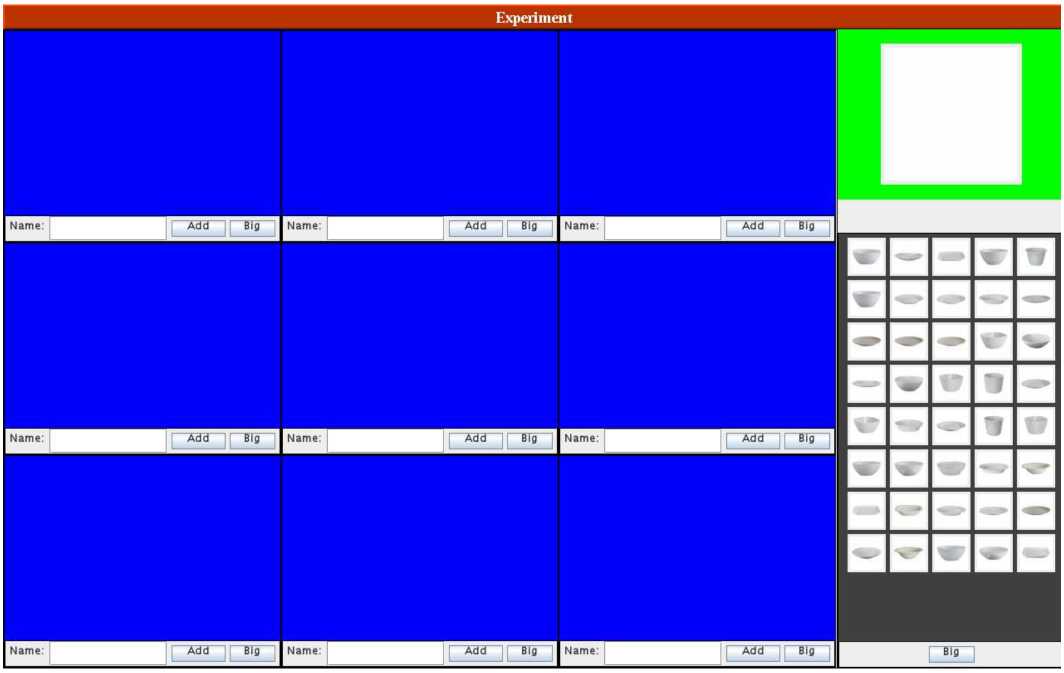
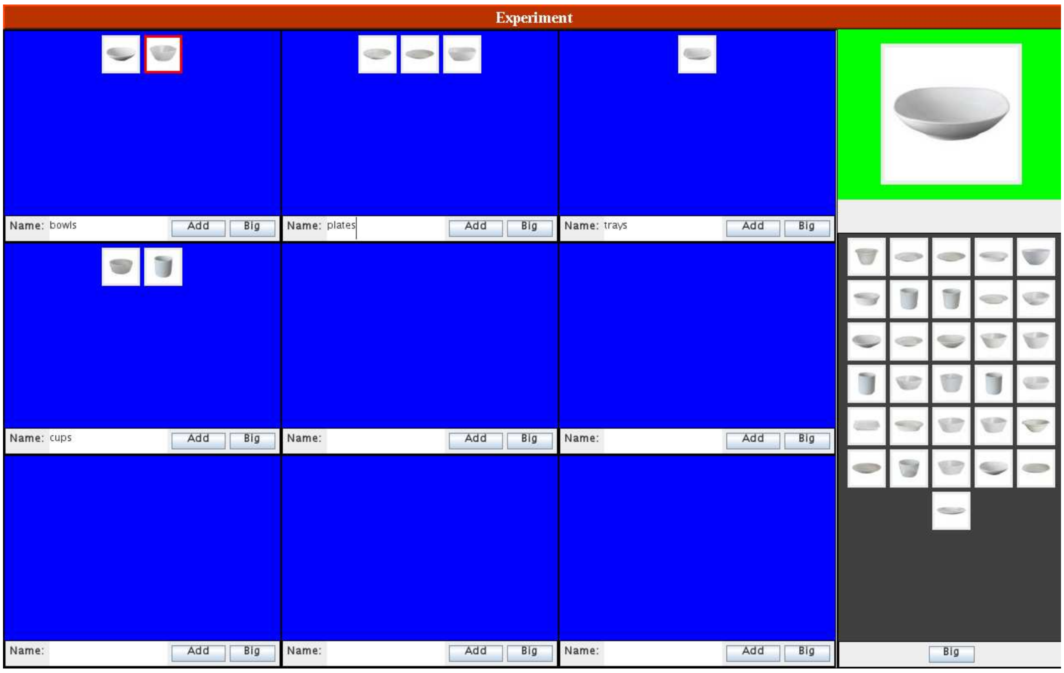


Figure B.1: Screenshot of the program interface at the beginning of a task.



Name: bowls

Name: plates

Name: trays

Name: cups

Name:

Name:

Name:

Name:

Name:

Figure B.2: Screenshot of the program interface mid-way through a task.

APPENDIX C

Experiment 3 materials

C.1 Instructions

Below are the instructions provided to participants, for the interface in general, the training task, the sequence of joint tasks, the similarity tasks, and the individual categorisation tasks. They are divided up by condition where appropriate. Note that interface instructions also came with a snapshot of the categorisation interface as it appears at the beginning of the training task.

Interface description

The main tasks in this experiment involve sorting pictures of everyday objects into categories. In each task, you will sort the items into categories using the interface shown at right.

Your job in each task will be to put each item from the black pool of items in the bottom right into one of the four blue category boxes on the left, and to label the categories. You do not need to use all of the boxes, but please use at least 2, with a maximum of 7 items per category (the program will warn you if you try to put more). There are 6 basic actions you can perform, described below.

A paper copy of these instructions has also been provided to you. Please feel free to refer to it at any time.

Actions

Naming a category Before putting items into a category, you must give the category a name. This is done by typing a name in the place provided beneath each category.

Adding an item to a category First click on the item, which will move it into the green focus box in the top right and magnify it for convenience. Then click on the 'Add' button underneath the category of your choice.

Magnifying a category You can magnify an entire category (or the pool) by clicking on the 'Big' button beneath it. This will show you all the items in that category in a larger size.

Changing the focus item If you have put an item in the focus box but change your mind and want to categorise a different item, simply click on another item. The original item will return to where it came from, and the new item will appear in the focus box.

Renaming a category You can rename a category at any time by simply editing the name you gave in the place below the category.

Changing the category of an item First click on the item, which will move it into the green focus box, and then click on the 'Add' button underneath the category you want to switch to.

Training task

The first task is a short practice task. Please put all the items you will see in the black pool of items into the blue category boxes, and name the categories.

Since the main purpose of this practice task is for you to become fully acquainted with the interface, please try out each of the six operations at least once (i.e., naming a category, adding an item to a category, magnifying a category, changing the focus item, renaming a category, and change the category of an item). The program will remind you at the end if you forget any of them.

Joint categorisation tasks

Okay, it's time to start the real thing.

There will now be 10 tasks which involve sorting pictures of shapes into categories. In these tasks, you will work with a partner. Your partnership will get a score for each task, and **the partnership with the highest total scores will receive a bonus 5 pounds each**.

The goal in each task is to **form categories that are as similar as possible to those of your partner**. The order and names of the categories do not matter for the score: what is important is **the item groupings**. Note that, in each task, your partner will be categorising the same items as you, but will initially see them in a different order on the screen.

Control condition

At the end of each task, the screen will show you your **partnership score** for that task. However, **you will NOT see your partner's category groupings or category names**.

After 20 seconds, the next task will begin.

Please remember that you **CANNOT** speak with the other participant.

Good luck!

Silent condition

At the end of each task, the screen will show you your **partnership score** for that task. **You will also see your partner's category groupings and category names beneath your own.** After 20 seconds, the next task will begin.
Please remember that you **CANNOT** speak with the other participant.
Good luck!

Talking condition

At the end of each task, the screen will show you your **partnership score** for that task. **You will also see your partner's category groupings and category names beneath your own.** After 20 seconds, the next task will begin.
This time, you MAY speak to your partner: feel free to do so when you feel it would help.
Good luck!

Similarity tasks

Interface

The next part of the experiment consists of similarity judgments. You will see pairs of pictures, one pair at a time, and are asked to judge pair based on how similar you find the two pictures, on a scale of 1 (very dissimilar) to 7 (very similar).

Training

There will now be 2 pairs of pictures for practice. Please judge the degree of similarity between each pair.

Main sequence

There will now be 60 pairs of pictures. Please judge the degree of similarity between each pair.
Please remember that you **CANNOT** speak with the other participant.

Individual categorisation tasks

> Great, you're done judging similarity. Now there will be a categorisation task. Please put the items into categories, and name the categories. The interface will be similar to the one you used earlier, except that this time there will be more pictures. Please use at least 2 categories, with a maximum of 20 items per category.
> Please remember that you **CANNOT** speak with the other participant.

C.2 Questionnaire

The participants were asked the same demographic questions as in Experiment 2. The feedback questions on the questionnaire were as follows:

1. What do you think was the purpose of the experiment?
2. On what basis did you categorise the pictures?
3. Did this basis change between the different tasks?
4. How did you try to agree with your partner on categories for the pictures?
5. On what basis did you assess the similarity between pairs of pictures?
6. When you looked at an item, did you mentally label it with an English word?
7. How familiar were the pictures to you? (options: very familiar, familiar, neutral, unfamiliar, very unfamiliar)
8. How natural-looking were the pictures? (options: very natural, natural, neutral, unnatural, very unnatural)
9. Do you have any comments about the program interface?
10. Do you have any other comments about the experiment?

C.3 Program interface

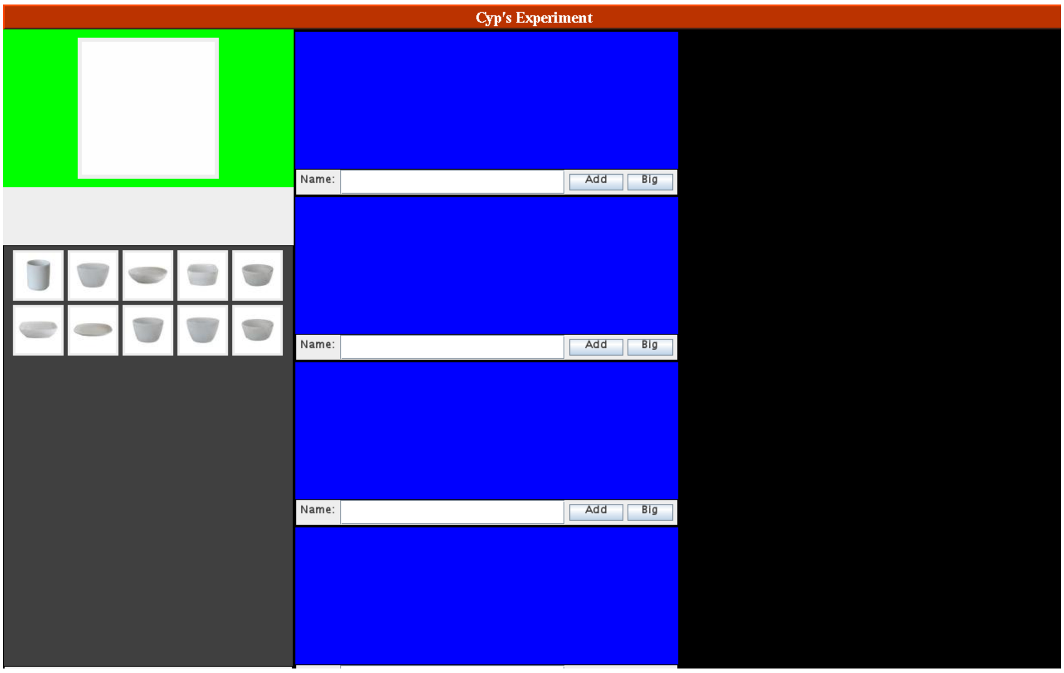
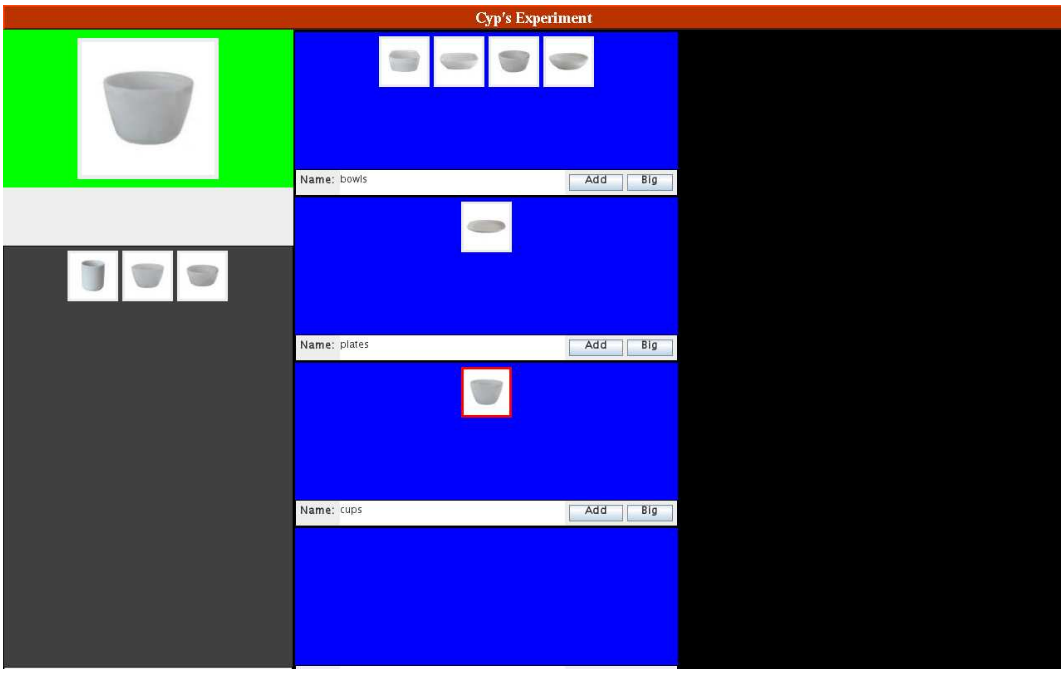


Figure C.1: Screenshot of the program interface at the beginning of a task.

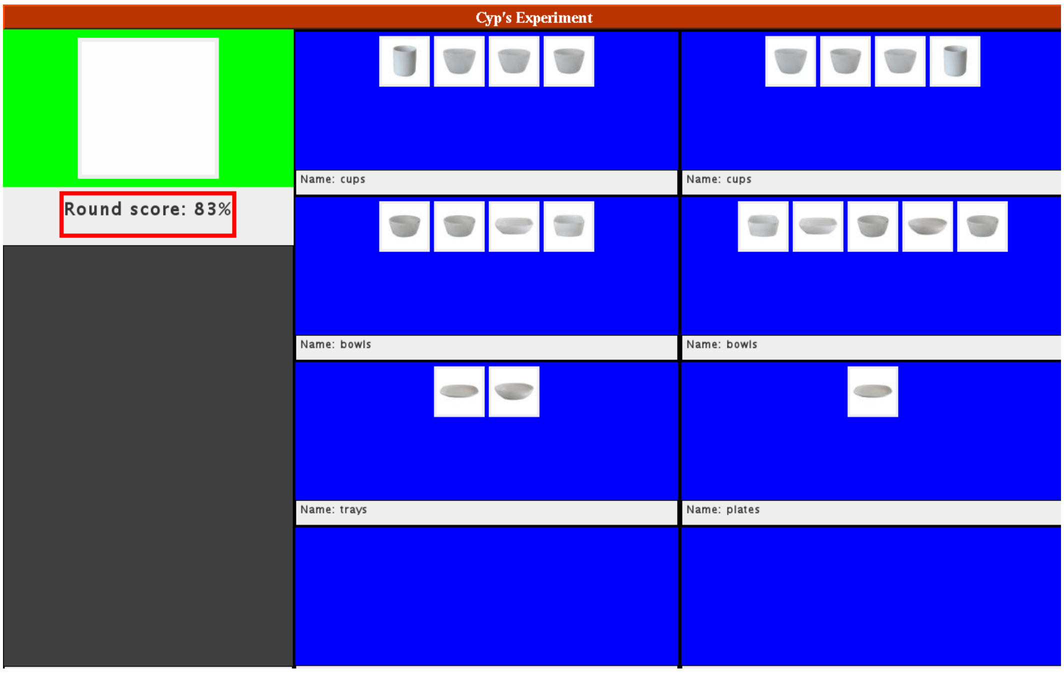


Name: bowls

Name: plates

Name: cups

Figure C.2: Screenshot of the program interface mid-way through a task.



Round score: 83%

Name: cups

Name: cups

Name: bowls

Name: bowls

Name: trays

Name: plates

Figure C.3: Screenshot of the program interface at the end of the task. This example illustrates what happens in the *silent* and *talking* conditions, since both kinds of feedback are provided.

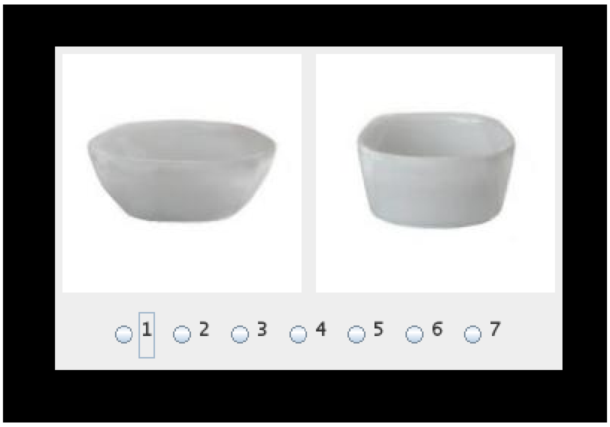
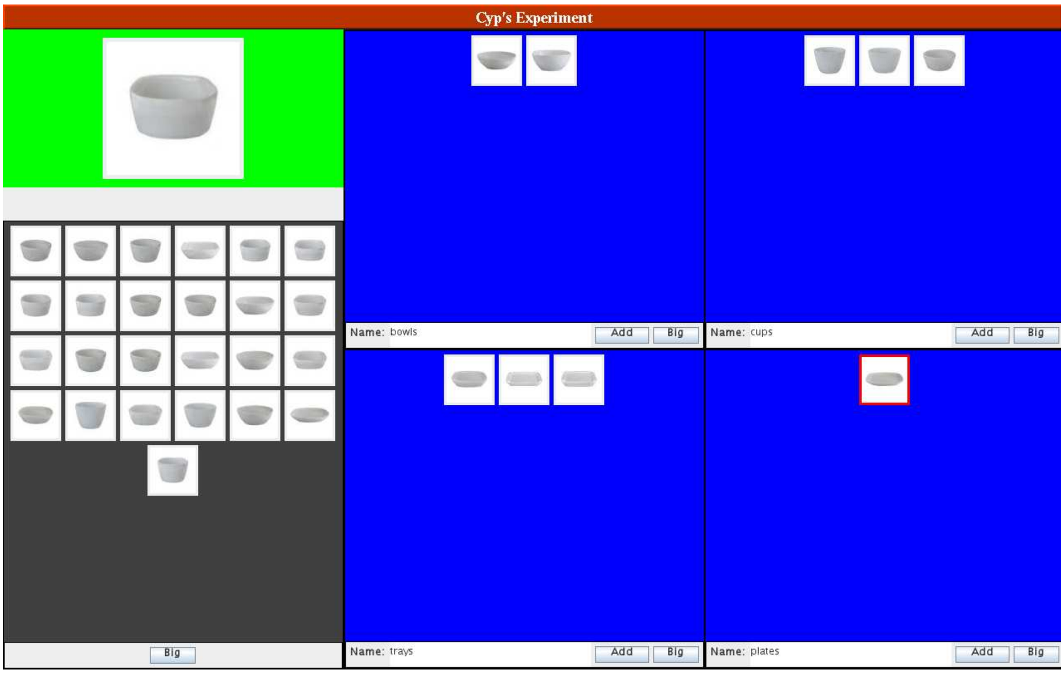


Figure C.4: A screenshot of the program during the similarity judment tasks. Participants rated silmilarity betwen item pairs on a scale of 1 (very dissimilar) to 7 (very similar).


Cyp's Experiment

Name: bowls

Name: cups

Name: trays

Name: plates

Figure C.5: A screenshot of the program at the beginning of the individual catetegorisation task.

APPENDIX D

Experiment 4 materials

D.1 Instructions

Below are the instructions provided to participants, first for the interface in general, then for the training task, and finally prior to the sequence of joint tasks. They are divided up by condition where appropriate. Note that interface instructions also came with a snapshot of the categorisation interface as it appears at the beginning of the training task.

Interface description

In this experiment, you will sort pictures into categories using the interface shown at right. A paper copy of these instructions has also been provided to you: please feel free to refer to it at any time.

In each task, you need to put each of the eleven pictures in the black area on the left into one of the two category boxes in the top right. You also need to name the categories. You can put a maximum of eight items in one category. The actions you can perform with the interface are described below.

Actions

Naming a category Before putting items into a category, you must give the category a name. This is done by typing a name in the place provided above each category. Category names must be in lower-case, and be valid words in a dictionary (the program will check and remind you otherwise).

Putting an item in a category Click on the item to categorise it. Click with the LEFT MOUSE BUTTON to put the item in the category on the left side, and with the RIGHT MOUSE BUTTON to put it in the category on the right side.

Renaming a category You can rename a category at any time by simply editing the name you gave it previously.

Changing the category of an item You can change your category decisions by clicking on the item with the appropriate mouse button, just as you would for initially categorising it.

Confirming your categories Once you have categorised all the items, a 'Done' button will appear on the left. When you are happy with your categories (i.e., you don't want to rename or recategorise anything), click on the 'Done' button to finish the task.

Training task

The first task is for practice. Please put all the items you will see on the left into the blue category boxes, and name the categories.

Since the main purpose of this practice task is for you to become fully acquainted with the interface, please try out each of the operations at least once (i.e., naming a category, placing an item in a category, renaming a category, changing an item's category, and confirming your categories). The program will remind you at the end if you forget any of them.

Joint categorisation tasks

Okay, it's time to start the real thing.

There will now be 30 tasks which involve sorting pictures of shapes into categories. In these tasks, you will work with a partner. Your partnership will get a score for each task, and **the partnership with the highest total scores will receive a bonus 5 pounds each**.

The goal in each task is to **form categories that are as similar as possible to those of your partner**. The order and names of the categories do not matter for the score: what is important is **the item groupings**. Note that, in each task, your partner will be categorising the same items as you, but will initially see them in a different order on the screen.

Neither condition

At the end of each task, the screen will show you your **partnership score** for that task. However, **you will NOT see your partner's category groupings or category names**.

After 20 seconds, the next task will begin.

Good luck!

Groupings condition

At the end of each task, the screen will show you your **partnership score** for that task.

You will also see your partner's category groupings beneath your own, but you will NOT see your partner's category names. After 20 seconds, the next task will begin.

Good luck!

Labels condition

> At the end of each task, the screen will show you your **partnership score** for that task. **You will also see your partner's category names beneath your own, but you will NOT see your partner's category groupings.** After 20 seconds, the next task will begin.
>
> Good luck!

Both condition

> At the end of each task, the screen will show you your **partnership score** for that task. **You will also see your partner's category groupings and category names beneath your own.** After 20 seconds, the next task will begin.
>
> Good luck!

D.2 Questionnaire

The participants were asked the same demographic questions as in Experiments 2 and 3. The feedback questions on the questionnaire were as follows:

1. What do you think was the purpose of the experiment?
2. On what basis did you categorise the shapes?
3. Did this basis change between the different tasks?
4. How did you try to agree with your partner on categories for the shapes?
5. When you looked at a shape, did you mentally label it with an English word?
6. Do you have any comments about the program interface?
7. Do you have any other comments about the experiment?

D.3 Program interface

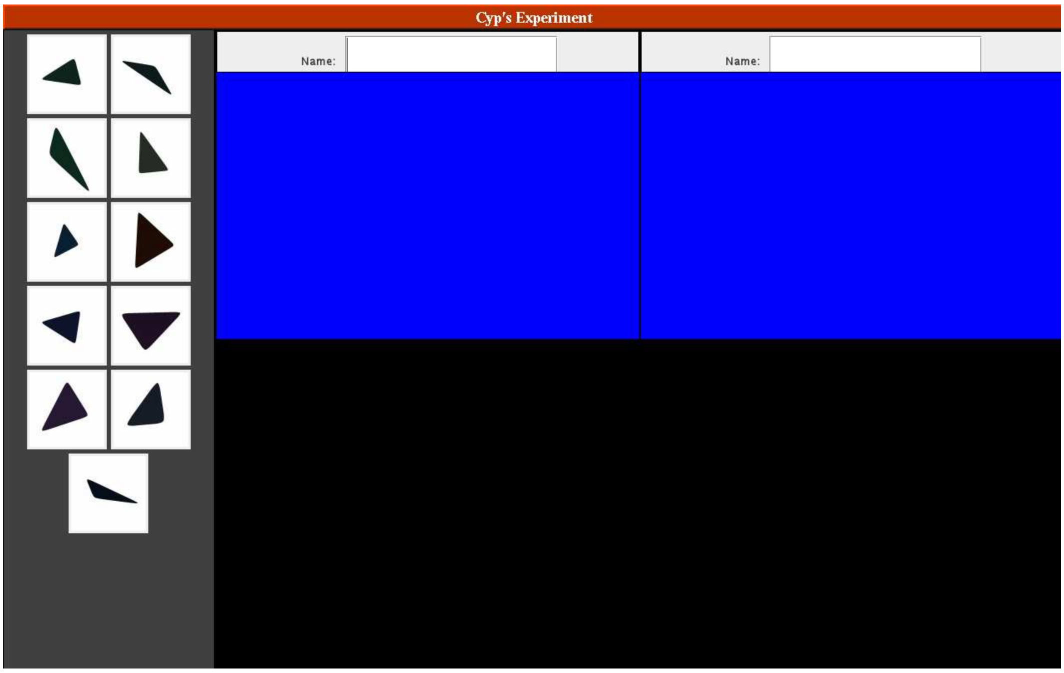


Figure D.1: Screenshot of the program interface at the beginning of a task.

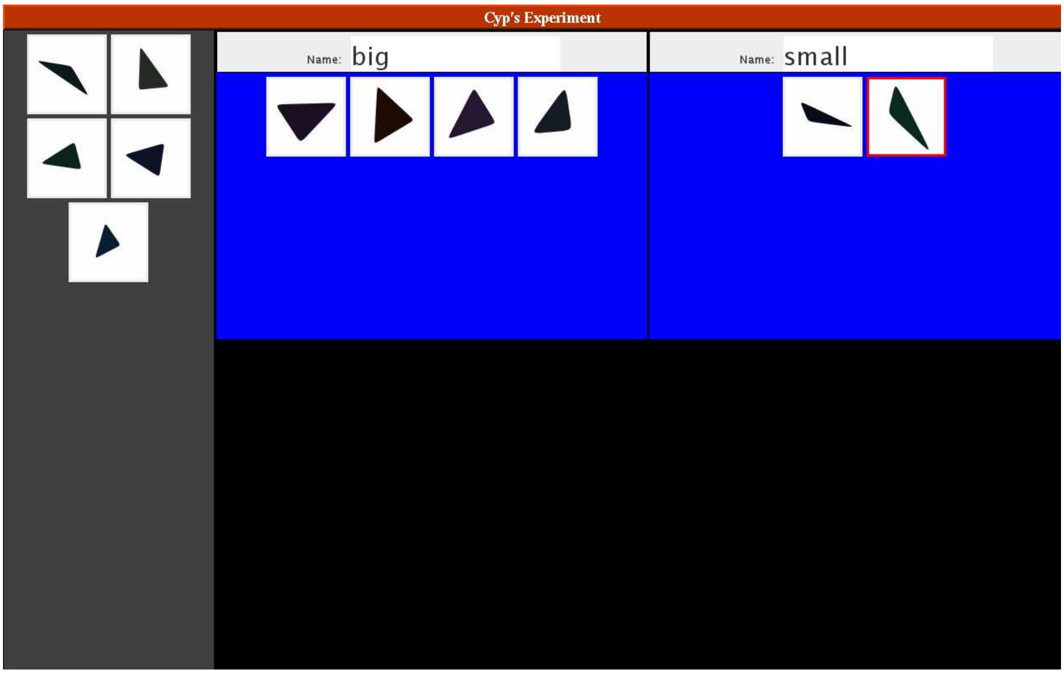


Figure D.2: Screenshot of the program interface mid-way through a task.

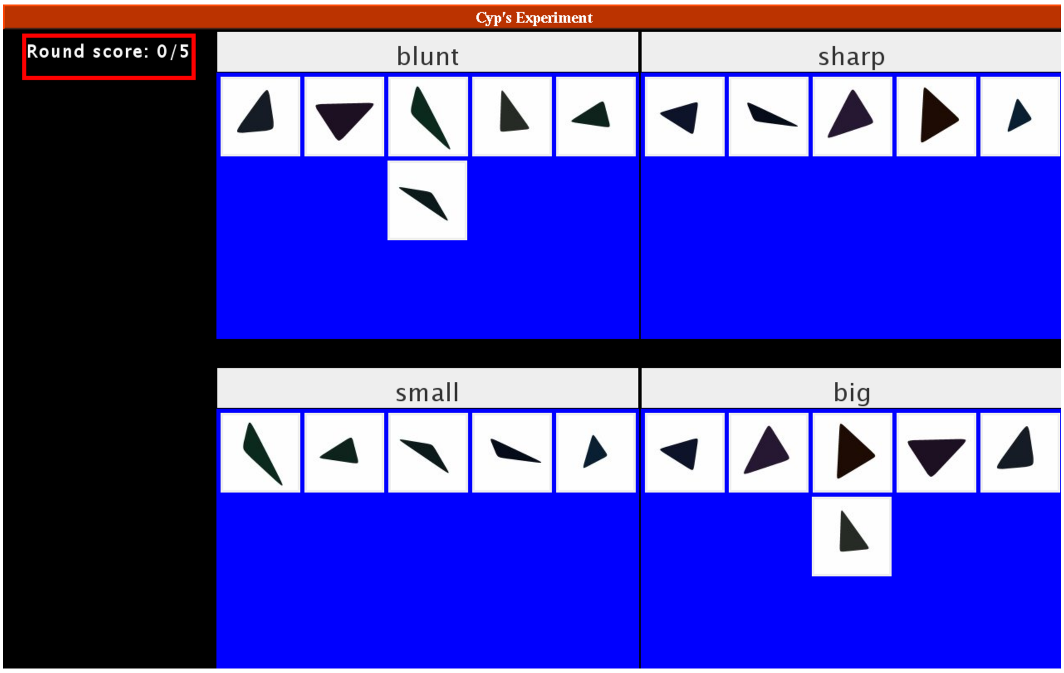


Figure D.3: A screenshot of the program interface when feedback is being shown to a participant. This example illustrates what happens in the *both* condition, since both kinds of feedback are provided.

APPENDIX E

Triangle generation program

Below is the script that was written to generate sets of triangle-like stimuli and used for pilots and Experiment 4.

```
#!/usr/bin/perl

### Generate sets of triangle-like shapes based on random values
### generated within certain parameter ranges.

use strict;
use Math::Polygon;

my $size = 600;             # the width and height of the images
my $total = 50;             # the number of images to create
my $dir = "/home/cyp/shapes";   # where to put the shapes

## The set of ranges from which to select random values for the
## colour and corner parameters, of the form (redmin, redmax,
## greenmin, greenmax, bluemin, bluemax, cornermin, cornermax,
## controlmin, controlmax).  The first six specify the ranges
## for the RGB values.  The 7th and 8th set the range for the
## "size" of the corner, and the 9th and 10th set the range for
## the "pointedness" parameter (called ’corner’ and ’control’
## here for technical reasons’).  The variable can actually
## contain multiple arrays, in which case images will be
## generated for each configuration.  Here the default values
## are the ones that were used for the stimuli in Experiment 4.
my @parameters =
  ([0.0, 0.2, 0.0, 0.2, 0.0, 0.2, 0.1, 0.3, 0.7, 1.0]);
```

```
## create a random parameter value between $min and $max
sub make_parameter {
  my ($min, $max) = @_;
  if ($min == $max) {
    return $min;
  } else {
    return ($min + (rand($max-$min)));
  }
}

## create a random x or y coordinate value between 0 and $size
sub make_point_value {
  return rand($size);
}

## find a point along the triangle edge from the corner
sub get_shifted_point {
  my ($z, $za, $ratio) = @_;
  return ($z + (($za - $z) * $ratio));
}

## find a control point for drawing the Bezier curves
sub get_shifted_control {
  my ($z, $za, $zb, $ratio) = @_;
  my ($zm) = ($za + $zb) / 2.0;
  return ($zm + (($z - $zm) * $ratio));
}

## generate one set of $total triangles for each parameter set
for (my $j=0; $j < @parameters; $j++) {

  ## get the parameter ranges from @parameters
  my @reds = ($parameters[$j][0],$parameters[$j][1]);
  my @greens = ($parameters[$j][2],$parameters[$j][3]);
  my @blues = ($parameters[$j][4],$parameters[$j][5]);
  my @corners = ($parameters[$j][6],$parameters[$j][7]);
  my @controls = ($parameters[$j][8],$parameters[$j][9]);

  my $start = ($j * $total) + 1;
  my $end = ($j+1) * $total;
  for (my $i=$start; $i<=$end; $i++) {

    ## generate random parameter values within the ranges
    ## specified
```

```
my ($red,$green,$blue) = (make_parameter(@reds),
                          make_parameter(@greens),
                          make_parameter(@blues));
my ($corner_ratio) = make_parameter(@corners);
my ($control_ratio) = make_parameter(@controls);

my ($x1,$x2,$x3,$y1,$y2,$y3,$ok);

while (!$ok) {

  ## generate random vertices
  my ($x1r, $y1r, $x2r, $y2r, $x3r, $y3r) =
    (make_point_value(), make_point_value(),
     make_point_value(), make_point_value(),
     make_point_value(), make_point_value());

  ## centre the triangle

  my $xdelta = ($size / 2.0) - (($x1r + $x2r + $x3r) / 3.0);
  $x1 = $x1r + $xdelta;
  $x2 = $x2r + $xdelta;
  $x3 = $x3r + $xdelta;

  my $ydelta = ($size / 2.0) - (($y1r + $y2r + $y3r) / 3.0);
  $y1 = $y1r + $ydelta;
  $y2 = $y2r + $ydelta;
  $y3 = $y3r + $ydelta;

  ## make sure the shifted triangle still fits in the image
  if (($x1 > 0 & $x1 < $size)
      & ($x2 > 0 & $x2 < $size)
      & ($x3 > 0 & $x3 < $size)
      & ($y1 > 0 & $y1 < $size)
      & ($y2 > 0 & $y2 < $size)
      & ($y3 > 0 & $y3 < $size)) {
    $ok = 1;
  }
}

## manipulate the corners

my $x12 = get_shifted_point($x1,$x2,$corner_ratio);
my $x13 = get_shifted_point($x1,$x3,$corner_ratio);
my $x1c = get_shifted_control($x1,$x12,$x13,$control_ratio);
my $x21 = get_shifted_point($x2,$x1,$corner_ratio);
```

```
    my $x23 = get_shifted_point($x2,$x3,$corner_ratio);
    my $x2c = get_shifted_control($x2,$x21,$x23,$control_ratio);
    my $x31 = get_shifted_point($x3,$x1,$corner_ratio);
    my $x32 = get_shifted_point($x3,$x2,$corner_ratio);
    my $x3c = get_shifted_control($x3,$x31,$x32,$control_ratio);

    my $y12 = get_shifted_point($y1,$y2,$corner_ratio);
    my $y13 = get_shifted_point($y1,$y3,$corner_ratio);
    my $y1c = get_shifted_control($y1,$y12,$y13,$control_ratio);
    my $y21 = get_shifted_point($y2,$y1,$corner_ratio);
    my $y23 = get_shifted_point($y2,$y3,$corner_ratio);
    my $y2c = get_shifted_control($y2,$y21,$y23,$control_ratio);
    my $y31 = get_shifted_point($y3,$y1,$corner_ratio);
    my $y32 = get_shifted_point($y3,$y2,$corner_ratio);
    my $y3c = get_shifted_control($y3,$y31,$y32,$control_ratio);

    my $p= Math::Polygon->new([$x12,$y12],[$x13,$y13],
                              [$x31,$y31],[$x32,$y32],
                              [$x23,$y23],[$x21,$y21],
                              [$x12,$y12]);
    my $area = $p->area();
    if ($area < 20000) {
      ## if the triangle is too small, make a new one.
      $i--;
    } else {
      ## create a Postscript output file
      my $outfile = "$dir/" . sprintf("%04d",$i) . ".eps";
      open(OUTFILE, ">$outfile") || die "Die!";
      print OUTFILE "%!PS-Adobe-3.0 EPSF-3.0\n";
      print OUTFILE "%%BoundingBox: 0 0 $size $size\n";
      print OUTFILE "\n";
      print OUTFILE "/red {$red} def\n";
      print OUTFILE "/green {$green} def\n";
      print OUTFILE "/blue {$blue} def\n";
      print OUTFILE "\n";
      print OUTFILE "/x12 {$x12} def\n";
      print OUTFILE "/x13 {$x13} def\n";
      print OUTFILE "/x1c {$x1c} def\n";
      print OUTFILE "/x21 {$x21} def\n";
      print OUTFILE "/x23 {$x23} def\n";
      print OUTFILE "/x2c {$x2c} def\n";
      print OUTFILE "/x31 {$x31} def\n";
      print OUTFILE "/x32 {$x32} def\n";
      print OUTFILE "/x3c {$x3c} def\n";
      print OUTFILE "\n";
      print OUTFILE "/y12 {$y12} def\n";
      print OUTFILE "/y13 {$y13} def\n";
      print OUTFILE "/y1c {$y1c} def\n";
      print OUTFILE "/y21 {$y21} def\n";
      print OUTFILE "/y23 {$y23} def\n";
      print OUTFILE "/y2c {$y2c} def\n";
      print OUTFILE "/y31 {$y31} def\n";
      print OUTFILE "/y32 {$y32} def\n";
      print OUTFILE "/y3c {$y3c} def\n";
      print OUTFILE "\n";
      print OUTFILE "red green blue setrgbcolor\n";
      print OUTFILE "newpath\n";
      print OUTFILE "x12 y12 moveto\n";
      print OUTFILE "x1c y1c x1c y1c x13 y13 curveto\n";
      print OUTFILE "x31 y31 lineto\n";
      print OUTFILE "x3c y3c x3c y3c x32 y32 curveto\n";
      print OUTFILE "x23 y23 lineto\n";
      print OUTFILE "x2c y2c x2c y2c x21 y21 curveto\n";
      print OUTFILE "closepath\n";
      print OUTFILE "fill\n";
      close OUTFILE;

      ## convert to JPEG format
      my $outimg = "$dir/" . sprintf("%04d",$i) . ".jpg";
      ‘convert $outfile -crop 600x600+12+192 $outimg‘;

    }

  }

}
```

APPENDIX F

Published papers

Bibliography

- Laskowski, C. (2008), The emergence of a lexicon by prototype-categorising agents in a structured infinite world, *in* A. D. M. Smith, K. Smith and R. Ferrer i Concho, eds, 'The evolution of language (EVOLANG 7)', World Scientific Press, London, pp. 195–202.
- Laskowski, C. and Pickering, M. (2010), How important are words for conceptual coordination?, *in* A. D. M. Smith, M. Schouwsrtra, B. de Boer and K. Smith, eds, 'The evolution of language (EVOLANG 8)', World Scientific Press, London, pp. 435–436.

THE EMERGENCE OF A LEXICON BY PROTOTYPE-CATEGORISING AGENTS IN AN INFINITE WORLD

CYPRIAN LASKOWSKI

Language Evolution and Computation Research Unit
University of Edinburgh, Edinburgh, EH8 9LL, UK
cyp@ling.ed.ac.uk

Over the last decade, computational models and simulations have been used to explore whether words could have emerged in the earliest stages of language evolution through a process of self-organisation in a population. In this paper, a new model of this family is presented, with two major differences from previous models. First, the world consists of an infinite number of objects, while remaining easily manipulable. Second, the agents' categories are based on prototypes, and their structure reflects the environments in which they are acquired and used. Simulation results reveal that, as in previous models, coherent lexicons still generally emerge, but they are sensitive to certain model conditions, including the world structure.

1. Introduction

Words pose an enigma for language evolution, for they are both fundamental and complex. On one hand, words constitute the basic building blocks of language, and on the surface they appear to be simply pairings of form and meaning. In fact, it makes little sense to speak of linguistic structure or its evolution without presupposing the existence of words, and their emergence is thus considered to constitute one of the earliest stages of language evolution (Jackendoff, 1999). On the other hand, words are distinguished by a set of properties that are not found together in any other animal species' signals: they are learned, arbitrary, referential and numerous. As such, words are unique to humans and cannot simply be taken as the very starting point of language evolution. Indeed, the evolutionary emergence of words is an unresolved puzzle.

Moreover, as for other aspects of language, the origins of words must be explained on at least two levels, biological and cultural. The biological level concerns questions of individual cognitive potential and linguistic preadaptations, such as a conceptual capacity. This can be partially investigated by comparing human and animal cognition, and assessing the extent to which animals can learn human words (Deacon, 1997). However, even if we pin down the necessary prerequisites, it is far from clear how the first words actually came into existence within a population of such individuals. Thus, at the cultural level, we must explain: how did hominins first start using words and agree on their meanings?

Since animals do not spontaneously invent words, while humans already have them, it is difficult to address these questions with direct empirical methods. However, Steels (1997) designed a simple computational model and showed that a coherent lexicon could emerge through a process of self-organisation. In particular, a population of individuals equipped with certain biological preadaptations gradually converged on a coherent lexicon by engaging in local communicative interactions about objects in a shared environment. Other models have since been developed to explore these issues further, and have generally yielded similar results, despite sometimes significant modifications (e.g., using robotic agents; Vogt, 2000). Further work is required, however, to assess whether the simulation results are contingent on idealisations that are inevitably implicit in such models.

This paper uses a new computational model to explore two representational issues, relating to the agents' world and their categories, respectively. I will first motivate and describe the model. Then I will present some simulation results, and finally discuss the relevance of the findings and possibilities for future work. For a detailed description of the model and simulation results, see Laskowski (2006).

2. A new model

2.1. *Changes relative to previous models*

The current model differs in two important respects from previous work. First, previous models have tended to represent the agents' world with a finite number of predefined objects (Steels, 1997; Smith, 2003b), so that agents encounter the same objects many times. Moreover, besides the notable exception of robot-based models (e.g., Vogt, 2000) the agents always perceive a given object identically. However, in the real world, we never perceive exactly the same stimulus twice, since there is a virtually unlimited variety of objects, and the appearance of the same object varies across situations. As a result, in the current model, the world consists of an infinite number of objects. Nevertheless, the distribution of objects in the world is not (necessarily) completely random. In fact, the world's structure is easily manipulable via parameters, making it possible to explore the effects of different kinds of structures on simulation results.

Second, previous models have not generally used psychologically plausible representations of agents' categories. Many models, for example, have used discrimination trees (Steels, 1997; Smith, 2003a). Such structures are efficient and simple, but are implementations of the classical theory of categorisation, which are considered obsolete (Murphy, 2002). Some models (e.g., Belpaeme, 2002) have addressed this by basing agents' categories on prototype theory (Rosch, 1978), which is still recognised as one of the leading psychological theories of concepts (Murphy, 2002). The representation used in the current model is based on that of Belpaeme (2002), but aims to be more sensitive to the context of category acquisition and usage.

2.2. *The world*

As in previous models (Smith, 2003a), the agents' world is represented with an N-dimensional space, where each dimension can be thought to represent a perceptual feature (e.g., colour, shape, size). Objects are defined as points in this space, whose dimension values are real numbers between 0 and 1 that identify the extent to which the objects have the corresponding features. Every agent-world interaction occurs in a context, which is a random subset of objects taken from the world. However, in contrast to previous models, each time a context is needed it is generated from scratch, and thus consists of entirely new objects. Therefore, an agent never sees the same object twice.

At the same time, however, the real world is not (necessarily) completely random, but has structure. The world is "clumpy" (Smith, 2003b), in the sense that, *within* dimensions, some values are generally more likely than others (e.g., animals usually have even numbers of legs). Also, the world is "correlated", so that values *across* dimensions tend to correlate to some extent (e.g., things that fly tend to have feathers, and vice versa). Consequently, rather than generating an entirely random vector each time an object is needed, the objects in this model are generated pseudo-randomly in accordance with probability distribution functions defined by the model's real-valued "clumpiness" and "correlation" parameters.

2.3. *Categories*

Following Steels (1997), agents are equipped with sensory channels which detect object dimension values directly and map them onto their perceptual space (which thus have the same general N-dimensional structure as the world). An agent's categories are superimposed onto the perceptual space, allowing for object categorisation.

In the current model, category structure is based on prototype theory (Rosch, 1978), so that categories have central members and graded membership. Categories are defined as Gaussian functions over the conceptual space which assign a degree of membership (a real number between 0 and 1) to every possible object. The category's prototype is the point of maximum membership (1), and the rate at which membership decreases as one moves away from the prototype depends on the category's sensitivity to each dimension. Formally, the category membership of an object o in a category c is given by a Gaussian function,

$$membership_c(o) = \left(\prod_{i=1}^{N} e^{-\frac{1}{2}\left(\frac{o_i - p_i}{s_i}\right)^2} \right)^{1/N} \tag{1}$$

where i identifies a dimension, with o_i being the object value, p_i the prototype value, and s_i the sensitivity.

This representation is based on that of Belpaeme (2002), with one important difference. In his model, the category's dimension sensitivities were all rigidly set to one default value, so that every dimension was equally important both across and within dimensions. However, this is not the case in the real world: for example, the shape of a screwdriver (but not a traffic light) is far more relevant than its colour. Consequently, in this model, the dimension sensitivities are not fixed, and depend on the contexts in which categories are acquired and used.

Although the category representation is relatively plausible psychologically, it makes categorisation of objects more complicated. Rather than identifying the category in whose space an object falls, it is necessary to find the category which best fits the object (i.e., the category for which the membership function yields the highest value). Moreover, a minimum threshold is defined (as a model parameter) so that an object can only be potentially considered as a member of a category if its degree of membership is above this threshold. Figure 1 shows an example of such a "candidate category" in two-dimensional space for a particular object.

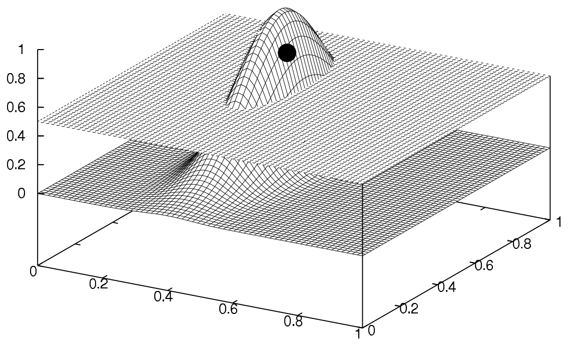


Figure 1. Category membership in 2 dimensions: $membership_c(o)$, the category membership function for an agent's category c in a conceptual space of two dimensions, with $p_0 = 0.4$, $s_0 = 0.05$, $p_1 = 0.6$, and $s_1 = 0.1$. The plane shows the value of the minimum membership threshold, and the dot indicates the object being categorised: since the dot is above the plane, this is a candidate category for the object.

Each category is also associated with a list of words and association strengths. Words themselves are atomic tokens with no internal structure. The word with the highest association strength is the best or most "natural" word for that category, and is the word that the agent will typically use when communicating about the category. The list can also be empty, in which case the category has not been lexicalised.

2.4. *Category development*

Agents develop and adapt their category systems through interactions with the world. Each interaction takes the form of a discrimination game (Steels, 1997), in which an agent is exposed to a context of objects, attempts to find a distinct category for one of the objects (called the topic), and adapts its category system accordingly. Over many discrimination games in different environments, an agent's category system gradually grows and adapts to the structure of the world.

Discrimination games have 3 basic possible outcomes: the creation of a new category, the splitting off of a subcategory, or the adjustment of an existing category. If the agent has no candidate categories for the topic object, then it will create a new category, whose prototype is set to the topic object, and whose initial dimension sensitivities are a function of how similar the other context objects were to the topic in the different dimensions. Otherwise, it will check whether any of its candidate categories are sufficiently discriminating as not to match any of the other context objects. If there are no such categories, then it takes the most refined candidate category (i.e., the one with the most sensitive dimensions), and creates a subcategory of it which is identical with it except for being more sensitive in the dimension in which the topic differs the most from the other context objects. If there were discriminating categories, then the topic is categorised with the one for which it has the highest membership, and this category's prototype and dimension sensitivities are adjusted slightly to fit the topic better.

2.5. *Lexical development*

The other kind of formal game that the agents engage in is the guessing game Steels (1997), which is actually built on top of the discrimination game. While discrimination games involve only one agent and do not involve any linguistic exchange, the guessing game is a communicative episode involving two agents and a shared environment. The "speaker" agent utters a word for one of the context objects (the topic), and the "hearer" agent guesses which object the speaker was referring to. The game is a success if and only if the hearer guesses correctly.

In each guessing game, a speaker and a (different) hearer are chosen from the population at random, and a new shared context of objects is generated. The speaker chooses a topic object at random, categorises it (via a discrimination game), and utters the word in its lexicon with the highest association score for that category. If the speaker has no word for that category, it randomly invents a new word. The hearer must find the best match between the word heard, a context object, and a word-category pair from its own lexicon. It first identifies all of its categories which have an association for the word. If there are no such categories, the game fails. Otherwise, it considers each possible category-object pair from these categories and the context objects, and determines the combination for which category membership is highest. If the resulting membership is below the minimum membership threshold, then the game fails. Otherwise, the hearer guesses the object from that pair. If this object is the topic, the game succeeds. Otherwise, the speaker points out the topic to the hearer (non-linguistically), and the hearer performs a discrimination game on it. Upon completion of the game, both agents independently update their lexicons, adjusting specific word-category association strengths in accordance with the results of the game.

3. Simulations

Simulations were run within this model with three questions in mind. First, would agents converge on a coherent lexicon, despite the more complex world and category representations used in this model? Second, do the simulation results depend substantially on the specific world structure used? Third, assuming that agents did converge, how stable would the results be if one varied specific parameters, such as population and context size?

Each simulation consisted of a large number of guessing games in a fixed population of agents, who all began with empty category systems and no lexicons. The guessing games were analysed in sets of 100 called epochs, and the average success rate (the ratio of successful to total number of guessing games) was tracked for each epoch. The first set of simulations explored whether this model would work at all in the simplest cases, with population and context sizes of 2 in a 1-dimensional world. After 200 epochs, regardless of how clumpy the world was (dimension correlation does not apply of course in a one dimensional world), communicative success in the final epoch averaged at around 99% over 100 simulations, despite the fact that the agents ended up with large category systems.

In a 3-dimensional world, the final communicative success was still very high, despite extreme manipulations of world structure. Four kinds of world were tested: "random" (dimensions were completely uncorrelated and non-clumpy), "correlated" (highly correlated dimensions but completely non-clumpy), "clumpy" (completely uncorrelated but highly clumpy), and "structured" (highly clumpy and correlated). Final communicative success after 200 epochs was still very high for all four world structure types, ranging from 96% for the random world and 99% for the structured world. Agents ended up with around 300 categories, except for the totally random world, where they tended to have over 500 categories.

In order to explore the scalability of the results and their potential dependence on a particular world structure, further sets of simulations were conducted. In each set, one of the main model parameters was manipulated, starting with a base case of a 3-dimensional world, 2 context objects, and 2 agents. Results showed that communicative success was significantly affected by manipulations of these variables, but the extent of the impact depended on the world structure. For instance, world dimensionality had a very large impact, such that in an 8-dimensional world, final communicative success rates tended to stay below a dismal 25% in the random and clumpy worlds. However, they were still in the high 90's in the correlated and structured worlds. Manipulations of the context size had similar impacts, although less drastic. Context sizes of 64 objects still yielded approximately an 80% final success rate in the correlated and structured worlds, but context sizes of only 16 objects resulted in rates below 50% for both random and clumpy worlds. The effects of population size changes did not follow the same pattern, however. Although higher population sizes corresponded with lower communicative success rates, the communicative success rate reached at least around 75% in all four world structures even with as many as 128 agents, and was best in the clumpy world (around 90%). Moreover, the communicative success rate curves also varied in clear ways between the four world types examined. In worlds with correlated dimensions (i.e., the correlated and structured worlds), communicative success rose very quickly (e.g., to about 75% with 128 agents), but then flattened out. In contrast, worlds with clumpy dimensions started off more slowly, but their communicative success rate curves did not flatten out as dramatically, so eventually they obtained higher success rates (at least in the clumpy world).

4. Discussion

Despite the use of a more complex model, in which agents never saw exactly the same object twice and their categories had a continuous prototype structure, simulation results were generally in line with those of previous work. Under a variety of conditions, populations of agents converged onto coherent lexicons after engaging through repeated communicative episodes in shared environments. Although each agent had an independent category system and lexicon which started out empty, communication success rates managed to reach high levels, often close to 100%. These results, then, add support to the idea that a population of hominins equipped with certain cognitive preadaptations could have grounded and developed a large system of learned, arbitrary, referential words through a series of local interactions (Steels, 1997). More specifically, they show that the general results of previous models cannot be dismissed on the grounds that they used psychologically implausible category representations.

However, the simulation results were sensitive to more complex conditions, as manifested by manipulations of the model's parameters. As in previous models, communicative success dropped in simulations in which the context size, population size, or world dimensionality was increased. Although this is not surprising, in some cases the effects were very drastic, and highly dependent on the world's structure. Moreover, even in successful simulations, the world structure sometimes influenced the rate of convergence. These patterns show that the simulation results do not easily scale up to larger systems, and thus must be treated cautiously. In particular, the world structure can have large consequences for whether a coherent lexicon will emerge and how long it will take. This points to the need for future models to choose their world representations carefully and justify their choices.

Returning to the bigger picture, how exactly do these results relate to language evolution? To answer this, we need to revisit the hypothesis and clearly separate what exactly is being given a priori in this model, as opposed to what appears to be emerging (Steels, 2006). We started by asking whether self-organisation was able to explain how a population of hominins could have "invented" a lexicon. However, it's important to keep in mind that this hypothesis is framed within an implicitly substantial environmental and cognitive infrastructure. We have already seen that the environment that the agents are exposed to can play a crucial role in determining the outcomes of the simulations. The extent of the cognitive prerequisites have not, however, been substantially manipulated here. Agents are instead consistently endowed with unrealistically powerful and facilitating faculties, including perfect word production and perception, powerful joint attention, limitless motivation for communication regardless of success, perfect and equal perception of objects, and perfect ability to use and interpret non-linguistic referential methods. What the simulation results of this model have done is to verify the internal consistency of the argument that, *given such abilities*, and under simple conditions, a lexicon could have emerged through a process of self-organisation, even if the world and category representations are made more complex in the way described. However, this work cannot address the question of whether the differences between the idealisations and the real phenomena are significant enough to give misleadingly optimistic results. In order to arrive at that, more work is needed, including integration with empirical experimental work with both humans and animals, as well as further modelling developments and explorations.

References

Belpaeme, T. (2002). *Factors influencing the origins of colour categories.* Unpublished doctoral dissertation, Vrje Universiteit Brussel.

Deacon, T. (1997). *The symbolic species: the coevolution of language and the brain.* New York: Norton.

Jackendoff, R. (1999). Possible stages in the evolution of the language capacity. *Trends in Cognitive Sciences*, *3*, 272–279.

Laskowski, C. (2006). *Prototype categorisation and the emergence of a lexicon in an infinite world.* Unpublished master's thesis, University of Edinburgh. (http://www.lel.ed.ac.uk/homes/cyp/dissertation/dissertation.pdf)

Murphy, G. L. (2002). *The big book of concepts.* Cambridge, MA: MIT Press.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. B. LLoyd (Eds.), *Cognition and categorization* (pp. 27–48). Hillsdale, NJ: Lawrence Erlbaum.

Smith, A. D. M. (2003a). *Evolving communication through the inference of meaning.* Unpublished doctoral dissertation, Theoretical and Applied Linguistics, School of Philosophy, Psychology and Language Sciences, University of Edinburgh.

Smith, A. D. M. (2003b). Intelligent meaning creation in a clumpy world helps communication. *Artificial Life*, *9*, 559–574.

Steels, L. (1997). Constructing and sharing perceptual distinctions. In M. van Someren & G. Widmer (Eds.), *Proceedings of the European conference on machine learning* (pp. 4–13). Berlin: Springer-Verlag.

Steels, L. (2006). How to do experiments in artificial language evolution and why. In A. Cangelosi, A. D. M. Smith, & K. Smith (Eds.), *The evolution of language: proceedings of the 6th international international conference* (pp. 323–332). London: World Scientific.

Vogt, P. (2000). *Lexicon grounding on mobile robots.* Unpublished doctoral dissertation, Vrije Universiteit Brussel.

HOW IMPORTANT ARE WORDS FOR CONCEPTUAL COORDINATION?

CYPRIAN LASKOWSKI, MARTIN PICKERING

Language Evolution and Computation Research Unit, University of Edinburgh,
3 Charles Street, Edinburgh, EH8 9AD, United Kingdom
cyp@ling.ed.ac.uk

In attempting to explain how words first emerged in language evolution, it is often assumed that words map passively onto a pre-existing static conceptual system (e.g., Hurford, 2007). However, human thought is very flexible: we can conceptualise the same referent in many different ways, depending on the context and our goal. For instance, we may conceptualise a lion as a dangerous beast, a thrilling sight, an unusual dinner, etc. Moreover, concepts may have different scope for different people. As a result, one of the main functions of words may be to provide an efficient way to share and coordinate conceptualisations with others. Indeed, to the extent that people do conceptualise things differently, it is not easy to imagine how such gaps could be bridged without the help of words and language, and previous work has shown that verbal labels can enhance category learning (Lupyan, Rakison, & McClelland, 2007). Either way, looking into this issue could contribute to our understanding of how much the advent of public symbols transformed human cognition (Deacon, 1997). The question under empirical investigation here is thus the importance of words to conceptual coordination.

Before this issue can be addressed experimentally, there is an immediate methodological challenge. In particular, in order to assess the role of words in conceptual coordination, we need a way of getting at experimental participants' concepts without relying on words. The solution that will be adopted here is the use of free classification tasks (Malt, Sloman, Gennari, Shi, & Wang, 1999): participants partition a set of items into groups, and these groups are assumed to be referential snapshots of their concepts.

I present here an experiment within the free classification framework which pits the importance of words against that of referential information. Pairs of native English speakers conducted a sequence of thirty free classification tasks involving a fluid domain of triangle-like stimuli. In each task, participants had to individually sort a set of eleven stimuli into two categories, and to label their categories. Their goal was to partition the stimuli into the same (or as similar as possible) two groups as their partner (irrespective of the labels used). Participants could not interact or communicate freely during the experiment, but they did receive feedback at the end of each task. All participants were shown their partnership's joint task score, and, depending on the condition that they were assigned to, either their partner's category groupings, category labels, both or neither (thus the experiment had a 2x2 between-pairs design).

The results revealed several patterns. First, although label agreement within pairs correlated with higher task scores, there were plenty of exceptions, where the participants used the same labels but differed in their category groupings, or achieved identical groupings despite using different labels. Second, averaging across the tasks, grouping feedback resulted in significantly higher scores, and these were higher still if accompanied by label feedback as well. On the other hand, label feedback on its own did not result in higher scores. Third, although there was a lot of fluctuation in scores even within pairs, they tended to go up over time, except in the groupings-only condition, where they stayed about the same. Due to these last two patterns, by the end of the experiment, the scores in the *both* condition were significantly higher than in the other three conditions. Thus it was only in the condition with both kinds of feedback that pairs started off with relatively high scores **and** improved over time.

Together, these results suggest that rich referential information is more useful than words for conceptual coordination, but that high levels of coordination can only be achieved when both types are available. Of course, it would be absurd to suggest that early hominins huddled together and explicitly sorted things into categories to coordinate their concepts. However, the current results shed light on the extent to which language may have revolutionised human cognition. Language does not seem to be crucial for conceptual coordination, but it does enhance it.

References

Deacon, T. (1997). *The symbolic species: the coevolution of language and the brain.* New York: Norton.

Hurford, J. R. (2007). *The origins of meaning: language in the light of evolution.* Oxford University Press.

Lupyan, G., Rakison, D. H., & McClelland, J. L. (2007). Language is not just for talking: labels facilitate learning of novel categories. *Psychological Science*, *18*, 1077–1083.

Malt, B. C., Sloman, S., Gennari, S., Shi, M., & Wang, Y. (1999). Knowing versus naming: similarity and the linguistic categorization of artifacts. *Journal of Memory and Language*, *40*, 230–262.

References

Ackrill, J. L., ed. (1963), *Aristotle: categories and de interpretatione*, Clarendon Press, Oxford.

Allen, S. W. and Brooks, L. R. (1991), 'Specializing the operation of an explicit rule', *Journal of Experimental Psychology: General* **120**, 3–19.

Ameel, E., Malt, B. C., Storms, G. and Van Assche, F. (2009), 'Semantic convergence in the bilingual lexicon', *Journal of Memory and Language* **60**, 270–290.

Ameel, E., Malt, B. and Storms, G. (2008), 'Object naming and later lexical development: from baby bottle to beer bottle', *Journal of Memory and Language* **58**, 262–285.

Ameel, E., Storms, G., Malt, B. C. and Sloman, S. A. (2005), 'How bilinguals solve the naming problem', *Journal of Memory and Language* **53**, 60–80.

Armstrong, S. L., Gleitman, L. R. and Gleitman, H. (1983), 'On what some concepts might not be', *Cognition* **13**, 263–308.

Athanasopoulos, P. (2009), 'Cognitive representation of colour in bilinguals: the case of Greek blues', *Bilingualism: Language and Cognition* **12**, 83–95.

Atran, S. (1989), 'Basic conceptual domains', *Mind and Language* **4**, 7–16.

Baayen, R. H. (1994), 'Productivity in language production', *Language and Cognitive Processes* **9**, 447–469.

Baayen, R. H. (2008), *Analyzing linguistic data: a practical introduction to statistics*, Cambridge University Press, Cambridge.

Bailenson, J. N., Shum, M. S., Atran, S., Medin, D. L. and Coley, J. D. (2002), 'A bird's eye view: biological categorization and reasoning within and across cultures', *Cognition* **84**, 1–53.

Balaban, M. T. and Waxman, S. R. (1997), 'Do words facilitate object categorization in 9-month-old infants?', *Journal of Experimental Child Psychology* **64**, 3–26.

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G. and Newlands, A. (2000), 'Controlling the intelligibility of referring expressions in dialogue', *Journal of Memory and Language* **42**, 1–22.

Barr, D. J. and Keysar, B. (2002), 'Anchoring comprehension in linguistic precedents', *Journal of Memory and Language* **46**, 391–418.

Barsalou, L. W. (1983), 'Ad hoc categories', *Memory and Cognition* **11**, 211–227.

Barsalou, L. W. (1987), The instablity of graded structure: implications for the nature of concepts, *in* U. Neisser, ed., 'Concepts and conceptual development', Cambridge University Press, Cambridge, pp. 101–140.

Barsalou, L. W. (1990), On the indistinguishability of exemplar memory and abstraction in category representation, *in* T. K. Srull and R. S. Wyer, eds, 'Advances in social cognition', Vol. 3, Lawrence Erlbaum, Hillsdale, NJ, pp. 61–88.

Barsalou, L. W. (1999), 'Perceptual symbol systems', *Behavioral and Brain Sciences* **22**, 577–660.

Barsalou, L. W. (2005), 'Continuity of the conceptual system across species', *Trends in Cognitive Sciences* **9**, 309–311.

Behrend, D. A., Scofield, J. and Kleinknecht, E. E. (2001), 'Beyond fast mapping: young children's extensions of novel words and facts', *Developmental Psychology* **37**, 698–705.

Bellugi, U., Marks, S., Bihrle, A. M. and Sabo, H. (1988), Dissociation between language and cognitive functions in Williams syndrome, *in* D. Bishop and K. Mogford, eds, 'Language development in exceptional circumstances', Churchill Livingstone, Edinburgh, pp. 177–189.

Berlin, B. (1992), *Ethnobiological classification: principles of categorization of plants and animals in traditional societies*, Princeton University Press, Princeton, NJ.

Berlin, B. and Kay, P. (1969), *Basic color terms: their universality and evolution*, University of California Press, Berkeley.

Bermúdez, J. L. (2003), *Thinking without words*, Oxford University Press, New York.

Bierwisch, M. and Schreuder, R. (1991), From concepts to lexical items, *in* W. J. M. Levelt, ed., 'Lexical access in speech production', Elsevier, Amsterdam, pp. 23–60.

Billman, D. and Krych, M. (1998), Path and manner verbs in action: effects of "skipping" and "exiting" on event memory, *in* 'Procedings of the 20th annual conference of the cognitive science society', Erlbaum Associates, Hillsdale, NJ, pp. 156–161.

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

Bornstein, M. H. (1985), 'On the development of color naming in young children: data and theory', *Brain and Language* **26**, 72–93.

Boroditsky, L. (2001), 'Does language shape thought?: Mandarin and English speakers' conceptions of time', *Cognitive Psychology* **43**, 1–22.

Boroditsky, L., Schmidt, L. A. and Phillips, W. (2003), Sex, syntax, and semantics, *in* D. Gentner and S. Goldin-Meadow, eds, 'Language in mind: advances in the study of language and thought', MIT Press, Cambridge, MA, pp. 61–80.

Boster, J. S. (2005), Categories and cognitive anthropology, *in* H. Cohen and C. Lefebvre, eds, 'Handbook of categorization in cognitive science', Elsevier, Amsterdam, pp. 91–118.

Bowerman, M. and Levinson, S. C., eds (2001), *Language acquisition and conceptual development*, Cambridge University Press, Cambridge.

Branigan, H. P. (2004), 'Full alignment of some but not all representations in dialogue. Commentary on Pickering and Garrod', *Behavioral and Brain Sciences* **27**, 296–297.

Branigan, H. P., Pickering, M. J. and Cleland, A. A. (2000), 'Syntactic co-ordination in dialogue', *Cognition* **75**, B13–25.

Brennan, S. E. and Clark, H. H. (1996), 'Conceptual pacts and lexical choice in conversation', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **22**, 1482–1493.

Brennan, S. E. and Hanna, J. E. (2009), 'Partner-specific adaptation in dialog', *Topics in Cognitive Science* **1**, 274–291.

Brown, P. M. and Dell, G. S. (1987), 'Adapting production to comprehension: the explicit mention of instruments', *Cognitive Psychology* **19**, 441–472.

Brown, R. and McNeill, D. (1966), 'The "tip of the tongue" phenomenon', *Journal of Verbal Learning and Verbal Behavior* **5**, 325–337.

Brown, R. W. and Lenneberg, E. H. (1954), 'A study in language and cognition', *Journal of Abnormal Psychology* **49**, 454–462.

Brown-Schmidt, S. (2009), 'Partner-specific interpretation of maintained referential precenents during interactive dialog', *Journal of Memory and Language* **61**, 171–190.

Brown-Schmidt, S., Gunlogson, C. and Tanenhaus, M. K. (2008), 'Addressees distinguish shared from private information when interpreting questions during interactive conversation', *Cognition* **107**, 1122–1134.

Burke, D. M., MacKay, D. G., Worthley, J. S. and Wade, E. (1991), 'On the tip of the tongue: what causes word finding failures in young and older adults?', *Journal of Memory and Language* **30**, 542–579.

Burnham, R. W. and Clark, U. R. (1955), 'A test of hue memory', *Journal of Applied Psychology* **39**, 164–172.

Carey, S. (2004), 'Bootstrapping and the origin of concepts', *Daedalus* **133**, 59–68.

Carey, S. (2009), *The origins of concepts*, Oxford University Press, Oxford.

Carmichael, L. C., Hogan, H. P. and Walters, A. A. (1932), 'An experimental study of the effect of language on the reproduction of visually perceived form', *Journal of Experimental Psychology* **15**, 73–86.

Carruthers, P. (2002), 'The cognitive functions of language', *Behavioral and Brain Sciences* **25**, 657–674.

Chen, J. (2007), 'Do Chinese and English speakers think about time differently? Failure of replicating Boroditsky (2001)', *Cognition* **104**, 427–436.

Choi, S. and Bowerman, M. (1991), 'Learning to express motion events in English and Korean: the influence of language-specific lexicalization patterns', *Cognition* **41**, 83–121.

Chomsky, N. (1968), *Language and mind*, Harcourt Brace & World, Inc., New York.

Chomsky, N. (1986), *Knowledge of language: its nature, origin, and use*, Praeger, New York.

Chomsky, N. (1988), *Language and problems of knowledge*, MIT Press, Cambridge, MA.

Chomsky, N. (2000), *On nature and language*, Cambridge University Press, New York.

Churchland, P. (1989), *A neurocomputational perspective: the nature of mind and the structure of science*, MIT Press, Cambridge, MA.

Churchland, P. M. (1998), 'Conceptual similarity across sensory and neural diversity: the Fodor/Lepore challenge answered', *Journal of Philosophy* **1**, 5–32.

Clark, A. (1998), Magic words: how language augments human computation, *in* P. Carruthers and J. Boucher, eds, 'Language and thought: interdisciplinary themes', Cambridge University Press, Cambridge, MA, pp. 162–183.

Clark, E. V. (1987), The principle of contrast: a constraint on language acquisition, *in* B. MacWhinney, ed., 'Mechanisms of language acquisition', Erlbaum, London, pp. 264–293.

Clark, H. H. (1996), *Using language*, Cambridge University Press, Cambridge.

Clark, H. H. and Brennan, S. A. (1991), Grounding in communication, *in* L. B. Resnick, J. M. Levine and S. D. Teasley, eds, 'Perspectives on socially shared cognition', APA Books, Washington, pp. 127–149.

Clark, H. H. and Krych, M. A. (2004), 'Speaking while monitoring addressees for understanding', *Journal of Memory and Language* **50**, 62–81.

Clark, H. H. and Marshall, C. R. (1981), Definite reference and mutual knowledge, *in* A. K. Joshe, B. Webber and I. A. Sag, eds, 'Elements of discourse understanding', Cambridge University Press, Cambridge, pp. 10–63.

Clark, H. H. and Wilkes-Gibbs, D. (1986), 'Referring as a collaborative process', *Cognition* **22**, 1–39.

Clopper, C. G. (2008), 'Auditory free classification: methods and analysis', *Behavior Research Methods* **40**, 575–581.

Cohen, H. and Lefebvre, C., eds (2005), *Handbook of categorization in cognitive science*, Elsevier, Amsterdam.

Cohen, R., Kelter, S. and Woll, G. (1980), 'Analytic competence and language impairment in aphasia', *Brain and Language* **10**, 331–347.

Cohen, R., Woll, G. and Ehrenstein, W. H. (1981), 'Recognition deficits resulting from focussed attention in aphasia', *Psychological Research* **43**, 391–405.

Costa, A., Pickering, M. J. and Sorace, A. (2008), 'Alignment in second language dialogue', *Language and cognitive processes* **23**, 528–556.

Cramer, H. (1946), *Mathematical methods of statistics*, Princeton University Press, Princeton, NJ.

Croft, W. and Cruse, D. A. (2004), *Cognitive linguistics*, Cambridge University Press, Cambridge.

Cruse, D. A. (1986), *Lexical semantics*, Cambridge University Press, Cambridge.

Dale, R., Dietrich, E. and Chemero, A. (2009), 'Explanatory pluralism in cognitive science', *Cognitive Science* **33**, 739–742.

Dale, R. and Spivey, M. J. (2005), 'From apples and oranges to symbolic dynamics: a framework for conciliating notions of cognitive representation', *Journal of Experimental and Theoretical Artificial Intelligence* **17**, 317–342.

Damasio, H., Grabowski, T., Tranel, D., Hichwa, R. D. and Damasio, A. R. (1996), 'A neural basis for lexical retrieval', *Nature* **380**, 499–505.

Daoutis, C. A., Pilling, M. and Davies, I. R. L. (2006), 'Categorical effects in visual search for colour', *Visual Cognition* **14**, 217–240.

Davidoff, J., Davies, I. and Roberson, D. (1999), 'Colour categories in a stone-age tribe', *Nature* **398**, 203–204.

Davidson, D. (2001), *Inquiries into truth and interpretation*, 2nd edn, Oxford University Press, Oxford.

Davidson, D. (2004), What thought requires, *in* 'Problems of rationality', Oxford University Press, Oxford, pp. 135–149.

Davidson, N. S. and Gelman, S. A. (1990), 'Inductions from novel categories: the role of language and conceptual structure', *Cognitive Development* **5**, 151–176.

Dawson-Tunik, T. L. (2006), The meaning and measurement of conceptual development in adulthood, *in* C. Hoare, ed., 'Handbook of adult learning and development', Oxford University Press, Oxford, pp. 433–454.

de Saussure, F. (1916/1983), *Course in general linguistics*, Duckworth, London.

Deacon, T. (1997), *The symbolic species: the coevolution of language and the brain*, Norton, New York.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G. and Harshman, R. (1990), 'Indexing by latent semantic analysis', *Journal of the American Society for Information Science* **41**, 391–407.

Dell, G. S. (1986), 'A spreading-activation theory of retrieval in sentence production', *Psychological Review* **93**, 283–321.

Dennett, D. (1995), *Darwin's dangerous idea*, Simon and Schuster, New York.

Dewar, K. and Xu, F. (2007), 'Do 9-month-old infants expect distinct words to refer to kinds?', *Developmental Psychology* **43**, 1227–1238.

Dixon, R. M. W. (1980), *The languages of Australia*, Cambridge University Press, Cambridge.

Drivonikou, G. V., Kay, P., Regier, T., Ivry, R. B., Gilbert, A. L., Franklin, A. and Davies, I. R. L. (2007), Further evidence that Whorfian effects are stronger in the right visual field than the left, *in* 'Proceedings of the National Academy of Sciences', Vol. 104, pp. 1097–1102.

Dummett, M. (1981), *Frege: philosophy of language*, 2nd edn, Duchworth, London.

Edwards, K. (2010), 'Concept referentialisms and the role of empty concepts', *Mind and Language* **25**, 89–118.

Ellen, R. (2006), *The categorical impulse: essays in the anthropology of classifying behaviour*, Berghahn Books, New York.

Enfield, N. J., Majid, A. and van Staden, M. (2006), 'Cross-linguistic categorisation of the body: introduction', *Language Sciences* **28**, 137–147.

Evans, G. (1982), *The varieties of reference*, Oxford University Press, Oxford.

Evans, N. and Levinson, S. C. (2009), 'The myth of language universals: language diversity and its importance for cognitive science', *Behavioral and Brain Sciences* **32**, 429–492.

Feist, M. I. and Gentner, D. (2007), 'Spatial language influences memory for spatial scenes', *Memory and Cognition* **35**, 283–296.

Flavell, J. H. (1963), *The developmental psychology of Jean Piaget*, D. Van Nostrand, Princeton, NJ.

Fodor, J. A. (1975), *The language of thought*, Hassocks Harvester Press, New York.

Fodor, J. A. (1983), *The modularity of mind: an essay on faculty psychology*, MIT Press, Cambridge, MA.

Fodor, J. A. (1998), *Concepts: where cognitive science went wrong*, Clarendon Press, Oxford.

Fodor, J. A. and Lepore, E. (1992), *Holism: a shopper's guide*, Blackwell, Oxford.

Franklin, A., Clifford, A., Williamson, E. and Davies, I. (2005), 'Color term knowledge does not affect categorical perception of color in toddlers', *Journal of Experimental Child Psychology* **90**, 114–141.

Franklin, A. and Davies, I. R. L. (2004), 'New evidence for infant colour categories', *British Journal of Developmental Psychology* **22**, 349–377.

Franklin, A., Drivonikou, G. V., Bevis, L., Davies, I. R. L., Kay, P. and Regier, T. (2008), 'Categorical perception of color is lateralized to the right hemisphere in infants, but to the left hemisphere in adults', *Proceedings of the National Academy of Sciences* **105**, 3221–3225.

Franklin, A., Wright, O. and Davies, I. R. L. (2009), 'What can we learn from toddlers about categorical perception of color? Comments on Goldstein, Davidoff, and Robeson', *Journal of Experimental Child Psychology* **102**, 239–245.

Frege, G. (1892/1948), 'Sense and reference', *The Philosophical Review* **57**, 209–230.

Frege, G. (1903/2004), Concepts, *in* B. Aarts, D. Denison, E. Keizer and G. Popova, eds, 'Fuzzy grammar: a reader', Oxford University Press, Oxford, p. 33.

Fulkerson, A. L. and Waxman, S. R. (2007), 'Words (but not tones) facilitate object categorization: evidence from 6- and 12-month olds', *Cognition* **105**, 218–228.

Galantucci, B. (2005), 'An experimental study of the emergence of human communication systems', *Cognitive Science* pp. 737–767.

Galati, A. and Brenna, S. E. (2010), 'Attenuating information in spoken communication: for the speaker or for the addressee?', *Journal of Memory and Language* **62**, 35–51.

Gärdenfors, P. (2000), *Conceptual spaces: the geometry of thought*, MIT Press, Cambridge, MA.

Garrod, S. and Anderson, A. (1987), 'Saying what you mean in dialogue: a study in conceptual and semantic co-ordination', *Cognition* **27**, 181–218.

Garrod, S. and Clark, A. (1993), 'The development of dialogue co-ordination skills in schoolchildren', *Language and Cognitive Processes* **8**, 101–126.

Garrod, S. and Doherty, G. (1994), 'Conversation, coordination and convention: an empirical investigation of how groups establish linguistic conventions', *Cognition* **53**, 181–215.

Garrod, S. and Pickering, M. J. (2009), 'Joint action, interactive alignment, and dialog', *Topics in Cognitive Science* **1**, 292–304.

Garrod, S. and Sanford, A. (1977), 'Interpreting anaphoric relations: the integration of semantic information while reading', *Journal of Verbal Learning and Verbal Behavior* **16**, 77–90.

Gauker, C. (2007), 'A critique of the similarity space theory of concepts', *Mind and Language* **22**, 317–345.

Geeraerts, D. (1993), 'Vagueness's puzzles, polysemy's vagaries', *Cognitive Linguistics* **4**, 223–272.

Gelman, R. and Gallistel, C. R. (2004), 'Language and the origin of number concepts', *Science* **306**, 441–443.

Gelman, S. A., Coley, J. D., Rosengran, K. S., Hartman, E. and Pappas, A. (1998), 'Beyond labeling: the role of maternal input in the acquisition of richly structured categories', *Monographs of the Society for Research in Child Development* **63**, 1–148.

Gelman, S. A. and Markman, E. M. (1986), 'Categories and induction in young children', *Cognition* **23**, 183–209.

Gennari, S., Sloman, S. A., Malt, B. C. and Fitch, T. (2002), 'Motion events in language and cognition', *Cognition* **83**, 49–79.

Gentner, D. (2006), Why verbs are hard to learn, *in* K. Hirsh-Pasek and R. Golinkoff, eds, 'Action meets word: how children learn verbs', Oxford University Press, Oxford, pp. 544–564.

Gentner, D. and Boroditsky, L. (2001), Individuation, relativity and early word learning, *in* M. Bowerman and S. Levinson, eds, 'Language acquisition and conceptual development', Cambridge University Press, Cambridge, pp. 215–256.

Gentner, D. and Goldin-Meadow, S. (2003*a*), Whither Whorf, *in* D. Gentner and S. Goldin-Meadow, eds, 'Language in mind: advances in the study of language and thought', MIT Press, Cambridge, MA, pp. 3–14.

Gentner, D. and Goldin-Meadow, S., eds (2003*b*), *Language in mind: advances in the study of language and thought*, MIT Press, Cambridge, MA.

Gergle, D., Kraut, R. E. and Fussell, S. R. (2004), 'Language efficiency and visual technology: minimizing collaborative effort with visual information', *Journal of Language and Social Psychology* **23**, 1–27.

Gilbert, A. L., Regier, T., Kay, P. and Ivry, R. B. (2006), Whorf hypothesis is supported in the right visual field but not the left, *in* 'Proceedings of the National Academy of Sciences', Vol. 103, pp. 489–494.

Giles, H., Coupland, N. and Coupland, J. (1992), Accommodation theory: communication, context and consequences, *in* H. Giles, N. Coupland and J. Coupland, eds, 'Contexts of accommodation', Cambridge University Press, Cambridge, pp. 1–68.

Gleitman, L. and Papafragou, A. (2005), Language and thought, *in* K. Hollyoak and R. Morrison, eds, 'Cambridge handbook of thinking and reasoning', Cambridge University Press, Cambridge, pp. 633–661.

Gliga, T., Volein, A. and Csibra, G. (2010), 'Verbal labels modulate perceptual object processing in 1-year-old children', *Journal of Cognitive Neuroscience* **22**, 2781–2789.

Goldin-Meadow, S. and Wagner, S. M. (2005), 'How our hands help us learn', *Trends in Cognitive Sciences* **9**, 234–241.

Goldstein, J., Davidoff, J. and Roberson, D. (2009), 'Knowing color terms enhances recognition: further evidence from English and Himba', *Journal of Experimental Child Psychology* **102**, 219–238.

Goldstein, K. (1948), *Language and language disturbances*, Grune and Stratton, New York.

Goldstone, R. L. (1994), 'Influences of categorization on perceptual discrimination', *Journal of Experimental Psychology: General* **123**, 178–200.

Goldstone, R. L. (1998), 'Perceptual learning', *Annual Review of Psychology* **49**, 585–612.

Goldstone, R. L. (1999), Similarity, *in* R. A. Wilson and F. C. Keil, eds, 'MIT encyclopedia of the cognitive sciences', MIT Press, Cambridge, MA, pp. 763–765.

Goldstone, R. L., Lippa, Y. and Shiffrin, R. M. (2001), 'Altering object representations through category learning', *Cognition* **78**, 27–43.

Gopnik, A. and Meltzoff, A. N. (1997), *Words, thoughts, and theories*, MIT Press, London.

Gordon, P. (2004), 'Numerical cognition without words: evidence from Amazonia', *Science* **306**, 496–499.

Gravlee, C. C. (2004), 'Ethnic classification in southeastern Puerto Rico: the cultural model of "color"', *Social Forces* **83**, 949–970.

Grice, H. P. (1975), Logic and conversation (from the William James lectures, Harvard University, 1967), *in* P. Cole and J. Morgan, eds, 'Syntax and semantics 3: speech acts', Academic Press, New York, pp. 41–58.

Gumperz, J. and Levinson, S. C., eds (1996), *Rethinking linguistic relativity*, Cambridge University Press, Cambridge.

Hahn, U. and Ramscar, M. (2001*a*), Introduction: similarity and categorization, *in* U. Hahn and M. Ramscar, eds, 'Similarity and categorization', Oxford University Press, Oxford, pp. 1–11.

Hahn, U. and Ramscar, M., eds (2001*b*), *Similarity and categorization*, Oxford University Press, Oxford.

Hall, D. G. and Waxman, S. R., eds (2004), *Weaving a lexicon*, MIT Press, London.

Hampton, J. A. (1979), 'Polymorphous concepts in semantic memory', *Journal of Verbal Learning and Verbal Behavior* **18**, 441–461.

Hampton, J. A. (1989), 'Concepts and correct thinking', *Mind and Language* **4**, 35–42.

Hanna, J. E. and Brennan, S. E. (2007), 'Speakers' eye gaze disambiguates referring expressions early during face-to-face conversation', *Journal of Memory and Language* **57**, 596–615.

Hanna, J. E., Tanenhaus, M. K. and Trueswell, J. C. (2003), 'The effects of common ground and perspective on domains of referential interpretation', *Journal of Memory and Language* **49**, 43–61.

Harnad, S. (2005), Cognition is categorization, *in* H. Cohen and C. Lefebvre, eds, 'Handbook of categorization in cognitive science', Elsevier, Amsterdam, pp. 20–43.

Harnad, S., ed. (1987), *Categorical perception: the groundwork of cognition*, Cambridge University Press, Cambridge.

Harris, H. D., Murphy, G. L. and Rehder, B. (2008), 'Prior knowledge and exemplar frequency', *Memory and Cognition* **36**, 1335–1350.

Haslam, C., Wills, A. J., Haslam, S. A., Kay, J., Baron, R. and McNab, F. (2007), 'Does maintenance of colour categories rely on language? Evidence to the contrary from a case of semantic dementia', *Brain and Language* **103**, 251–263.

Hatfield, E., Cacioppo, J. T. and Rapson, R. L. (1994), *Emotional contagion*, Cambridge University Press, Cambridge.

Haun, D. B. M., Call, J., Janzen, G. and Levinson, S. C. (2006), 'Evolutionary psychology of spatial representations in the Hominidae', *Current Biology* **16**, 1736–1740.

Haun, D. B. M., Rapold, C., Call, J., Janzen, G. and Levinson, S. C. (2006), 'Cognitive cladistics and cultural override in Hominid spatial cognition', *Proceedings of the National Academy of Sciences* **103**, 17568–17573.

Hauser, M. D. and Carey, S. (2003), 'Spontaneous representations of small numbers of ojbects by rhesus macaques: examinations of content and format', *Cognitive Psychology* **47**, 367–401.

Hauser, M. D., Chomsky, N. and Fitch, W. T. (2002), 'The faculty of language: what is it, who has it, and how did it evolve?', *Science* **298**, 1569–1579.

Heider, E. R. (1972*a*), 'Probabilities, sampling, and the ethnographic method: the case of Dani colour names', *Man* **7**, 448–466.

Heider, E. R. (1972*b*), 'Universals in color naming and memory', *Journal of Experimental Psychology* **93**, 10–20.

Heider, E. R. and Olivier, D. C. (1972), 'The structure of the color space in naming and memory for two languages', *Cognitive Psychology* **3**, 337–354.

Heit, E. (1998), 'Influences of prior knowledge on selective weighting of category members', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**, 712–731.

Hellige, J. B. (1993), *Hemispheric assymetry: what's right and what's left*, Harvard University Press, Cambridge, MA.

Hespos, S. J. and Spelke, E. S. (2004), 'Conceptual precursors to language', *Nature* **430**, 453–456.

Hickerson, N. (1971), 'Review of Berlin and Kay (1969)', *International Journal of American Linguistics* **37**, 257–270.

Hodges, J. R., Patterson, K., Oxbury, S. and Funnell, E. (1992), 'Semantic dementia: progressive fluent aphasia with temporal lobe atrophy', *Brain* **115**, 1783–1806.

Horton, W. S. (2007), 'The influence of partner-specific memory associations on language production: evidence from picture naming', *Language and Cognitive Processes* **22**, 1114–1139.

Horton, W. S. and Gerrig, R. J. (2002), 'Speakers' experiences and audience design: knowing when and knowing how to adjust utterances to addressees', *Journal of Memory and Language* **47**, 589–606.

Horton, W. S. and Gerrig, R. J. (2005*a*), 'Conversation common ground and memory processes in language production', *Discourse Processes* **40**, 1–35.

Horton, W. S. and Gerrig, R. J. (2005*b*), 'The impact of memory demands on audience design during language production', *Cognition* **96**, 127–142.

Horton, W. S. and Keysar, B. (1996), 'When do speakers take into account common ground?', *Cognition* **59**, 91–117.

Hubert, L. and Arabie, P. (1985), 'Comparing partitions', *Journal of Classification* **2**, 193–218.

Hunt, E. and Agnoli, E. (1991), 'The Whorfian hypothesis: a cognitive psychology perspective', *Psychological Review* **98**, 377–389.

Hurford, J. R. (1987), *Language and number: the emergence of a cognitive system*, Basic Blackwell, Cambridge, MA.

Hurford, J. R. (1999), The evolution of language and languages, *in* R. Dunbar, C. Knight and C. Power, eds, 'The evolution of culture', Edinburgh University Press, Edinburgh, pp. 173–193.

Hurford, J. R. (2007), *The origins of meaning: language in the light of evolution*, Oxford University Press.

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., Scherer, S. W. and Lee, C. (2004), 'Detection of large-scale variation in the human genome', *Nature Genetics* **36**, 949–951.

Imai, M. and Mazuka, R. (2007), 'Language-relative construal of individuation constrained by universal ontology: revisiting language universals and linguistic relativity', *Cognitive Science* **31**, 385–413.

Imai, S. and Garner, W. R. (1965), 'Discriminability and preference for attributes in free and constrained classification', *Journal of Experimental Psychology* **69**, 596–608.

Jackendoff, R. (1983), *Semantics and cognition*, MIT Press, Cambridge, MA.

Jackendoff, R. (1989), 'What is a concept, that a person may grasp it?', *Mind and Language* **4**, 68–102.

Jacobson, N. S. and Truax, P. (1991), 'Clinical significance: a statistical approach to defining meaningful change in psychotherapy research', *Journal of Consulting and Clinical Psychology* **59**, 12–19.

James, W. (1890/1981), *The principles of psychology*, Harvard University Press, Cambridge, MA.

January, D. and Kako, E. (2007), 'Re-evaluating evidence for linguistic relativity: reply to Boroditsky (2001)', *Cognition* **104**, 417–426.

Johnson-Laird, P. N. (1983), *Mental models: towards a cognitive science of language, inference, and consciousness*, Cambridge University Press, Cambridge.

Johnson, M. (1987), *The body in the mind: the bodily basis of meaning, imagination, and reason*, University of Chicago Press, Chicago.

Juslin, P., Jones, S., Olsson, H. and Winman, A. (2003), 'Cue abstraction and exemplar memory in categorization', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **29**, 924–941.

Karmiloff-Smith, A. (1992), *Beyond modularity: a developmental perspective on cognitive science*, MIT Press, Cambridge, MA.

Katz, J. J. and Fodor, J. A. (1963), 'The structure of a semantic theory', *Language* **39**, 170–210.

Kay, J. and Ellis, A. (1987), 'A cognitive neuropsychological case study of anomia', *Brain* **110**, 613–629.

Kay, P., Berlin, B., Maffi, L. and Merrifield, W. (1997), Color naming across languages, *in* C. L. Hardin, ed., 'Color categories in thought and language', Cambridge University Press, Cambridge, pp. 21–56.

Kay, P. and Kempton, W. (1984), 'What is the Sapir-Whorf hypothesis?', *American Anthropologist* **86**, 65–79.

Kay, P. and Regier, T. (2003), 'Resolving the question of color naming universals'.

Kay, P. and Regier, T. (2006), 'Language, thought and color: recent developments', *Trends in Cognitive Sciences* **10**, 51–54.

Keil, F. C. (1992), *Concepts, kinds and cognitive development*, MIT Press, Cambridge, MA.

Keil, F. C. (1998), 'Words, moms, and things: language as road map to reality. Commentary in Gelman, S., Coley, J., Rosengren, K., Hartman, E., and Pappas, A.', *Monographs of the Society for Research in Child Development* **63**, 149–157.

Keysar, B., Barr, D. J., Balin, J. A. and Brauner, J. S. (2000), 'Taking perspective in conversation: the role of mutual knowledge in comprehension', *Psychological Science* **11**, 32–38.

Keysar, B. and Henly, A. S. (2002), 'Speakers' overestimation of their effectiveness', *Psychological Science* **13**, 207–212.

Kirby, S., Cornish, H. and Smith, K. (2008), 'Cumulative cultural evolution in the laboratory: an experimental approach to the origins of structure in human language', *Proceedings of the National Academy of Sciences* **105**, 10681–10685.

Koenig, P., Smith, E. E., Glosser, G., DeVita, C., Moore, P., McMillan, C., Gee, J. and Grossman, M. (2005), 'The neural basis for novel semantic categorization', *Neuroimage* **24**, 369–383.

Krauss, R. M. and Weinheimer, S. (1966), 'Concurrent feedback, confirmation, and the encoding of referents in verbal communication', *Journal of Personality and Social Psychology* **4**, 343–346.

Kraut, R. E., Fussell, S. R. and Siegel, J. (2003), 'Visual information as a conversational resource in collaborative physical tasks', *Human-Computer Interaction* **18**, 13–49.

Kripke, S. (1972), *Naming and necessity*, Harvard University Press, Cambridge, MA.

Kronmüller, E. and Barr, D. J. (2007), 'Perspective-free pragmatics: broken precedents and the recovery-from-preemption hypothesis', *Journal of Memory and Language* **56**, 436–455.

Lakoff, G. (1987), *Women, fire and dangerous things: what categories reveal about the mind*, University of Chicago Press, Chicago.

Lakoff, G. and Johnson, M. (1980), *Metaphors we live by*, The University of Chicago Press, Chicago.

Lamme, V. A. F., Super, H. and Spekreijse, H. (1998), 'Feedforward, horizontal, and feedback processing in the visual cortex', *Current Opinion in Neurobiology* **8**, 529–535.

Landau, B. (2000), 'Concepts, the lexicon and acquisition: Fodor's new challenge', *Mind and Language* **15**, 319–326.

Langacker, R. W. (1987), *Foundations of cognitive grammar, volume 1: Theoretical prerequisites*, Stanford University Press, Stanford, CA.

Lantz, D. and Stefflre, R. (1964), 'Language and cognition revisited', *Journal of Abnormal and Social Psychology* **69**, 472–481.

Lenneberg, E. H. (1961), 'Color naming, color recognition, color discrimination: a reappraisal', *Perceptual and Motor Skills* **12**, 375–382.

Levelt, W. J. M. (1989), *Speaking: from intention to articulation*, MIT Press, Cambridge, MA.

Levelt, W. J. M., Roelofs, A. and Meyer, A. S. (1999), 'A theory of lexical access in speech production', *Behavioral and Brain Sciences* **22**, 1–75.

Levinson, S. C. (1997), From outer to inner space: linguistic categories and non-linguistic thinking, *in* J. Nuyts and E. Pederson, eds, 'Language and conceptualization', Cambridge University Press, Cambridge, pp. 13–45.

Levinson, S. C. (2003), *Space in language and cognition*, Cambridge University Press, Cambridge.

Lewis, D. K. (1969), *Convention: a philosophical study*, Harvard University Press, Cambridge, MA.

Li, P. and Gleitman, L. (2002), 'Turning the tables: language and spatial reasoning', *Cognition* **83**, 265–294.

Liberman, A. M., Harris, K. S., Hoffman, H. S. and Griffith, B. C. (1957), 'The discrimination of speech sounds within and across phoneme boundaries', *Journal of Experimental Psychology* **54**, 358–368.

Lindsey, D. T. and Brown, A. M. (2006), 'Universality of color names', *Proceedings of the National Academy of Sciences* **103**, 16608–16613.

Lloyd-Jones, T. J. and Humphreys, G. W. (1997), 'Categorizing chairs and naming pears: category differences in object processing as a function of task and priming', *Memory and Cognition* **25**, 606–624.

Lockridge, C. B. and Brennan, S. E. (2002), 'Addressees' needs influence speakers' early syntactic choices', *Psychonomic Bulletin and Review* **9**, 550–557.

Loewenstein, J. and Gentner, D. (2005), 'Relational language and the development of relational mapping', *Cognitive Psychology* **50**, 315–353.

Lucy, J. A. (1992), *Grammatical categories and cognition: a case study of the linguistic relativity hypothesis*, Cambridge University Press, Cambridge.

Lucy, J. A. and Gaskins, S. (2001), Grammatical categories and the development of classification preferences: a comparative approach, *in* M. Bowerman and S. C. Levinson, eds, 'Language acquisition and conceptual development', Cambridge University Press, Cambridge, pp. 257–283.

Lucy, J. and Shweder, R. (1979), 'Whorf and his critics: linguistic and nonlinguistic influences on color memory', *American Anthropologist* **81**, 581–615.

Lupyan, G. (2008*a*), 'The conceptual grouping effect: categories matter (and named categories matter more)', *Cognition* **108**, 566–577.

Lupyan, G. (2008*b*), 'From chair to "chair": a representational shift account of object labeling effects on memory', *Journal of Experiment Psychology: General* **137**, 348–369.

Lupyan, G. (2009), 'Extracommunicative functions of language: verbal interference causes selective categorization impairments', *Psychonomic Bulletin & Review* **16**, 711–718.

Lupyan, G., Rakison, D. H. and McClelland, J. L. (2007), 'Language is not just for talking: labels facilitate learning of novel categories', *Psychological Science* **18**, 1077–1083.

Machery, E. (2005), 'Concepts are not a natural kind', *Philosophy of Science* **72**, 444–465.

Malt, B. C. (2006), Opening the black box on language, culture, and thought, *in* 'Proceedings of the 5th international conference of the cognitive sciences', Lawrence Erlbaum Associates, Mahwah, NJ, pp. 60–61.

Malt, B. C. and Sloman, S. (2003), 'Linguistic diversity and object naming by non-native speakers of English', *Bilingualism: Language and Cognition* **6**, 47–67.

Malt, B. C. and Sloman, S. (2004), 'Conversation and convention: enduring influences on name choice for common objects', *Memory and Cognition* **32**, 1346–1354.

Malt, B. C., Sloman, S. and Gennari, S. (2003), 'Universality and language specificity in object naming', *Journal of Memory and Language* **49**, 20–42.

Malt, B. C., Sloman, S., Gennari, S., Shi, M. and Wang, Y. (1999), 'Knowing versus naming: similarity and the linguistic categorization of artifacts', *Journal of Memory and Language* **40**, 230–262.

Mandler, J. M. (2000), 'Perceptual and conceptual processes in infancy', *Journal of Cognition and Development* **1**, 3–36.

Mandler, J. M. (2004), 'Thought before language', *Trends in Cognitive Sciences* **8**, 508–513.

Mandler, J. M., Bauer, P. J. and McDonough, L. (1991), 'Separating the sheep from the goats: differentiating global categories', *Cognitive Psychology* **23**, 263–298.

Markman, A. B. and Makin, V. S. (1998), 'Referential communication and category acquisition', *Journal of Experimental Psychology: General* **127**, 331–354.

Martens, M. A., Wilson, S. J. and Reutens, D. C. (2008), 'Research review: Williams syndrome: a critical review of the cognitive, behavioral, and neuroanatomical phenotype', *Journal of Child Psychology and Psychiatry* **49**, 576–608.

Martin, A. (2007), 'The representation of object concepts in the brain', *Annual Review of Psychology* **58**, 25–45.

Matthews, C., Kirby, S. and Cornish, H. (2010), The cultural evolution of language in a world of continuous meanings, *in* A. D. M. Smith, M. Schouwstra, B. de Boer and K. Smith, eds, 'The evolution of language (EVOLANG 8)', World Scientific Press, pp. 451–452.

McDonough, L., Choi, S. and Mandler, J. M. (2003), 'Understanding spatial relations: flexible infants, lexical adults', *Cognitive Psychology* **46**, 229–259.

Medin, D. L. and Aguilar, C. M. (1999), Categorization, *in* R. A. Wilson and F. C. Keil, eds, 'The MIT encyclopedia of the cognitive sciences', MIT Press, Cambridge, MA, pp. 104–106.

Medin, D. L., Dewey, G. L. and Murphy, T. D. (1983), 'Relationships between item and category learning: evidence that abstraction is not automatic', *Journal of Experimental Psychology: Learning, Memory and Cognition* **9**, 607–625.

Medin, D. L., Goldstone, R. L. and Gentner, D. (1993), 'Respects for similarity', *Psychological Review* **100**, 254–278.

Medin, D. L., Lynch, E. B., Coley, J. D. and Atran, S. (1997), 'Categorization among tree experts: do all roads lead to Rome?', *Cognitive Psychology* **32**, 49–96.

Medin, D. L., Lynch, E. B. and Solomon, K. O. (2000), 'Are there kinds of concepts?', *Annual Review of Psychology* **51**, 121–147.

Medin, D. L. and Schaffer, M. M. (1978), 'Context theory of classification learning', *Psychological Review* **85**, 207–238.

Medin, D. L. and Schwanenflugel, P. J. (1981), 'Linear separability in classification learning', *Journal of Experimental Psychology: Human Learning and Memory* **7**, 355–368.

Meila, M. (2007), 'Comparing clusterings – an information based distance', *Journal of Multivariate Analysis* **98**, 873–895.

Meila, M. and Heckerman, D. (2001), 'An experimental comparison of model-based clustering methods', *Machine Learning* **42**, 9–29.

Mervis, C. B. and Rosch, E. (1981), 'Categorization of natural objects', *Annual Review of Psychology* **32**, 89–115.

Metzing, C. and Brennan, S. E. (2003), 'When conceptual pacts are broken: partner-specific effects on the comprehension of referring expressions', *Journal of Memory and Language* **49**, 201–213.

Millikan, R. G. (2000), *On clear and confused ideas: an essay about substance concepts*, Cambridge University Press, Cambridge.

Munnich, E., Landau, B. and Dosher, B. A. (2001), 'Spatial language and spatial representation: a cross-linguistic comparison', *Cognition* **81**, 171–207.

Murphy, G. L. (2002), *The big book of concepts*, MIT Press, Cambridge, MA.

Murphy, G. L. and Kaplan, A. S. (2000), 'Feature distribution and background knowledge in category learning', *Quarterly Journal of Experimental Psychology A: Human Experimental Psychology* **53A**, 962–982.

Murphy, G. L. and Medin, D. L. (1985), 'The role of theories in conceptual coherence', *Psychological Review* **92**, 289–316.

Murray-Smith, R. D., Ramsay, A., Garrod, S., Jackson, M. and Musizza, B. (2007), Gait alignment in mobile phone conversations, *in* A. D. Cheok and L. Chittaro, eds, 'Proceeding of MobileHCI 2007', Vol. 309 of *AMC International Conference Proceeding Series*, Singapore, pp. 214–221.

Nakamura, H., Nakanishi, M., Hamanaka, T., Nakaaki, S. and Yoshida, S. (2000), 'Semantic priming in patients with Alzheimer and semantic dementia', *Cortex* **36**, 151–162.

Neiworth, J. J. and Wright, A. A. (1994), 'Monkeys (*Macaca mulatta*) learn category matching in nonidentical same-different task', *Journal of Experimental Psychology: Animal Behavior Processes* **20**, 429–435.

Nelson, K. (1996), *Language in cognitive development: emergence of the mediated mind*, Cambridge University Press, Cambridge.

Nosofsky, R. M. and Johansen, M. K. (2000), 'Exemplar-based accounts of "multiple-system" phenomena in perceptual categorization', *Psychonomic Bulletin and Review* **7**, 375–402.

Nosofsky, R. M. and Palmeri, T. J. (1997), 'An exemplar-based random walk model of classification learning', *Psychological Review* **104**, 266–300.

Nosofsky, R. M., Palmeri, T. J. and McKinley, S. C. (1994), 'Rule-plus-exception model of classification learning', *Psychological Review* **101**, 53–79.

Nuyts, J. and Pederson, E., eds (1997), *Language and conceptualization*, Cambridge University Press, Cambridge.

Ogden, C. K. and Richards, I. A. (1923), *The meaning of meaning*, Kegan Paul, Trench, Trubner & Co., Ltd., London.

Palmeri, T. J. and Blalock, C. (2000), 'The role of background knowledge in speeded perceptual categorization', *Cognition* **77**, B54–B57.

Papafragou, A., Massey, C. and Gleitman, L. (2002), 'Shake, rattle, 'n' roll: the representation of motion in language and cognition', *Cognition* **84**, 189–219.

Pardo, J. S. (2006), 'On phonetic convergence during conversation interaction', *Journal of the Acoustical Society of America* **119**, 2382–2393.

Pazzani, M. J. (1991), 'Influence of prior knowledge on concept acquisition: experimental and computational results', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **17**, 416–432.

Peacocke, C. (1992), *A study of concepts*, MIT Press, Cambridge, MA.

Peirce, C. S. (1932), *Collected papers of Charles Sanders Peirce*, Vol. 2. Elements of logic, Harvard University Press, Cambridge, MA. Edited by C. Hartshorne, P. Weiss, and A. W. Burks.

Pepperberg, I. M. (1999), *The Alex studies: cognitive and communicative abilities of grey parrots*, Harvard University Press, Cambridge, MA.

Pica, P., Lerner, C., Izard, V. and Dehaene, S. (2004), 'Exact and approximate arithmetic in an Amazonian indigenous group', *Science* **306**, 499–501.

Pickering, M. J. and Garrod, S. (2004), 'Towards a mechanistic psychology of dialogue', *Behavioral and Brain Sciences* **27**, 169–190.

Pinheiro, J. C. and Bates, D. M. (2000), *Mixed-effects models in S and S-PLUS*, Springer Verlag, New York.

Pinker, S. (1994), *The language instinct*, Morrow, New York.

Pinker, S. and Jackendoff, R. (2005), 'The faculty of language: what's special about it?', *Cognition* **95**, 201–236.

Pinker, S. and Prince, A. (1987), 'On language and connectionism: analysis of a parallel distributed processing model of language acquisition', *Cognition* **28**, 73–194.

Plunkett, K., Hu, J.-F. and Cohen, L. B. (2008), 'Labels can override perceptual categories in early infancy', *Cognition* **106**, 665–681.

Prinz, J. (2002), *Furnishing the mind: concepts and their perceptual basis*, MIT Press, Cambridge, MA.

Prinz, J. J. (2005), The return of concept empiricism, *in* H. Cohen and C. Lefebvre, eds, 'Handbook of categorization in cognitive science', Elsevier, Amsterdam, pp. 679–695.

Putnam, H. (1975), *Mind, language and reality: philosophical papers, ii*, Cambridge University Press, Cambridge.

Putnam, H. (1981), *Reason, truth and history*, Cambridge University Press, Cambridge.

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical Association* **66**, 846–850.

Ratneshwar, S., Barsalou, L. W., Pechmann, C. and Moore, M. (2001), 'Goal-derived categories: the role of personal and situational goals in category representations', *Journal of Consumer Psychology* **10**, 147–157.

Ravin, Y. and Leacock, C., eds (2000), *Polysemy: an overview*, Oxford University Press, Oxford.

Regier, T. and Kay, P. (2009), 'Language, thought and color: Whorf was half right', *Trends in Cognitive Sciences* **13**, 439–446.

Regier, T., Kay, P. and Cook, R. S. (2005), 'Focal colors are universal after all', *Proceedings of the National Academy of Sciences* **102**, 8386–8391.

Richardson, D. C. and Dale, R. (2005), 'Looking to understand: the coupling between speakers' and listeners' eye movements and its relationship to discourse comprehension', *Cognitive Science* **29**, 1045–1060.

Richardson, D. C., Dale, R. and Kirkham, N. Z. (2007), 'The art of conversation is coordination: common ground and the coupling of eye movements during dialogue', *Psychological Science* **18**, 407–413.

Richardson, M. J., Marsh, K. L., Isenhower, R. W., Goodman, J. R. L. and Schmidt, R. C. (2007), 'Rocking together: dynamics of intentional and unintentional interpersonal coordination', *Human Movement Science* **26**, 867–891.

Rips, L. J. (1989), Similarity, typicality, and categorization, *in* S. Vosniadou and A. Ortony, eds, 'Similarity and analogical reasoning', Cambridge University Press, Cambridge, MA, pp. 21–59.

Rivera, S. M. and Zawaydeh, A. N. (2007), 'Word comprehension facilitates object individuation in 10- and 11-month-old infants', *Brain Research* **1146**, 146–157.

Rivera, S. and Zawaydeh, A. (2006), 'Word comprehension facilitates object individuation in 10- and 11-month-old infants', *Brain Research* **1146**, 146–157.

Roberson, D. and Davidoff, J. (2000), 'The categorical perception of colors and facial expressions: the effect of verbal interference', *Memory and Cognition* **28**, 977–986.

Roberson, D., Davidoff, J. and Braisby, N. (1999), 'Similarity and categorisation: neuropsychological evidence for a dissociation in explicit categorisation tasks', *Cognition* **71**, 1–42.

Roberson, D., Davidoff, J., Davies, I. R. L. and Shapiro, L. R. (2004), 'The development of color caetgories in two languages: a longitudinal study', *Journal of Experimental Psychology: General* **133**, 554–571.

Roberson, D., Davidoff, J., Davies, I. R. L. and Shapiro, L. R. (2005), 'Color categories: evidence for the cultural relativity hypothesis', *Cognitive Psychology* **50**, 378–411.

Roberson, D., Davies, I. and Davidoff, J. (2000), 'Color categories are not universal: replication and new evidence from a stone-age culture', *Journal of Experimental Psychology: General* **129**, 369–398.

Roberson, D., Davies, I. R. L., Corbett, G. and Vandervyver, M. (2005), 'Freesorting of colors across cultures: are there universal grounds for grouping?', *Journal of Cognition and Culture* **5**, 349–386.

Roberson, D., Pak, H. and Hanley, J. R. (2008), 'Categorical perception of colour in the left and right visual field is verbally mediated: evidence from Korean', *Cognition* **107**, 752–762.

Romney, A. K., Batchelder, W. H. and Brazill, T. (1995), Scaling semantic domains, *in* T. Indow and R. D. Luce, eds, 'Geometric representations of perceptual phenomena: papers in honor of Tarow Indow on his 70th birthday', Lawrence Erlbaum, Mahwah, NJ, pp. 267–294.

Romney, A. K., Boyd, J. P., Moore, C. C., Batchelder, W. H. and Brazill, T. J. (1996), 'Culture as shared cognitive representations', *Proceedings of the National Academy of Sciences* **93**, 4699–4705.

Romney, A. K., Moore, C. C., Batchelder, W. H. and Hsia, T. (2000), 'Statistical methods for characterizing similarities and differences between semantic structures', *Proceedings of the National Academy of Sciences* **97**, 518–523.

Room, A. (1981), *Room's dictionary of distinguishables*, Routledge & Kegan Paul, Boston.

Rosch, E. (1975), 'Cognitive representations of semantic categories', *Journal of Experimental Psychology: General* **104**, 192–233.

Rosch, E. (1978), Principles of categorization, *in* E. Rosch and B. B. LLoyd, eds, 'Cognition and categorization', Lawrence Erlbaum, Hillsdale, NJ, pp. 27–48.

Rosch, E. and Mervis, C. B. (1975), 'Family resemblances: studies in the internal structure of categories', *Cognitive Psychology* **7**, 573–605.

Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M. and Boyes-Braem, P. (1976), 'Basic objects in natural categories', *Cognitive Psychology* **8**, 382–439.

Rosch, E., Simpson, C. and Miller, R. S. (1976), 'Structural bases of typicality effects', *Journal of Experimental Psychology: Human Perception and Performance* **2**, 491–502.

Ross, B. H. and Murphy, G. L. (1999), 'Food for thought: cross-classification and category organization in a complex real-world domain', *Cognitive Psychology* **38**, 495–553.

Ross, B. H., Perkins, S. J. and Tenpenny, P. L. (1990), 'Reminding-based category learning', *Cognitive Psychology* **22**, 460–492.

Ross, D. (1951), *Plato's theory of ideas*, Oxford University Press, Oxford.

Roth, E. M. and Shoben, E. J. (1983), 'The effect of context on the structure of categories', *Cognitive Psychology* **15**, 346–378.

Rothbart, M., Davis-Stitt, C. and Hill, J. (1997), 'Effects of arbitrarily placed category boundaries on similarity judgments', *Journal of Experimental Social Psychology* **33**, 122–145.

Rothbart, M. and Lewis, S. (1988), 'Inferring category attributes from exemplar attributes: geometric shapes and social categories', *Journal of Personality and Social Psychology* **55**, 861–872.

Rousselet, G. A., Fabre-Thorpe, M. and Thorpe, S. J. (2002), 'Parallel processing in high-level categorization of natural images', *Natural Neuroscience* **5**, 629–630.

Rumelhart, D. E. and McClelland, J. L., eds (1986), *Parallel distributed processing: explorations in the microstructure of cognition*, Vol. 2, MIT Press, Cambridge, MA.

Russell, B. (1905), 'On denoting', *Mind* **14**, 479–493.

Savage-Rumbaugh, S., Sevcik, R. A., Brakke, K. E. and Rumbaugh, D. M. (1990), Symbols: their communicative use, comprehension, and combination by bonobos (pan paniscus), *in* L. P. Lipsitt and C. Rovee-Collier, eds, 'Advances in infancy research', Ablex Publishing Co., Norwood, NJ, pp. 221–278.

Schafer, G. and Plunkett, K. (1999), 'What's in a name? Lexical knowledge drives infants' visual preferences in the absence of referential input', *Developmental Science* **2**, 187–194.

Schelling, T. C. (1960), *The strategy of conflict*, Harvard University Press, Cambridge, MA.

Schober, M. F. (2004), 'Just how aligned are interlocutors' representations? Commentary on Pickering and Garrod', *Behavioral and Brain Sciences* **27**, 209–210.

Schober, M. F. (2005), Conceptual alignment in conversation, *in* B. F. Malle and S. D. Hodges, eds, 'Other minds: how humans bridge the divide between self and others', Guilford Press, New York, pp. 239–252.

Schober, M. F. and Clark, H. H. (1989), 'Understanding by addressees and overhearers', *Cognitive Psychology* **21**, 211–232.

Schober, M. F., Conrad, F. G. and Fricker, S. S. (2004), 'Misunderstanding standardized language in research interviews', *Applied Cognitive Psychology* **18**, 169–188.

Schyns, P. G., Goldstone, R. L. and Thibaut, J. P. (1998), 'The development of features in object concepts', *Brain and Behavioral Sciences* **21**, 17–54.

Sebanz, N., Bekkering, H. and Knoblich, G. (2006), 'Joint action: bodies and minds moving together', *Trends in Cognitive Sciences* **10**, 70–76.

Segal, G. (2000), *A slim book about narrow content*, MIT Press, Cambridge, MA.

Sera, M. D., Elieff, C., Burch, M. C., Forbes, J. and Rodríguez, W. (2002), 'When language affects cognition and when it does not: an analysis of grammatical gender and classification', *Journal of Experimental Psychology: General* **131**, 377–397.

Seyfarth, R. M. and Cheney, D. L. (1990), *How monkeys see the world*, University of Chicago Press, Chicago.

Shepard, R. N. (1987), 'Toward a universal law of generalization for psychological science', *Science* **237**, 1317–1323.

Shintel, H. and Keysar, B. (2007), 'You said it before and you'll say it again: expectations of consistency in communication', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **33**, 357–369.

Shintel, H. and Keysar, B. (2009), 'Less is more: a minimalist account of joint action in communication', *Topics in Cognitive Science* **1**, 260–273.

Shockley, K., Santana, M. V. and Fowler, C. A. (2003), 'Mutual interpersonal postural constraints are involved in cooperative conversation', *Journal of Experimental Psychology: Human Perception and Performance* **29**, 326–332.

Slobin, D. I. (1996), From "thought and language" to "thinking for speaking", *in* J. Gumperz and S. C. Levinson, eds, 'Rethinking linguistic relativity', Cambridge University Press, Cambridge, pp. 70–96.

Slobin, D. I. (2003), Language and thought online: cognitive consequences of linguistic relativity, *in* D. Gentner and S. Goldin-Meadow, eds, 'Language in mind: advances in the study of language and thought', MIT Press, Cambridge, MA, pp. 157–191.

Sloman, S. A. (1996), 'The empirical case for two systems of reasoning', *Psychological Bulletin* **119**, 3–22.

Sloutsky, V. M., Lo, Y. and Fisher, A. V. (2001), 'How much does a shared name make things similar? Linguistic labels, similarity, and the development of inductive inference', *Child Development* **72**, 1659–1709.

Smith, A. D. M. (2005), Mutual exclusivity: communicative success despite conceptual divergence, *in* M. Tallerman, ed., 'Language origins: perspectives on evolution', Oxford University Press, Oxford, pp. 372–388.

Smith, J. D. and Minda, J. P. (1998), 'Prototypes in the mist: the early epochs of category learning', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **24**, 1411–1436.

Smith, L. B. and Samuelson, L. K. (1997), Perceiving and remembering: category stability, variability and development, *in* K. Lamberts and D. Shanks, eds, 'Knowledge, concepts and categories', Psychology Press, London, pp. 161–196.

Snowden, J. S., Goulding, P. J. and Neary, D. (1989), 'Semantic dementia: a form of circumscribed cerebral atrophy', *Behavioural Neurology* **2**, 167–182.

Solomon, K. O., Medin, D. L. and Lynch, E. (1999), 'Concepts do more than categorize', *Trends in Cognitive Sciences* **3**, 99–104.

Spinozzi, G. (1996), 'Categorization in monkeys and chimpanzees', *Behavioural Brain Research* **74**, 17–24.

Spinozzi, G. and Langer, J. (1999), 'Spontaneous classification in action by a human-enculturated and language-reared bonobo (*pan paniscus*) and common chimpanzee (*pan troglodytes*)', *Journal of Comparative Psychology* **113**, 286–296.

Stalnaker, R. C. (1978), Assertion, *in* P. Cole, ed., 'Syntax and semantics', Academic Press, New York, pp. 315–332.

Steels, L. (1997), 'The synthetic modeling of language origins', *Evolution of Communication* **1**, 1–34.

Stefflre, V., Castillo, V. and Moreley, L. (1966), 'Language and cognition in Yucatan: a cross-cultural replication', *Journal of Personality and Social Psychology* **4**, 112–115.

Studdert-Kennedy, M., M., L. A., Harris, K. S. and Cooper, F. S. (1970), 'Motor theory of speech perception: a reply to Lane's critical review', *Psychological Review* **77**, 234–249.

Sweetser, E. E. (1999), Compositionality and blending: semantic composition in a cognitively realistic framework, *in* G. Redeker and T. Janssen, eds, 'Cognitive linguistics: foundations, scope and methodology', Mouton de Gruyter, Berlin, pp. 129–162.

Tallon-Baudry, C. and Bertrand, O. (1999), 'Oscillatory gamma activity in humans and its role in object representation', *Trends in Cognitive Sciences* **3**, 151–162.

Talmy, L. (1983), How language structures space, *in* H. Pick and L. Acredolo, eds, 'Spatial orientation: theory, research, and application', Plenum Press, New York, pp. 225–282.

Talmy, L. (1985), Lexicalization patterns: semantic structure in lexical forms, *in* T. Shopen, ed., 'Language typology and syntactic description, volume 3: grammatical categories and the lexicon', Cambridge University Press, Cambridge, pp. 57–149.

Tanaka, J. W. and Taylor, M. (1991), 'Object categories and expertise: is the basic level in the eye of the beholder?', *Cognitive Psychology* **23**, 457–482.

Taylor, J. R. (1995), *Linguistic categorisation: prototypes in linguistic theory*, 2nd edn, Clarendon Press, Oxford.

Tomasello, M. (2000), *The cultural origins of human cognition*, Harvard University Press, Cambridge, MA.

Tomasello, M., Carpenter, M., Call, J., Behne, T. and Moll, H. (2005), 'Understanding and sharing intentions: the origins of cultural cognition', *Behavioral and Brain Sciences* **28**, 675–735.

Tomasello, M., Carpenter, M. and Liszkowski, U. (2007), 'A new look at infant pointing', *Child Development* **78**, 705–722.

Tootell, R. B. H., Silverman, M. S., Hamilton, M. S., Switkes, E. and De Valois, R. L. (1988), 'Functional anatomy of macaque striate cortex, v. spatial frequency', *The Journal of Neuroscience* **8**, 1610–1624.

Tversky, A. (1977), 'Features of similarity', *Psychological Review* **84**, 327–352.

Twisk, J. and Proper, K. (2004), 'Evaluation of the results of a randomized controlled trial: how to define changes between baseline and follow-up', *Journal of Clinical Epidemiology* **57**, 223–8.

Uttal, W. R. (2003), *The new phrenology: the limits of localizing cognitive processes in the brain*, MIT Press, Cambridge, MA.

Voiklis, J. K. (2008), A thing is what we say it is: referential communication and indirect category learning, Master's thesis, Columbia University.

Vosniadou, S. and Brewer, W. F. (1992), 'Mental models of the earth: a study of conceptual change in childhood', *Cognitive Psychology* **24**, 535–585.

Warrington, E. K. (1975), 'The selective impairment of semantic memory', *Quarterly Journal of Experimental Psychology* **27**, 635–657.

Waxman, S. R. and Braun, I. (2005), 'Consistent (but not variable) names as invitations to form object categories: new evidence from 12-month-old infants', *Cognition* **95**, B59–B68.

Waxman, S. R. and Markow, D. B. (1995), 'Words as invitations to form categories: evidence from 12- to 13-month-old infants', *Cognitive Psychology* **29**, 257–302.

Weiskopf, D. A. (2009), 'The plurality of concepts', *Synthese* **169**, 145–175.

Whittlesea, B. W. A., Brooks, L. R. and Westcott, C. (1994), 'After the learning is over: factors controlling the selective application of general and particular knowledge', *Journal of Experimental Psychology: Learning, Memory, and Cognition* **20**, 259–274.

Whorf, B. L. (1956), *Language, thought and reality: selected writings of Benjamin Lee Whorf*, MIT Press, Cambridge, MA. Edited by J. B. Carroll.

Wilkes-Gibbs, D. and Clark, H. H. (1992), 'Coordinating beliefs in conversation', *Journal of Memory and Language* **31**, 183–194.

Wills, A. J. and McLaren, I. P. L. (1998), 'Perceptual learning and free classification', *The Quarterly Journal of Experimental Psychology* **51B**, 235–270.

Winawer, J., Witthoft, N., Frank, M., Wu, L., Wade, A. and Boroditsky, L. (2007), 'Russian blues reveal effects of language on color discrimination', *Proceedings of the National Academy of Sciences* **104**, 7780–7785.

Wittgenstein, L. (1953), *Philosophical investigations*, Blackwell, Oxford.

Woolford, E. (1984), 'Universals and rule options in kinship terminology: a synthesis of three formal approaches', *American Ethnologist* **11**, 771–790.

Wright, W. D. and Pitt, F. H. G. (1934), 'Hue-discrimination in normal colour-vision', *Proceedings of the Physical Society* **46**, 459–473.

Xu, F. (2002), 'The role of language in acquiring object kind concepts in infancy', *Cognition* **85**, 223–250.

Xu, F., Cote, M. and Baker, A. (2005), 'Labeling guides object individuation in 12-month-old infants', *Psychological Science* **16**, 372–377.

Yendrikhovskij, S. N. (2001), 'A computational model of colour categorization', *Color Research and Application* **26**, S235–S238.

Zaki, S. R. (2004), 'False prototype enhancement effects in dot pattern categorization', *Memory and Cognition* **32**, 390–398.

Zwaan, R. A. and Radvansky, G. A. (1998), 'Situation models in language comprehension and memory', *Psychological Bulletin* **123**, 162–185.