# Investigating and developing beginner learners' decoding proficiency in second language French:

## An evaluation of two programmes of instruction

Robert Woore
Exeter College / Department of Education

Thesis submitted for the degree of DPhil, University of Oxford
Michaelmas Term, 2011

**Acknowledgments**

So many people have provided me with invaluable support during the completion of this thesis that it would be impossible to fit all the names onto this page! However, I must say a very big THANK YOU to some people in particular.

First of all, I am hugely grateful to all the teachers involved in this project, and to their Heads of Department and other colleagues who supported them. They put up with my frequent visits to their classrooms and allowed their lessons to be disrupted whilst their students completed the Reading Aloud Test and other tasks. Additionally, six teachers delivered programmes of decoding instruction to their classes over a period of around thirty lessons. Without this enormous commitment from them, the project would have been impossible to complete.

Second, I owe a huge debt of thanks to my supervisor, Professor Ernesto Macaro, for all his help and support over the five years of this project. His comments and advice were invaluable: they always hit the nail on the head and helped me to develop my thinking in so many ways. Thank you, Ernesto, for your insights, commitment, patience and the generosity with which you gave your time.

Third, I would like to thank my colleagues at OUDE for their comments and advice as well as for just being very kind and supportive people. In the final year of the project in particular, many members of the PGCE team helped me to create extra time for writing up. Trevor Mutton, as MFL course leader and overall course leader, was especially helpful in this respect. I also owe him a big thank you for acting as a moderator in my qualitative data analysis. He freely gave up a lot of his time to help me, despite being having so many other things to do.

Fourth, I would like to thank my family and friends for their help and support: in particular, my wife Marianne (whom I had not even met when I embarked on this project and who spent five years living with it!). Thank you, Marianne, for all your kindness and patience. My daughter Florence, now aged two and a half, did not exactly help me complete this thesis punctually, but she provided the most wonderful and enjoyable distraction imaginable and I am sure I could not have done it so well without her. Finally, throughout this long, part-time project, my parents also provided infinite moral support and encouragement as well as practical help in countless ways (especially related to childcare!). Thank you, Mum and Dad: I definitely could not have completed this project without you.

I would also like to thank the ESRC for their financial support for this project.

# Abstract

Second language (L2) decoding – the sub-lexical process of mapping the graphemes of an alphabetic writing system onto the phonemes they represent – is argued to underpin various aspects of L2 learning, particularly vocabulary acquisition. Recently, second language acquisition research has shown increased interest in decoding, consistently finding evidence for L1-to-L2 transfer effects on learners' processing mechanisms and outcomes. Correspondingly, studies conducted in Modern Foreign Language (MFL) classrooms in English secondary schools – an under-researched context – have found that beginner learners of French tend to (a) pronounce L2 words according to English decoding conventions and (b) make poor progress in this aspect of L2 learning. Recent official guidance for MFL teachers has addressed this problem by advocating an explicit focus on decoding, but there is a lack of convincing evidence (both in the MFL context and more widely) that explicit L2 decoding instruction can be effective. The current study therefore trialled two programmes of French decoding instruction for beginner MFL learners, delivered in ten- to fifteen-minute segments over around thirty lessons. Three intact secondary school classes followed a phonics-based approach; three classes from another school followed a programme in which learners were encouraged to derive the pronunciations of French graphemes from 'source words' in a memorized poem; and six classes in two other schools received no explicit decoding instruction. Participants (N=186) completed pre- and post-tests of French decoding; a sub-sample (N=15) also completed task-based self-report interviews. The two intervention groups made significantly more progress than the comparison group in terms of the number of graphemes pronounced 'acceptably', although the magnitude of the difference between the groups was small. Compared to the comparison group, the two intervention groups also appeared to show different and more extensive patterns of change in their realizations of individual graphemes, even where their pronunciations were still not 'acceptable'. Finally, self-report data generally revealed little change in participants' strategic reasoning, either in the intervention or comparison group. Together, these findings suggest that explicit instruction can improve beginner learners' proficiency in decoding L2 French, but that their progress may follow a longer and more complex trajectory than simply moving directly from 'incorrect' to 'correct' forms. Further research is required to assess the effects (if any) of a given improvement in decoding proficiency on other language-learning outcomes; and to design and evaluate alternative programmes of instruction.

# Contents

ACKNOWLEDGMENTS ..... 2
ABSTRACT ..... 3
CONTENTS ..... 4
LIST OF TABLES ..... 10
LIST OF FIGURES ..... 11
LIST OF ABBREVIATIONS ..... 12

Chapter 1: Introduction ..... 13
1.1 Background ..... 13
1.2 The Modern Foreign Languages context ..... 14
1.3 Research into L2 writing systems ..... 18

Chapter 2: Literature Review ..... 20
2.1 Introduction ..... 20
Part I: The nature and importance of L2 decoding ..... 23
2.2 Definitions ..... 23
2.2.1 Introduction ..... 23
2.2.2 Graphemes ..... 23
2.2.3 Word recognition and decoding ..... 24
2.3 L2 decoding, reading and vocabulary learning ..... 27
2.3.1 L2 decoding and reading comprehension ..... 27
2.3.1.1 L1 reading ..... 27
2.3.1.2 L2 reading ..... 30
2.3.1.3 Decoding in L2 reading comprehension ..... 32
2.3.2 L2 decoding and vocabulary acquisition ..... 34
2.3.3 Summary of Part I ..... 37
Part II: L2 word recognition, decoding and transfer ..... 39
2.4 How do the French and English writing systems differ? ..... 39
2.4.1 Dimensions of difference between writing systems ..... 39
2.4.2 Grapheme-phoneme correspondences (GPC) in French and English ..... 42
2.4.3 ‘Orthographic Depth’ of French and English ..... 43
2.5 Cross-linguistic differences in L1 word recognition ..... 47
2.5.1 Introduction ..... 47
2.5.2 Word recognition: L1 English ..... 49
2.5.2.1 Dual route models ..... 49

2.5.2.2 Connectionist models ..... 50
2.5.3 Word recognition and its development in different alphabetic writing systems ..... 52
2.5.3.1 Introduction ..... 52
2.5.3.2 Orthographic Depth Hypothesis ..... 54
2.5.3.3 Psycholinguistic Grain Size Theory ..... 57
2.5.4 Summary ..... 61
2.6 Word recognition and decoding in an L2 ..... 62
2.6.1 Introduction ..... 62
2.6.2 Empirical studies ..... 69
2.6.3 Empirical studies of word recognition and decoding in L2 French ..... 75
2.6.4 Phonology in L2 Decoding ..... 86
2.6.5 Summary ..... 88
Part III: The role of explicit instruction ..... 91
2.7 L2 instruction and the development of L2 decoding proficiency ..... 91
2.7.1 Introduction ..... 91
2.7.2 Literacy instruction in L1 English ..... 91
2.7.3 Policy and practice in MFL classrooms ..... 95
2.7.3.1 Little emphasis on L2 decoding ..... 95
2.7.3.2 Renewed official guidance ..... 99
2.7.4 Approaches to explicit L2 decoding instruction ..... 101
2.7.5 Summary ..... 105
Part IV: Overall summary and Research Questions ..... 107

Chapter 3: Research design ..... 112
3.1 Overview ..... 112
3.2 Issues in research design ..... 114
3.2.1 Randomized Control Trials in Educational Research ..... 114
3.2.2 Initial plans for research design ..... 116
3.2.2.1 Introduction ..... 116
3.2.2.2 Sampling: general considerations ..... 116
3.2.2.3 Allocating groups to conditions ..... 118
3.2.2.4 Hawthorne control procedures ..... 121
3.2.2.5 Planned random allocation of groups to conditions ..... 123
3.3 Plans refracted through the prism of reality ..... 125
3.4 Describing the sample ..... 127
3.4.1 Schools ..... 127

3.4.2 Teachers ..... 130
3.4.2.1 Teacher Information Questionnaire ..... 130
3.4.2.2 TIQ Findings ..... 132
3.4.3 Classes ..... 135
3.5 Intervention design ..... 137
3.5.1 Content ..... 137
3.5.2 Format ..... 139
3.6 Main data collection ..... 143
3.6.1 Introduction ..... 143
3.6.2 Baseline information on sample students ..... 144
3.6.2.1 School-held data ..... 144
3.6.2.2 Background Information Questionnaire ..... 146
3.6.2.3 L2 proficiency measures ..... 147
3.6.3 The Reading Aloud Test ..... 150
3.6.3.1 Format of the RAT ..... 150
3.6.3.2 Administration ..... 154
3.6.3.3 Processing RAT data ..... 155
3.6.3.4 Validity ..... 166
3.6.3.5 Reliability ..... 167
3.6.4 Self-report interviews ..... 170
3.6.4.1 Rationale and validity ..... 170
3.6.4.2 Administration ..... 174
3.6.4.3 Analysis ..... 175
3.7 Fidelity to condition ..... 178
3.7.1 Methods ..... 178
3.7.2 Findings ..... 182
3.7.2.1 Non-intervention teaching ..... 182
3.7.2.2 Decoding segments ..... 184
3.7.2.3 Intervention ‘healthcheck’ meetings ..... 185
3.8 Ethics ..... 187

Chapter 4 Findings I: Reading Aloud Test scores ..... 190
4.1 Introduction ..... 190
4.2 Preliminaries ..... 191
4.2.1 Missing data ..... 191
4.2.1.1 Inaudible pronunciations ..... 191

4.2.1.2 Omission of RAT items ..... 192
4.2.1.3 Non-Linear decoding ..... 193
4.2.2 Fidelity to condition ..... 197
4.2.3 Intra- and inter-rater reliability ..... 198
4.2.4 Summary ..... 199
4.3 Comparing intervention and comparison groups ..... 200
4.3.1 Introduction ..... 200
4.3.2 Outliers and extreme values ..... 201
4.3.3 'Correct' realizations (*score1*) ..... 203
4.3.3.1 ANOVA ..... 204
4.3.3.2 Non-parametric tests ..... 208
4.3.4 'Acceptable' realizations (*score2*) ..... 209
4.3.4.1 ANOVA ..... 211
4.3.4.2 Non-parametric tests ..... 214
4.3.4.3 Comparing *score1* and *score2* ..... 215
4.4 Comparing the two interventions ..... 217
4.4.1 Introduction ..... 217
4.4.2 Two-way mixed ANOVA ..... 218
4.4.3 Non-parametric tests ..... 219
4.4.4 Summary ..... 220
4.5 Class-level differences ..... 220
4.6 Analysis by grapheme type ..... 227
4.6.1 Introduction ..... 227
4.6.2 Consonants ..... 231
4.6.3 Vowels ..... 234
4.6.4 Nasal vowels ..... 235
4.7 Summary ..... 236

Chapter 5: Realizations of grapheme tokens ..... 239
5.1 Introduction ..... 239
5.2 Grapheme token <ç> ..... 242
5.2.1 Realizations produced by the whole sample at t1 and t2 ..... 242
5.2.2 Changes over time ..... 245
5.2.3 Summary ..... 247
5.3 Realizations of <ou> ..... 248
5.3.1 Realizations produced by the whole sample at t1 and t2 ..... 248

5.3.2 Analysis of grapheme token <ou> in *ougrien* ..... 251
5.3.3 Changes over time ..... 255
5.3.4 Summary ..... 258
5.4 Realizations of <oin> ..... 259
5.4.1 Realizations produced by the whole sample at t1 and t2 ..... 259
5.4.2 Similarities and differences between the groups ..... 263
5.4.3 Time 1 ..... 265
5.4.4 Changes at t2 ..... 266
5.4.5 Comparing less frequent realizations ..... 267
5.4.5.1 Realizations beginning with [w] ..... 269
5.4.5.2 Realizations including nasality ..... 271
5.4.5.3 Monosyllabic versus bisyllabic realizations ..... 273
5.4.6 Summary ..... 274
5.5 Summary ..... 276

Chapter 6: Self-report interviews ..... 279
6.1 introduction ..... 279
6.2 Overview of approaches to decoding ..... 279
6.2.1 Introduction ..... 279
6.2.2 Decoding strategies ..... 281
6.2.2.1 Dividing words into smaller units ..... 281
6.2.2.2 References to known words ..... 283
6.2.2.3 Trying out and evaluating possible pronunciations ..... 286
6.2.3 Knowledge bases ..... 287
6.2.3.1 Knowledge about specific GPC ..... 288
6.2.3.2 Origins of knowledge ..... 292
6.2.3.3 'Feeling for French' ..... 293
6.2.4 Metacognition and affective reactions ..... 294
6.3 Summary of changes in participants' approaches to decoding ..... 295
6.4 Case studies ..... 298
6.4.1 Case study 1: OK (comparison group) ..... 299
6.4.1.1 Background ..... 299
6.4.1.2 Approaches to decoding at t1 ..... 299
6.4.1.3 Approaches to decoding at t2 ..... 306
6.4.1.4 Change between t1 and t2 ..... 311
6.4.2 Case study 2: EH (Intervention B) ..... 313

6.4.2.1 Background ..... 313
6.4.2.2 Approaches to decoding at t1 ..... 314
6.4.2.3 Approaches to decoding at t2 ..... 322
6.4.2.4 Change between t1 and t2 ..... 328
6.4.3 Case study 3: LV (intervention A) ..... 330
6.4.3.1 Background ..... 330
6.4.3.2 Approaches to decoding at t1 ..... 331
6.4.3.3 Approaches to decoding at t2 ..... 339
6.4.3.4 Relationship between the self-report interviews at t1 and t2 ..... 345

Chapter 7: Discussion and limitations ..... 348
7.1 Research Questions 1 and 2 ..... 348
7.1.1 Non-Linear Decoding ..... 348
7.1.2 Main analysis ..... 350
7.1.3 Comparing intervention and comparison groups ..... 351
7.1.4 Comparing Programmes A and B ..... 354
7.1.5 Interpreting the effects of the instruction ..... 355
7.1.6 Analysis of individual GPC ..... 359
7.2 Research Question 3 ..... 364
7.2.1 Realizations of <ç> ..... 367
7.2.2 Overall realizations of <ou> ..... 369
7.2.3 Realizations of <ou> in ougrien ..... 370
7.2.4 Realizations of <oin> ..... 372
7.2.4 Summary ..... 375
7.3 Research Question 4 ..... 376
7.3.1 Strategies ..... 378
7.3.2 Knowledge bases ..... 380
7.3.3 Metacognition ..... 382
7.3.4 RAT scores of SRI participants ..... 383

Chapter 8: Conclusions ..... 385
8.1 The problem ..... 385
8.2 An attempted solution ..... 387
8.3 Future directions ..... 394

References ..... 397

# List of Tables


| | | |
|---|---|---|
| Table 2.1 | French Diacritics | 41 |
| Table 2.2 | Probabilities for various realizations of English <ou> | 45 |
| Table 3.1 | Participating teachers' native languages and amount of teaching experience | 133 |
| Table 3.2 | Numbers of participants in each school and class with various characteristics | 137 |
| Table 3.3 | Problematic pronunciations of RAT items and illustrations of the linear Mapping Principle. | 160 |
| Table 3.4 | Intervention classes' coverage of planned content in the programmes of instruction | 186 |
| Table 3.5 | Approximate number of minutes reportedly spent on decoding segments in the period preceding each 'healthcheck' interview | 187 |
| Table 4.1 | Mean numbers and percentages of grapheme tokens pronounced inaudibly per participant | 192 |
| Table 4.2 | Mean numbers and percentages of graphemes coded as NLD (Non-Linear Decoding) per participant | 194 |
| Table 4.3 | Numbers of students in each class obtaining NC levels 3-5 in the reading component of English SAT | 196 |
| Table 4.4 | Intra-rater reliability figures | 198 |
| Table 4.5 | Inter-rater reliability figures | 199 |
| Table 4.6 | Results for *score1* (number of target graphemes realized 'correctly') | 206 |
| Table 4.7 | Results for *score2* (number of target graphemes realized 'acceptably') | 212 |
| Table 4.8 | Mean numbers and percentages of target grapheme tokens whose realizations were 'only acceptable' (i.e. *score2-score1*) | 216 |
| Table 4.9 | Results for *score2* (number of target graphemes realized 'acceptably') | 218 |
| Table 4.10 | Results for *change2* (difference in the number of target grapheme tokens pronounced 'acceptably' between t1 and t2) | 223 |
| Table 4.11 | Numbers of participants obtaining negative *change2* values | 226 |
| Table 5.1 | Percentages of participants in each group producing various realizations of <ç> at each time point | 245 |
| Table 5.2 | Numbers of distinct realizations of <ou> in each group at t1 and t2 | 253 |
| Table 5.3 | Realizations of <ou> produced by four or more participants at each time point, ordered by descending frequency | 254 |
| Table 5.4 | Percentage of participants in each group producing frequent realizations of <ou> at each time point | 257 |
| Table 5.5 | Numbers of distinct realizations of <oin> produced by each group at t1 and t2 | 261 |
| Table 5.6 | Attested realizations of <oin> produced by four or more participants at each time point, ordered by descending frequency | 262 |
| Table 5.7 | Percentage of participants in each group producing various frequent realizations of <oin> at each time point | 266 |
| Table 6.1 | Table 6.1 Summary of differences in participants' strategic reasoning between t1 and t2. | 296 |

# List of Figures


| | | |
|---|---|---|
| Figure 3.1 | Overview of research design | 113 |
| Figure 3.2 | Outline of ‘core’ segment plans for Programmes A and B | 140 |
| Figure 4.1 | Histograms of numbers of cases of NLD at t1 and t2 | 195 |
| Figure 4.2 | Box and whisker plots showing the t1 and t2 distributions of *score1* and *score2* for the participating sample as a whole | 203 |
| Figure 4.3 | Estimated marginal means of *score1* for intervention and comparison participants | 206 |
| Figure 4.4 | Interaction graph for effects of time and condition on *score2* | 213 |
| Figure 4.5 | Mean percentages of target grapheme tokens whose realizations were ‘acceptable’, ‘correct’ and ‘only acceptable’ | 217 |
| Figure 4.6 | Interaction graph for effects of time and condition on *score2* | 219 |
| Figure 4.7 | Distributions of *change2* for each class. | 224 |
| Figure 4.8 | Changes in mean percentages of ‘acceptable’ realizations of grapheme types | 230 |
| Figure 5.1 | Attested realizations of <ç> in *caleçon* at t1 and t2 and percentages of participants in each group producing each one | 244 |
| Figure 5.2 | Percentage of all participants producing the eight most frequent realizations of <ou> in the four RAT items at t1. | 250 |
| Figure 5.3 | Percentage of all participants producing the eight most frequent realizations of <ou> in the four RAT items at t2. | 251 |
| Figure 5.4 | Attested realizations of <ou> in *ougrien* at t1 and t2 and percentages of participants in each group producing each one | 252 |
| Figure 5.5 | Attested realizations of <oin> in *goinfrez* at t1 and t2 and percentages of participants in each group producing each one | 260 |
| Figure 5.6 | Percentage of participants in each group realizing <oin> with and without initial [w] | 271 |
| Figure 5.7 | Percentage of participants in each group realizing <oin> with and without nasality | 272 |
| Figure 5.8 | Percentage of participants in each group producing mono- and bisyllabic realizations of <oin> | 274 |

# List of Abbreviations

| | |
|---|---|
| CAT | Cognitive Abilities Test |
| GCSE | General Certificate of Secondary Education (public examination taken by pupils at the end of compulsory schooling in England, Wales and Northern Ireland) |
| GPC | Grapheme-Phoneme Correspondences |
| ITE | Initial Teacher Education |
| KS2 | Key Stage 2 (pupils aged 7-11) |
| KS3 | Key Stage 3 (pupils aged 11-14) |
| KS4 | Key Stage 4 (pupils 14-16) |
| L1 | First Language |
| L2 | Second Language |
| LMP | Linear Mapping Principle |
| MFL | Modern Foreign Languages |
| NC | National Curriculum for England, Wales and Northern Ireland |
| NLD | Non-Linear Decoding |
| Ofsted | Office for Standards in Education, Children's Services and Skills (government body which regulates schools in England) |
| RAT | Reading Aloud Test |
| SAT | Standard Assessment Test (statutory National Curriculum assessments in England) |
| SLA | Second Language Acquisition |
| SRI | Self-Report Interview |
| Y7 | Year 7 (pupils aged 11-12) |
| Y8 | Year 8 (pupils aged 12-13) |
| Y9 | Year 9 (pupils aged 13-14) |

# Chapter 1: Introduction

## 1.1 Background

*"These words get more ridiculous as we go along"*
*"Some of these don't even look like real words. I reckon you're just trying to trick me"*
*"They're like really weird words, like they don't exist"*
*"I don't know because most of them look like gibberish basically"*

These comments were made by students in an English secondary school towards the end of what was for some their first, for others their third year of learning French as a second language (L2). The students were reading aloud unfamiliar French words as part of a study into their proficiency in 'decoding' the L2 writing system, in the sense of unlocking the sounds represented by the written code (Woore, 2006). The formal data gathered by the study consisted of participants' audio-recorded pronunciations of the unfamiliar words; however, some participants also made incidental comments, including those quoted above, which happened to be captured by the audio recording equipment. These 'accidental' data eloquently encapsulate the sense of confusion and alienation which these participants experienced when trying to pronounce unfamiliar written words in French, despite having age-expected literacy levels in their first language (L1) and despite the fact that they had studied French for up to three years. In other cases, participants did not appear to find decoding the French words problematic, but simply pronounced them as if they were English pseudo-words, that is, according to English symbol-sound correspondences. It is these problems encountered by L1-literate, English-speaking learners when decoding L2 French which formed the central impetus for this thesis.

## 1.2 The Modern Foreign Languages context

The current study emanated from and is rooted in Modern Foreign Language (MFL) classrooms in UK schools. Amongst UK-based researchers, there is increasing interest in this context as a distinctive one within the field of second language acquisition (SLA) research. For example, the Centre for Applied Language Research at the University of Southampton recently ran a workshop entitled "Research Perspectives on Modern Foreign Languages in UK schools" (Courtney et al., 2011). This follows observations that most SLA research has been conducted in very different settings, predominantly focussing on learners of L2 English in Higher Education contexts, especially North American ones (e.g. Macaro, 2003). The under-representation of the MFL context in the SLA literature may have contributed to a sense of disjuncture between research findings on the one hand and MFL classroom practice on the other (Macaro, 2003). One purpose of the current study is to contribute to knowledge of L2 learning in this under-researched setting and thus to provide a more direct link between research and practice.

A greater understanding of UK MFL classrooms is urgently needed, because language learning does not appear to be flourishing in this context. Whilst the establishment of an entitlement to MFL lessons at Key Stage 2 (DfES, 2002) has led to a large expansion in the number of language learners at this level (Cable et al., 2010; Wade, Marshall & O'Donnell, 2009), there has been a sustained decline at all levels where language learning is optional. Since 2002, this includes the post-14 context (reversing the position in 1996, when it was made compulsory). According to CILT's (2010) latest annual survey of UK language learning (based on a questionnaire sent to a representative sample of 2000 schools, albeit with a response rate of 37%), the percentage of students taking a

language GCSE has fallen from 78% in 2001 to 43% in 2010. There have also been falling numbers post-16 (Fisher, 2001; Hawkins, 2005) and in Higher Education (Nuffield, 2000; Watts, 2003).

Given the low numbers opting to study languages, it is unsurprising that various studies have found low levels of motivation amongst school-level MFL learners. For example, Stables and Wikeley (1999) found that French and German ranked joint last in terms of learners' enjoyment of school subjects; MFL was also the subject which Year 9 and 10 pupils most wished they could drop. The situation appears not to have improved over the following years, as various studies have documented (e.g. Williams, Burden & Lanvers, 2002; Coleman, Galaczi & Astruc, 2007). Thus, making MFL study optional post-14 cannot be said to have *caused* the decline in numbers of students, as seems to have been argued in some quarters (see Evans, 2007); it has simply opened the floodgates to the natural consequences of students' low motivation for the subject.

However, the picture appears to be more complex than one of uniformly low motivation. Dobson (1997), drawing on evidence from Ofsted inspections of secondary schools[1], noted a decrease in learners' motivation after a period of initial enthusiasm in Year 7 (Y7). A decade later, Coleman, Galaczi and Astruc (2007)'s survey of over 10,000 MFL learners aged 11-14 produced similar findings. Thus, in Chambers' (1993) terms, we are dealing here with the *de-motivation* of pupils who are initially motivated, rather than simply with pupils who are, from the outset, *unmotivated*. Indeed, since MFL became an entitlement in English primary schools, fears have been expressed that this pattern of rapidly declining enthusiasm may simply be reproduced at a younger age (Sharpe, 2001).

[1] Ofsted (Office for Standards in Education, Children's Services and Skills): the government body which regulates schools in England.

Clearly, negative attitudes towards languages and cultures other than English, which appear to be prevalent in UK society, are likely to be important determinants of low motivation for MFL learning (Coleman, 2009); thus, the issue of low recruitment can only be understood (and countered) by "looking beyond the school gates" (ibid.:112). However, not only does this fail to help teachers in dealing with their students' low motivation, it also provides an inadequate explanation for the *decline* in motivation observed amongst beginner learners after a period of initial enthusiasm. Macaro (2008:106), who has extensive experience as an MFL teacher and teacher educator, suggests that a crucial way of improving beginner learners' motivation is to ensure that they make "rapid and substantial progress" early in their L2 learning career. This reflects Graham's (2010) call for greater emphasis, in the MFL context, on students' expectations of success within the expectancy-value component of motivation developed by Eccles and colleagues (Wigfield & Eccles, 2000). The key question is then: what can teachers do to increase beginner learners' progress in MFL?

Recently, the idea that decoding proficiency may play an important – and previously neglected – role in MFL learning has gathered momentum, in large part through empirical work and its subsequent dissemination by Erler (e.g. 2002, 2004, 2007). It will be argued in this thesis (section 2.3) that decoding is a foundation skill which facilitates many other aspects of L2 learning, contributes to learner autonomy and is involved in the complex picture of learners' motivation. At the same time, there is an increasing body of evidence (reviewed in section 2.6.3) that MFL learners in Key Stage 3 (KS3) have poor decoding proficiency and make little progress in this aspect of their learning – despite the publication of official guidance emphasizing the need for explicit teaching of symbol-

sound correspondences (DfES, 2003, 2005; DCSF, 2009). There also appears to be a lack of research evidence to show that explicit decoding instruction is effective in L2 learning generally, and in the MFL context in particular: for example, no intervention studies are mentioned in the reviews of L2 writing system research by Cook and Bassetti (2005b) or Koda (2007).

In response to these arguments, the current study adopts a quasi-experimental, pre-/post-test design to evaluate two programmes of explicit instruction in French decoding, aimed specifically at this population of learners. The thesis makes no attempt to suggest that decoding provides 'the' answer to the problems in MFL teaching. However, it aims to contribute to the systematic investigation of one potentially important aspect of language learning, building on Erler's previous work as well as that of Macaro and Erler (2008b) and Woore (e.g. 2007, 2009, 2010).

The focus on French is justified within the MFL context by the continuing predominance of this language in UK schools, despite calls for a diversification of language provision over a number of years (see Phillips & Filmer-Sankey, 1993). It is offered at some level by 99% of state-maintained secondary schools (CILT, 2010); by comparison, the next most frequently-taught languages are Spanish (offered by 76% of schools), German (67%) and Italian (17%). Similarly, French is offered by nine out of ten of those primary schools providing MFL lessons at KS2 (Wade, Marshall & O'Donnell, 2009). As Pachler (2002) has argued, this in turn is likely to increase pressure on secondary schools to offer French, thus reinforcing its predominance.

## 1.3 Research into L2 writing systems

Although the current study is situated in the UK MFL context, it also makes a wider contribution to research into the learning and use of L2 writing systems – "any writing system other than the system that the person learnt to read and write for their first language" (Cook & Bassetti, 2005b:25f). This emergent, cross-disciplinary field of research has recently gained greater prominence, as reflected in the publication of Cook & Bassetti's (2005a) overview and in the albeit short-lived appearance of the journal *Writing Systems Research* (Cook, Vaid & Bassetti, 2009). Within this area of inquiry, the current study deals with a specific, sub-lexical component of L2 reading: decoding, that is, the assembly of a word's pronunciation from the written representations of its component sounds.

Following a well-established tradition within L2 writing system research, the current study draws upon the concept of cross-linguistic transfer to explore the ways in which knowledge of an L1 writing system may facilitate or hinder L2 decoding; these issues are considered within the broader, explanatory framework of cognitive processing theories, as reviewed by writers such as Ellis (2008). Under the broad heading of 'L1-to-L2 transfer', many studies over the past twenty-five years have investigated the interaction between writing systems based on different units of linguistic representation. Particular attention has been paid to pairs of writing systems such as Chinese on the one hand – in which written forms principally represent meaning-based units – and alphabets such as English on the other, in which the written symbols represent sounds. This research paradigm is exemplified by the seminal work of Koda (e.g. 1989, 1990, 1999). By contrast, the current study contributes to the smaller body of research into learners whose

L1 and L2 writing systems have much in common – they share the same representational units (i.e. the writing systems are both sound-based), the same directionality (both are written left-to-right) and the same basic script (both use the Roman alphabet) – but have different symbol-to-sound correspondences and different levels of 'orthographic depth', which is a measure of the consistency with which a language's written symbols are mapped onto its sounds (section 2.4.3).

The current study also makes a distinctive contribution in that (a) the participants are young beginners rather than proficient adult learners and (b) their L2 is French – a language which, according to the *Institut Français de Vienne* (2008), is currently being learnt as an L2 by eighty-two million people worldwide. By contrast, the predominant L2 in writing system research – as in the SLA field more generally – is, by a considerable margin, English (Cook & Bassetti, 2005).

# Chapter 2: Literature Review

## 2.1 Introduction

What are the challenges involved in learning to decode written L2 text? Which pedagogical approaches are most likely to help learners meet these challenges? The current chapter addresses these over-arching questions with particular reference to the under-researched context of beginner learners of French in English MFL classrooms. The review has four parts.

Part I considers the role of L2 decoding proficiency within the broader enterprise of L2 learning. First, section 2.2 defines decoding as one strand of the lower-level processing involved in word recognition; section 2.3 then situates it within the wider field of L2 reading. Although a case could be made for the importance of decoding as a foundation skill underpinning various aspects of instructed L2 learning, the emphasis in this literature review is placed on reading, the traditional home of decoding research. However, 'reading' is interpreted not only in terms of the comprehension of connected text: the role of decoding in generating phonological forms for *unfamiliar* (or non-securely known) L2 words is also considered, which is argued to play a fundamental role in vocabulary acquisition and therefore in L2 learning more generally.

Part II of the literature review addresses various issues falling within a broad framework of cross-linguistic transfer, provisionally defined as "the influence resulting from similarities and differences between the target language and any other language that has

been previously … acquired" (Odlin, 1989:27). Transfer, which has been the dominant framework for L2 writing system research (Cook & Bassetti, 2005), is viewed in this thesis from a cognitive processing perspective, whereby learners' perception and processing of written text become 'tuned' or 'entrenched' through experience of their L1 (Ellis, 2008; Littlemore, 2009). It follows that, where there are differences between an L1 and L2 writing system, beginner learners' L1-based processing 'habits' will not be optimally tuned for dealing with written text in the L2: in other words, they will process written text differently than would an experienced L1 reader of the language. Conversely, where the two writing systems overlap, beginner learners' L1-based processing may have a facilitative effect on their processing of the L2.

Based on these arguments, Part II of the literature review addresses the following four questions, both from a broad cross-linguistic perspective and with specific reference to the French and English writing systems:

1. What is the nature of the French and English writing systems and how do they differ? (Section 2.4)
2. What is the nature of lower-level processing (such as word recognition and decoding) in different L1 writing systems? (Section 2.5)
3. Do differences in the lower-level processing of written L1 text affect the processing of written L2 text, and if so, how? (Section 2.6)

From a cognitive processing perspective, beginner decoders of French who are already (more or less) literate in English will need to 'retune' their processing to accommodate those aspects of the L2 writing system which differ from the L1. Part III of the literature review considers the effects of different instructional approaches on this process. First,

section 2.7.2 briefly considers evidence for the importance of explicit phonics instruction in developing L1 English decoding proficiency. Section 2.7.3 offers a critical discussion of recent policy initiatives and practice in MFL classrooms, framing a rationale for the current study. Section 2.7.4 then reviews previous intervention studies in L2 decoding, which (at least for learners of L2s other than English) appear to be sparse.

Finally, Part IV of the literature review outlines the research questions of the current study.

# Part I: The nature and importance of L2 decoding

## 2.2 Definitions

### 2.2.1 Introduction

The process referred to in the current study as 'decoding' is associated with some terminological confusion in the literature, perhaps reflecting the fact that it is approached from various disciplinary backgrounds, such as psycholinguistics, second language acquisition and developmental psychology. The literature review therefore begins by defining decoding and other key terms.

### 2.2.2 Graphemes

As will be explored in greater depth in section 2.4.1, alphabetic writing systems such as French and English are particular types of phonographic (sound-based) writing systems in which the consecutive, segmental sounds of the language are represented in linear sequence by written symbols. However, sounds are not represented in all their phonetic detail, but rather at the phonemic level, where a *phoneme* is a "minimal sound unit … capable of contrasting word meaning" (Katamba, 1989:21).

The Roman alphabet used in English and French comprises twenty-six individual letters; however, both languages have more phonemes than this: around forty-four in British English (Hawkins, 1984) and thirty-six in French (Grevisse, 1988). Therefore, there

cannot be a simple, one-on-one mapping of letters to phonemes. Rather, the units of written representation are *graphemes*, defined by Berndt, Reggia and Mitchum (1987:1) as "letters or letter clusters corresponding to single phonemes". However, this definition does not extend straightforwardly to languages other than English: for example, French graphemes may contain diacritics as well as letters, as in words such as *façonné* (/fasɔne/, 'manufactured')[2]. Elsewhere, the term 'grapheme' may also be used more broadly, "essentially as a synonym for 'written symbol'" irrespective of its relationship to a language's phonemes (Cook & Bassetti, 2005:4). This reflects the fact that some writing systems do not represent language at the phonemic level (section 2.4.1). In the current study, however, *graphemes* always denote written units corresponding to phonemes.

### 2.2.3 Word recognition and decoding

Decoding, as understood in the current study, can be considered as a sub-process of word recognition, which comprises "the processes involved in obtaining both phonological codes (or pronunciations) and context appropriate lexical meanings from a visual display of words" (Koda, 1996:451). As will be explored in section 2.5, certain models of word recognition propose that accessing the pronunciation of a known word may proceed via different 'routes': (a) by 'sight recognition', matching the graphic shape of the whole word directly to a whole phonological form held in long-term memory; (b) by computing

[2] Following conventional usage, angled brackets denote graphemes (e.g. <g>) and forward slashes indicate phonemes (e.g. /g/), transcribed using the phonetic alphabet of the IPA (2005). Where phonetic detail beyond the phonemic level is required, or where sounds in participants' output cannot be unambiguously identified as belonging to a particular phoneme, square brackets are used (e.g. [g$^{h}$], indicating an aspirated realization of the phoneme /g/). French or English words cited as examples are *italicized* when there is no need to distinguish between spoken and written forms. French words are followed by their phonemic representation and English translation in inverted commas (except in the case of words used in the Reading Aloud Test, for which see Appendix 2.1).

the sound of the word using sub-lexical correspondences between written symbols and sounds. The latter route must also be used to generate pronunciations for words with no pre-stored phonological representations, such as unknown words and 'pseudowords' (possible, but not actual, words of a language). In the current study, 'decoding' refers exclusively to this sub-lexical process of print-to-sound conversion.

In the wider literature, there is some lack of clarity in the way 'decoding' (as defined above) is referred to. First, several alternative terms are used to denote the same process. The most frequent alternative is '(phonological) recoding', as used in the title of Share's (1995) influential account of its role in L1 reading acquisition as well as in many other publications in psycholinguistics (e.g. Papagno, Valentine & Baddeley, 1991) and SLA (e.g. Koda, 1990). Other terms are also occasionally used, such as 'phonological coding' (e.g. Baddeley, Eldridge & Lewis, 1981) and 'cipher reading' (e.g. Hoover & Gough, 1990). However, following various recent SLA studies involving French (e.g. Erler, 2002, 2004, 2006; Walter, 2008), the current study uses the term 'decoding'. This is partly because 'decoding' has greater intuitive clarity: the sounds of a language are 'encoded' in a phonographic writing system; the reader then 'decodes' the written symbols to access the sounds. This analogy is made explicit in various references to 'cracking the code' of a writing system (e.g. Hickey, 2005).

The second problem is that different terms (especially 'decoding' and 'recoding') sometimes appear to be used interchangeably by the same author, either between different studies (e.g. Koda, 1990 versus Koda, 1998) or even within the same study (e.g. Frith, Wimmer & Landerl, 1998). In the case of review articles, these switches in usage

appear to be influenced by the terminology used in the original studies cited (e.g. Cook & Bassetti, 2005).

Third, both 'decoding' and 'recoding' are sometimes used with a broader meaning, including other processes involved in what was defined above as 'word recognition'. For example, Share (1995:155) defines recoding as "the ability to translate printed words independently [i.e. without reliance on context] into their spoken equivalents", which in the case of known words could include accessing pre-stored pronunciations based on a direct sight recognition process. Similarly, in Koda (1998), decoding seems to refer to both sub-lexical print-to-sound conversion and direct accessing of lexical phonology. Finally, some writers (e.g. Koda, 1998; Jorm & Share, 1983) include the pronunciation of Chinese characters in their use of the term 'decoding', which given the logographic nature of the writing system cannot involve systematic print-to-sound conversion[3].

To summarize, in the current study 'decoding' always refers to the sub-lexical process of print to sound conversion, as occurs when pronouncing unfamiliar written words or pseudowords in an alphabetic writing system, as well as (arguably) in pronouncing known words (section 2.5). To refer to the sight recognition of known words, 'word recognition' is used, subdivided where necessary into 'lexical access' (i.e. accessing meaning from written forms) and 'word naming' (i.e. generating appropriate pronunciations).

[3] As Perfetti and Zhang (1991:637) note, some Chinese characters ('phonograms') do include "phonologically useful components (phonetic radicals) that can guide pronunciation", albeit unreliably.

## 2.3 L2 decoding, reading and vocabulary learning

This section considers the role of decoding in L2 reading comprehension (section 2.3.1) and vocabulary learning (section 2.3.2). There is not space to review all the various aspects of L2 learning to which decoding may contribute. Nonetheless, it is hoped to provide some indication of the importance of L2 decoding – and, therefore, of finding ways to help learners develop their proficiency in this area.

### 2.3.1 L2 decoding and reading comprehension

Since research on L2 reading comprehension has been strongly influenced by research on monolingual L1 reading, principally of English (Bernhardt, 2000; Koda, 2004),this section begins by briefly reviewing L1-based models of reading (section 2.3.1.1).

#### 2.3.1.1 L1 reading

Stuart (1995:126) suggests that, "for skilled fluent readers, the print seems a transparent window through which we look to the meaning of the text". Despite this apparent effortlessness, however, fluent reading depends upon rapid and coordinated processing at multiple levels. Following the model used in Koda's (2007:4) review, which is broadly similar to various other models (e.g. Grabe & Stoller, 2002; Kintsch & Rawson, 2005), reading is assumed to involve three main types of process: (i) word recognition (which she here terms "decoding"), in which linguistic information is extracted from print; (ii) the linguistic processes of text-information building, described as "integrating the

extracted information into phrases, sentences and paragraphs"; and (iii) 'reader-model construction', in which the information derived from the text is integrated with the reader's prior knowledge and experiences. It has been argued that only the first of these components is specific to reading (e.g. Hoover & Gough, 1990).

However, during the past forty years, the importance of lower level processes such as word recognition in reading comprehension has been a matter of controversy. An influential school of thought beginning in the 1970s held that fluent reading comprehension was essentially a "top-down" process (Goodman, 1976; Smith, 1971). For Goodman, skilled reading was a "psycholinguistic guessing game", in which readers used the linguistic (and paralinguistic) context, together with their world knowledge, to predict the words which would occur next; the 'bottom-up' processes of recognizing individual words on the basis of their written forms were used only sparsely, in order to confirm the reader's top-down predictions. A greater reliance on orthographic information was argued to be characteristic of weaker readers.

However, following arguments by figures such as Stanovich (1980) and Perfetti (1985), it is now generally accepted that – in contrast to top-down models – skilled reading is characterized by the rapid, context-free recognition of words based on their written forms. As Ehri (2005) notes, the bottom-up activation of both word meanings and pronunciations is so fast that there is no time for systematic contextual inferencing to be involved in word recognition. Eye movement research also provided counter-evidence to top-down models, confirming that almost all content words in a text are directly fixated by the reader (e.g. Just & Carpenter, 1980). However, as will be discussed in

section 2.5, the actual mechanisms involved in bottom-up word recognition remain controversial, particularly with regard to the involvement of decoding in this process.

According to models such as Stanovich's (1980) interactive compensatory model, top-down processes may nonetheless play a role in reading comprehension, stepping in when word recognition skills are deficient; in other words, readers may infer the meaning of a word from context if they are unable to identify it by bottom-up means – the inverse of Smith's (1971) view. However, according to limited capacity processing models (e.g. LaBerge & Samuels, 1974), using an alternative, top-down route to identify words incurs a cost: it is slow and requires attentional resources, which are then unavailable for the higher-level processes involved in textual meaning construction and interpretation. Fluent reading therefore relies on the automatization of lower-level word recognition processes (Daneman & Carpenter, 1980, 1983) through the repeated processing of similar instances. By contrast, higher-level tasks such as meaning construction and interpretation are less predictable and therefore not susceptible to automatization.

Turning to the acquisition of L1 reading proficiency, decoding is widely assumed to play a pivotal role, as reflected in the description of it as a "*sine qua non* of reading acquisition" in the title of a widely-cited article by Share (1995). This is because it provides beginner-readers with a "self-teaching" mechanism, allowing them to sound out unknown written forms to access the large bank of words which they already know orally (approximately fifteen thousand words for a typical beginning reader: Bee, 2000). In turn, successfully recognizing a word which has been sounded out provides positive feedback on the decoding process itself, further helping children to learn the language's GPC.

#### 2.3.1.2 L2 reading

Whilst insights from L1 reading research have provided a basis for understanding L2 reading, there has also been growing recognition of the distinctiveness of this process (Bernhardt, 2000; Koda, 2007), deriving from at least two factors: (a) L2 readers' lower levels of linguistic proficiency and (b) their co-existing knowledge of two (or more) languages.

A major strand of research in the 1980s and 1990s saw L2 reading proficiency as being determined by L1 reading proficiency on the one hand and L2 linguistic proficiency on the other, and sought to establish which of the two was the more important determinant (e.g. Alderson, 1984; Carrell, 1991; Lee & Schallert, 1997). This approach gave rise to the widespread view that there is a 'threshold' of L2 linguistic knowledge which learners must attain in order to access the reading comprehension skills developed in their L1. However, much of this research fails to distinguish between two distinct strands of 'L2 proficiency': (a) proficiency in word recognition and (b) proficiency in comprehending the linguistic material made available by these word recognition processes. Based on this argument, it is possible to see L2 reading as comprising the same three components as identified by Koda (2007) in respect of L1 reading: (a) word recognition; (b) linguistic processes of text-information building and (c) reader-model construction. For beginner L2 readers, both of these first two components (word recognition and linguistic processing) are likely to be inefficient and/or inaccurate.

Applying the interactive compensatory model to L2 reading, top-down processes such as contextual inferencing may make up for some deficiencies in word recognition.

However, as argued above, this leaves fewer attentional resources for higher-level processing such as meaning construction and interpretation. Interesting in this respect is Chamot and El-Dinary's (1999) investigation of the reading strategies deployed by young L2 learners in immersion classrooms: they found that lower proficiency L2 learners tended to focus their attention at the individual word level, relying "more on phonetic decoding during reading than on any other strategy" whereas "high-rated students focused more on using background knowledge and inferencing to understand a text" (p.332). However, one interpretation of their results (which they do not appear to consider) is that higher-proficiency learners may be able to focus more on meaning *precisely because* their word recognition is more efficient and therefore less effortful. This view implies that one priority for L2 reading instruction is to develop learners' knowledge (and fluent recognition) of an appropriate vocabulary base, as advocated by researchers such as Coady, 1997.

A further source of complexity in L2 reading is the reader's concurrent knowledge of (at least) two language systems. As Bernhardt (2000:802) says, the interaction between the reader's L1- and L2-based knowledge constitutes "the absolute essence of the L2 reading experience". Bernhardt also argues that this interaction may facilitate or hinder L2 reading, depending on the relationship between a given pair of languages: for example, the "mapping" of L1 German onto L2 Spanish is considered "convenient", that of Hindi onto English "inconvenient" and that of Thai onto Finnish "really inconvenient". However, rather than characterizing these cross-linguistic relationships in global terms, it would be helpful to consider the individual components of linguistic proficiency separately. Thus, L2 word recognition may be expected to be facilitated by similarities in the two languages' writing systems; the linguistic processes of meaning construction

may be facilitated by similarities in the languages' morphosyntactic and lexical systems; and the 'higher-level' processing involved in interpreting a text may be facilitated by shared textual conventions or cultural background. It is the relationship between the processing of words in the L1 and L2, based on similarities and differences between the languages' writing systems, which forms the central concern of this thesis.

#### 2.3.1.3 Decoding in L2 reading comprehension

Turning to decoding more specifically (rather than the wider process of word recognition), this may be used as a conscious strategy in L2 reading, just as it is in the acquisition of L1 reading: words which are known orally but unfamiliar in written form may be 'sounded out' in order to access their meaning. Indeed, Macaro and Erler's (2008a) survey of learners' reading strategies, conducted with a large, representative sample of KS3 students (N=1735) in English MFL classrooms, found that 66% of respondents agreed with the statement, "When I can't read a French word easily, I sound it out in my head". However, since learners in instructed contexts may be introduced to written L2 text at the very outset of language learning, they may not have a large bank of oral language knowledge on which to draw, as L1 learners do. Therefore, as Grabe and Stoller (2002:43) argue, "one benefit of developing accurate letter-sound correspondences ... is lost": sounding out words may be a blind alley in terms of accessing meaning, if the words are not already known orally. Data gathered as part of a vocabulary learning task (originally planned as an additional strand of the current project) illustrate this point: in a test requiring English translations of written French words, one Y7 French learner guessed that *morceau* (/mɔʁso/, 'piece') meant 'more so', *colline* (/kɔlin/, 'hill') meant 'clean' and *la langue* (/la lɔ̃g/, 'the tongue') meant 'along'.

It therefore appears that this learner was able to *decode* the French words (i.e. generate sounds for them) with some accuracy, but did not know the meanings of the resulting phonological forms, associating them instead with similar-sounding English words.

A more common problem in the MFL context is the inability to generate accurate pronunciations of French words in the first place (section 2.6.3), with learners tending to decode L2 written material according to L1 grapheme-phoneme correspondences (Erler, 2002). This clearly makes it less likely that sounding out a written form will successfully lead to a French word being recognized, even if the learner *does* hold an accurate phonological representation of it. Confusion between L2 lexical items may also ensue: Walter (2008) provides the hypothetical example of an L1 French learner of L2 English, whose L1-based decoding of *white* as /wiːt/ makes it indistinguishable from the phonological representation for *wheat*.

Finally, various L2 reading researchers suggest that accurate decoding is important not only for lexical access, but also because reliable phonological representations of written words facilitate the operation of verbal working memory (e.g. Koda, 1990; Erler, 2006; Walter, 2004, 2007, 2008). Drawing on Baddeley and Hitch's (1974) multiple-component model of working memory, they argue that items are held in the phonological loop during the processing of linguistic units at the clause or sentence level, until the formulated meanings are integrated into higher levels of processing. Further, this has been argued by some to apply not only to phonographic writing systems, but also (counter-intuitively) to logographic systems such as Chinese, since "visually presented materials are more efficiently processed when converted into their phonological forms" (Koda 1990:395f). Conversely, reviewing the multiple-component working memory

model twenty-five years on, Baddeley and Logie (1999) argue that there is little evidence for the involvement of the phonological loop in 'normal' reading, although it may be important in particularly demanding reading situations, such as those involving complex syntax or "obscure" semantics. For example, Baddeley, Eldridge and Lewis (1981) – in a study often cited in relation to this argument – convincingly demonstrate that interfering with the phonological loop (via articulatory suppression) impairs participants' ability to detect anomalies in connected text, but not their ability to process the semantics of correct sentences. However, it is conceivable that beginner L2 reading – where readers' processing of vocabulary and syntax may be inaccurate and/or inefficient – represents exactly the kind of demanding reading situation referred to by Baddeley and Logie. Therefore, irrespective of the arguments concerning fluent L1 reading comprehension, phonological working memory (and hence, decoding) may be important in L2 reading.

### 2.3.2 L2 decoding and vocabulary acquisition

It was argued above that a conscious strategy of decoding unfamiliar written words (or 'sounding them out') may often be a blind alley for beginner L2 readers, due to their low levels of existing (oral) vocabulary knowledge. However, the other side of this coin is that reading may offer plentiful opportunities for learning new vocabulary.

On the one hand, this argument has been applied to incidental vocabulary learning through extensive reading, both in L1 (e.g. Paribakht & Wesche, 1996; Nagy, Herman & Anderson, 1985; Nagy, Anderson & Herman, 1987) and in L2 (e.g. Day & Bamford, 1998; Coady, 1997). However, it can be argued that both research into L2 reading

instruction and MFL pedagogy (section 2.7.3) have tended to emphasize the use of contextual cues to infer the *meanings* of unknown words as part of the reading comprehension process (a precursor to incidental learning), whereas learning a word involves learning its phonological form as well. Rather than depending on the teacher to provide the pronunciations of new words encountered when reading, proficient L2 decoders can derive the pronunciations themselves from the written forms (at least in sound-based writing systems such as French). Decoding proficiency therefore enhances learner autonomy, providing a self-teaching mechanism for vocabulary acquisition.

The same arguments can also be applied to intentional vocabulary learning, as for example when students attempt to memorize a list of words which the teacher has presented orally in class. Away from the classroom, learners who are poor decoders lack the means to recover the words' pronunciations, should they forget what they heard in class. This problem is highlighted in preliminary data gathered as part of the current study (not reported systematically in this thesis for space reasons): for a sample of 330 Y7 learners from five comprehensive schools, (a) 'saying the target words aloud' was the most frequently reported strategy when trying to memorize words at home; yet (b) inaccurate decoding when doing so led participants to learn incorrect phonological forms, even though they had previously been introduced to the correct phonological forms in class. This caused problems when the participants were subsequently tested on aural recognition of the items. For example, one participant complained: "I don't think I did very well because I didn't recognise how Miss said them… I thought they were pronounced differently". This participant's report of competing phonological forms of the same word will be seen to echo Erler's (2002) findings with the same population of learners (section 2.6.3).

Hamada and Koda (2008) also investigated the role of decoding in an L2 vocabulary learning task. Based on a sample of L1 Korean and L1 Chinese learners of L2 English, they found that greater decoding efficiency – as reflected in (a) greater typological similarity between the L1 and L2 writing systems and (b) greater regularity of L2 word spelling – led to better learning and retention of English pseudowords. However, individual differences in decoding proficiency and their relationship with vocabulary learning were not measured, making their central claim potentially open to misinterpretation. Further work is clearly required within what the authors claim to be a "new line of inquiry", uniting the fields of decoding (traditionally treated instead as part of reading) and word learning.

Evidence from psycholinguistics also emphasizes the central importance of phonological working memory in learning novel phonological forms, as in the case of acquiring new words (Baddeley, Gathercole & Papagno, 1998). This is held to apply to both L1 and L2 learning. For example, an experiment by Papagno, Valentine and Baddeley (1991) showed that articulatory suppression (which interferes with the operation of the phonological loop) impaired participants' learning of associations between familiar L1 words and words in an unfamiliar L2, although their learning of L1-L1 word pairs was unaffected. This was found to occur both when the words were presented aurally and when they were presented in written form (transliterated according to L1 orthographic conventions). However, what Papagno and colleagues do not explore is how the operation of the phonological loop is affected by L2-specific phonology: clearly, participants would have decoded the transliterated L2 items according to L1 phonological patterns; the outcome measures also involved writing rather than speaking

the L2 words, using L1 phoneme-grapheme correspondences. Inevitably, there is also no consideration of L2-based decoding (which would have been impossible in the context of the study). Nonetheless it is clear that accurate decoding is a fundamental prerequisite for successfully acquiring new L2 words encountered in written form.

### 2.3.3 Summary of Part I

Established, interactive models of L1 reading comprehension provide a useful framework for conceptualizing L2 reading comprehension, where top-down processes may provide an alternative route to word recognition when bottom-up processes are inaccurate and/or inefficient. However, this places a large burden on the limited capacity attentional system. Fluent L2 word recognition is therefore essential to the development of L2 reading comprehension proficiency.

As in the case of L1 literacy acquisition, decoding (in the sense of a conscious 'sounding out' strategy) can provide access to words which readers already know orally, although its effectiveness may be undermined for beginner learners by their lack of existing vocabulary knowledge. In the MFL context, a further problem is likely to be posed by learners' lack of L2 decoding proficiency in the first place, so that sounded out words may not be recognized even if readers *do* know them orally. Finally, accurate decoding may support the operation of phonological working memory – which has been argued to play an important role in L2 reading comprehension – by generating well-differentiated phonological forms for L2 words.

Beyond the traditional concern with reading as the comprehension of connected text, beginner L2 learners in instructed contexts may also encounter many unfamiliar words in written form, both incidentally (as in the case of extensive reading) and in the context of deliberate attempts to learn new vocabulary. Accurate decoding can play a central role in developing learners' autonomy in this area, allowing them to generate accurate phonological forms for written words whose spoken forms they either do not know, or do not know securely. This frees them from relying on the teacher (or other spoken model) for the pronunciations of new words. By contrast, poor decoding proficiency may lead them to generate incorrect phonological forms for L2 words, which then fail to match the correct forms heard previously or subsequently. Psycholinguistic research has also underlined the central importance of phonological working memory in acquiring new words both in L1 and L2.

# Part II: L2 word recognition, decoding and transfer

Part II of the literature review is concerned with processes of cross-linguistic transfer in the processing of L2 writing systems. First, in order to contextualize the subsequent argument, section 2.4 reviews ways in which writing systems differ, beginning with a broad cross-linguistic perspective (drawing on the comprehensive overview by Cook and Bassetti, 2005), then focussing more narrowly on alphabetic systems, particularly French and English. Sections 2.5 and 2.6 then consider (a) how the differences between writing systems have been argued to affect L1 word recognition processes; and (b) how this L1-tuned processing may affect L2 word recognition. Finally, section 2.6.4 briefly considers the nature of learners' L2 phonemic representations and how these affect decoding outcomes.

## 2.4 How do the French and English writing systems differ?

### 2.4.1 Dimensions of difference between writing systems

There are various dimensions of contrast between writing systems, of which the most relevant for the purposes of this review are (a) the linguistic units represented by the written forms; (b) the script used to represent these linguistic units; (c) the degree of consistency (within sound-based systems) with which written symbols map onto sounds; and (d) within systems sharing the same representational units and the same script, the symbol-sound correspondences themselves.

In terms of representational units, a major distinction separates meaning-based ('morphemic', 'logographic') writing systems on the one hand and sound-based ('phonographic') ones on the other. Logographic systems include Chinese, where characters directly represent morphemes rather than their pronunciations[4]. In phonographic systems, by contrast, written symbols represent the sounds of the spoken language. Phonographic systems differ in terms of the size of the phonological units that are represented: syllables in some (such as Japanese *kana*, used alongside the morphemic *kanji*) and phonemes in others. Within phonemic systems, there is then a further distinction between (a) consonantal systems, such as Hebrew, in which written symbols mainly represent consonants (leaving vowels to be added by the reader) and (b) alphabetic systems, as in both French and English, in which all phonemic segments of the spoken language (both vowels and consonants) are represented.

According to Coulmas's (1999:9) comprehensive *Encyclopaedia of Writing Systems*, a phonemic writing system[5] is "characterized by a systematic mapping relation between its signs (graphemes) and the minimal units of speech (phonemes)". Such systems also follow a complementary principle of linearity (Cook & Bassetti, 2005:6), stating that the linear order of graphemes on the page corresponds to the chronological sequence of phonemes in the spoken language. Together, these two principles form the basis for decoding: written words are "timechart[s] of sound" (Diack, 1965). In both French and English, the sequence of graphemes flows from left to right; in other languages, this directionality is reversed (e.g. Hebrew).

[4] See footnote 3.
[5] This use of the term 'phonemic' follows Cook & Bassetti (2005); Coulmas uses 'alphabetic' here.

Phonographic writing systems also differ in terms of their script, "the graphic form of a writing system" (Coulmas, 1999:454). For example, French and English both use the Roman (as opposed to, say, Greek or Cyrillic) alphabet, although French supplements this with various diacritics (Table 2.1). English does not use diacritics except in loan words (e.g. *café*, *naïf*).

**Table 2.1 French diacritics**

| **Diacritic** (English name) | **Example** |
|---|---|
| acute accent | *dés* (/de/, 'dice') |
| grave accent | *dès que* (/dɛkø/, 'as soon as') |
| circumflex | *tempête* (/tɒ̃pɛt/, 'storm') |
| diaeresis | *ouïr* (/wiʁ/, 'to hear') |
| cedilla | *caleçon* (/kalsɔ̃/, 'underpants') |

Finally, even in writing systems which share the same representational units, the same script and the same directionality, there are differences in orthography (literally, 'correct writing'), "the set of rules for using a script in a particular language" (Cook & Bassetti, 2005:3). Thus the French and English writing systems, despite their many commonalities, differ in certain grapheme-phoneme correspondences (GPC). Space does not permit a full description of the two languages' GPC; these can be found in Berndt, Reggia and Mitchum (1987) for English and Appendix 1.2 for French. However, it is important to note that there are both similarities and differences between the two orthographies. This can be analysed on two levels: (a) the way strings of letters are segmented to form individual graphemes; (b) the mappings between the resulting graphemes and the phonemes they represent. The following paragraphs consider each of these in turn.

### 2.4.2 Grapheme-phoneme correspondences (GPC) in French and English

The grapheme inventories of both French and English contain not only single-letter graphemes (e.g. <a>) but also digraphs (e.g. <au>) and trigraphs (e.g. <eau>). However, the languages differ in terms of (a) the frequency of different di- and trigraphs: for example, <eau> is more frequent in French than in English; and (b) the way letter-strings are segmented into individual graphemes: for example, <gn> often functions as a single digraph in French (representing /ɲ/, as in *agneau*, /aɲo/, 'lamb') but almost always as two separate graphemes in English (representing /gn/, as in *magnet*). English orthography also includes 'split digraphs', where the pronunciation of a vowel preceding a consonant is affected by an <e> which follows that consonant. This is illustrated by the English words *rat* and *rate*. By contrast, in the identically-spelt pair of words in French, the vowels are pronounced identically: *rat* (/ʁa/, 'rat'); *rate* (/ʁat/, '(s/he) misses'). On the other hand, French orthography is complicated by the presence of many word-final consonantal graphemes with no phonemic realization, referred to hereafter as 'Silent Final Consonants' or 'SFC' (e.g. *petit*, /pəti/, 'small')[6].

In terms of the mappings between individual graphemes and phonemes in French and English, the relationship between the two writing systems is again characterized by both similarity and difference. In respect of consonants in particular, there is considerable overlap: for example, in both languages, <f> represents /f/ in most contexts. Nonetheless, some consonantal GPC do differ: for example, <th> represents /đ/ or /θ/ in English (as in *the theatre*) but /t/ in French (as in *thé*, /te/, 'tea' and *théâtre*, /teɑtʁ/,

[6] In connected text, interlexical 'liaison' may occur, whereby SFC are nonetheless sounded if the following word begins with a vowel (e.g. *petit animal*, /pətitanimal/, 'small animal'). However, the current study considers only the decoding of single, decontextualized words.

'theatre'). There are also many differences in respect of vowels (Sprenger-Charolles, 2004); to give one example of many which could be cited, <ou> is realized as /aʊ/ in the English word *scout* but as /u/ when the same word is borrowed into French. This example (like *théâtre* above) also illustrates how the presence of many cognates in French and English – words with the same or similar spelling and meaning – may be helpful for L2 learners in terms of accessing meaning but confusing in terms of decoding.

### 2.4.3 'Orthographic Depth' of French and English

On a more abstract level, alphabetic writing systems also differ in their phonological "transparency" (Cook & Bassetti, 2005:7), that is, the extent to which they adhere to the 'ideal' principle of one-on-one mapping between written symbols and sounds. A widespread alternative term is 'orthographic depth', defined by Katz and Frost (1992:150) as follows:

> *An orthography in which the letters are isomorphic to phonemes in the spoken word (completely and consistently) is orthographically shallow. An orthography in which the letter-phoneme relation is substantially equivocal is said to be deep (e.g., some letters have more than one sound and some phonemes can be written in more than one way (...).*

Frost (2005), in a later review of research on word recognition in different writing systems, sees orthographic depth as the product of two factors, regularity and consistency, following the distinction drawn by Glushko (1979). *Regular* words can be computed by grapheme-to-phoneme mapping 'rules' (e.g. English *haze*), whereas irregular ones cannot (e.g. *have*); *consistency*, on the other hand, is determined by the

number of a word's orthographic "neighbours" (i.e. words with similar spelling) which are also pronounced similarly. Thus, the word *wade* is both regular and consistent, because all other words with the spelling body <ade> rhyme with it (e.g. *shade, fade*)[7]. By contrast, *wave* is regular but inconsistent, because it has at least one irregular neighbour which does not rhyme with it (*have*). Deep orthographies have a higher proportion of irregular and/or inconsistent words compared to shallow ones. As will be explored in section 2.5, the issue of orthographic depth is important because it is associated with differences in word recognition processes.

Coltheart (1978, in Lange & Content, in preparation) argued that there was considerable regularity in English GPC, predicting that rules mapping each grapheme onto its dominant phonological realization could correctly derive the pronunciation of 80-95% of monosyllabic words – a prediction subsequently borne out in Lange and Content's empirical work. However, monosyllabic words of course form only a portion of English lexis. Further, Lange and Content noted that irregular words (for which GPC 'rules' generate incorrect pronunciations) tended to be high frequency items.

A contrasting, and much more widespread view is put by Nunes (2004), who emphasizes the complex, hierarchical nature of English print-to-sound mappings and the high number of irregular and inconsistent words. Thus, as noted above, graphemes such as <ou> are realized differently depending on their orthographic context. This variability can be expressed in terms of the probabilities with which the grapheme maps onto its various realizations (Berndt, Reggia & Mitchum, 1987). Table 2.2 shows the findings of

[7] The 'spelling body' is the orthographic unit corresponding to the phonological 'rime', the portion of a syllable comprising the vowel and any subsequent consonants.

Berndt and colleagues in respect of the pronunciation of <ou>, based on empirical analysis of a large corpus of (American) English words.

**Table 2.2 Probabilities for various realizations of English <ou>, adapted from Berndt, Reggia and Mitchum (1987)**

| grapheme | pronunciation | probability | example |
|---|---|---|---|
| <ou> | ə | .480 | enormous |
| | aʊ | .324 | out |
| | ɐ | .042 | touch |
| | ɔː | .041 | four |
| | u | .041 | you |
| | ʊ | .035 | could |
| | ɚ | .031 | glamour |
| | w | .001 | bivouac |

In English, this complexity is bidirectional: it applies not only when moving from print to sound (i.e. orthography to phonology or 'O→P'), but also when moving from sound to print (P→O). For example, the sound /u/ may be represented in various other ways besides <ou> (as illustrated in words such as *to, two, too, tutti frutti, shoe, flue, Crewe*). However, this thesis is concerned only with O→P mappings.

More recently, one strand of an influential study by Treiman et al. (1995) showed that the consistency of English print-to-sound mappings is much higher if orthographic units larger than individual graphemes are considered. Treiman and colleagues analysed spelling-sound correlations in around 1,300 monosyllabic English words with a CVC phonological structure (e.g. *heap*). They found that (a) in line with previous findings, the pronunciation of vowels was more variable than that of consonants; but (b) "the uncertainty in the pronunciation of a vowel is reduced if the following consonant is taken into account" (p.130), that is, by considering print-to-sound mappings at the rime level. This can be illustrated with the grapheme <a>, which has different, but consistent pronunciations in the following three groups of words: (a) *cat, bat, fat*; (b) *car, bar, far*;

and (c) *call, ball, fall*. Peereman and Content (1998) have subsequently claimed that, in monosyllabic words, the consistency with which an English vowel grapheme is pronounced increases from 48% to 78% when considered in the context of the surrounding letters.

Turning to the French writing system, some classifications place it with English in the category of deep orthographies (e.g. Defior, 2004; Seymour, 2005; Seymour, Aro & Erskine, 2003) whilst others consider it to be much shallower (e.g. Grabe & Stoller, 2002). These differing views may be explained by an asymmetry in the language's P→O (phonology to orthography) and O→P (orthography to phonology) mappings: whilst the latter are "globally quite predictable" (Lange & Content, 1999:3), P→O mappings are highly unpredictable. For example, Ziegler, Jacobs & Stone's (1996) analysis of monosyllabic French words found that only 12% had spelling bodies with more than one possible pronunciation, whereas almost 80% had phonological rimes with more than one possible spelling. They concluded that, in popular terms, French is "easy to read but difficult to spell" (p.507) – though given L2 learners' difficulties with French decoding (section 2.6.3), this might be better phrased as a comparative: 'French is *easier* to read than to spell'. This asymmetry in French orthography is illustrated by the phonological form /vɛʁ/, which is shared by several real French words (e.g. *vers*, 'towards'; *vert*, 'green'; *ver*, 'worm'; *verre*, 'glass' and *verres*, 'glasses') as well as numerous pseudowords (e.g. **vair*,**vère*,**verds*); however, each of these orthographic forms has only one possible pronunciation, /vɛʁ/.

Since this thesis is concerned solely with O→P mappings (which form the basis for decoding), the French writing system is considered (relatively) 'shallow', and certainly

shallower than English. As Sprenger-Charolles and Casalis (1995:41) argue, though the language's GPC may be numerous, they are "highly regular". Compared to English, there is therefore less need to look to larger orthographic units (such as spelling bodies) in order to find consistency in print-to-sound mappings. This will be shown to have important implications for the processes of word recognition in French as compared to English (section 2.5).

## 2.5 Cross-linguistic differences in L1 word recognition

### 2.5.1 Introduction

Corresponding to the different characteristics of writing systems discussed above, cross-linguistic differences have also been postulated both in the processes of fluent L1 word recognition and in the acquisition of this skill; this review concentrates mainly on the former, since the young adolescent participants in the current study have considerable experience and fluency in L1 word recognition. An understanding of these L1-based word recognition mechanisms will form the basis for an explanatory account of L1-to-L2 transfer in section 2.6. Whilst French and English are the languages directly implicated in the current study, a broader cross-linguistic perspective is also taken in order to support the development of this argument.

Of particular interest in this section is the extent to which word recognition involves phonological decoding as opposed to the direct mapping of written forms onto meanings (with phonological forms being generated 'post-lexically'). This question has generated

vast research activity in relation to English alone (Lupker, 2005) but few conclusive answers (Van Orden & Kloos, 2005). In particular, an abundance of human performance data, for which theories of word recognition must account, has been gathered from various sources: studies of impaired reading caused by dyslexia or aphasia; studies of children's early literacy development; and experiments investigating the speed and accuracy with which 'normally functioning' readers name and recognize various categories of words and pseudowords. However, findings sometimes appear to be contradictory. Further, the results of experimental studies are often difficult to compare because of differences in the tasks used (e.g. word naming versus lexical decision). Finally, as Lupker (2005) notes, not all of the available human performance data may ultimately prove relevant to understanding word recognition in fluent reading.

Cross-linguistic comparisons of L1 word recognition complicate the picture further. Here, it becomes crucial to disentangle genuine processing differences on the one hand from, on the other hand, differences resulting from (a) the nature of the participants or (b) the characteristics of items used in word naming or lexical decision tasks (e.g. the words' frequency, phonological complexity or similarity to other items).

Inevitably, this review includes only a small proportion of the relevant primary studies; more comprehensive overviews of word recognition research are provided by Underwood and Batt (1996) and Lupker (2005) – relating to English – and Ziegler and Goswami (2005) relating to cross-linguistic evidence.

Finally, it should be noted that word recognition research has focussed disproportionately on English, as various researchers working on other languages have

complained (e.g. Muljani, Koda & Moates, 1998; Seymour, 2005). This review therefore begins by briefly reviewing theories of English word recognition (section 2.5.2) before considering additional writing writing systems in section 2.5.3.

### 2.5.2 Word recognition: L1 English

#### 2.5.2.1 Dual route models

An influential conceptualization of English word recognition has been provided by 'dual route' models (e.g. Coltheart, 1978 and Coltheart, Curtis, Atkins & Haller, 1993, summarized in Coltheart, 2005). As will be seen in section 2.6.2, these have provided the predominant framework for L2 word recognition research (Cook & Bassetti, 2005). According to dual route models, the pronunciations of words with regular spellings (e.g. *yak, pin* and *ruff*) are 'assembled' (i.e. decoded) via a sub-lexical, print-to-sound conversion mechanism, whereas the pronunciations of irregularly-spelt 'exception' words (e.g. *yacht, pint* and *rough*) are retrieved intact from the mental lexicon. Similar models have been proposed for other aspects of linguistic processing, such as English speakers' ability to produce both regular and irregular past tense forms (e.g. *hatch – hatched* versus *catch – caught*), the former being generated by rules and the latter being retrieved as pre-learnt instances (e.g. Pinker, 1994). In dual route models of word recognition, the sub-lexical pathway also accounts for the naming of pseudowords (e.g. *yan, piff* and *ruk*).

Dual route models have been argued to account for a wide range of human performance data (Coltheart, 2005). However, as Plaut et al. (1996) note, the distinction between

'regular' and 'exception' words on which they depend is problematic, since even 'exception' words contain some regular GPC: for example, three of the four graphemes in the word *pint* have 'regular' realizations, whilst the single irregular grapheme <i> is pronounced similarly in other words such as *pine* and *wild*.

#### 2.5.2.2 Connectionist models

Increasingly, dual-route accounts of English word recognition have been challenged by connectionist models (e.g. Seidenberg & McClelland, 1989; Plaut et al., 1996). These are computational simulations of human word recognition falling within the broader field of cognitive processing theories. They hold that the same underlying mechanisms are involved not only in all aspects of language learning and use, but in other areas of human cognition more broadly (Ellis, 2008; Littlemore, 2009). This contrasts with the nativist assumption of a separate language processing system.

Within connectionist models, both 'regular' and 'exception' words are processed similarly, rather than relying on separate pathways. Orthographic knowledge is represented not as a series of 'rules' and 'items', but rather as "an elaborate matrix of correlations among letter patterns, phonemes, syllables and morphemes" (Seidenberg & McClelland, 1989:525). Further, they are self-organizing, 'learning' the relevant associations between print and sound through exposure to a 'training set' of stimuli. The more frequently a particular pattern of activation is produced through contact with the training set (e.g. by a particular combination of letters or print-sound correspondence), the more the connections within that pattern of activation are strengthened. The models are therefore sensitive to the probabilistic relations between orthographic and

phonological units (section 2.4.2), in contrast to the all-or-none mapping rules used in the sub-lexical pathway of dual route models. This has been argued to better reflect human performance data (e.g. Lange & Content, in preparation). As Ellis (2008:374) argues in relation to cognitive processing approaches more broadly,

> [t]he systematicities of language competence, at all levels of analysis (…) emerge from learners' implicit tallying of the constructions in their usage history and the implicit distributional analysis of this lifetime sample of language input.

Connectionist models of word recognition are arguably compatible with findings from the field of L1 literacy development. There is no space to review this research here; a recent and comprehensive overview is provided by Ehri (2005). However, in view of the argument to be made in section 2.6, it is important to note that the final phase of L1 literacy acquisition is generally said to be characterized by "highly developed automaticity and speed in identifying unfamiliar as well as familiar words" (Ehri & McCormick, 1998:156f). Most words encountered in this phase are already included in readers' 'sight vocabularies', causing them to be recognized rapidly, effortlessly and involuntarily, as demonstrated in image-word interference tasks such as those used in the classic experiments by Stroop (1935). Print-to-sound decoding is also automatized, accounting for the ease with which pseudowords can be processed, as Baddeley, Eldridge and Lewis (1981:439) demonstrate with the following sentence:

> Iff yew kann sowned owt thiss sentunns, ewe wil komprihend itt.

As argued in section 2.3.1, automaticity in word recognition is important because it releases attentional capacity for higher-level comprehension processes. Within connectionist models, automatization is interpreted as the result of repeated experiences

of a given association, leading to strong connections which are activated involuntarily upon contact with the relevant stimulus. However, as will be explored in the following section, the problem for L2 readers is that these connections are the result of a "lifetime sample" of L1 input, and will therefore be inappropriate for processing L2 input wherever the two systems differ.

### 2.5.3 Word recognition and its development in different alphabetic writing systems

#### 2.5.3.1 Introduction

In order to make the case that learners' processing of written L2 text is influenced by processing mechanisms which have been tuned through exposure to L1 input, it is necessary to consider the processes and acquisition of word recognition in other languages besides English. To what extent do these processes differ between writing systems generally, and between French and English in particular?

Though research into L1 word recognition and its acquisition has predominantly been conducted within an English-speaking context, the English writing system is only one amongst many others which vary in terms of their representational units, script and orthographic depth (section 2.4). Even within the subset of alphabetic writing systems, it was argued that English shows an unusual degree of inconsistency in its GPC. This has led various researchers to question whether English-based models of word recognition are applicable to other writing systems (e.g. Muljani, Koda & Moates, 1998; Ziegler & Goswami, 2005; Seymour, 2005). For example, the influential dual route approach was developed specifically to account for readers' ability to process both 'regular' and

'exception' words; yet other, less 'opaque' orthographies have fewer irregularly-spelled words, in some cases none at all (Katz & Frost, 1992). Indeed, even some of the tests used to investigate the psychological processes of reading depend upon the particular characteristics of English: for example, Baddeley, Eldridge and Lewis' (1981) attempt to create items "phonemically but not visually similar to target word[s]" (p.443) could not be envisaged in a fully transparent orthography – as Goswami, Gombert and De Barrera (1998) discovered when trying to create similar items in Spanish.

At the very least, models of literacy development in different alphabetic languages must account for cross-linguistic variations in the *rate* at which L1 reading proficiency is acquired (e.g. Frith, Wimmer & Landerl, 1998; Wimmer & Goswami, 1994; Goswami, Gombert & De Barrera, 1998; Bruck, Genesee & Caravolas, 1997; Seymour, Aro & Erskine, 2003). In all these studies, the accuracy with which primary school children read words and pseudowords aloud was highest in the case of shallow writing systems (such as Spanish and German), lowest for the deepest writing system (English), and intermediate in the case of writing systems falling between these two poles (e.g. French). These results persisted even after controlling for confounding variables such as reading age and methods of literacy instruction.

In addition, *qualitative* differences have been postulated both in the processes of fluent word recognition and in the acquisition of this skill. Again, these differences are claimed to reflect the extent and reliability of the phonological information provided by the writing system. Thus, it is often argued that word recognition in logographic writing systems such as Chinese relies more on visual processing than it does in phonographic orthographies (Koda,1990:393) – although there is some counter-evidence suggesting

that phonological activation is an intrinsic component of word recognition even in Chinese (e.g. Perfetti & Zhang, 1995; Scholfield & Chwo, 2005). Further, amongst alphabetic writing systems, processing differences have been claimed to emerge as a consequence of differing levels of orthographic depth. Two principal models have been proposed: (a) the *Orthographic Depth Hypothesis* (ODH), which has provided a framework for much L2 writing system research (see section 2.6.2); and more recently (b) *Psycholinguistic Grain Size Theory* (PGST), which will be argued to account for some of the errors produced by participants in the current study.

#### 2.5.3.2 Orthographic Depth Hypothesis

The ODH (Katz and Frost, 1992) draws upon dual route models of word recognition. It emanated from various comparative studies of fluent L1 readers, such as the influential work by Katz and Feldman (1983) and Frost, Katz and Bentin (1987) – although it has not been universally supported by empirical evidence (e.g. Seidenberg, 1985; Baluch & Besner, 1991). The studies by Katz and colleagues found that lexical factors, such as word frequency and semantic priming, affected word naming speed in very deep orthographies (English and Hebrew) but not in very shallow ones (Serbo-Croatian). The ODH therefore proposes that the sub-lexical route is preferred in shallow orthographies – in which phonological forms can be computed at low processing cost due to the reliability of symbol-sound correspondences – but that the lexical route is preferred in deep orthographies, where computing pronunciations sub-lexically is inefficient due to the unreliability of print-sound correspondences.

Katz and Frost (1992) reject the possibility that readers of very deep or shallow orthographies might rely *exclusively* on one route or the other; they argue instead for a 'weak' version of the ODH, whereby orthographies differ in terms of the *relative* involvement of the two routes. They nonetheless assume that (a) phonological decoding is, to some extent, involved in word recognition in all orthographies; and (b) as in developmental models such as that of Ehri and McCormick (1998), phonological decoding of a given word may be replaced by direct visual recognition after repeated exposures to it. Nonetheless, "[a] higher replacement criterion will obtain in a shallow orthography where assembled phonology is easy to generate than in a deeper orthography" (Katz & Frost, 1992:158), leading to a greater involvement of direct (lexical) word recognition in deep orthographies.

The ODH has also been invoked to account for the different rates of literacy acquisition in beginner readers with different L1 backgrounds (see previous section). Whilst a full account of developmental processes falls outside the scope of this review, most participants in the current study having already acquired considerable fluency in L1 word recognition, the remainder of this section briefly reviews some findings which will contribute to the analysis of learners' errors in the current study (chapter 7).

An example is Frith, Wimmer and Landerl's (1998) study of 7- to 12-year-old readers of L1 German (shallow orthography) and English (deep orthography). In addition to the finding that English children generally read words and pseudowords less accurately than their German counterparts in the same school grade, the nature of the errors produced by the English participants was also interesting. Their pronunciations of pseudowords were often characterized by deviations from the sequence of graphemes (e.g. reading

"twiwana" for "tewanu"; "antina" for "utina"); they also exhibited many more word substitution errors than the German-speakers. Frith and colleagues suggest that their English participants were not only less proficient at decoding than the German children, but also tended to use this sub-lexical mechanism less, relying instead on the direct recognition of familiar words based on their orthographic form (i.e. 'sight reading'). This exactly echoes the conclusions reached a few years previously by Wimmer and Goswami (1994) in their comparative study of early reading in English and German.

The errors reported by Frith and colleagues also recall an incidental finding in Moustafa's (1995) study of seventy-five first-grade readers of L1 English: participants read many pseudowords incorrectly, realizing them instead as familiar real words in 40% of cases. Within developmental models of L1 English reading (e.g. Ehri, 1995), these kinds of substitution errors are exactly as would be predicted for the early 'partial alphabetic' phase: here, children forge connections between a word's more salient (e.g. initial and final) letters and the sounds they represent, ignoring other letters (such as those in the middle). When reading unfamiliar words, they may make guesses influenced by these partial phonological cues and any contextual information available, resulting in errors where they mistake the target word for another known word sharing similar salient letters (e.g. reading *bird* for *bond*). By contrast, Wimmer and Hummer (1990) found that the errors made by fifty-six beginner-readers of L1 German "consisted mainly of nonwords beginning with the first letter(s) of the target" (p.364), suggesting that participants predominantly used "an alphabetic strategy" (that is, sub-lexical decoding). In other words, there is evidence that the course of early reading development varies according to the orthographic depth of the writing system concerned, and that English children (who have to cope with a particularly deep orthography) may

have to pass through an additional phase of 'partial alphabetic' word recognition, or at least spend longer in this phase before progressing to the 'full alphabetic' phase in which linear decoding is mastered.

#### 2.5.3.3 Psycholinguistic Grain Size Theory

More recently, Ziegler and Goswami (2005, 2006) have proposed the *Psycholinguistic Grain Size Theory* (PGST), synthesizing a large body of empirical findings (including many of their own) from fields of study which have traditionally been distinct, such as (a) pre-reading development of phonological awareness in children with different L1s; (b) the development of word recognition in different writing systems; and (c) the nature and causes of developmental dyslexia in different languages. PGST builds on and extends the ODH; indeed, even some of the original proponents of the ODH have commented that PGST "seems to present an improved and modern alternative" (Frost, 2006:439).

Underlying PGST is the belief that beginner readers face three distinct problems, whose nature varies across writing systems and which therefore – together with the nature of the explicit literacy instruction received – influence the development of word recognition skills in language-specific ways:

(a) *The availability problem*. Children with different L1s differ in terms of their phonological awareness, specifically in terms of the size of phonological units of which they are consciously aware (e.g. syllables; onsets and rimes). Therefore, pre-

readers of different languages will have different sizes of phonological unit available for mapping onto the corresponding orthographic units.

(b) *The consistency problem*. Sound-symbol mappings may be inconsistent: a single orthographic unit may have more than one phonological form, and vice versa. The extent of this inconsistency (i.e. orthographic depth) differs cross-linguistically (section 2.4.3).

(c) *The granularity problem*. The smaller the 'grain size' of the orthographic units involved in print-to-sound mapping, the fewer individual mappings there will be to learn. For example, though English spelling bodies have more consistent pronunciations than individual graphemes (making them easier to work with when developing print-to-sound mappings), there are many more possible phonological rimes in the language than there are phonemes; and, in turn, there are many more possible words than there are rimes.

Like the ODH, PGST predicts differences among writing systems both in the processes of fluent word recognition and in the course of their development. However, according to PGST, these differences do not derive from the degree of involvement of phonological and lexical 'routes'; rather, they reflect the 'grain size' of the orthographic units involved in phonological decoding when learning to read, ranging from the smallest (graphemes) through intermediate (body/rime units and syllables) to the largest (whole words). The preferred grain size in a given writing system reflects the most efficient solution to the three problems outlined above. Thus, in a phonologically transparent (shallow) orthography such as Spanish, it is most efficient to use the smallest grain size, which is

individual graphemes and phonemes (subject to the development of phonological awareness at this level of detail). Since Spanish has few or no exceptions to its GPC, there is no advantage to using units of a larger grain size: this would involve learning a greater number of individual orthography-to-phonology mappings, yet with no payback in terms of increased consistency. By contrast, in a deep orthography such as English, GPC are often an unreliable basis for generating pronunciations and greater consistency can be achieved at the larger body/rime grain size; however, this entails the learning of a greater number of individual print-sound mappings. Ziegler and Goswami (2005) also point out that some English words have no orthographic neighbours at all (e.g. *choir*, *people*, *yacht*), in which case the most efficient unit of print-to-sound mapping may be the whole word. Following Shallice, Warrington and McCarthy (1983), they therefore conclude that "[t]he reduced reliability of small grain sizes in relatively inconsistent orthographies may well lead children to develop reading strategies at more than one grain size"[8] (p.11). This will take longer than the corresponding task of learning the (smaller) set of individual GPC in a transparent orthography, thus explaining the different rates of reading development in writing systems of differing orthographic depth.

The significance of this for the current study is that L1 English readers, whose processing of written words has developed in response to a very deep (inconsistent) writing system, may be expected to process L2 French text using larger (non-optimal) grain sizes than L1 readers of French would do. This suggests the hypothesis that L1-English readers may show inappropriate contextual variations in their realizations of

[8] In word recognition research, the term 'strategies' often designates not only beginner-readers' approaches to sub-lexical decoding, but also the subconscious processes involved in fluent (adult) reading, for which the term 'processes' is preferred in this thesis.

some French graphemes, particularly where those graphemes show contextual variation in their English realizations (e.g. <ou>: section 2.4.2). By contrast, fluent L1 readers of French would realize these graphemes identically in all contexts, due to the greater consistency of French GPC (although this difference is relative rather than absolute: there *is* some contextual variation in the realization of French graphemes, as shown in Appendix 1.2).

PGST appears to have received broad support from prominent researchers in various disciplines concerned with word recognition, as evidenced by the series of invited responses published in a single edition of *Developmental Science* (e.g. de Jong, 2006; Pugh, 2006; Paulesu, 2006; Wimmer, 2006). However, whilst these authors agree that beginner readers use different grain sizes according to the phonological transparency of their L1 writing system, all suggest that processing mechanisms converge as word recognition becomes more fluent. In particular, it is argued that – even in transparent orthographies – the recognition of familiar words moves towards a greater reliance on the 'direct', visual recognition of whole words (as demonstrated for L1 French by Sprenger-Charolles, Siegel & Bonnet, 1998). This view coincides with developmental models in which increasing reading skill is associated with the development of a 'sight vocabulary' of words, accessed automatically (Ehri & McCormick, 1998). Indeed, Goswami and Ziegler (2006) concede this point, arguing that "as learning proceeds, large grain size couplings between orthography, phonology and semantics will emerge in all orthographies and will be accessed automatically" (p.452), thus providing fast access to meaning.

In L1 English, the progression from smaller to larger grain sizes has also been observed in sub-lexical decoding, as required by pseudoword naming tasks. For example, Brown and Deavers (1999) compared thirty more skilled and thirty less skilled young readers in terms of their realizations of two classes of pseudowords: (a) those with 'regular consistent' rimes (such as *deld*); and (b) those with 'irregular consistent' rimes, such as *dalk*, which shared orthographic bodies with 'exception' words (e.g. *talk*). Skilled readers were more likely to pronounce 'irregular consistent' pseudowords in ways which rhymed with their irregular real-word counterparts (e.g. *dalk* rhymed with *talk*), whereas less skilled readers tended to pronounce these words according to regular GPC (e.g. *dalk* rhymed with *talc*).

### 2.5.4 Summary

Decoding appears to be crucial for developing L1 word recognition – and, hence, reading – in all alphabetic languages. However, the processes involved differ according to the 'orthographic depth' of the writing system. In shallow orthographies, beginner readers are able to exploit regularities in print-sound mappings at the smallest 'grain size', i.e. between individual graphemes and phonemes. By contrast, readers of deeper orthographies such as English must use regularities at various different grain sizes, particularly at the larger body/rime level. However, this entails learning a larger set of individual mappings, because there are a greater number of distinct orthographic bodies than there are distinct graphemes. Decoding proficiency is therefore acquired more slowly than in shallow orthographies. Further, as will be explored in the following section, L1 English readers approaching a 'shallower' (i.e. phonologically more consistent) L2, such as French, may be expected to show incorrect contextual variation in

their realizations of some individual graphemes as a result of their large grain size processing.

This 'Psycholinguistic Grain Size' model offers an elegant re-formulation of its forerunner, the Orthographic Depth Hypothesis, which emerged from dual route models of word recognition. Dual route models postulate distinct processing mechanisms for 'regular' and 'exception' words, the former being generated by symbol-to-sound mapping rules and the latter being looked up as pre-stored wholes. However, it has been argued that connectionist models – which form part of a wider, cognitive processing approach – offer a more satisfactory alternative to dual route accounts. Connectionist models use common computational principles to process all types of words, 'learning' the probabilistic relationships between written symbols and sounds through repeated experiences of processing individual items. The resulting processing mechanisms are inevitably tuned to the reader's L1 (assuming s/he starts out as a monolingual L1 reader), because they have been shaped by that writing system. This will be argued to have considerable implications for the processing of L2 writing systems.

## 2.6 Word recognition and decoding in an L2

### 2.6.1 Introduction

This section addresses the central question of how L2 decoding proceeds and how it is acquired. As will be argued below, the mechanisms differ from those involved in learning to decode in an L1. L1 beginner-readers can be likened to a *tabula rasa* as far

as word recognition is concerned: they begin the process without even the basic awareness that written words represent speech, although their L1 oral proficiency is relatively well developed. Beginner L2 readers are likely to be in the opposite position: they have limited oral proficiency to draw upon, but are a *tabula repleta* (Ellis, 2008) as far as processing mechanisms are concerned, having already developed fluent L1 word recognition. As various commentators have noted, the interaction between this existing L1 knowledge and the L2 writing system lies at the heart of L2 word recognition and decoding (e.g. Cook & Bassetti, 2005; Koda, 1996).

Traditionally, the effect of L1 knowledge on L2 learning has been conceptualized as an issue of 'transfer' (Odlin, 1989), which may be either 'positive' or 'negative' – also referred to as 'facilitation' and 'interference' (e.g. Koda & Reddy, 2008) – depending on the relationship between the two languages. This view can be traced back to Lado's (1957) Contrastive Analysis Hypothesis, according to which structural similarities between the L1 and L2 are associated with faster or more successful L2 learning, whilst cross-linguistic differences cause problems.

Clearly, L1-to-L2 transfer is not the only source of learners' errors in L2 learning: these may also result from factors such as the overgeneralization of L2 rules or the effects of teaching materials (Selinker, 1972); or they may reflect universal developmental trajectories independent of L1 influence (Dulay & Burt, 1974). Whilst L1-to-L2 transfer has, correspondingly, received less emphasis in many areas of SLA research, it has remained the dominant framework for studying L2 writing systems (Cook & Bassetti, 2005). Thus, following the Contrastive Analysis tradition, Koda (2007:27) recently argued that "L1-induced facilitation can be predicted, with great accuracy, through

finely-tuned cross-linguistic analysis" of a learner's L1 and L2 writing systems. 'Reverse transfer' between writing systems has also been explored (e.g. D'Angiulli, Siegel & Serra, 2001), though it is not discussed further here, since the outcome measures in the current study relate only to the L2.

Cognitive processing approaches to second language acquisition (section 2.5.2) have provided a new lens through which to interpret the issue of cross-linguistic transfer, "complement[ing] and extend[ing] earlier approaches to contrastive analysis" (Littlemore, 2009:6). These suggest that both perceptual and processing mechanisms become "tuned", "entrenched" (Ellis, 2008) or "primed" (Littlemore, 2009) through experience of L1 exemplars. It was argued above that repeated experiences of particular form-meaning associations will result in the strengthening of those connections to the point that they are activated automatically. 'Transfer' can thus be seen as the automatic activation of L1 connection patterns triggered by L2 input (Koda, 2007). This activation, occurring without any attention or intent on the part of the learner, will be helpful where the two systems 'overlap' (e.g. where the mappings between particular graphemes and phonemes are the same) but unhelpful where they differ.

However, for cognitive processing theories, the outcome of the learning process is "a dynamic, ever-changing state, rather than a static entity" (Koda, 2007:10): processing mechanisms will continue to develop through exposure to further input. This implies that, as the network gains experience of processing a new (L2) writing system, it may gradually become more tuned to that system (in addition to the L1 system) and therefore more efficient at processing it. This pattern has indeed been observed in L2 learners. For example, various empirical studies of high-proficiency ESL learners have recorded

faster responses to high-frequency words than to low-frequency words in lexical decision and naming tasks (e.g. Muljani, Koda & Moates, 1998; Akamatsu, 2002; Wang & Koda, 2005). Others have found L2 learners to respond more quickly to pseudowords with 'regular' as opposed to 'irregular' English spelling patterns (e.g. Brown & Haynes, 1985; Hamada & Koda, 2008). These frequency and regularity effects can only result from experience of the L2, since they depend upon repeated exposures to L2-specific orthographic patterns. Finally, Segalowitz and Segalowitz (1993) showed that adult L2 learners' responses in a lexical decision task grew faster (and, they argued, became automatized) through repeated exposures to items over the course of a study involving thirty sequential trials.

However, it has also been argued that L2 input alone may not be sufficient to develop L2-appropriate associations: for example, Ellis (2008) suggests that L1-entrenched perceptual mechanisms may simply continue to become further entrenched upon exposure to L2 input, unless a conscious intervention disrupts this process. Section 2.6.3 will present some empirical evidence to support this claim in the case of learners of L2 French. In the current study, the programmes of explicit instruction in L2 decoding are designed to support this kind of conscious intervention.

Returning to a learner's early L2 reading, how might the nature and extent of the 'overlap' between L1 and L2 writing systems (leading to 'positive transfer') be defined? Conversely, in what ways may the two systems differ, leading to 'negative transfer'? First, at a fundamental level, it may be assumed that prior acquaintance with *any* writing system will confer certain advantages: beginning L2 readers already understand certain underlying concepts. For example, they know that there is a systematic relationship

between writing and speech and that written text has directionality. Beyond this basic level of facilitation, a given pair of writing systems may overlap or differ on various dimensions, some of which (based on the analysis of writing systems provided in section 2.4) are summarized below:

(a) *Representational units*. If both L1 and L2 writing systems are phonographic, the L2 learner already knows that spoken words can be segmented into smaller phonological units; and that these, in turn, are systematically represented by corresponding orthographic units. Further, learners with an alphabetic L1 writing system will have developed phonological awareness at the fine-grained level of individual phonemes, whereas learners with a different (syllabic or logographic) L1 background may not (Morais et al., 1979; Bryant, 1993; Defior, 2004; see also section 2.7.2). This may in turn be expected to facilitate the learning of GPC in an alphabetic L2 writing system.

(b) *Directionality*. L1 readers of a left-to-right writing system such as English may find it easier to read an L2 writing system written in the same direction (e.g. Russian) than one written in a different direction (e.g. Arabic).

(c) *Script*. Learners' recognition of L1 written symbols will have become automatized through practice (LaBerge & Samuels, 1974), thus facilitating their learning of an L2 writing system using the same script. They will not need to learn to recognize a new set of written symbols, as would, say, an L1 English learner of Russian or Arabic.

(d) *Orthographic depth*. Odlin (1989) suggests that moving between writing systems of different orthographic depth may lead to asymmetrical levels of difficulty, because –

other things being equal – it is easier to learn a shallow orthography than a deep one. However, as outlined in section 2.5.3, Psycholinguistic Grain Size Theory would predict difficulties for learners moving both (a) from a shallower to a deeper orthography and (b) from a deeper to a shallower orthography. This is because their L1-tuned processing will be based on inappropriate grain sizes for dealing with the L2. For example, beginner readers whose L1 is English (deep orthography) may process written text in L2 French (shallower orthography) using unnecessarily large grain sizes, incorrectly leading to variations in their realizations of some graphemes according to their orthographic context.

(e) *Grapheme inventory*. Different writing systems using the same script may group written symbols into graphemes in similar or different ways (section 2.4.2): for example, <gn> usually forms a single digraph in French but not in English. Learners may 'parse' L2 letter-strings according to the grapheme inventory of the L1, resulting inevitably in incorrect L2 GPC. The French writing system also differs from English in its systematic use of diacritics. These are likely to pose different difficulties for L1 English learners: the problem here will not be one of automatic, L1-tuned processing, since learners have not had repeated experiences of processing them as part of their L1 input; rather, learners may be expected to have problems in noticing the diacritics and, subsequently, in knowing how to deal with them.

(f) *Print-to-sound mappings*. In two writing systems using the same script, some individual print-to-sound mappings may be similar (section 2.4.2): for example, <f> almost always represents /f/ in both French and English. This GPC may therefore be predicted to pose little difficulty for learners moving between the two languages,

since automatic L1-based decoding will yield a correct outcome. By contrast, many other GPC differ between the two languages: for example, <th> is realized as /θ/ or /đ/ in English but as /t/ in French. This is especially true in the case of vowels. Of course, the fact that some French and English GPC are the same (or very similar) in turn depends upon the overlap between the two languages' grapheme and phoneme inventories: thus, the GPC <f>→/f/ is similar in French and English only because both languages contain (a) the grapheme <f> and (b) the phoneme /f/, realized identically or near-identically. As will be explored in section 2.6.4, other L2 phonemes have no direct L1 counterpart.

(g) *Orthographic structure of words*. Within writing systems using the same script, certain sequences of graphemes may occur frequently in one language but less so in the other; this is in turn a function of (a) the phonological structure of words in the two languages and (b) their grapheme inventories. For example, many English words contain the rime /ɪŋ/, represented by the spelling body <ing>, whereas this phonological structure and spelling body are much less frequent in French. If (following Psycholinguistic Grain Size Theory) rime-level units are important in L1 English word recognition and decoding, the fact that an L2 contains many unfamiliar spelling bodies will result in dysfluency. By contrast, where sequences of graphemes are the same in the two languages, this may encourage automatic, L1-based processing. Therefore, whilst the existence of many cognate words in French and English may be helpful in terms of accessing meaning, it is likely to be particularly disruptive in terms of decoding. This is illustrated by the English word *nation* (/neɪʃən/) and its French homograph *nation* (/nasjõ/), which has the same meaning but is pronounced differently.

It is clear from the above list that, whilst some aspects of the overlap between an L1 and L2 writing system may be facilitative for the beginning L2 decoder, others may not be: some similarities are misleading, making it more likely that automatic, L1-based processing will be triggered inappropriately.

### 2.6.2 Empirical studies

These issues of L1-to-L2 transfer in word recognition and decoding have attracted increased research attention over the last twenty-five years. This follows a period of relative neglect (Koda, 1998), perhaps reflecting an emphasis on top-down views of reading or an assumption that lower-level processing skills would unfurl naturally in the wake of learners' increasing L2 proficiency.

As elsewhere in the SLA field, there has been an overwhelming predominance of writing system research on learners of English as an L2 (especially adult learners in an ESL context). Further, studies have focussed largely on the differing ways that L2 English words are processed by readers of phonographic L1 writing systems on the one hand and logographic ones on the other. Support has repeatedly been found for the view that learners from these differing L1 backgrounds tend to rely on different processing mechanisms for generating phonological forms from written L2 words (e.g. Brown & Haynes, 1985; Akamatsu, 1999; Koda, 1989, 1990, 1999; Muljani, Koda & Moates, 1998), with these mechanisms in turn being determined by the nature of the L1 writing system (as predicted by the Orthographic Depth Hypothesis). Thus, it is claimed that "alphabetic readers rely heavily on intraword analysis in obtaining a word's phonology"

whereas morphographic readers, whose L1 writing system does not systematically encode phonological information, "retrieve phonological information lexically through whole-word (or morpheme) activation" (Koda, 1999:53). A 'congruity effect' has also been found, whereby L2 English word recognition is facilitated in readers whose L1 writing system is 'congruent' (alphabetic) rather than 'non-congruent' (non-alphabetic) (e.g. Hamada & Koda, 2008). This is readily interpretable within a cognitive processing framework, as has been explicitly discussed in some of the more recent studies (e.g. Muljani, Koda & Moates, 1998).

A smaller number of studies have produced contrasting findings. For example, Akamatsu (2002) found evidence that three groups of ESL learners with contrasting L1 backgrounds (Chinese, Japanese and Persian) used similar word recognition procedures, consistent with the 'universal view' outlined by Baluch and Besner (1991). However, Akamatsu's participants – all graduate students at an English-speaking university – were clearly very proficient L2 readers. Any L1-based differences in their L2 word processing may therefore have been obscured as a result of their extensive experience of reading English.

Investigations of L1-to-L2 transfer effects amongst writing systems using the same script, as in the current study, have been less frequent. Of interest here is the effect of learning an L2 writing system which differs from the L1 system in terms of (a) orthographic depth and (b) its specific symbol-sound mappings, as in the case of French and English. Some of the studies mentioned above, investigating the processing of L2 English, have included participants with another Roman alphabetic L1 as part of a three-way comparison of learners with different L1 backgrounds. For example, participants

with L1 Spanish as well as Arabic and Japanese participated in studies conducted by two researchers in the mid- to late eighties (Brown & Haynes, 1985; Koda, 1989, 1990), apparently independently of each other (the Brown and Haynes study not being referred to in the later papers by Koda). However, in all three cases, the main emphasis is placed on the relative advantage conferred by an alphabetic as opposed to a logographic L1 background when reading L2 English. Indeed, Brown and Haynes report only a subset of their findings, omitting most of those relating to the L1 Spanish learners.

As an illustration of these three studies, Koda (1990) investigated "[t]he use of L1 reading strategies in L2 reading" amongst a sample of adult ESL learners comprising around twenty participants from each L1 background (Japanese, Arabic, Spanish; plus an English control group). Having controlled for English proficiency via a cloze test (although it appears that the lack of significant differences between the groups' scores may have been influenced by low sample sizes), participants read extended passages of English describing either (a) five invented cocktails or (b) five imaginary breeds of fish. The names of these novel items were either English pseudowords (e.g. *sunpi*) or Sanskrit symbols (e.g. □), both of which were unfamiliar to all participants. Outcome measures were (a) the time taken to read the passage and (b) scores on a recall test requiring participants to match the novel items to English descriptions of them.

As expected, both overall test scores and reading speed were lower in the Sanskrit condition, interpreted as an indication that "phonological inaccessibility seriously interfered with the reading process regardless of orthographic backgrounds" (p.401). However, increases in reading time were significantly greater for the Arabic, Spanish and English participants compared to the Japanese participants, suggesting that

"phonographic readers were handicapped by phonological inaccessibility significantly more than morphographic readers" (p.402). This is argued to reflect the fact that the Japanese participants were accustomed to generating phonological forms from unfamiliar graphic shapes (since they have to do this in L1 whenever they encounter unfamiliar or invented *kanji* characters), whereas alphabetic readers depend upon orthographic information to generate phonological forms. It was further found that, within the phonographic group, English participants were the most seriously impaired by Sanskrit symbols, followed by the Spanish and finally the Arabic participants, apparently reflecting the degree of overlap between the L1 and L2 writing systems (Spanish being Roman alphabetic and Arabic being phonographic but written in a different script). Overall, Koda claims to have found "strong empirical evidence of L1 orthographic influence on cognitive strategies used in L2 reading … demonstrate[ing] that cognitive transfer does indeed occur in the L2 reading process" (p.404).

However, it might be objected that the task used in this study (involving the reading of a mixed-script text) is an artificial one, somewhat removed from the experience of 'ordinary' reading; indeed, this is frequently the case in L2 writing system research, as for example in Akamatsu's (1999, 2005) tasks involving 'cAsE aLtErNaTiOn'. Koda's final recall test also appears to measure memory as well as reading comprehension. Further, she gathers no information concerning the conscious strategies which learners may have used (and which may have differed from one L1 group to another) to help them complete the task.

A subsequent investigation of word recognition by learners from contrasting L1 backgrounds, co-authored by Koda (Muljani, Koda & Moates, 1998), provides a little

more emphasis on the interaction between L1 and L2 writing systems which share the same script. Participants (N=32) were again ESL learners, half L1 Indonesian (Roman alphabetic) and half Chinese (logographic). The relatively small sample size seems to be rather characteristic of studies in the field. A lexical decision task was performed using words and pseudowords which were either (a) high or low frequency (in the case of pseudowords, this applied to the real words on which they were based) and (b) composed of either 'congruent' or 'incongruent' spelling patterns, where a 'congruent' pattern was permissible in both Indonesian and English (e.g. *garden*, having a CVC syllable structure) and an 'incongruent' pattern was permissible in English only (e.g. *branch*, having a CCVCC structure).

As predicted, Indonesian participants were faster than Chinese participants at naming all categories of item, presumably because of the greater overlap between their L1 and L2 writing systems. All participants also showed a frequency effect, reflecting their considerable experience of processing the L2 writing system (most being graduate students at a US university). Most importantly, however, congruent as opposed to incongruent spelling patterns generally facilitated word recognition for the Indonesian but not the Chinese participants, suggesting that their familiarity with these letter-combinations from the L1 helped them to process the items more quickly. This finding is explained within a connectionist framework: the congruent spellings activated connections which were already well established through experience of the L1 orthography.

Other studies have adopted a different paradigm, focussing on learners with a shared L1 and analysing their reading errors in an L2 which uses the same script as the L1, but

differs in orthographic depth. Hickey (2005), for example, documents typical word recognition errors made by young (third grade) L1 English learners of L2 Irish Gaelic, drawing on data gathered previously by the author. (The sample and procedures for data collection are not stated in this publication). Gaelic orthography is shallower than English but has complex GPC; it also includes many letter combinations unfamiliar to English readers (e.g. word-initial <mh>, <bp>, <ng> and <bhf>) as well as the unfamiliar acute accent used as a length marker on vowels. As such, it may be likened to French in terms of both (a) its orthographic depth and graphemic complexity and (b) its relationship to the English writing system.

Hickey reports that learners in this population tend to mis-read L2 items as more familiar L1 or L2 words, apparently based on similarities in their global shape or on the recognition of salient (e.g. word-initial) letters. For example, <[a] choileán> (/kwilʲaːn/, '[his] puppy') is read as L2 <ceol> (/kʲɔːl/, 'music'); <siad> (/sʲəd/, 'they') as English *said*. In the context of a bilingual connectionist network in which both L1 and L2 lexical items are interactively activated (e.g. Van Heuven, 2005), this can be interpreted as the activation of familiar words which have a similar pattern of salient letters to the target items. Another common error type identified by Hickey involves transposing letters to give more familiar (L1-like) sequences, recalling the difficulties experienced by participants in Muljani, Koda and Moates (1998) when processing "incongruent" letter strings: for example, <[ar an] tsráid> (/traːd/, '[on the] street') is read as the pseudoword /trasɪd/, as if the letter-string were <trasid>.

Overall, the L2 word recognition errors identified by Hickey do not seem to reflect sub-lexical decoding at the individual GPC level; rather, they tend to be substitution errors or

'deviations from the grapheme sequence', similar to those reported in early L1 English reading by Frith, Wimmer and Landerl (1998) and Moustafa (1995). These are consistent with Ehri's (1995, 2005) 'partial alphabetic' phase. However, the young learners described by Hickey (2005) are at an earlier stage of L1 reading development than the participants in the current study; further, they differ in that their early L1 and L2 literacy is developing simultaneously.

### 2.6.3 Empirical studies of word recognition and decoding in L2 French

Although L2 English has received the bulk of attention in L2 writing system research, the past ten years have also seen increased interest in L2 French word recognition, specifically amongst beginner learners in English MFL classrooms (the same population as the current study). Lynn Erler has been particularly influential in this field, beginning with her doctoral thesis (2002) and continuing with various publications (2004, 2006, 2007, 2008; Macaro & Erler, 2008b); indeed, Cook and Bassetti's (2005) recent review of L2 writing system research cites Erler (2002) as the only example of a study involving a phonologically deeper L1 and shallower L2. Most of Erler's published articles have appeared in either *Language Learning Journal* or its sister publication *Francophonie*, which have (in the past) been aimed predominantly at UK-based teaching practioners rather than international researchers. Nonetheless, Erler's work is discussed in detail, since of all the available literature on L2 writing systems it has the most direct relevance to the current study.

Erler (2002) examined Y7 students' French decoding proficiency as part of a wider, exploratory investigation into their experiences of L2 reading. The sample (N=359) was

drawn from fourteen teaching groups (guarding against teacher effects) in two English comprehensive schools, giving a wide spectrum of L2 attainment. Participants completed pen-and-paper tests of L2 decoding, requiring them to (a) identify rhymes, (b) identify matching onset sounds, (c) match cognate forms to their English translations (avoiding 'false friend' distractors) and (d) translate into English a set of classroom instructions, presumed to be familiar to participants in aural but not written form (e.g. *asseyez-vous*, /asejevu/, 'sit down'). Participants scored poorly: for example, the mean rhyme test score was 3 out of 15. Erler concluded that "most pupils had little idea after one year of learning French about spelling-sound rules for principal vowel sounds in French, and silent final consonants" (p.172). However, the tests included heterogeneous items which were likely to be either (a) known to participants in both spoken and written form (e.g. *chat*, /ʃa/, 'cat'), (b) known in spoken form only (e.g. the classroom instructions quoted above) and (c) entirely unknown (e.g. *tablier*, /tablje/, 'apron'). It is thus difficult to know to what extent participants were 'decoding' the words in the sense used in this thesis (i.e. by a process of sub-lexical assembly), and to what extent they were retrieving pre-stored pronunciations of familiar words based on sight recognition.

A second strand of Erler's (2002) investigation of decoding involved a timed reading aloud task. A sub-sample (N=23) read nine words, chosen to exemplify various vowel sounds and the SFC 'rule' (section 2.4.2). These words, presented individually on cards, were "likely" to be unknown to participants, giving a purer measure of decoding as understood in the current study. The findings corroborated those from the pen-and-paper tasks: overwhelmingly, participants were unable to use French GPC knowledge to pronounce the words correctly. As an illustration, only two participants correctly pronounced the vowel in *se* (/sə/, 'herself/himself'), despite undoubtedly having had

frequent exposures to similar written forms (e.g. *je*, /ʒə/, 'I'; *le*, /lə/, 'the'). This suggests that – as argued by Ellis (2008) – L2 input alone may be insufficient to develop accurate L2 print-sound connections (section 2.6.1).

Erler's participants compensated for their lack of L2 decoding proficiency by using English print-sound mappings; sometimes, items were also read in terms of orthographically similar English words (e.g. English *says* or *sees* for French *ses*, /se/, 'his/her'). Further, this reliance on English was sometimes commented on explicitly (e.g. *fougère* "looked like forgery"). Altogether, twenty of the twenty-three participants referred to English at least once when discussing the French words. English-resembling items (e.g. *faire, mouvement*) were also reported to be easier to pronounce, even though in reality they were read in an anglicized way. By contrast, unfamiliar letter combinations which could not readily be related to English words were seen as problematic: for example, one participant said that the letters in *lieu* "were in a weird order so I didn't know how to pronounce them". This perception of the L2 orthography as 'odd' (with reference to the L1 'norm') can be interpreted within a cognitive processing framework: letter combinations which do not occur in the L1 cannot be processed automatically (because they have seldom or never been experienced before) and so will attract conscious attention. It is therefore unsurprising that Erler's participants found diacritics, absent from English orthography, particularly problematic. The perception of French orthography as 'weird' in comparison to English is of course ironic given the greater consistency of French GPC.

A further interesting finding from Erler's (2002) reading aloud task was that, when French words did resemble English ones, five out of seven participants who were

particularly slow to name them were also those who pronounced them most accurately. Erler concludes that these slow-but-accurate decoders may have been consciously trying to overcome automatized L1-based processing in order to produce more accurate French pronunciations. However, this presupposes that the learners can identify their initial (L1-based) pronunciations as incorrect in the first place. Several of Erler's participants made comments suggesting that they were indeed aware of a mismatch between (a) their own decoding of familiar L2 words and (b) the correct pronunciations they had heard or rehearsed in class, as in the following example:

> *There's ... a thing that's going on in my head trying to pronounce it, but when you hear it [in your head], it doesn't sound anything like what you're hearing [from Miss in lesson]*

This comment about the problem of dual phonological forms recalls those made by learners on the vocabulary learning task described in section 2.3.2.

Given the deficiencies shown by Erler's (2002) participants in pronouncing even basic French words, she concluded that their condition was consistent with a clinical definition of phonological dyslexia in L1. Subsequent research (Erler, 2007, 2008) – using a similar rhyme judgment task but with a larger sample (N=1735) of students in Years 7, 8 and 9, argued to be representative of KS3 learners of French in England – obtained similar results: only around 3% of participants scored at better-than-chance level. Further, though there were significant differences between the scores of the different year groups, these were extremely small in magnitude, suggesting a lack of progress in L2 decoding over time (although longitudinal data would be required to make definite claims about participants' progress). However, Woore (2006) noted some weaknesses in the rhyme test as a measure of decoding. For example, it again included both familiar

and unfamiliar words. Further, participants may have used a range of other strategies besides L2 decoding in order to complete the test, such as judging items on their visual similarity (Erler, 2007); indeed, for participants aware of their own lack of French decoding proficiency, the use of alternative strategies would be an entirely rational choice. Finally, correct answers could be generated on some test items simply by treating the words as if they were English. Nonetheless, the findings presented in Erler (2008) can be seen as a reliable indication of participants' poor L2 decoding proficiency, since they would certainly not have scored so poorly had they been proficient decoders of French.

A further important strand of Erler's work has been to highlight the possible causal link between L2 decoding and motivation for language-learning, at least amongst MFL learners. Given the widespread deficiencies in basic literacy skills such as word identification and decoding, "some pupils are bound to take offence and become disaffected learners" (Erler, 2002:308). As in L1 reading, this may lead to a downward spiral or "Matthew effect" (Stanovich, 1986), where poor readers do not enjoy reading, read less and so lose opportunities to develop fluency in word recognition; this leads them to enjoy reading even less, thus perpetuating a negative cycle.

Erler also argues that French orthography may be particularly prone to engender negative feelings in comparison to more transparent systems such as German or Spanish. This recalls a related finding by Saito, Horwitz and Garza (1999) that L2 reading-related anxiety was higher amongst L1 English learners of French than of Russian (although the study had some sampling problems, in that a higher proportion of the L2 French learners were studying the language compulsorily compared to the L2 Russian learners). Saito

and colleagues explain their finding by the fact that, although French shares the same alphabet as English, the differences in print-to-sound mappings may be "disconcerting to many students" (p.217): in other words, the problem may again be one of misleading similarity between the writing systems.

Erler and Macaro (2012) investigated further the link between basic literacy and motivation in L2 French, using the same large, representative sample of MFL learners as reported in Erler (2007, 2008). They found that students who (a) achieved higher scores on rhyme judgment and syllable segmentation tasks and (b) reported higher levels of self-efficacy on decoding-related tasks were significantly more likely to express a wish to continue learning French post-14 (when it becomes optional). The L2 literacy tests used in this study – which included the same rhyme test as that used by Erler (2007, 2008) – imply a broader definition of 'decoding' than in the current study and had some problems in their design (see above and Woore, 2006). Nonetheless, it is striking that participants' combined test scores predicted around 8% of the variance in their reported intention to continue learning French. As noted in Woore (2009), this figure is surprisingly high, given the many factors which may impinge on MFL learners' motivation for the subject (e.g. Coleman, Galaczi, & Astruc 2007).

Building on Erler's work, a further series of studies into L2 (mainly French) decoding amongst beginner MFL learners has been conducted by Woore (2003, 2004, 2006, 2007, 2009, 2010), in the form of Masters dissertations and related publications in *Language Learning Journal*. In one strand of this work, Woore (2006) replicated previous findings of deficient word-level processing amongst KS3 learners of French. The study used a Reading Aloud Test comprising fifty-three individually-presented French words, chosen

to include at least three 'tokens' of each of fifty-one French graphemes. This was argued to offer a purer measure of decoding (as understood here) than Erler's pen-and-paper tests, since (a) all words were unfamiliar to participants, forcing them to generate pronunciations via sub-lexical assembly and (b) participants could not gain positive scores by using alternative strategies, as on the rhyme judgment task. The participants (N=188), selected randomly from five teaching groups in each of Y7 and Y9 in one secondary school, performed poorly on the test, pronouncing around half of all graphemes acceptably. Informal observations by the researcher suggested that they were overwhelmingly reading the words as if they were English pseudowords; their positive scores may therefore be taken as a rough indication of the extent of overlap between the French and English writing systems. Woore (2006) also found that, although Y9 participants obtained significantly higher scores than the Y7s, the difference was extremely small, strongly suggesting that MFL learners in the sample school were making little if any progress in French decoding.

This view was subsequently confirmed by Woore (2009), who re-tested the Y7 participants from the previous study one year later. Mean scores at the two time points were almost identical. These findings of poor L2 decoding proficiency and poor progress in this skill amongst MFL learners are consistent with recent comments by Ofsted (2011:24) – the English schools inspectorate – based on visits to ninety secondary schools of diverse character: "[a] large majority of the students [are] still reaching the end of KS4 with no real idea, for instance, of the effect of an accent on an individual letter or what a cedilla denotes in French".

However, informal observations made in the course of Woore's (2009) study also suggested that, whilst there was no increase in the number of graphemes realized acceptably, at least some Y8 learners pronounced words differently than they had one year previously. This raises the possibility that 'progress' in decoding may not involve a simple, linear progression from incorrect to correct outcomes; rather, it may follow a more complex trajectory, recalling 'U-shaped' models of developing L2 proficiency (McLaughlin, 1990). In other words, some participants seemed to have changed in their patterns of decoding, and in that sense to have made 'progress', even though they did not produce a higher proportion of 'correct' realizations on the Reading Aloud Test. Further research is required to explore this possibility more systematically.

A second strand of Woore's work has involved analysing participants' decoding errors. Since the Reading Aloud Test used in the various studies mentioned above is scored at the level of individual GPC, it can be used to diagnose areas of particular difficulty for participants and thus to test the predictions of a 'contrastive analysis' approach. It was argued above that L1 English learners should experience 'positive transfer' in cases where graphemes have similar realizations in French and English (so that automatically activated, L1-based connections will deliver the correct L2 output), whereas greater difficulties would be expected where the languages' GPC differ. Woore (2006) found that this was indeed the case. The graphemes most often pronounced correctly (in over 90% of cases) were single-letter consonant graphemes, whose phonological realizations are similar in both languages: < b d f g l m n p t >. By contrast, vowel graphemes (where most of the differences between the two writing systems lie) were realized correctly less than 50% of the time. So too were 'Silent Final Consonants' (a feature of French but not English orthography) as well as several complex consonantal graphemes, which either do

not exist as single graphemes in English or have contrasting pronunciations (<ç gn gu qu th>). A logical further step in this error analysis would be to examine the nature of participants' realizations of those problematic graphemes. As outlined above, a prediction based on cognitive processing theories and Psycholinguistic Grain Size Theory (section 2.5.3) is that English readers – having adapted to the inconsistency of GPC in their L1 – may pronounce certain vowel graphemes differently depending on their orthographic context. This will often yield incorrect outcomes in the shallower French writing system due to the greater consistency of its GPC.

However, in an earlier, small-scale project, Woore (2003) found that Y7 students did not always decode L2 French words in a linear, grapheme-by-grapheme fashion at all. For example, on an earlier version of the Reading Aloud Test, *épatante* (/epatɒ̃t/, 'marvellous') was realized by the five participants as [ɛpɔnəti], [ɛfɔ̃ti], [ɛpɔ̃ntət], [ɛpatɐni] and [epatɒ̃]. Apart from the near-correct [epatɒ̃], all these pronunciations seem to reflect deviations from the grapheme sequence, like those produced by the young L1 English readers in Frith, Wimmer and Landerl (1998). Further, in an investigation of L2 German decoding by sixty MFL learners in Y7, Woore (2004) found that two participants often read L2 items as familiar English words (or else omitted them altogether). For example, one participant read the German words *schmusen* (/ʃmuzən/, 'to smooch') as *school, Quelle* (/kvɛlə/, 'source') as *question* and both *Bund* (/bʊnt/, 'alliance') and *Bussi* (/bʊsi/, 'kiss') as *bird*. This seems to reflect an erroneous sight recognition process based on salient letters, again recalling the substitution errors documented by Frith, Wimmer and Landerl (1998) and Moustafa (1995). However, it was argued that these two participants, who had low L1 reading ages, may process

unfamiliar L2 words using qualitatively different mechanisms from the phonological decoding employed (even if imperfectly) by their peers.

Finally, in a third strand of research, Woore (2010) investigated learners' conscious reasoning when decoding French words. This exploratory study aimed to build on Erler's (2002) suggestion that L2 learners who pronounce written French words more accurately may do so through a conscious effort to move beyond automatic, L1-based decoding. A sample of mixed-attaining MFL students (N=12), towards the end of their first year of learning French at secondary school, completed individual, task-based self-report interviews in which they read unknown French words aloud. As expected, many of the pronunciations produced in the interviews reflected, wholly or partly, English symbol-sound mappings (e.g. *goinfrez*, /ɡwɛ̃fʁe/, '[you] eat greedily', pronounced [ɡɔɪnfrɛz]). However, all participants also drew upon a range of strategies in an attempt to make their pronunciations sound 'more French'. Often, this involved focussing on the realizations of particular graphemes which were perceived as problematic, such as letters bearing diacritics or those arranged in combinations unfamiliar from the L1. Further, there was a high degree of consistency in the strategies reported by the different participants. The most frequent examples were as follows:

(a) *breaking words up* into smaller units and working on the pronunciation of one section at a time;
(b) *appealing to conscious knowledge* of particular symbol-sound mappings;
(c) *drawing analogies to known words* to support the pronunciation of target words with similar spellings;

(d) *trying out possible pronunciations* by saying them aloud and evaluating whether they 'sounded French'.

However, whilst all of these strategies may potentially be helpful in supporting L2 decoding, in most cases they were found to "misfire" for various reasons, leading to incorrect outcomes. First, having divided words into smaller units, participants usually decoded the resulting word-parts according to English GPC. Moreover, some individual digraphs, such as <gn>, were incorrectly subdivided into separate graphemes, making correct pronunciations of them highly unlikely, if not impossible. Second, the conscious knowledge on which participants drew was often incorrect, being based on English GPC: for example, one participant said that he "knew that O I sort of made an /ɔɪ/ sound", as in English *moist*, whereas in French it is in fact pronounced /wa/. Third, the analogy strategy was often undermined by participants' inaccurate knowledge of the 'source' words on which they drew. For example, one participant attempted to use the familiar phrase *qui s'appelle* (/kisapɛl/, 'who is called') to work out the pronunciation of the first three letters of *quignon* (/kiɲɔ̃/, 'crusty end of a baguette'), but mispronounced the source phrase in an anglicized way as [kwisapɛl]. Finally, participants sometimes repeatedly tried out possible pronunciations of words, yet had no secure basis for evaluating their output beyond a sense that the word should somehow sound 'different than it would in English'.

Woore (2010:15) concluded that these participants could be "compared to people struggling to solve a logic puzzle where the information provided is insufficient to complete the task and where the criteria against which to measure successful completion of the puzzle are lacking". Therefore, whilst strategic reasoning clearly has the potential

to deliver improved L2 decoding outcomes, it can only do so if underpinned by secure knowledge of L2 symbol-sound mappings, of the L2 grapheme inventory and of any L2 words used as sources of analogy.

### 2.6.4 Phonology in L2 Decoding

The discussion up to this point has examined the challenges of acquiring L2-specific GPC based mainly on the similarities and differences between the L1 and L2 writing systems. However, before moving on to discuss the role of instruction in helping learners overcome these challenges (Part III), it is necessary to consider briefly the issue of L2 phonology and its relationship to L2 decoding. This issue has been largely or completely ignored in much L2 writing system research. However, given that decoding involves converting the written symbols of a language into its corresponding phonological forms, fully accurate (native speaker-like) decoding outcomes can be achieved only insofar as the reader can produce accurate L2 phonological forms in the first place. For example, the French grapheme <an> cannot be realized correctly as the nasal /ɔ̃/ if the reader cannot produce this phoneme accurately.

However, 'correct' pronunciations of at least some L2 phonemes may be difficult for learners to achieve. Although there may be some exceptions (Bongaerts, 1999), it is widely believed that there is a 'Critical Period' for acquiring native-like phonological representations (Birdsong, 1999; Scovel, 1969). People who begin L2 learning after the end of this period (assumed to be somewhere between age 6 and puberty: Long, 1990) will inevitably have a 'foreign accent'. Under this view, the participants in the current study (aged 11-12) fall towards the end of this period.

The problem for late-onset learners appears to lie not simply in *producing* L2 sounds, but in perceiving them (Flege, 1999; Kuhl, 1994). Kuhl, for example, proposes that speech perception becomes language-specific early in life. Through experience of L1 input, the perceptual system 'learns' which sounds are relevant for distinguishing meaning in that language (i.e. its phonemes) and which are not, creating phonetic prototypes or 'Native Language Magnets' to which other, proximate sounds are 'attracted' – an example of 'perceptual learning' as discussed by Ellis (2008). 'Mature', L1-tuned systems will therefore perceive L2 sounds in terms of the L1 magnets. Phonetic differences between the L2 pronunciation of a phoneme and its L1 equivalent may then be lost to the listener: for example, native speakers of English may fail to perceive the difference in voice onset timing between French voiceless plosives (pˀ, tˀ, kˀ) and their aspirated English counterparts ($p^h$, $t^h$, $k^h$). Further, L2 phonemes which do not exist in the L1 may be perceived as exemplars of the acoustically closest L1 phonemes. For example, French has a lip-rounded high front vowel /y/, whereas in English only back vowels are rounded. English speakers may therefore identify French /y/ as /u/ (Rochet, 1995, cited in Flege, 1999); and they may produce this vowel either indistinguishably from /u/ or with non-native colouring (e.g. /ju/) (Levy & Law, 2010). French nasal phonemes (absent in English) may also be 'unpacked' by English-speakers (Čubrović, 2002) to give a combination of L1 oral vowel plus nasal consonant: for example, /ɔ̃/ may be represented (and therefore pronounced) as /ɒn/. This is exemplified in various French-into-English loanwords, such as *entrée*, whose pronunciation is listed in the Concise Oxford English Dictionary (1990) as /ɒntɪeɪ/.

This means that the accuracy of L2 decoding may be constrained by the nature of learners' L2 phonological representations, even where the correct symbol-to-sound mappings are in place. For example, French <an> may be realized as /ɒn/ (rather than the correct /ɔ̃/) by an L1-English decoder who (a) knows the correct GPC but (b) has a non-native-like representation of the L2 phoneme. In the current study, an attempt is therefore made to distinguish systematically between 'correct' (native-like) and 'acceptable' (L1-influenced) phonological forms.

### 2.6.5 Summary

L2 writing system research has predominantly been conducted within a framework of L1-to-L2 transfer. Following a contrastive analysis approach, similarities between the L1 and L2 systems are seen as facilitative ('positive transfer'), whereas differences will create difficulties for the learner ('negative transfer'). Some researchers have claimed that pairs of writing systems such as English and French (both Roman alphabetic) can be considered broadly 'convenient' compared to typologically more distant pairs, such as English (alphabetic) and Chinese (logographic). More recently, cognitive processing models of learning have provided an explanatory framework for L1-to-L2 transfer, which is seen as the triggering by L2 input of automatized, L1-tuned processing mechanisms.

Empirical research in this area has focussed mainly on the effects of moving from a logographic L1 system to an alphabetic L2 (usually English). Learners with a logographic L1 have repeatedly been found to rely more on visual word recognition mechanisms than readers from phonographic L1 backgrounds. This has often been

interpreted in terms of the Orthographic Depth Hypothesis and only more recently in terms of connectionist models.

A smaller number of studies have focussed on the effects of moving from one alphabetic writing system to another, where the L1 and L2 may differ in terms of orthographic depth and, in the case of systems using the same script, in terms of their actual symbol-to-sound mappings. These studies have generally supported the view that similarities between writing systems at various levels have a facilitative effect for the L2 learner.

Recent studies have also been conducted in the English MFL context, looking specifically at L2 French decoding by KS3 learners who are (in most cases) already fluent L1 decoders. Various studies have found that these learners (a) have poor French decoding proficiency and (b) make poor progress in this aspect of their learning. Participants have overwhelmingly relied on English symbol-sound correspondences to decode French words, resulting in incorrect, anglicized pronunciations. Therefore, simply classifying the English and French writing systems as a 'convenient' pairing is misleading. From a cognitive processing perspective, this is because the shared script promotes the triggering of automatic, L1-based processing mechanisms by L2 input. Whilst this sometimes leads to correct outcomes, as in the case of graphemes with similar realizations in French and English, it also generates problems. First, some GPC (especially vowels) differ in the two languages. English L1 readers are, furthermore, accustomed to greater variation in the pronunciation of vowels according to their orthographic context; they may therefore base their decoding on larger orthographic 'grain sizes', a strategy which is often inappropriate for French due to the greater consistency in its GPC. Second, some groups of letters, which form single graphemes in

one language but not the other, will tend to be processed automatically according to L1 conventions, making incorrect outcomes almost inevitable. Third, where the L2 uses combinations of letters which do not occur in the L1, learners will be unaccustomed to processing them and may therefore perceive them as 'odd' in comparison to the L1 'norm'.

These automatic, L1-tuned processing mechanisms are likely to be difficult for learners to suppress. Exposure to L2 input alone may not suffice to establish new, L2-specific connections, unless a conscious effort is made to interrupt and override L1-based processing. There is some evidence, gathered in the MFL context, of more successful beginner-decoders making this kind of conscious intervention in their L2 decoding. However, it is also clear from self-report data that conscious interventions can only succeed insofar as they are based on secure knowledge of the L2.

Finally, fully accurate L2 decoding requires accurate representations of L2 phonemes. However, learners may perceive L2 phonemes in terms of the nearest L1 categories, which are established early in life and are difficult (some would say impossible) to alter subsequently. This means that some degree of L1 'colouring' may be expected even in learners who successfully acquire L2-specific GPC.

# Part III: The role of explicit instruction

## 2.7 L2 instruction and the development of L2 decoding proficiency

### 2.7.1 Introduction

This section examines the role of reading instruction in the development of L2 decoding proficiency, with an emphasis on English MFL classrooms at KS3. First, in order to contextualize this discussion, section 2.7.2 briefly reviews the teaching of early literacy in L1 English. This has been the focus of considerable research attention – perhaps unsurprisingly so, given the unusual degree of complexity of the English writing system (section 2.4.2). Second, section 2.7.3 considers recent pedagogical approaches to the teaching of L2 reading in MFL classrooms. Finally, section 2.7.4 reviews the limited evidence concerning the effectiveness of explicit L2 decoding instruction for learners already literate in their L1.

### 2.7.2 Literacy instruction in L1 English

The teaching of early literacy in L1 English has been subject to long-standing controversy, based on the rival positions of phonics-based instruction and other approaches such as 'whole language' methods (Graves & Dykstra, 1997; Hurry, 2004; Brook, 1999). The former involves the discrete teaching of letters and graphemes on the one hand and the individual sounds they represent on the other; children learn to blend these sounds in order to decode written words. The assumption is that children are able

to benefit from conscious rules explicitly taught to them. Whole language approaches, by contrast, emphasize the reading of genuine texts for the purpose of accessing their meaning. A range of cues may be used to identify words, including contextual inferencing. The assumption is that children will work out the necessary print-to-sound correspondences for themselves. These contrasting approaches can be seen to reflect the competing 'top-down' and 'bottom-up' views of skilled reading, as outlined in section 2.3.1.1.

Stahl, Duffy-Hester and Stahl (1998), in their review of different approaches to teaching phonics, argue that in reality phonics-based methods and whole language approaches co-exist in the practice of skilled practitioners. However, when viewed in polarized terms, the two approaches are mutually incompatible. Advocates of the phonics approach argue that focussing on contextual cues at the expense of a word's orthographic form reduces opportunities to learn and consolidate print-sound relations. On the other hand, whole-language theorists claim that the abstract learning of decontextualized rules has limited value, since children are unable to apply them in practice to their own reading. Further, curriculum pressures are such that time must not be wasted teaching something that children can simply acquire by themselves. These arguments are equally relevant to L2 reading development, where even less curriculum time is available.

More recently, evidence from large-scale quasi-experimental trials of different teaching methods – including some carried out in the UK (e.g. Sylva & Hurry 1996; Hurry, Sylva & Riley, 1999; Sylva et al., 1999) – has led to a consensus that explicit phonics teaching is important for children in the early stages of learning to read English. Phonics-based methods have also been positively evaluated in contexts with high numbers of children

with English as an Additional Language (e.g. Stuart, 1999). Hurry (2004), in her review of instructional methods, notes that many studies in this area are flawed: for example, some lack random assignment of groups to experimental and comparison conditions; and there is considerable variation within the 'same' instructional approach as implemented in different studies. Nonetheless, "from sheer weight of numbers, the evidence for the value of explicit phonics instruction is overwhelming" (p.561).

This conclusion was echoed by Rose (2006:4) in his *Independent Review of the Teaching of Early Reading*:

> [T]he systematic approach, … generally understood as 'synthetic' phonics, offers the vast majority of young children the best and most direct route to becoming skilled readers and writers.

Synthetic phonics – in which symbol-sound correspondences are taught in isolation and decoded sounds are then blended together or 'synthesized' to derive the pronunciations of whole words – was subsequently advocated in English primary schools by both the previous and current governments (DfES, 2006b; DfE, 2010).

In theoretical terms, the importance of explicit phonics teaching may reflect the fact that young children require help in acquiring the 'alphabetic principle' (section 2.4.1): namely, that sub-lexical orthographic units (in linear sequence) represent the individual sounds of the spoken language. This is linked to the finding that young pre-readers do not yet possess phonological awareness at the level of individual phonemes, which is a prerequisite for learning grapheme-phoneme mappings. Rather, it is the process of learning to read an alphabetic writing system itself which develops their phonological awareness at this finest-grained level (Morais et al., 1979; Bryant, 1993).

By contrast, there is some evidence to suggest that, once the underlying principle of grapheme-phoneme mapping has been grasped, the acquisition of further GPC may be able to occur implicitly. For example, Byrne and Fielding-Barnsley (1991) found that a group of children who had received explicit instruction in six grapheme-phoneme mappings performed better than a comparison group in a post-test requiring them to identify these phonemes from their written representations; but they were also significantly better at identifying six new phonemes which had not been included in the training programme. This suggests that beginner readers of an alphabetic L2, already literate in an alphabetic L1, may be able to transfer their knowledge of the alphabetic principle to the new language, hence reducing the need for explicit instruction. On the other hand, recent studies of young MFL learners found no evidence that they were able to acquire L2 GPC implicitly (section 2.6.3).

Finally, though recent policy documents have focussed on 'synthetic phonics', other approaches to systematic phonics instruction have also been proposed and evaluated (Stahl, Duffy-Hester & Stahl, 1998). These include analogy-based approaches, as advocated by Goswami (1995) and Moustafa (1995), in which learners are taught to use "the spelling-sound pattern of one word, such as *beak*, as a basis for working out the spelling-sound correspondence of a new word, such as *peak*" (Goswami, 1995:139). Gaskins et al. (1988) trialled an analogy-based approach with several cohorts of poor L1 English readers (N=275) who had not successfully learned to decode under a synthetic phonics regime. In daily sessions of 15-20 minutes, learners (a) acquired a secure sight vocabulary of around 120 words then (b) practised using these 'source' words to decode unknown words with similar spelling patterns. Significant improvements were recorded

on word and pseudoword reading tests, although there was no control group with which to compare these results. More generally, Stahl, Duffy-Hester and Stahl (1998) report mixed results for experimental evaluations of analogy-based teaching programmes in L1. However, in relation to L2 decoding, it is interesting that the use of an 'analogy strategy' was reported frequently by Woore's (2010) participants when attempting to decode unfamiliar French words.

### 2.7.3 Policy and practice in MFL classrooms

#### 2.7.3.1 Little emphasis on L2 decoding

Erler (2002) argued that her participants' poor French decoding proficiency (section 2.6.3) resulted from a pedagogical approach which – reflecting the requirements of the *National Curriculum* (DfEE, 1995, 1999) and GCSE examinations[9] – provided limited opportunities for L2 reading. When reading tasks *were* completed, these usually involved short texts (comprising single words, phrases and sentences) used to test comprehension. This echoes similar conclusions by both Dobson (1997), following a review of Ofsted inspection data from a broad range of English secondary schools, and Grenfell (1992, 1995), drawing on considerable experience as a teacher, teacher educator and researcher in the MFL context.

More specifically, Erler (2002) noted a lack of emphasis on the development of decoding proficiency. She found that beginner-learners had "little opportunity to read anything beyond what ha[d] already been introduced and practised orally" (p.62), reflecting the

[9] Public examination completed at age 16.

model of reading progression enshrined in the *National Curriculum*'s KS3 assessment criteria ('Level Descriptors'). Mitchell (2003), in a persuasive critique of these assessment criteria, argues that they are structured around an implicit preoccupation with accuracy: thus, if students only read L2 words which they already know orally, there is less danger of mispronunciations arising as a result of defective decoding. However, this approach also has disadvantages: if beginner-learners read texts composed only of familiar material, this (a) restricts the type of text to which they have access, (b) prevents them learning new words through reading and (c) denies them opportunities to improve their decoding of unfamiliar written forms. Further, even if learners did develop accurate knowledge of L2 symbol-sound mappings, they would be unlikely to read in sufficient quantity for their decoding to become automatized. As Grabe and Stoller (2002:76) summarize: "Students learn to read by reading a lot, yet reading a lot is not the emphasis of most [L2] reading curricula".

The neglect of word-level processing documented by Erler (2002) may have resulted from an emphasis on top-down processes in reading, perhaps a delayed reflection of the top-down theories prominent in L1 reading research and pedagogy some twenty years earlier. Indeed, Erler finds evidence of "a persistent conceptualization in the UK context that phonological processing is something which less proficient readers engage in, and which more proficient readers do not do" (p.61). For example, Mitchell's (2002) practical advice for MFL teachers on "Developing Reading Skills" focuses entirely on higher-level comprehension; there is no mention of 'bottom-up' processes such as decoding. For Mitchell, the "main aim" of L2 reading instruction "must ... be a feeling of increased confidence when faced with a potentially daunting task ... an awareness that not every word needs to be understood ... and that visual images can help with

comprehension" (p.100). However, whilst such compensatory strategies are important in helping L2 readers access texts above their current linguistic level, if this is *all* that reading pedagogy emphasizes, then fluent word-level processing may not have the chance to develop. Thus, MFL students' deficiencies in L2 decoding may (at least partly) be an artefact of the pedagogical approach taken in their classrooms.

***Implicit acquisition of L2 decoding***

As in L2 reading research up to the 1980s (Koda, 1996), there may also have been an assumption that decoding skills would 'look after themselves', being acquired implicitly (and inevitably) as wider L2 proficiency increased. This view may have been encouraged, in the case of pairs of languages such as French and English, by the high degree of overlap between the two writing systems. It is interesting in this respect that commentators on other pairs of Roman alphabetic writing systems have made similar complaints about teachers' neglect of L2 decoding (e.g. Hickey, 2005).

In support of the 'implicit acquisition' position, it was argued (section 2.6.4) that the need for explicit phonics instruction in L1 is linked to the fact that children need help in acquiring the basic 'alphabetic principle'; but that knowledge of this principle (once acquired) can then be transferred to a second alphabetic writing system, allowing L2-specific GPC to be acquired implicitly. This of course depends on the extent to which phonemic awareness, developed in relation to the L1, can be transferred to L2. There is both theoretical and empirical support for this hypothesis (e.g. Koda, 2007 and Durğunoglu, Nagy & Hancin-Bhatt, 1993, respectively).

If correct, these arguments would suggest that explicit instruction may be less necessary for learning to decode in a second alphabetic writing system than it is in the acquisition of early L1 English literacy. However, it was argued above that learners whose L1 and L2 writing systems share the same script face different problems, associated with L1-entrenched perceptual and processing mechanisms. Within a cognitive processing framework, Ellis (2008) and Littlemore (2009) argue that explicit, form-focussed L2 instruction may help learners to overcome this problem through conscious attention. More generally, there is a body of evidence in the SLA sphere pointing to the effectiveness of form-focussed instruction (embedded within a wider communicative approach), particularly in helping learners acquire those linguistic features "in which there is a misleading similarity between the L1 and L2" (Spada, Lightbown & White, 2005:199; see also Spada & Lightbown, 1999, 2008). However, most research in this area has focussed on the acquisition of grammatical features rather than on lower-level reading processes such as decoding.

Nonetheless, there is some evidence that L2 decoding can be acquired implicitly: for example, van Berkel (2005) claims that Dutch learners of L2 English quickly gain proficiency in the L2 writing system, despite its highly complex and inconsistent nature, even in the absence of systematic instruction. However, as argued in Chapter 1, MFL learning in England takes place in a very different social and educational context, where learners' motivation for learning an L2 is likely to be considerably lower and where curriculum time is severely limited. There is convincing evidence that beginner MFL learners do not acquire L2 decoding proficiency implicitly, or at least do so too slowly to be detected in the studies conducted to date (section 2.6.3). Erler (2002) therefore concluded that MFL learners need systematic, explicit instruction in French decoding,

echoing a previous call by Siddons (2001). At face value, French would appear to lend itself well to this kind of treatment, since its complex GPC are nonetheless highly regular (section 2.4.2). However, whether explicit decoding instruction can be effective in this context, and if so, what form it should take, are questions remaining to be answered.

#### 2.7.3.2 Renewed official guidance

Subsequent to the appearance of Erler's (2002) thesis, and mostly after the initial planning phase of the current project (2005-2006), governmental guidance for MFL teachers began to place greater emphasis on the development of L2 decoding proficiency, perhaps influenced by developments in L1 literacy teaching policy. For example, a series of official publications in the 2000s stated that developing learners' knowledge of symbol-sound correspondences should be an explicit aim of MFL instruction[10].

First, the *Key Stage 3 Framework for Languages* (DfES, 2003) specified that Y7 learners should be taught "[t]he alphabet, common letter strings and syllables, sound patterns, accents and other characters", such as French "-qu-, -oi-, -ant, -on, other standard vowel sounds, accents, cedilla"; then, in Year 8, they should learn "[s]ome common exceptions to the usual patterns of sounds and spellings". Y9 learners are then expected to move on from decoding to focus on meaning. However, there was an absence of guidance on the specific teaching methods which might be employed. There was also a lack of evidence to suggest that it was realistic to expect pupils to master regular symbol-sound mappings in Y7 and then progress to the more difficult 'exceptions' in Y8, especially given that

[10] The status of these documents is in doubt due to the current government's education reforms.

learners typically receive only around two hours' MFL teaching per week. Nonetheless, this assumption concerning the rate of learners' progress was retained when the *Key Stage 3 Framework* was revised (DCSF, 2009) to provide greater alignment with its KS2 counterpart (see below).

Next, the *Key Stage 2 Framework for Languages* (DfES, 2005) appeared, supporting the introduction of MFL teaching in primary schools, a central plank of the previous government's *National Languages Strategy* (DfES, 2002). This again included the recommendation that students should be encouraged to "[r]ead aloud unknown words by applying rules of the sound/spelling system they have learned". Indeed, from the very outset of children's L2 learning in Y3 (age 7), they should be expected to "pronounce accurately the most commonly used characters, letters and letter strings". Perhaps influenced by the *Primary Framework for Literacy* in L1 English (DfES, 2006b), it again seems to be assumed that learners will quickly become proficient decoders (Y6 students should be able to "[r]ead aloud with confidence, enjoyment and expression") and thus move on to focus on higher-level processing and comprehension. However, the extent to which these ambitious goals are being realized on a national scale is questionable, particularly since recent evaluations of primary MFL provision have noted a relative neglect of literacy in favour of oracy (Ofsted, 2011; Wade, Marshall & O'Donnell, 2009; Cable et al., 2010).

Finally, a revised *National Curriculum* (QCA, 2007b) stipulated that MFL lessons should include the teaching of "the interrelationship between sounds and writing in the target language", which "includes underpinning principles such as common letter strings". However, the assessment criteria – 'Level Descriptors' – used to assess

students' progress, published as part of this document, remain substantially unchanged from the earlier version critiqued by Erler (2002) and Mitchell (2003). In respect of reading, they still contain only sporadic references to decoding and word naming. Further, the ability to read *unfamiliar* items aloud – requiring decoding proficiency as defined in this thesis – is not mentioned until Level 5, expected to be reached by most students at the end of Year 9.

Despite the prominence attached to decoding in this official guidance for MFL teachers, it is striking that much of the evidence of poor decoding proficiency cited above (e.g. Erler, 2008; Woore, 2006, 2009) was gathered several years *after* the launch of the Key Stage 2 and 3 Frameworks. This may be because further time is required for the impact of these documents to be felt. However, given the importance that measuring and auditing students' performance has assumed in the current educational climate (Fielding, 1999), an additional concern is that the various recommendations concerning the teaching of decoding may have been overshadowed by the need to meet the requirements of the assessment framework. As argued above, this places little emphasis on decoding and offers no clear pathway for its development.

### 2.7.4 Approaches to explicit L2 decoding instruction

At the time of planning the current study (just after the publication of the original *Key Stage 3 Framework*), there was an absence of specific guidance for MFL teachers on the form that explicit L2 decoding instruction might take in MFL classrooms – although the *Framework* seemed implicitly to follow a phonics-based approach, since it specified individual GPC which learners should know. Subsequently, suggestions for teaching

activities have appeared in various official documents, such as the *Key Stage 2 Framework* (DfES, 2005), the sample schemes of work for primary MFL teaching (QCA, 2007a) and the revised *Key Stage 3 Framework* (DCSF, 2009). However, the effectiveness of the methods proposed – and indeed of explicit instruction in L2 decoding more generally – remains to be convincingly demonstrated by empirical studies.

There have been some positive trials, both in other L2 learning contexts and in MFL classrooms. To give two examples, Yen's (2004) Masters dissertation and Liaw (2003) both found that – at least in an elementary school context – programmes of explicit phonics instruction led to improved L2 word-decoding skills in Taiwanese EFL classrooms. The former study involved systematic coverage of forty-two English sounds over a three-month period, based on the commercial *Jolly Phonics* scheme (Jolly Learning, 2011). However, both studies had small samples (N=37 and N=41 respectively) and only Yen (2004) included a comparison group. Yen also recorded anecdotal evidence that learning to sound out English words "helped EFL pupils conquer their fear of this foreign language" (p.85); however, these students were starting out from a logographic L1 writing system (Chinese), a rather different situation than that faced by learners such as those in the current study, whose L1 and L2 are both alphabetic.

In the MFL context, Woore (2004, 2007) conducted a small-scale, quasi-experimental evaluation of a programme of L2 decoding instruction for Y7 pupils (N=59), albeit in German, a shallower writing system than French (section 2.4.2). The programme of instruction included explicit coverage of individual GPC likely to pose difficulties for L1 English learners (e.g. <äu> → /ɔɪ/), but also an additional strand building on the analogy-

based approach described in Goswami (1995), Moustafa (1995) and Gaskins et al. (1988). This involved participants memorizing, in both oral and written form, short poems designed to exemplify various problematic GPC. The memorized poems then provided a bank of 'source words' which could be used to determine the pronunciations of unknown written forms with similar spellings, through the operation of a chain of conscious strategies comprising the following stages (adapted from Woore, 2007):

1. Identify, in the unknown *target word*, a *target grapheme* which is problematic to decode;
2. Decide to use the poems as a source of analogy;
3. Search the written forms of the poems (initially on paper but with the eventual aim of doing so 'mentally') for a *source word* which contains the target grapheme;
4. Recall the pronunciation of the source word;
5. Segment the sound and spelling of the source word in order to determine the pronunciation of the target grapheme;
6. Transfer the pronunciation of the target grapheme to the target word, blending it with the other sounds in the word.

The principle behind the 'analogy strategy' is that the pronunciation of individual graphemes can be determined without the need to recall pre-learnt GPC 'rules'. It was assumed that the poems would be easier to memorize than a large body of individual GPC.

The decoding instruction in Woore (2004, 2007) was conducted over a ten-week period (eighteen lessons) in the form of 'starter activities', that is, in short segments at the beginnings of lessons (similar to the discrete decoding segments recommended for early

L1 literacy instruction by the *Primary Framework for Literacy*: DfES, 2006b). The intervention group performed significantly better than the comparison group on a post-test requiring them to read unknown German words aloud, despite there being no significant differences between the groups' pre-test scores (although teacher effects were acknowledged as a potentially confounding variable). Both groups also performed equally well on school-wide, end-of-term assessments in reading and listening comprehension, suggesting that the decoding instruction had not hindered their progress in other aspects of L2 learning.

However, whilst the difference between the two groups' post-test scores was statistically significant, it was only small. Supported by evidence from interviews with a sub-sample of participants, it was suggested that this may have been because participants' attempts to use the analogy strategy were undermined by inadequate knowledge of the source words in the poems (particularly their written forms). It was therefore recommended that any similar programmes of instruction in future should devote more time to memorizing the source words. However, it is unclear to what extent (if any) the memorized poem and analogy strategy actually contributed to participants' improved decoding outcomes over and above the effects of other aspects of the instruction programme, such as the explicit instruction in individual GPC.

In retrospect, it is also notable that – whilst participants had the analogy strategy modelled for them and were encouraged to use it when decoding new words encountered in class – they did not receive a systematic, 'staged' programme of strategy instruction such as that proposed by Macaro (2001) and exemplified in Macaro and Erler (2008a). This involves (a) awareness raising and strategy modelling, (b) scaffolded practice of

using the strategy and (c) the progressive removal of scaffolding to develop independence in strategy use. This may be another reason why participants made less effective use of the analogy strategy at post-test than might have been expected. The effectiveness of L2 decoding instruction based on the analogy strategy – as compared to a phonics-based approach involving the teaching of individual symbol-sound mappings, which are then used to sound out words – therefore merits further exploration.

### 2.7.5 Summary

Erler (2002) argued that the deficiencies in French decoding found amongst her participants resulted from a neglect of decoding (and of reading more generally) in MFL classrooms. She consequently called for greater emphasis on explicit decoding instruction at KS3. Cognitive processing theory provides some justification for the potential effectiveness of explicit instruction in developing learners' L2 decoding proficiency, by helping them overcome their automatic, L1-tuned processing.

More recently, official guidance for MFL teachers has caught up with policy and practice in the teaching of early literacy in L1 English, recommending that the development of learners' knowledge of L2 GPC be an explicit aim of instruction. However, investigations of beginner MFL learners have so far provided no evidence to suggest that this guidance has produced the hoped-for improvements in decoding proficiency. On the other hand, it may be that participants in these studies did in fact made progress in French decoding, but that this progress was not linear and so was not discernible in terms of increased numbers of 'correct' realizations.

At the time of planning the current study, there was an absence of specific guidance on the form that a programme of explicit L2 decoding instruction might take in MFL classrooms. Subsequently, various suggestions have appeared in official publications. However, the available empirical evidence concerning the effectiveness of these proposed teaching methods – and indeed of L2 decoding instruction more generally – is sparse. A previous, small-scale study by Woore (2004, 2007) evaluated a programme of instruction in a Y7 German classroom, based partly on the use of an 'analogy strategy' (involving the use of 'source words' in a memorized poem to derive the pronunciations of unfamiliar words with similar spellings) alongside more traditional, synthetic phonics-type activities. The treatment group showed a significant but small advantage over the comparison group in terms of their progress in German decoding, warranting a larger-scale trial of this approach. However, it was unclear exactly what contribution the analogy strategy had made to participants' outcomes.

# Part IV: Overall summary and Research Questions

To summarize, the literature review has suggested that L2 decoding may play an important role in L2 reading and in L2 learning more broadly, especially through its contribution to vocabulary acquisition. However, L2 decoding appears to be heavily influenced by transfer from L1. This may have both facilitative and inhibitory effects according to the nature and extent of the overlap between the two writing systems. Within a cognitive processing framework, it has been argued that transfer can be explained as the triggering of automatic, L1-based processing mechanisms by L2 input. This may lead to correct outcomes where the GPC of the two languages are similar, but incorrect ones where they differ. Further, this automatic, L1-tuned processing may involve not only individual symbol-to-sound mappings, but also (a) the way in which letters are segmented into graphemes in the first place and (b) the 'grain size' of orthographic units involved in decoding.

It was argued that conscious intervention may be effective in overriding this L1-tuned processing, allowing L2 words to be decoded according to L2 GPC. This would be the first step to 'retuning' the system to accommodate L2 input, where the ultimate goal is the achievement of automaticity through practice. Automatic word recognition is in turn an important component of fluent reading comprehension.

In the particular case of French and English, the two writing systems overlap on various dimensions, leading to claims that they are a 'convenient' pairing in terms of acquiring L2 decoding proficiency. However, there are also a number of differences between them: for example, the same letter string may be segmented differently into graphemes in

the two languages; some individual GPC are different, especially in the case of vowels; and there is more contextual variation in the way individual graphemes are pronounced in English than in French. These differences can be classed as 'misleading similarities', because the orthographic representations themselves contain no indication that there may be differences in the way they are decoded. This seems likely to encourage the inappropriate triggering of L1-tuned processing, generating incorrect (L1-based) pronunciations.

Recent research amongst beginner learners of French in English MFL classrooms has indeed found a reliance on L1 GPC. Further, most learners in this context appear to make little (or very slow) progress in French decoding over their first few years of MFL lessons. There is some evidence to suggest that these learners can successfully interrupt the operation of automatic, L1-tuned decoding mechanisms through conscious reasoning; however, this did not usually lead to correct pronunciations, because learners' knowledge of French GPC (and of the language more broadly) was deficient. It is possible that explicit instruction may help learners to (a) develop their knowledge of the L2 decoding system and (b) use it more effectively to override L1-based decoding.

In early literacy instruction in L1 English, a large body of empirical research has contributed to a consensus that phonics teaching – in which individual sound-symbol correspondences are taught explicitly and discretely – provides the most effective way of helping children learn to read. This may have influenced an increasing emphasis, in official guidance for MFL teachers, on explicit instruction in L2 decoding, following a period of apparent neglect of this skill. However, the theoretical rationale for phonics teaching is different in the two cases: in L1, it is linked to the need to establish the

alphabetic principle and to develop phonological awareness at the finest-grained level of individual phonemes; L2 readers, by contrast, have already acquired the alphabetic principle and phonemic awareness, but explicit instruction may help them overcome automatic L1-based processing. However, there is an absence of convincing empirical evidence to show that explicit decoding instruction can be effective in an L2.

***Research Questions***

To address this gap in the literature, the current study evaluates two programmes of L2 French decoding instruction in MFL classrooms: (a) a phonics-based approach; and (b) a programme which additionally aims to develop learners' use of the analogy strategy based on familiar 'source words' in a memorized poem, as in Woore (2004, 2007). The study is driven by the following overarching research question: *How effective are two programmes of French decoding instruction, delivered in secondary school MFL classrooms?*

However, following the arguments advanced in this literature review, it may be necessary to address this question in several different ways. At the simplest level, the effectiveness of the programmes can be measured by the extent of any increase in correct decoding outcomes between the different groups of participants. However, it was argued in section 2.6.3 that beginner-learners' progress in L2 decoding may be more complex than simply moving directly from incorrect to correct outcomes. Indeed, given the interpretation of L1-to-L2 transfer adopted in this thesis, one initial indication of progress might be a lesser reliance on L1-based processing, even though the outcomes may still not be correct: in other words, the nature of learners' errors may change.

Finally, it may be that increases in learners' decoding proficiency arise not through conscious appeal to a strengthened knowledge of French GPC, but instead through improvements in their implicit knowledge; or it may be that their knowledge of French GPC has improved but is not yet reflected in their decoding outcomes, because their L2-specific decoding processes have not yet had the opportunity to become automatized.

Based on these arguments, the specific research questions for the current study are:

1. *Do the programmes of instruction lead to more accurate decoding of L2 French?*
2. *Is one of the programmes more effective than the other?*
3. *Do the programmes of instruction lead to any other changes in participants' realizations of French graphemes?*
4. *Do the programmes of instruction lead to changes in learners' strategic reasoning when decoding French words?*

These questions are addressed in the 'real world' context of MFL classrooms, generally underrepresented in SLA research (chapter 1). The aim is to produce findings which MFL teachers perceive to be as relevant as possible when making decisions about their own practice, thus addressing the sense of disjuncture observed in this context between research findings on the one hand and classroom practice on the other (Macaro, 2003). Whilst it is acknowledged that teachers will always need to "make subtle judgments in unique situations" (Hagger & McIntyre, 2000:487), it is hoped that this thesis will contribute to evidence-informed practice in MFL classrooms. As Atkinson (2000:322) argues:

> The possibility of finding final answers to pedagogical problems through educational research is questionable at best; but the possibility of

promoting and extending critical discourse among both researchers and teachers, and of opening up channels for debate and consideration of a range of solutions to classroom problems, will remain fruitful as long as educational research continues to exist.

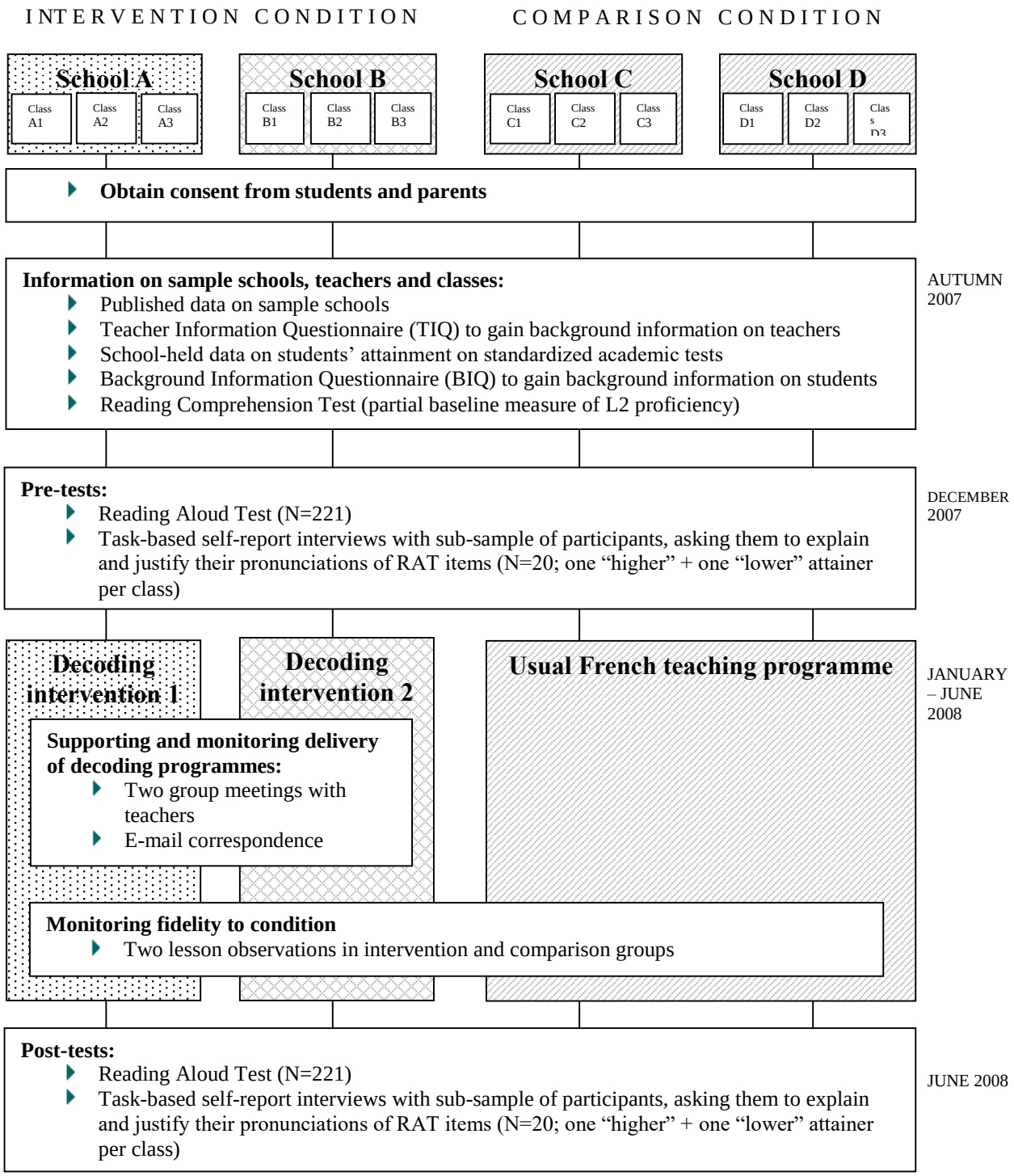
# Chapter 3: Research design

## 3.1 Overview

To address the research questions, a quasi-experimental, non-equivalent control group design (Campbell & Stanley, 1963) was adopted involving three classes in each of four schools (labelled A-D). The three classes in school A followed the programme of decoding instruction based on the 'poem approach' (Programme A); those in school B followed the synthetic phonics-based programme (Programme B); and those in schools C and D formed a comparison group, who followed their usual programme of French teaching and received no systematic instruction in decoding. Tests of decoding proficiency, administered before and after the period of instruction (henceforth t1 and t2), measured changes in the proportion of graphemes decoded accurately by each group of participants. Their realizations of individual graphemes at each time point were also analysed. Finally, task-based self-report interviews at t1 and t2 investigated the strategic reasoning of a sub-sample of intervention and comparison participants when decoding French words. Delayed post-tests were also originally planned, in order to assess the extent to which any improvements in decoding proficiency caused by the programmes of instruction were sustained over time; however, for various reasons, these were not implementable in practice (section 3.3).

Figure 3.1 provides a schematic overview of the implemented research design.

**Figure 3.1 Overview of research design**



## 3.2 Issues in research design

### 3.2.1 Randomized Control Trials in Educational Research

Following Gorard (2001:2), large-scale randomized control trials (RCTs) were considered to represent the 'gold standard' for evaluating specific interventions. There have been various calls for greater use of RCTs in educational research (e.g. Torgerson, 2001), but this can be problematic on several counts.

The first problem concerns the unit of randomization. To evaluate a medical intervention (e.g. a new drug), individual research participants can be randomly allocated to experimental or control groups. It is therefore likely that any potentially confounding variables will be evenly distributed between the two conditions (the "simple elegance of randomization": Torgerson, 2001:323). Certain educational interventions may also be evaluated using this model of individual allocation to condition. Examples include Butler's (1988) study of different feedback types and, in the SLA sphere, Fukkink, Hulstijn and Simis's (2005) investigation of whether training in L2 word recognition improved reading comprehension. The training in this latter case was delivered individually via computer within normal lessons. However, whilst individualized, computer-based programmes of L2 decoding instruction are conceivable – and indeed have already been planned (Zoe Handley, personal communication, 27.09.10) – the current study wished to remain rooted in the specific context of MFL classrooms, where ready access to computers for all students may not be guaranteed. A programme of instruction was therefore preferred which could be delivered under 'normal' conditions on a whole-class basis.

Clearly, when evaluating interventions designed for use in a whole-class context, randomly allocating participants to experimental and control conditions is impossible, because school students are already grouped into classes according to the institution's own organizational principles. Shuffling students into new groupings for research purposes would therefore be disruptive. A solution to this problem – exemplified by Moore, Graham and Diamond (2003) in their evaluation of a sex education programme – is the "cluster RCT", where whole classes (or whole schools) are randomly allocated as intact units to intervention or control conditions. As discussed in more detail below, the current study partially follows this cluster model in allocating participants to conditions.

A second question when transferring the RCT model from the medical to the educational context concerns the nature of the interventions. In medical research, a drug under trial is administered in identical form to all participants. By contrast, educational interventions delivered by human teachers will inevitably vary from one context to another. The current study adopted a pragmatic response to this difficulty: whilst attempts were made to standardize the decoding instruction between the various participating classrooms (section 3.5), it was considered inevitable that there would be some variation in how the programmes were implemented in individual classes – just as there would be if the programmes were used 'for real' in a non-research context. Lesson observations and interviews with teachers were therefore used (section 3.7) to monitor 'fidelity to condition' in each class and to document any variations in implementation, so that these could be taken into account in the analysis.

### 3.2.2 Initial plans for research design

#### 3.2.2.1 Introduction

Inevitable resource limitations on the planning of the current study, together with additional issues arising during its implementation, made Gorard's (2001) "perfect piece of research" in the form of a large-scale RCT unattainable. Various compromises were required as a result of the attempt to situate the study in MFL classrooms in the "real world" (Robson, 2002). It is worth noting that many classroom intervention studies, even in the published SLA literature, suffer from similar limitations, yet these are not always made as transparent as they could be. To take one example, Erlam's (2003) study of the effects of different types of grammar instruction demonstrates several of the problems discussed later in this section, including the fact that a small sample of three intact classes from one secondary school were assigned to three different instructional conditions and were all taught by the researcher herself. However, though these details *are* discernible in Erlam's account of her methods, the limitations they impose are not highlighted and the final conclusions seem to be phrased more strongly than they would warrant. In the current study, it is hoped that a greater degree of methodological transparency will make it easier for MFL teachers to draw critically upon the findings as a source of evidence to inform their own practice.

#### 3.2.2.2 Sampling: general considerations

An initial compromise in planning the current study concerned the sample size. Given sufficient resources, it would have been desirable to produce findings which could be

generalized statistically to the wider population, as in the case of Macaro and Erler's (2008b) investigation into the relationship between basic French literacy skills and motivation for learning the language. However, the size of their achieved sample (N= 1735), based on a stratified random draw of secondary schools purchased from the NFER and argued to be representative of KS3 French learners in England, lay beyond the scope of the current study.

Considerations of possible sample size were also intertwined with and constrained by other aspects of the study design. The first was the nature of the planned decoding instruction. Interventions differ in terms of the resources needed to implement and support them. In Moore, Graham and Diamond's (2003) intervention study, a large sample of twenty-four secondary schools was possible, because participating teachers only had to attend a short training course and then deliver a single sex education lesson. By contrast, the current study asked teachers to integrate decoding instruction into their lessons over a period of at least nineteen weeks, in addition to attending initial training and on-going support meetings with the researcher. Considerable resources were invested simply in sustaining this level of intervention. This in turn imposed limitations on both the size and geographical spread of the sample.

Second, the nature of the tests of decoding proficiency used in the current study also restricted the sample size. Clearly, pen-and-paper decoding tests (e.g. Macaro & Erler, 2008b) can be administered quickly to large numbers of participants on a whole-class basis. However, the current study used an individualized Reading Aloud Test (section 3.6.3). Though arguably a more valid measure of decoding proficiency, this is also considerably more labour-intensive both to administer and analyse.

These constraints on the size and characteristics of the sample were discussed with research colleagues both informally and in seminars. It was considered that a sample of four local schools, each providing up to three participating classes, represented a realistic maximum size for the current project. Whilst this would not permit findings to be generalized statistically to the wider population of KS3 learners, it would be sufficiently large to support statistical comparisons between the groups. Further, whilst it will be seen that all four schools are in some senses uncharacteristic of the wider national population in England, they are arguably not unrepresentative of many other English schools in non-inner city settings.

#### 3.2.2.3 Allocating groups to conditions

A second difficulty encountered during the planning of the current study concerned the unit of allocation of groups to the intervention and comparison conditions. Assuming a maximum of four participating schools (see previous section), two possibilities arose:

(a) *One school, three conditions* (hereafter 'one-to-many aproach'). Three classes would be recruited in each school, one assigned to each of the two instructional programmes and one to the comparison condition.

(b) *One school, one condition* (hereafter 'one-to-one approach'). All participating classes within each school would be allocated to the same condition: those in the first school to Programme A, those in the second to Programme B and those in the remaining schools to the comparison condition.

Following the example of previous SLA intervention studies such as Macaro and Mutton (2009), the current study adopted the latter (one-to-one) approach. This decision was taken for two reasons. First, most importantly, the alternative (one-to-many) approach runs the risk of programme contamination through contact between teachers and/or students in the different conditions. Second, it was felt that working as part of a larger team (as in the one-to-one approach) would allow colleagues to provide mutual support in the context of an extended intervention requiring considerable commitment.

Despite these advantages of the one-to-one model for allocating groups to conditions, there is a risk of school effects: i.e. any improvements in one group's outcomes may result not from the treatment they received but rather from some particular features of the school they attended. Two precautions were taken to mitigate this threat. First, a (partially successful) attempt was made to match participating schools in terms of relevant variables (see section 3.4.1). Second, school effectiveness research suggests that school effects are less important determinants of student achievement than teacher effects (Scheerens & Bosker, 1997; Sammons, 2007; Luyten, 2003) and that they operate mainly indirectly, by influencing classroom-level factors such as teaching style (Kyriakides et al., 2010; Hill & Rowe, 1996). Applied to the particular sphere of L2 decoding, this matches intuitive assumptions that it is mainly what goes on in MFL lessons which will determine a student's proficiency in reading French words aloud. The current study therefore planned to use teachers who would be matched as far as possible in terms of relevant characteristics, although again this was not achieved in reality (section 3.4.2). The MFL teaching received by each participating class was also investigated through a combination of teacher questionnaires and lesson observations (section 3.7), to help identify confounding variables.

It should also be noted that the threat of teacher effects – where differences in the groups' decoding outcomes may result from non-relevant variation in the teaching they received, rather than from the particular programmes of instruction they followed – is in any case equally great under both the one-to-one and the one-to-many approaches, since both would involve three different teachers in each condition. The current study is not alone in the SLA field in facing this difficulty: for example, Arteaga, Herschensohn and Gess's (2003) quasi-experimental study of the effects of a programme of phonological awareness-raising also involved different teachers for different groups. Possible solutions to the problem would have been to (a) greatly increase the sample size, so that differences between individual teachers were averaged out over the sample as a whole or (b) use the same teacher for all classes. The former was impossible due to resource constraints; the latter could theoretically have been achieved by the researcher visiting all participating classes in all four schools to deliver the decoding segments. This approach, adopted in some published intervention studies, is however dependent upon the scale of the project: in Erlam (2005), for example, all participating classes received instruction from the researcher herself, but this instruction comprised only three forty-five minute lessons per group. In the current study, it would not have been viable for the researcher to deliver all thirty-eight intervention segments to all six classes. In any case, this would have reduced the study's ecological validity – one of the aims being to evaluate the decoding instruction as implemented by teachers in 'real' classrooms – and would have raised the further possibility of skewed findings due to experimenter bias.

### 3.2.2.4 Hawthorne control procedures

Many educational experiments explicitly address the possibility of Hawthorne effects by using an additional control group to assess the impact of one or more of the following variables from which this methodological artefact may derive (Adair, 2004): (a) special attention received by intervention participants from observers or researchers; (b) the novelty of the activities undertaken by intervention participants; and (c) participants' awareness that they are taking part in an experiment. For example, Kyriacou's (2008) evaluation of a programme of narrative writing instruction for primary pupils included both a 'no treatment' control group and a 'second treatment' control group, the latter following a programme of numeracy instruction whose contents were theoretically unrelated to the dependent variable.

However, the current study appeared to offer little grounds for suspecting any Hawthorne effect according to any of the three variables listed above. First, both intervention and comparison groups received equal attention from the researcher. This attention was limited to test administration at t1 and t2 (section 3.6) and two lesson observations (section 3.7); the programmes of instruction themselves were delivered entirely by the students' own class teachers during scheduled lesson time. This contrasts with Kyriacou's (2008) evaluation study: here, the fact that intervention but not comparison participants received small-group tuition from the researcher herself constituted a potential confounding variable.

Second, the researcher's experience of observing teachers and trainee teachers within the university's Initial Teacher Education (ITE) partnership suggests that there is nothing

unusual or novel in the *format* of the programme of instruction, which comprised short, form-focussed segments involving whole class teaching and pair-work (section 3.5). It therefore seems reasonable to assume that the only novel aspect of the programmes of instruction was their *content*: i.e. the explicit focus on L2 decoding.

Third, there is no reason to suppose that intervention participants were any more aware of participating in a research study than those in the comparison group. Rather than being marked out as a research-related activity, the intervention segments were treated as part of the classes' usual MFL teaching. Further, students' informed consent to participate in the study related only to the data collection procedures, in which all classes (intervention and comparison) participated equally.

In any case, Adair, Sharpe and Huynh's (1989) meta-analysis of those educational experiments which explicitly include Hawthorne control groups found no evidence for an overall Hawthorne effect, operationalized in terms of the three variables described above. In studies involving three-way comparisons between a treatment group, a no-treatment group and a Hawthorne control group, the "mean effect associated with Hawthorne manipulations was non-significant, and hence such groups essentially could be regarded as no different from a no-treatment control" (p.224); indeed, this pattern is exemplified in Kyricaou's (2008) study, mentioned above.

Nonetheless, Adair and colleagues caution against concluding that the Hawthorne effect does not exist; rather, they argue that a more likely source of the artefact lies in participants' "expectations of treatment outcomes or hypothesis awareness" (p.226). To guard against this phenomenon, attempts were made to discourage participants from

becoming aware of the focus of the research. First, as outlined above, efforts were made to keep the interventions separate from the data collection in participants' perception: the former were treated as part of pupils' routine MFL teaching, delivered by their own class teacher; the latter was conducted by the researcher and required consent. Further, the data collection included not only a Reading Aloud Test of decoding proficiency, but also several other pre- and post-test measures, including reading comprehension tests, vocabulary learning tasks and motivation questionnaires. (Some of these were originally planned as additional strands of the current study but are not included in this thesis for space reasons). This range of tests, it is argued, may have helped mask the true focus of the research from participants in both intervention and comparison groups.

#### 3.2.2.5 Planned random allocation of groups to conditions

Moore, Graham and Diamond (2003:679f) regret that a frequent research design in the educational context involves recruiting intervention classes or schools first and then seeking 'matched' groups, deemed similar in key respects, to form the comparison condition (e.g. Sylva et al., 1999). The problem with this design is that only the intervention groups, but not the comparison ones, have volunteered to participate in the intervention; this may reflect an underlying difference in ethos (e.g. in terms of openness to new ideas) which would be difficult to measure. A possible solution is to recruit all participating groups first and then to allocate them randomly to the intervention and comparison conditions. Initial plans were for the current study to adopt this model.

The sampling frame comprised local schools which – for practical reasons – lay within fifteen miles of the research base and were members of the university's ITE partnership,

on the basis that their already close working relationships with the university might make them more willing to take part in a project requiring sustained commitment from teachers. From the group of schools willing to participate, the intention was to (a) establish a 'pool' of matched schools, as similar as possible in terms of relevant characteristics, following a similar model to Graham and Macaro (2008); and (b) randomly select one of these matched schools to implement each of the intervention programmes and two further schools to form the comparison condition. A similar procedure was envisaged for selecting those teachers within the participating schools who would take part in the study.

As for the participating Y7 students themselves, it was anticipated that these would simply comprise whichever classes were assigned to those teachers selected for participation, with the proviso that all classes should (i) have similar weekly amounts of French tuition and (ii) be broadly similar in composition. In practice, this latter point was operationalized rather bluntly through a requirement that all participating classes should be of mixed sex and not setted or streamed according to previous attainment or aptitude scores[11]. Nonetheless, the characteristics of students in each class were documented via (a) a background information questionnaire, completed by the students themselves and (b) secondary data on their academic performance, supplied by the school. These arrangements would of course be contingent upon students' consent to participate in the project (section 3.8).

[11] 'Streaming' refers to grouping pupils according to their overall attainment, where the resulting groups operate across multiple subject areas. 'Setting' refers to grouping pupils according to their attainment in particular subject areas, where the resulting classes may differ from one subject to another. See Harlen and Malcolm (1999).

## 3.3 Plans refracted through the prism of reality

As indicated above, various practical obstacles arose when attempting to implement the planned research design. In some cases, major methodological changes were unavoidable. The most significant difficulty lay in recruiting schools to the project, a process which was eventually finalized only after the originally planned start date of the interventions. Only four MFL departments were both willing and suitable to participate, a number equal to the planned sample size. Further, not all of the original inclusion criteria were met, since school B operated a system of streaming in KS3, whereas Y7s were not setted in the other three schools. It was therefore decided to select, as the participating groups from school B, one (of two) lower stream classes and two (of three) higher-stream classes, with the aim of obtaining approximately the same range of overall prior attainment across the three classes as in the other schools. This decision was made (a) in consultation with school B's Head of MFL and (b) by comparing the academic attainment of the five available classes in school B with that of classes in the other schools, based on the limited set of comparable attainment data that was available (section 3.4).

A further compromise in terms of research design was that, following discussions with colleagues who had experience of working with these four schools (both on research projects and within the ITE partnership), it was decided to allocate schools to the intervention and comparison conditions purposively rather than randomly. Schools A and B were assigned to the intervention condition because their MFL departments had a track record of successful collaboration in university research projects. This made it more likely that teachers in these schools could be relied upon to implement the

programmes of decoding instruction systematically and faithfully, an important consideration given the sustained commitment required to deliver thirty-eight individual intervention segments (section 3.5). For this reason the advantages of randomly assigning groups to conditions were felt to be overridden. However, in terms of which intervention school (A or B) was assigned to which programme of instruction, this was decided at random with the following outcome: school A followed Programme A (the 'poem' approach); school B followed Programme B (the 'phonics' approach).

The non-random allocation of schools to conditions raises the possibility of pre-existing differences between the intervention and comparison schools, which would threaten validity. However, unlike in the model cited by Moore, Graham and Diamond (2003), all four schools had volunteered to take part in the project on equal terms: all were willing to participate in both the intervention and comparison conditions. Further, subsequent collaboration with schools C and D as part of the current study suggested that the precaution of purposive allocation had been unfounded.

The selection of participating teachers and classes was also highly constrained: the small size of some departments combined with unexpected timetabling arrangements and the fact that two teachers did not wish to take part removed almost all flexibility. In schools A and C, only three classes were available to fill the three required spaces; in schools B and D, four classes were available and one was discarded at random.

Finally, practical constraints disrupted the planned delayed post-tests (t3). One school (D) withdrew from this phase of the study, and the earliest mutually convenient point at which t3 tests could be scheduled in the other three schools was not until nine months

after the end of the intervention period. This was considered too late to provide valid delayed post test data, since participants had by this time received around six months' additional MFL teaching, in many cases in different teaching groups and with different teachers (thus introducing potentially confounding variables).

## 3.4 Describing the sample

Given the various constraints on selecting participants for the current study, it is important to document the characteristics of the final sample. This section therefore describes participating schools, teachers and classes and their relationship to the wider population.

### 3.4.1 Schools

The complexity of secondary schools defies the possibility of fully describing or even understanding them. Only an extended ethnographic approach – clearly beyond the scope of the current project – could get close to capturing the many nuances involved. However, in recent years, increasing amounts of more 'objective' data (including students' results in public examinations) have become publicly available. Though these data have been criticized for being subject to distortion and manipulation (e.g. Gillborn & Youdell, 2000), they do at least offer some common reference points for comparing schools, provided this is done with caution.

Two main sources of published data (accessed on line up to 18.03.2007) were used to formulate profiles of the four sample schools in the current study: (a) 'Performance Tables' provided by the (then) Department for Children, Schools and Families (DCSF) for the year 2006, the most recent year for which data was available; and (b) inspection reports published by Ofsted. Additional data was drawn from the schools' own websites and, in the case of information about the size and nature of their locations, from local government and town websites. Finally, information on the average national figures for English schools was retrieved to provide a point of comparison (DfES, 2006a). The complete set of data accessed, together with further information on the sources, appears in Appendix 5; a brief overview is provided below.

All four sample schools are mixed-sex, state comprehensive schools[12], accurately reflecting the majority of secondary schools in the national population. All include a 'Sixth Form' for students aged 16 to 18 years (beyond the age of compulsory schooling). The schools vary in size from over 2,000 students on roll (school A) to roughly 1,000 (schools B, C and D); by comparison, in 2006 most English secondary schools (44%) had between 1,000 and 1,500 students on roll across all age groups (DCC, 2009).

In terms of examinations results at the end of compulsory schooling (taken at age 16), schools A, C and D obtained broadly similar GCSE (or equivalent) grades, slightly below the England average. By contrast, the GCSE results achieved by students in school B were well below the national average. The schools' 'Contextual Value Added' scores – a published statistic argued by some (e.g. DCSF, 2007) but not others (e.g. Gorard, 2005) to provide a fairer indication of students' progress – suggest that students

[12] These are defined as "government funded, open-entry schools" (Macaro & Erler, 2008a:27).

in schools B and D make somewhat better progress than those in schools A and C, although the magnitude of the difference is small. This perhaps reflects the fact that, at the time of writing, the most recent Ofsted inspections produced similar judgments about the four schools' overall effectiveness: all were rated "good", the second of four possible judgments ranging from "excellent" to "unsatisfactory".

Turning to the broad characteristics of the schools' student populations, all four are below the national averages for England in terms of their numbers of (a) students with Special Educational Needs (SEN); (b) students from ethnic minority backgrounds; and (c) students with English as an Additional Language (EAL), whose L1 is not English. Schools A, C and D also have below average percentages of students eligible for free school meals – a widely-used though not uncontroversial proxy for family income (Hobbs & Vignoles, 2010) – whereas for school B the figure is about average. Indeed, according to school B's most recent Ofsted report, the socio-economic status of its student population is below the national average.

Together, these data suggest that the pupils in the sample schools do not fully represent the wider national population. This reflects the fact that the sample schools are not situated in large urban centres, where there are often higher proportions of students with SEN, EAL and low socio-economic status. However, there is no reason to suppose that the sample schools are untypical of other comprehensive schools in non-inner city areas. The low number of pupils in the sample with EAL can also be seen as advantageous for the current study, because beginner learners' L2 decoding is likely to be strongly influenced by their L1 literacy (section 2.6.1). It could therefore have been problematic

to compare the decoding outcomes of schools with different proportions of pupils with different L1s.

Finally, though sample schools were sought in which classes received similar amounts of French teaching, there was inevitably some variation in this respect. The total time spent in French lessons per week was approximately 130 minutes per class in school A, 120 minutes in school B, 100 minutes in school C and 158 minutes in school D. Participating teachers also felt that the timing of lessons could be important: for example, more productive learning might be expected on Monday mornings than Friday afternoons. However, controlling for such factors was beyond the scope of the current project.

The lower average socio-economic status and lower average examination grades of students in school B, as compared to those in schools A, C and D, potentially constitute a confounding variable when comparing the decoding outcomes of participants in the different conditions. On the other hand, the lower attainment data of school B means that, should participants in this school make more progress in L2 decoding than those in the comparison schools, this would strengthen the evidence for the effectiveness of the instruction they received.

### 3.4.2 Teachers

#### 3.4.2.1 Teacher Information Questionnaire

Since teacher effects represent a potentially confounding variable in the current study (because each class is taught by a different teacher), a 'Teacher Information

Questionnaire' (TIQ; Appendix 7.1) – originally intended as a tool for identifying a pool of 'matched' teachers – was used to document similarities and differences between participating teachers and their approaches to MFL teaching.

The TIQ had three sections. The first comprised factual questions about respondents' native language and number of years' teaching experience; the second (qualitative) and third (quantitative) sections investigated respondents' self-reported 'teaching style': i.e. their general approach to MFL teaching with Y7 classes. Of particular interest here was the amount of emphasis they placed on decoding. However, teachers' views on this topic were not sought explicitly, to avoid raising awareness of the focus of the research (especially in the comparison condition, which would have threatened validity). However, if respondents did routinely place emphasis on L2 decoding, they had chance to mention this under related topics such as reading and pronunciation.

As a means of investigating respondents' teaching style, the questionnaire has at least two potential drawbacks. First, respondents may be concerned to give the 'right' answer (particularly to a member of staff from their university ITE partner) rather than accurately reporting their own beliefs or practices. However, participating teachers were encouraged to be as honest as possible. Second, on the qualitative section of the TIQ, the broad scope and open-ended nature of the questions mean that differences between a given pair of responses may not necessarily reflect genuine differences between respondents: rather, it may simply be that they chose to emphasize different aspects of their practice within the broad area identified. This issue was magnified by the fact that respondents varied in the amount of detail given in their answers. However, there was

only one instance across all the completed questionnaires of an item being left blank[13]. Subsequent lesson observations (section 3.7) also provided some triangulation of TIQ responses.

The TIQ was piloted with two secondary school teachers who were not participants in the current study. This confirmed that the questions were clear and that the questionnaire did not take long to complete (assuming answers to the open-ended questions were kept brief). Teachers could choose to complete the TIQ either electronically or on paper. There was a 100% response rate.

#### 3.4.2.2 TIQ Findings

Table 2 summarizes teachers' responses to the factual questions in section 1 of the questionnaire[14]. Participants' levels of teaching experience varied widely, with some school-level patterns which would have been avoided if possible. On the other hand, the sample contains no Newly Qualified Teachers completing their first, probationary year in post (see DfEE, 1999). Further, all participating teachers with an L1 other than French were judged by the researcher to speak the language with similar, high levels of phonological accuracy; given additional time and resources, this might have been assessed more systematically.

[13] Respondent A2 did not respond to question 3, which is in any case strongly negatively related to question 2.
[14] To preserve anonymity, teachers are identified by the letter of their school (A-D) and a number (1-3).

**Table 3.1 Participating teachers' L1 and amount of teaching experience**

| Teacher | L1 | Years' teaching experience |
|---|---|---|
| A1 | English | 2 |
| A2 | English | 10 (including 3 part-time) |
| A3 | English | 26 (including 6 part-time) |
| B1 | French | 10 |
| B2 | English | 28 |
| B3 | English | 19 (including 7 part-time) |
| C1 | English | 2 |
| C2 | German | 3 |
| C3 | German | 3 |
| D1 | French | 6 |
| D2 | English | 2 |
| D3 | English | 4 |

Turning next to the third (quantitative) section of the TIQ – which asked teachers to indicate how often they included various kinds of activities in their lessons – responses showed a high level of agreement on most items. What might be seen as 'core' activities, including the four central skill areas of listening, speaking, reading and writing, tended to be used frequently by all respondents. On the whole, responses also showed reassuringly few school-level patterns, except perhaps that Assessment for Learning activities such as peer- and self-assessment (Black & Wiliam, 1998; Black & Jones, 2006) were reportedly more frequent in school D. In this section of the TIQ, there were also no strong differences between respondents in respect of those activities which are most obviously related to decoding proficiency, namely reading and speaking. Further, no respondent mentioned decoding when invited to comment on any other aspects of MFL teaching which they regularly emphasized.

Finally, on the second (qualitative) section of the questionnaire – which asked teachers open-ended questions about their views on various aspects of language teaching and about the kinds of activities they used most and least frequently – responses were again broadly similar. Few school-level patterns emerged, except for a tendency for teachers at school D to answer distinctively on some items: they reported greater use of both explicit grammar teaching and whole-class teaching.

References to decoding were made explicitly, or could be inferred, in a few of the responses to questions 7 and 8, which elicited teachers' views on teaching French pronunciation and reading respectively. (However, it should be noted that teachers in the intervention schools completed the TIQ after having learned of the focus of the programmes of instruction, which may have distorted their responses on this issue). On question 7, teacher A1 felt that "pronunciation (…) need[s] to be taught formally" because it "often differs from the written form"; B1 says she teaches students "the sound spelling link" in order to build their confidence and independence; and D1 "often refer[s] to pronunciation in lessons → last consonant often not pronounced → some groups of letters pronounced together". On question 8, whilst most responses emphasized the development of learners' comprehension strategies and their confidence when approaching written texts, there was at least one reference to reading as a vehicle for developing L2 decoding: teacher B2 stated that reading is "useful for embedding sound-spelling links". Teacher B3 may also have had decoding in mind when she noted that "we don't often read aloud in Year 7". It is interesting that most respondents, given free rein to comment on their priorities for L2 reading instruction, do not mention the issue of GPC, despite its prominence in recent official guidance for MFL teachers (section 2.7.3).

Since the references to decoding mentioned above are incidental, rather than being the result of direct questions on this topic, it is not possible to make firmer comparisons between the different teachers' views on this subject. However, the fact that some but not others mention explicit decoding instruction means that pre-existing differences within the sample in terms of teachers' approaches to L2 decoding cannot be ruled out. Notably, one of the three who reported engaging in explicit decoding instruction (teacher D1) is in a comparison school. This will be discussed further in section 3.7, which considers fidelity to condition in the intervention and comparison groups.

### 3.4.3 Classes

In terms of recruiting pupils to participate in the current study, a distinction was drawn between taking part in the programmes of decoding instruction on the one hand and completing tests as part of the data collection procedures on the other. Because the interventions were integrated into participating teachers' MFL lessons, all members of these teachers' classes (and members of comparison classes, who followed their usual MFL teaching) were involved *de facto*. This meant that the researcher had no control over the membership of these classes, which varied in size and composition (Table 3.2, Row 1). A subset of pupils in each teaching group (the majority) also participated in data collection. Background Information Questionnaire responses and secondary academic data were used to exclude from the sampling frames individuals who, on theoretical grounds, might threaten the validity of the findings: for example, students of francophone parentage or with certain specific learning difficulties. Students for whom personal or parental consent was withheld were also removed (section 3.8).

For the Reading Aloud Test (RAT) (section 3.6.3), the aim was to obtain longitudinal data by repeated testing of the same individuals. Participants in each class were drawn randomly to create a sequence which was followed when administering the tests. The achieved sample in each class depended on (a) how much lesson time was made available by the teacher at each round of testing (usually two or three whole lessons) and (b) whether facilities were available to test two participants simultaneously (each working individually in a separate location) or whether it was only possible to accommodate one at a time. At t2, the sample was reduced slightly as a result of withdrawn consent (N=4), absence from school (N=4), background noise obscuring recordings (N=3), poor behaviour (N=1) and administrative error by the researcher (N=1). School administration also dictated some movement of students between teaching groups; this mainly affected the 'lower stream' group B3, which therefore contributed only a small number of participants to the RAT sample. As shown in Table 3.2, Row 2, the overall number of participants who completed the RAT at both time points was 221, with an even gender balance (♀: N=110; ♂: N=111). Appendix 6.1 provides further details of the gender balance in each participating class and school.

For the task-concurrent self-report interviews conducted at t1 and t2 (section 3.6.4), a further sub-sample of participants was selected, comprising one high- and one low-attaining pupil from each participating class; these individuals were selected randomly from a pool of students nominated in each category by the respective class teacher (following Chamot and El-Dinary, 1999). Ideally, these teacher-nominated attainment levels would have been cross-referenced with participants' academic attainment data; however, these data were made available only after the beginning of data collection. The planned sample for the self-report interviews therefore comprised twenty-four

participants; it was felt that a sample larger than this would generate unmanageable quantities of data. Due to one case of withdrawal of consent at t2, two cases of absence during the interview period and one case of administrative error, the number of participants who completed the interviews at both time points was twenty (Table 3.2, Row 3). The gender balance was again roughly equal overall, though not at the level of individual classes or schools (Appendix 6.1).

**Table 3.2 Numbers of participants in each school and class who…**

| | **school A** | class A1 | class A1 | class A3 | **school B** | class B1 | class B2 | class B3 | **school C** | class C1 | class C2 | class C3 | **school D** | class D1 | class D2 | class D3 | **total** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| …participated in teaching | **84** | 28 | 29 | 27 | **93** | 31 | 32 | 30 | **75** | 25 | 25 | 25 | **63** | 23 | 22 | 18 | **315** |
| …completed RAT at t1 and t2 | **69** | 24 | 25 | 20 | **58** | 25 | 21 | 12 | **40** | 15 | 15 | 10 | **54** | 18 | 21 | 15 | **221** |
| …completed SRIs at t1 and t2 | **4** | 1 | 2 | 1 | **6** | 2 | 2 | 2 | **4** | 1 | 2 | 1 | **6** | 2 | 2 | 2 | **20** |

## 3.5 Intervention design

This section describes the two programmes of decoding instruction created for use in the current study: programme A ('poem' approach) and Programme B ('phonics' approach).

### 3.5.1 Content

To determine which French GPC would be included in the programmes of instruction, a list was first compiled of all GPC occurring in the language (Appendix 1.2), based on the "neutral form of pronunciation" embodied in "Supralocal French" (Coveney, 2001:3f),

the variety of the language most likely to be encountered by MFL learners. Since a manual search of the literature identified no existing inventory of French GPC, as comprehensive a list as possible was drawn up based on (a) the researcher's own knowledge of the language and (b) searches in a bilingual French-English dictionary (Atkins et al., 1998) and an on-line French rhyme dictionary (Barbéry, 2006). Each GPC in the resulting list was then considered individually as a candidate for inclusion in the decoding instruction, based on three principal factors: (a) the difficulty that it is likely to pose for L1 English learners, as indicated by a contrastive analysis of the English and French writing systems; (b) empirical evidence concerning its difficulty for learners in this population, as evidenced by the accuracy with which it was pronounced by participants in Woore (2006) (section 2.6.3); (c) the frequency of the GPC in written French, estimated on the basis of the researcher's knowledge of the language (no published source of this information having been identified). This was because knowledge of more frequently-encountered GPC would be more useful to learners than knowledge of infrequent ones. A detailed rationale for the inclusion or exclusion of each GPC appears in Appendix 1.2.

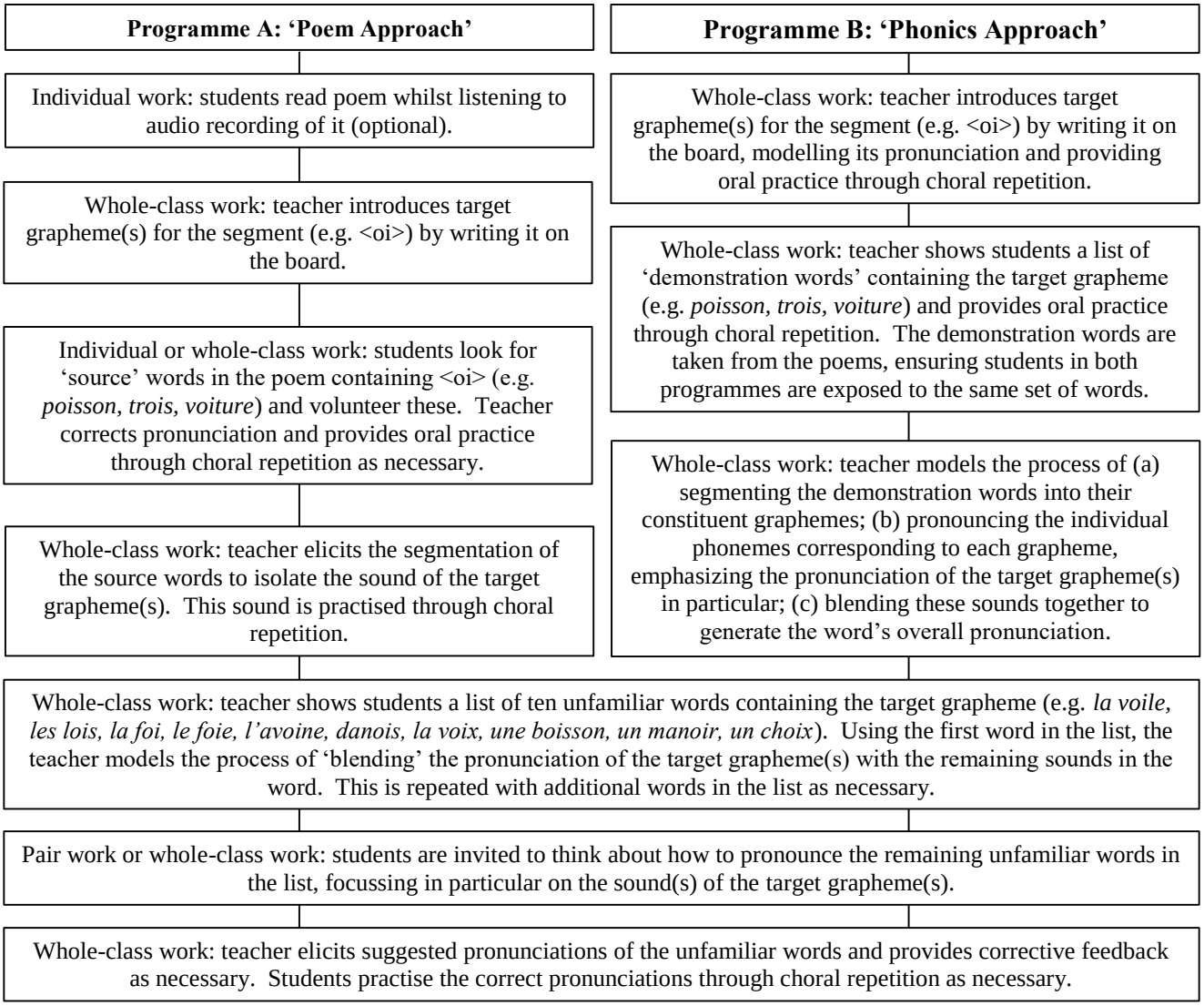
Overall, out of 122 French GPC, 55 are included in the decoding instruction, together with the Silent Final Consonant (SFC) 'rule'. Additionally, the instruction includes eleven 'grapheme groups' (Appendix 1.2.5). These were used to simplify the decoding instruction in cases where individual graphemes are realized differently according to their orthographic context: for example, <ai> is usually realized as /ɛ/ but is pronounced /a/ when followed by <ll>; rather than explain this rule, the grapheme group <aille> (with invariant pronunciation) was taught as a unit. Both programmes (A and B) covered the same GPC in the same sequence, to allow fair comparisons of their outcomes.

### 3.5.2 Format

The programmes of instruction were designed to be delivered in roughly ten- to fifteen-minute segments over the course of around thirty-eight lessons, spanning nineteen weeks; this was the maximum length of time available in the academic year, allowing for (a) the time-consuming process of pre- and post-testing and (b) the delays caused by sample recruitment problems (section 3.3). As in other decoding interventions evaluated in the literature (e.g. Gaskins et al., 1988 in L1 English; Liaw, 2003 and Yen, 2004 in EFL), participants were thus introduced to the target GPC gradually rather than in an intensive, potentially unmanageable block, and had repeated opportunities for practice. (In reality, the full set of thirty-eight segments was not completed by all classes: see section 3.7.2).

Figure 3.2 provides an overview of the planned activities in the 'core' segments in each programme (i.e. excluding those devoted solely to learning the poem in Programme A); further details and accompanying resources are included on the accompanying CD ROM in the form of the 'Teacher Packs' provided to all participating teachers in the intervention group (see section 3.7.1).

**Figure 3.2 Outline of 'core' segment plans for Programmes A and B**




Distilling the information in Figure 3.2, it can be said that, in both programmes, the principal components of each 'core' segment were as follows:

(a) The introduction of one or two (occasionally more) 'target' GPC. In Programme B, the target grapheme's written and spoken forms were initially presented by the teacher and exemplified in a list of sample words. In Programme A, participants

themselves searched the memorized poem for 'source words' containing the target grapheme and used these to work out the phoneme it represented.

(b) Opportunities for learners to work out the pronunciations of unfamiliar French words containing the target GPC, usually through paired discussion. Pairwork was considered to offer various benefits (cf. Pollard & James, 2004), not least the opportunity for all learners to engage actively with decoding and to try out pronunciations for themselves. In addition to the target grapheme(s), on which participants were instructed to focus in particular, the unfamiliar words contained various other graphemes which had either been covered previously or would be covered in future segments. This provided opportunities for repeated practice of various GPC, albeit to potentially varying degrees.

(c) The provision of explicit, form-focussed feedback by the teacher. Again this focussed on the target GPC, although by modelling the pronunciation of the words as a whole, the teacher inevitably also provided feedback on the additional GPC they contained.

The poem itself (in Programme A), which aimed to include at least one example of each of the target GPC, used a simple rhyme scheme to facilitate memorization. In a collaborative project with two sixth form students of French, who were taught at that time by the researcher, the poem was set to music (to increase learner motivation) and audio recorded so that it could be delivered identically in all intervention classes. Following Woore's (2004, 2007) suggestion that participants' use of the analogy strategy was undermined by inadequate knowledge of the poems' written forms (section 2.7.4),

the first six segments in Programme A were devoted entirely to memorizing the poem, using activities such as listening to an audio recording, simultaneously reading the text, reciting the text along with the recording and completing gap fill worksheets. Students also read the poem whilst listening to the audio recording at the beginning of subsequent segments. To keep the two programmes equal in length, an equivalent number of segments in Programme B were spent on revision activities, where participants recapped the GPC covered to date using various classroom games such as 'battleships', 'mystery pictures' and 'bingo'[15].

It is clear from the above descriptions that the two programmes have much in common. The key difference between them is the fact that, in Programme A only, participants are encouraged to refer to the 'source words' in the memorized poem as a mechanism for retrieving the pronunciations of target graphemes. By contrast, participants following Programme B rely on their teacher to tell them the pronunciations of the target graphemes and are expected to remember these as a series of individual GPC.

One aspect of the original rationale for Programme A was to develop learners' independent use of the analogy strategy, which involves deriving the pronunciations of unfamiliar words by referring to known source words with similar spellings. In retrospect, however, it is clear that – as in Woore (2004, 2007) – a more systematic approach to strategy instruction could have been adopted, for example by following a 'staged' approach such as that proposed by Macaro (2001) (see section 2.7.4). In the current study, whilst participants are encouraged to search the source words for instances of the target graphemes, they are not systematically supported in developing their

[15] Further details of all activities described in this paragraph can be found in the Teacher Packs (Appendix 1.1).

independent use of this strategy when reading. Further, irrespective of whether students participated successfully in each segment's initial phase of searching the poem for instances of the target grapheme, the teacher subsequently (a) confirmed for the class the pronunciation of this grapheme, (b) encouraged pupils to decode unfamiliar words containing it and (c) provided corrective feedback on pupils' pronunciations. These stages are identical to those in Programme B, arguably making the programmes far more similar than different. Participants following Programme A may therefore show less development in their use of the analogy strategy than might otherwise have been expected. On the other hand, since there is some evidence to suggest that learners from this population may use the analogy strategy spontaneously (Woore, 2010), the memorized source words in the poem may provide them with a more secure basis for doing so.

## 3.6 Main data collection

### 3.6.1 Introduction

This section describes the main data collection procedures, comprising (a) the collection of baseline information on participating students, (b) pre- and post-tests of French decoding using the Reading Aloud Test (RAT) and (c) task-based self-report interviews (SRIs). An additional strand of data collection – described separately in section 3.7 – was concerned with monitoring fidelity to condition in both intervention and comparison groups.

### 3.6.2 Baseline information on sample students

Background information on participants was sought for two reasons: (a) to contextualize their results on the decoding tests and (b) to exclude from analysis any individuals who might threaten the validity of between-group comparisons. This information was derived from three sources: school-held academic data (section 3.6.2.1); a Background Information Questionnaire (section 3.6.2.2); and an L2 reading comprehension test, which provided a partial measure of French proficiency (section 3.6.2.3).

#### 3.6.2.1 School-held data

Subject to consent from students, parents and headteachers (section 3.8), all sample schools provided access to the full set of academic data held in relation to participating students at the time of requesting this information (Autumn 2007). Unfortunately, this dataset included no specific information concerning attainment in French. National Curriculum Levels in the four MFL attainment targets (QCA, 2007b) – the only measures of MFL attainment at KS3 used on a national scale – had not yet been awarded; these are in any case based on teachers' subjective appraisal of students' work without any systematic process of moderating standards between or even within schools (Ofsted, 2011), resulting in data of questionable reliability. Neither had the sample schools received any systematic MFL attainment data from their feeder primary schools.

Between them, sample schools held data on twenty-three different attainment variables, including the following: (a) National Curriculum (NC) Levels obtained in statutory national assessments (commonly known as Standard Assessment Tests or 'SATs') in

English, Maths and Science, completed by all students at the end of KS2; (b) the specific Levels achieved on the reading and writing sub-components of the English SAT; (c) the points allocated to students on the basis of their SAT Levels; and (d) scores on the Cognitive Abilities Test (CAT), the most frequently used test of reasoning in the UK (Strand, 2004), completed at the start of Y7. The CAT comprises three batteries of sub-tests measuring verbal, quantitative and non-verbal reasoning (Deary et al., 2007).

Given the current study's focus on processing written words and learning new symbol-sound correspondences, the most relevant of these academic variables would appear, a priori, to be those relating to language and reading: (a) the NC Level obtained on the English SAT and more specifically on the reading component of this test; and (b) the verbal test score obtained on the CAT. The CAT scores have two advantages over the KS2 SAT results. First, they are more up-to-date (having been completed more recently). Second, as a scale variable, they are finer-grained than the blunt ordinal data provided by NC Levels.

An unexpected frustration was the lack of consistency between schools concerning the academic variables on which they held data, despite their being affiliated to the same Local Authority. CAT scores were missing for school B and SAT reading levels were missing for schools C and D. The only relevant variable on which all four schools held data was the National Curriculum Level obtained on the English SAT. This clearly restricts the range of comparisons that can be made. Further, there were a number of missing data in terms of both English SAT levels (forty-nine missing values) and CAT scores (around a dozen missing values); in the former case, this was because some

primary schools had not yet supplied the information to the secondary schools. However, the missing values were fairly evenly distributed between the schools.

It had been anticipated that students with particularly low CAT scores would be excluded from the sampling frame. Since they may have specific difficulties with L1 literacy, they may find it equally difficult to benefit from explicit instruction in L2 decoding. However, visual inspection of box-and-whisker plots of participants' verbal CAT scores and of their overall mean CAT scores revealed no outliers which were obvious candidates for exclusion. Since participants' English SAT levels fell into only three categories (Levels 3, 4 and 5), there were also no outliers on this variable.

#### 3.6.2.2 Background Information Questionnaire

To gather additional information about participants which might influence decoding outcomes, a Background Information Questionnaire ('BIQ': Appendix 6.2) was used to elicit their previous experiences of learning or using other languages, especially French. The BIQ, based on the version used successfully in Woore (2004) but expanded and refined, was piloted with a class of Y7 students in school A who were not participating in the current study. This resulted in a few minor changes to the wording and content.

Subject to participants' own and parental consent, the BIQ was administered by teachers at their convenience during MFL lessons in September 2007. Teachers gave full instructions on how to complete the questionnaire and supervised the filling-in process, answering any queries that arose. The small number of pupils who declined consent received French exercises from their teacher. If students were absent when classmates

completed the questionnaire, teachers tried to get them to complete it subsequently. Nonetheless, some slippage occurred compared to the original class lists; the final completion rate of the BIQ was 85.1% (N=280). Almost all submitted questionnaires were completed in full (i.e. almost no obligatory items were left unanswered).

On the basis of participants' BIQ responses, fourteen respondents – those of francophone parentage, previously domiciled in France or whose L1 was not English – were excluded from the data analysis. The use of a homogeneous sample in terms of participants' L1 background was preferred, in order to reduce confounding variables associated with the interaction of L1 and L2 literacy (section 2.6.1). Informal observations further suggested that the overwhelming majority of participants spoke with a similar (Southern English) accent.

#### 3.6.2.3 L2 proficiency measures

Given the absence of MFL-specific measures in participants' previous academic attainment data, baseline measures of various aspects of L2 proficiency would ideally have been administered. This would have allowed the effects of these variables on students' subsequent progress in L2 decoding to be controlled for. A measure of phonological awareness would also have been desirable, given the documented importance of phonological awareness in learning to decode in an L1 (section 2.7.2). However, the current study already involved a large burden of testing: beyond the instruments outlined in Figure 3.1, additional pre- and post-test measures of (a) reading comprehension, (b) vocabulary learning, (c) vocabulary learning strategies and (d) attitudes towards learning French were administered, originally planned as further

strands of the project but to be written up separately for space reasons. The implementation of additional tests was therefore considered impracticable.

To provide at least a partial baseline measure of L2 proficiency, however, participants' scores on the above-mentioned reading comprehension test at t1 were used. This measure has the drawback of not being totally independent of the study's target variable, since decoding proficiency may contribute to reading comprehension (section 2.3.1). However, comprehending written text clearly draws upon many other aspects of L2 knowledge and strategic behaviour, thereby offering a wider picture of proficiency in the language. Further, previous research has found learners in this population to have negligible L2-specific decoding proficiency (section 2.6.3), suggesting that this variable was unlikely to make much contribution to participants' reading comprehension scores.

In the absence of any standardized test of L2 reading comprehension suitable for this population of beginner learners, the t1 reading comprehension test (Appendix 4.1) was based on a text devised by Macaro and Erler (2008a), expanded slightly in length and complexity on the advice of participating teachers. The text was 111 words long; the test comprised three separate tasks, designed (as part of the planned additional strand of the project) to assess different aspects of reading comprehension. In the first stage, following the reading comprehension task used by Macaro and Mutton (2009), learners read the whole passage and wrote out in English anything they had understood. In the second and third stages, they answered a series of questions (in L1) relating to the first and second paragraphs of the text respectively. However, since the reading comprehension test in the current study served only to provide contextualizing

information on participants' baseline L2 proficiency, the scores on the three sections of the text were amalgamated to give a single, overall score.

Initial versions of the test were piloted with two Y7 classes from school A (N=55 in total) who were not part of the main sample. The test itself was then administered to all classes at a convenient time during the first two weeks after the half-term break in November 2007, i.e. after seven or eight weeks of secondary MFL lessons. Instructions were given by the researcher in person following a standardized script, allowing time for questions. The researcher and class teacher then worked together to ensure that the administration of the test proceeded smoothly and that students worked on their own in silence. For the few students absent on the day of the test, alternative arrangements were made; therefore, of the 221 participants who completed the RAT at t1 and t2, only twelve did not also complete the reading comprehension test. The missing cases were relatively evenly distributed between the intervention and comparison groups.

To ensure reliability in marking the reading comprehension tests, a detailed mark scheme was prepared (Appendix 4.2). The mark scheme was continuously updated as new answers were encountered during the marking process. Once the first round of marking was complete, all scripts were then reviewed in the light of the final version of the mark scheme to ensure consistency, with occasional changes being made. Inter- and intra-rater reliability figures were considered unnecessary, since all attested responses were explicitly included in the finalized mark scheme, removing the need for subjective judgments at this stage.

### 3.6.3 The Reading Aloud Test

Measuring L2 decoding proficiency in order to monitor its development is fundamental to the current study. However, a manual search of the literature revealed no standardized test of French decoding suitable for use with beginner learners in the English MFL context. The study therefore used a version of the Reading Aloud Test (RAT) first developed by Woore (2004) and subsequently extended, refined and evaluated (Woore, 2006). This section reviews (a) the development and format of the RAT (section 3.6.3.1); (b) its administration in the current study (section 3.6.3.2); (c) the processing and scoring of participants' RAT output (section 3.6.3.3); and (d) issues of validity and reliability (sections 3.6.3.4-5). For a more detailed account of the development and piloting of the RAT, see Woore (2006).

#### 3.6.3.1 Format of the RAT

The RAT involves reading aloud fifty-three real French words presented individually via *Powerpoint* in one of four different sequences, to control for ordering effects (Appendix 2.1). The computerized presentation – following the example of Sprenger-Charolles, Siegel and Bonnet's (1998) word reading test for young L1 French learners – was designed to be more engaging than the printed word lists used in Woore (2004). For clarity (since misperception of graphemes would undermine accurate decoding), university guidelines for the accessible presentation of written material were consulted (OUFMML, 2010): each word therefore appears alone in the centre of the screen in large, bold, black, sans serif, lower case type against a pale blue background. (For sample screen shots, see Appendix 2.2). Following comments made during piloting

(Woore, 2006), a small counter in the bottom right-hand corner of the screen indicates the number of items remaining.

Participants progressed through the RAT at their own pace, pressing the space bar to move on to the next item; this posed only very occasional problems when participants accidentally pressed the space bar twice and skipped an item. Robert Vanderplank (personal communication, 06.11.2007) questioned why participants were allowed unlimited time for each item, given the importance of *rapid* word recognition for reading comprehension (section 2.3.1). An alternative would have been to display each word for a fixed time interval. However, the view taken in this study is that conscious processing may be required in order to override the automatic L1-based processing of L2 words, providing a stepping-stone *en route* to automatic L2 decoding (section 2.6.1). If so, imposing a time limit may simply truncate the conscious processing before its benefits are obtained. It might also add to the stress potentially associated with L2 oral production tests (Horwitz, Horwitz and Cope, 1986), which might in turn negatively affect processes and outcomes. In short, what is being investigated here is not necessarily automatized print-to-sound decoding, but rather the prerequisite ability of learners to make accurate print-to-sound correspondences in the first place.

***Content***

For reasons discussed below (section 3.6.3.4), the RAT contained words unlikely to be known by KS3 learners, as judged on the basis of (a) the researcher's teaching experience with this age group and (b) consultation with participating teachers. For participants in the original piloting of the test, this was also confirmed through semi-

structured interviews (see Woore, 2006). Where possible, words which – in whole or part – closely resembled English words were avoided, in order to discourage automatic L1-based processing; this was because resemblance to English was identified by Erler (2002) as a factor affecting L2 French decoding. Word length (both orthographic and phonological) was also controlled for: all RAT words are bisyllabic[16]; all contain between six and eight letters; and almost all comprise four to six graphemes (two have seven). The use of bisyllables rather than monosyllables – chosen in order to reduce the length of the test whilst still including all the target graphemes (see below) – introduces some complications for anglophone learners, because of the tendency in English to reduce unstressed vowels to schwa (Hawkins, 1984). On the other hand, since this is a feature of English but not French pronunciation, the bisyllabicity of RAT items provides an opportunity to measure participants' non-application of the vowel reduction principle in their L2, though this is not specifically targeted in the decoding instruction programmes.

The original RAT (Woore, 2004) was designed to include at least three 'tokens' (or examples) of each of fifty-one GPC 'types'. This principle is weakened slightly in the current study: following the exclusion of some GPC at the scoring stage, together with the merging of some consecutive graphemes into unitary 'grapheme groups' (section 3.6.3.3), some GPC are represented by only two tokens and some grapheme groups occur only once. The original intention of having multiple tokens of each GPC type was to reduce the potential effects of 'idiosyncratic' realizations of individual items: these would be 'averaged out' at the scoring stage to indicate how accurately a given grapheme type was 'generally' pronounced by each participant. However, there are

[16] An exception is *caleçon*, which may also be realized trisyllabically in careful speech.

problems with this argument when applied to L1 English decoders. Since they may be expected to show incorrect contextual variation in their realizations of individual graphemes, especially vowels (section 2.6.1), it could be argued that it is inappropriate to talk about their 'average' pronunciation of a given grapheme type across different orthographic contexts. On the other hand, since this contextual variation in grapheme realization is less frequent in French than in English, learners who make progress in French decoding might be expected to show increasing consistency in their pronunciation of, say, the grapheme <a> across its various contexts. In this sense, calculating an average score for the GPC <a>→/a/ remains meaningful.

The smallest identifiable set of French words which met the various criteria described above was large, comprising fifty-three individual items. A shorter test could have been achieved by using pseudowords rather than real words, but this was considered undesirable for ethical and pedagogical reasons. There are also implications for test validity: for example, Papagno, Valentine and Baddeley (1991:342) found that participants performed less well at learning pseudowords than real words (even of a language they were "never likely to use"), ostensibly because they were less motivated by "learning nonsense". In any case, word naming tests of similar length to the RAT feature in various other studies, not only with older participants (e.g. Koda, 1998) but also with younger ones (e.g. Sprenger-Charolles, Siegel and Bonnet, 1998; Frith, Wimmer and Landerl, 1998). Moreover, piloting with Y7 pupils suggested that the length was acceptable (Woore, 2006).

Some participating teachers also felt that some RAT items, being very low frequency words, were "too difficult" for Year 7 participants. However, the issue of frequency is

separate from that of orthographic complexity. It is hard to argue that any of the test items are more difficult to decode than many words which Y7s routinely encounter in their early MFL lessons, such as *beaucoup* (/boku/, 'a lot') and *qu'est-ce que c'est?* (/kɛskəsɛ/, 'what's this?').

#### 3.6.3.2 Administration

The RAT was administered individually in a location which was as quiet and private as possible (within the constraints of the secondary school environment), such as an empty office, classroom or corridor. Instructions were presented on screen (Appendix 2.2) and simultaneously explained (and elaborated on) by the researcher; participants then had an opportunity to ask questions. In particular, the researcher wished to ensure that participants did not think that the RAT items were simply unknown *English* words. Participants were also reassured that (a) the results of the test would be anonymous and would not affect their school grades in any way; (b) they were not expected to *know* the RAT items, but should simply read them aloud as they thought they should sound; and (c) they would not receive feedback either during or after the test, but this did not imply that their pronunciations were either 'right' or 'wrong'. When participants were ready, they completed three practice items in the researcher's presence to check that they had correctly understood all instructions. To minimize the potential for distraction or anxiety, they were then left alone to complete the test, except for one further visit from the researcher after about five items to check that all was in order and that they were happy to continue (which they all were). Completion of the RAT was followed by a short 'de-brief', in which participants were thanked for their participation.

All tests were audio-recorded, initially using an analogue cassette recorder and desktop stand microphone but later (when these became available to the researcher) using portable digital recorders with built-in microphone. Both technologies yielded recordings of similar quality, but the digital option was preferred when possible, because of its greater convenience. Analogue recordings were subsequently digitized. All digital audio files were then 'tidied up': the delivery of instructions was removed and participants' names were replaced with identification numbers, in order to preserve anonymity and to ensure unbiased processing of the data.

At each round of testing (t1 and t2), the same version of the RAT was used. The possibility of a practice effect was considered negligible, given the presumed unfamiliarity of the words and the considerable time delay between t1 (December 2007) and t2 (June 2008); any practice effect would in any case apply equally to all groups. A slightly greater risk of a practice effect exists for the sub-sample of participants who completed the self-report discussions (section 3.6.4); however, as they were few in number (N=20) and drawn equally from all classes, the threat to validity was again considered minimal.

#### 3.6.3.3 Processing RAT data

Each participant's RAT output was (a) transcribed and (b) assigned two sets of scores on the basis of these transcriptions.

***Transcription***

Participants' pronunciations of RAT items were transcribed at the level of individual grapheme tokens. Clearly, the notion that their pronunciation of each word could be neatly compartmentalized into individual phonemes, each mapping onto a specific grapheme, is a simplification: processes such as phonetic assimilation cross phonemic boundaries (Kenstowicz 1994; Ladefoged 1993). In any case, participants did not always decode words in a simple linear manner (see below). Nonetheless, in previous studies using the RAT, most participants did decode most items in such a way that individual phonemes in their spoken output could be unambiguously mapped onto the corresponding graphemes in the words' written forms.

Not all of the 268 GPC tokens included in the RAT were transcribed as part of the current study. First, as when designing the programmes of instruction, some GPC types were omitted because of their similar realizations in French and English (e.g. <m>, <b>, <f>); previous research had found that participants from the same population pronounced these graphemes accurately over 90% of the time (section 2.6.3). Second, in the course of processing the data, certain individual GPC tokens were found to be problematic due to contextual assimilation effects. For example, participants tended to nasalize the <on> in *Hongrie* under the influence of the following <g> (as observed in English words such as *long*). Since this may lead to an over-estimation of the degree to which participants decoded <on> as a nasal vowel, this token (together with other similar cases) was omitted. Third, it was found impossible to transcribe some GPC tokens as isolated units due to their orthographic context. For example, the final three letters of *ralingue* can be interpreted as comprising two graphemes: <gu> and silent <e>. Participants in the

current study frequently pronounced this group of letters as /gjʉ/. However, when transcribing this pronunciation, it is impossible to choose between the following possible grapheme-to-phoneme mappings:

(a) <gu> → /g/; <e> → /jʉ/

(b) <gu> → /gj/; <e> → /ʉ/.

Moreover, we cannot assume that participants correctly segment the letter group into these two graphemes at all: they could equally assume that the constituent graphemes are {<g> + <ue>}, {<g> + <u> + <e>} or even simply {<gue>}. Without knowing something about the participants' processing mechanisms, it is therefore logically impossible to segment their spoken output reliably in terms of the GPC which led to its production: we cannot infer the process from the product. Participants' pronunciations of <gue> – and of nine other groups of grapheme tokens which posed similar problems – were therefore transcribed and scored as a single units, labelled 'grapheme groups'. (Note that these grapheme groups were different from those used in the programmes of instruction). Taking these issues into account, the pronunciations of roughly two thirds (174/268) of the GPC tokens in the RAT were transcribed, representing 48 GPC types (Appendix 2.3): eight nasal vowels; seventeen oral vowels; thirteen consonants (including the SFC 'rule'); and ten grapheme groups[17].

Transcriptions were based on the phonetic alphabet of the International Phonetic Association (IPA, 2005; Appendix 2.4). To ensure that the symbols (particularly those representing vowels) were used consistently, reference websites were consulted

[17] The mapping of these grapheme groups onto the sounds they represent can be designated using the plural formulation '*graphemes-phonemes* correspondences' (rather than the singular grapheme-phoneme correspondences). The abbreviation GPC can therefore be retained, but with an expanded interpretation.

(Appendix 2.4); these offered audio-recorded exemplifications of (a) IPA symbols and (b) British English, American English and French vowels. Example words in the researcher's own variety of English were also used as additional points of reference. Further, in order to facilitate transcription of the data, customized versions of the IPA vowel chart – one for monophthongs, one for diphthongs – were created (Appendix 2.5), containing those symbols used most frequently when transcribing participants' RAT output and accompanied by sample words exemplifying these sounds.

It should be noted that some symbols appearing in the charts in Appendices 2.4 and 2.5 may be found elsewhere with different phonetic values. This is because (a) some symbols have language-specific currencies and (b) there are sometimes disagreements between different authorities over the representation of particular sounds. For example, Atkins et al. (1998) use /ɔ/ and /ɔː/ to represent the vowels in French *sonner* (/sɔne/, 'to ring') and English *sauna* (/sɔːnə/) respectively; however, the difference between these two vowels is more than simply a matter of length, as the addition of the IPA lengthening diacritic [ː] would suggest. Similarly, the IPA table uses /ʌ/ to represent the unrounded, half-open, back cardinal vowel; but (as noted by Hawkins, 1984) this is also the standard representation of the vowel in English words like *cup*, which in many British varieties is closer to the IPA's /ɐ/. As a final example, Atkins et al. (1998) represent the nasal vowels in *grand* and *vin* as /ɑ̃/ and /ɛ̃/ respectively, whereas Coveney's (2001) meticulous articulatory investigation suggests that /ɒ̃/ and /æ̃/ would be more appropriate. This made it all the more important that the symbols in Appendices 2.4-5 were used consistently.

Participants' realizations of all relevant graphemes were recorded in a table (Appendix 2.7). Each distinct realization across the sample as a whole was then assigned a unique identification number, used to log the pronunciations in SPSS 15.0. This allowed patterns in the pronunciations of specific graphemes to be identified and comparisons to be drawn between (a) t1 and t2 and (b) participants in the different groups.

Several limitations to the transcription process must be acknowledged:

(1) There is an unavoidable tension inherent in using an IPA-based alphabet – which classifies sounds according to their place and manner of articulation – when transcribing audio files, since these lack the visual cues (e.g. the speaker's mouth movements) which have been shown to play an important role in speech perception (see McGurk & MacDonald, 1976; Massaro & Cohen 1990).

(2) The audio quality imposed some constraints on the level of detail that could be transcribed. For example, English voiceless stops characteristically have a later voice onset time than their French counterparts, represented in a narrow transcription as English [$p^h$ $t^h$ $k^h$] versus French [pˀ tˀ kˀ] (Kenstowicz, 1994). It would have been interesting to see whether participants' voice onset time changed over the course of the study; however, this distinction was too fine to hear accurately and consistently.

(3) Despite attempts to insulate the RAT sessions from unwanted noise, participants' output was occasionally obscured by the inevitable cacophony of the secondary school environment: bells ringing, a class filing past, doors slamming, pupils arguing or being

reprimanded. However, pronunciations were lost to background noise only rarely (0.7% of cases at t1; 0.6% at t2).

(4) As noted above, participants did not always pronounce RAT items in a way that enabled straightforward decisions about how the phonemes in the spoken output should be mapped onto the words' constituent graphemes. Examples are given in Table 3.3: as exemplified by the attested pronunciations in Row 2, at a descriptive level, phonemes were sometimes duplicated (examples A-B), omitted (C), added (D-E) or displaced (F).

**Table 3.3 Problematic pronunciations of RAT items and illustrations of the Linear Mapping Principle.**

| | | A | B | C |
|---|---|---|---|---|
| 1 | RAT item | gueulez | poitrine | chacun |
| 2 | attested pronunciation | gl**ʉ**lɛz | pɔɪntɹɪn | ʃan |
| 3 | segmented graphemes | gu eu l ez | p oi t r i n e | ch a c un |
| 4 | grapheme-phoneme mapping | gl **ʉ** l ɛz | p ɔɪ nt ɹ ɪ n - | **NLD** |

| | | D | E | F |
|---|---|---|---|---|
| 1 | RAT item | piochais | embruns | acanthe |
| 2 | attested pronunciation | pɪnətʃoʊs | ɛmbɹʊnəs | akaΘən |
| 3 | segmented graphemes | p i o ch ai s | em b r un s | a c an th e |
| 4 | grapheme-phoneme mapping | **NLD** | **NLD** | **NLD** |

Such cases posed serious obstacles to transcribing and scoring the pronunciations on a grapheme-by-grapheme basis; again, the underlying problem is that we hear only the product of the decoding but have no insight into the processes involved. For example, should the extraneous /n/ in example B be coded as part of the following consonant or as part of the preceding vowel? Or again, in example C, how should the missing /k/ be treated? The pronunciation may conceivably reflect the participant's belief that the <c> is 'silent' (like an SFC), but if we posit a GPC of <c> → / - / (indicating no phonological

realization), we are then left with an arbitrary choice between two options for coding the vowel phoneme:

1. {<a> → /a/} + {<un> → / n /}
2. {<a> → / - /} + {<un> → / an /}

The opposite problem arises in example D where an extra syllable occurs. When attempting a phoneme-to-grapheme mapping of this output, it would appear reasonable to interpret the /ɪnə/ as corresponding to the <io>, perhaps resulting from perceptual error; the remaining sounds seem to map straightforwardly (though in two cases incorrectly) onto the other constituent graphemes: <p> → /p/; <ch> → /tʃ/; <ai> → /oʊ/; <s> → /s/. However, this intuitively plausible interpretation rests upon a subjective assumption about the most *likely* phoneme-to-grapheme mapping for this pronunciation. Since the purpose of the current study is, precisely, to investigate how particular graphemes are decoded, transcribing grapheme-to-phoneme mappings based on what seems 'likely' would mean pre-judging that very issue. Thus, in example D, it is *logically* possible (however implausible) that the phoneme-to-grapheme mapping could proceed in various other ways, such as the following: <p> → /p/; <io> → /ɪ/; <ch> → /n/; <ai> → /ə/; <s> → /tʃoʊs/.

To avoid relying on subjective, *ad hoc* decisions about the most likely decoding outcomes in such cases, a solution was adopted which could be applied more objectively to the transcription of 'problem' pronunciations. This was termed the Linear Mapping Principle (LMP). The pronunciation of each RAT item was first segmented into constituent phonemes which were then processed sequentially, working from left to right

through the word's written form. Any number of consecutive consonant phonemes were mapped onto any number of consecutive consonant graphemes, and any number of consecutive vowel phonemes onto any number of consecutive vowel graphemes, beginning with whichever of the two (consonant or vowel) occupied the initial position in the word's written form. Where this process broke down (effectively because the phonological output contained too many or too few syllables), a grapheme-by-grapheme transcription of the word was not attempted: rather, a separate code ('999') was assigned to all the word's constituent graphemes to indicate 'Non-Linear Decoding' (NLD). Rows 3 and 4 in Table 3.3 illustrate the outcome of the LMP in respect of the problematic words discussed above.

The LMP is not a perfect solution. It suffers from at least four problems:

(1) It occasionally results in arbitrary decisions about the grapheme under which a given phoneme should be transcribed. In example B in Table 3.3 (*poitrine*), the mapping of the epenthetic /n/ to the grapheme <t> rather than to the vowel <oi> is arbitrary. In fact, informal observations suggested that the additional /n/ arose frequently in the case of certain vowel/consonant *combinations* (such as <oit>); however, this cannot be captured within the RAT's segmented transcription framework.

(2) Even if the violation of the LMP affects only part of a word, all of its constituent graphemes are excluded from the transcription. This is illustrated by example D in Table 3.3, where (as argued above) it is highly probable that the final syllable of the pronunciation (/tʃoʊs/) corresponds to the final three graphemes of the written form

(<chais>). If so, then the data on how this participant decodes these graphemes is lost, even though this may reflect a wider pattern.

(3) Some other potential insights into the decoding process are lost. For example, informal observations suggest that the epenthetic schwa in example E in Table 3.3 may reflect a wider pattern in the data, which may result from participants consciously 'sounding out' graphemes (as in /kə - a - tə/ to make English *cat*). Likewise, there are glimpses of a possible tendency by some participants to use French alphabetic letter names when sounding out words: for example, one realization of *huilage* as /aʃɥadʒ/ begins with the French name for 'H' (/aʃ/). This will be seen to echo a finding of the self-report interviews (chapter 6). Again, however, the possibility of collating and analyzing these occurrences is denied by applying the 'NLD' code to all the graphemes in these words.

(4) Conversely, apparently 'accidental' realizations *are* recorded in the data. This occurs in example A in Table 3.3, where the pronunciation of the initial <g> is transcribed as /gl/ even though this seems more likely to reflect a perceptual error than a 'real' belief that this is how <g> is pronounced.

Despite these limitations, the LMP was adopted as the basis for transcription. Its problems were felt to be outweighed by the fact that it avoids the need for subjective (and hence potentially inconsistent) decisions about which phonemes should be matched onto which graphemes. Further, the limitations of the LMP affect all participants equally.

***Scoring***

The second stage of processing participants' RAT output involved converting the transcribed pronunciations into scores. Section 2.6.4 suggested that developing accurate L2 decoding may include both (a) learning to apply L2-specific GPC and (b) developing more accurate phonological representations of L2 phonemes. An attempt was made to reflect this when scoring the RAT, moving beyond the simple, unitary scoring system of previous studies (e.g. Woore, 2006, 2009). Two distinct scores were therefore generated for vowel and nasal graphemes (but not for consonants, whose French and English pronunciations are generally more similar save for small differences such as degree of aspiration, to which the transcription was not sensitive):

(1) The first score ('score1') was derived from a stringent mark scheme, which awarded one point for every 'correct' realization of a grapheme, interpreted as being of native-like quality. For the purposes of scoring, 'native-like' was benchmarked against the 'supralocal' variety of French documented by Coveney (2001) (see section 3.5.1).

(2) The second score ('score2') derived from a more lenient mark scheme, awarding one point for every 'acceptable' realization of a grapheme. This was defined as being either (a) 'correct' or (b) interpretable as an anglicized pronunciation of the correct French phoneme. For example, under the more lenient mark scheme, participants received a point for the <eau> in *moineau* if they realized it either (a) as /o/, the correct form; or (b) as the diphthong /əʊ/ (and variant forms such as /ɐʊ/, /əʊ/ or /oʊ/), which in the researcher's experience are frequent, anglicized realizations of /o/ by MFL learners. In the case of nasal vowels, absent from the English phonemic inventory, the lenient mark

scheme drew upon Čubrović's (2002) concept of 'unpacking' (section 2.6.4): that is, credit was given for those combinations of oral vowel phoneme plus nasal consonant phoneme which often serve as anglicized equivalents of these sounds. For example, when scoring pronunciations of <an>, a point was awarded not only for the correct realization (/ɒ̃/) but also for the more 'English-sounding' /ɒn/. By contrast, no point was awarded for /an/, which would be generated by the application of English GPC.

Appendix 2.6 lists all pronunciations accepted under the stringent and lenient markschemes.

One criticism of this approach (where graphemes within each word are scored individually) is that it does not reflect naturalistic speech processing, where problems with the perception of individual phonemes may be compensated for by drawing upon the linguistic and non-linguistic context. Further, weighting all GPC equally in the scoring system fails to capture the differential effects on intelligibility that may be caused by mispronouncing different elements of a word. These arguments seem to favour a mark scheme based on a holistic judgment of pronunciation accuracy, similar to those used in various research studies (e.g. Bongaerts, 1999) and in national MFL examinations in England (e.g. AQA, 2008). However, whilst we may not necessarily comprehend spoken input one phoneme at a time, the alphabetic principle does involve the mapping of individual graphemes to phonemes in linear sequence (section 2.4.1). A mark scheme operating at the level of individual GPC therefore provides richer information than one based on holistic judgments: while the latter indicates to what extent a native-like pronunciation of a word has been achieved overall, the former is able to provide specific information about which GPC within the word were pronounced in a

non-native-like way. From a research perspective, this allows the RAT to provide empirical evidence concerning (a) which GPC tend to pose particular problems for learners (as in Woore, 2006) and (b) whether specific GPC have been more susceptible than others to improvement as a result of the instructional programmes.

#### 3.6.3.4 Validity

Woore (2006) provides a detailed evaluation of the validity of the RAT; the key arguments are summarized below.

First, the use of (presumably) unknown words makes the RAT effectively equivalent to pseudo-word reading tests, used in many empirical studies both in L1 (e.g. Johnston & Watson, 2005) and in L2 (Koda, 1998). Since readers have no pre-stored phonological form for unknown words and no sight-recognition template, sub-lexical decoding provides the only way to generate a pronunciation for them. Further, the use of individual, decontextualized words prevents interference from top-down word recognition processes or strategies (section 2.3.1). Assuming that the RAT items were indeed unknown to participants – which is highly likely, although not empirically proven – this gives the RAT good face validity as a measure of decoding proficiency as defined in this thesis.

On the other hand, reading aloud tests can be criticized as inherently impure measures of grapheme-to-phoneme conversion, because the outcomes are based on participants' spoken output: whilst incorrect pronunciations *may* reflect incorrect underlying phonemic representations, they could also arise through articulatory problems. A

prototype test intended to address this issue required participants to select, from a range of recorded pronunciations, which they thought was the correct realization of a given grapheme. However, tests following this model are based on participants' *perception* of L2 sounds and therefore fail to operationalize decoding as a *productive* skill. Further, offering participants a set of predetermined alternatives from which to select the correct pronunciation of a grapheme artificially constrains their decoding choices: what if the sound they would have produced does not feature amongst the available options?

Other possible alternatives to the RAT considered for the current study were Erler's (2007, 2008) pen-and-paper tests of rhyme judgment and syllable segmentation (section 2.6.3). However, Woore's (2006) empirical evaluation of these tests argued that, though they offered certain practical advantages such as greater ease and speed of administration and scoring, they were less valid than the RAT as measures of decoding, at least as defined in the current study (section 2.6.3).

#### 3.6.3.5 Reliability

On the grounds of both ethics and research design, test/re-test reliability calculations were considered unfeasible: asking participants to complete the RAT again would have involved considerable additional disruption to lessons and would have increased the risk of a practice effect. However, it was considered essential to demonstrate the reliability of transcribing and scoring, given the inevitable role of subjective judgments in this process. Both intra- and inter-rater reliability measures were therefore calculated.

For the intra-rating, a five per cent sub-sample of participants' RAT outputs at both t1 and t2 was selected via an on-line random number generator (www.random.org/integers) and was transcribed a second time by the researcher without reference to the original transcriptions. This gave second transcriptions of 22 participants' RAT outputs, comprising 4,092 grapheme tokens. The original and second transcriptions of each grapheme token were compared and the percentage of exact agreements calculated. In the case of vowels in particular, given that the transcription involved representing sounds which could fall anywhere within a three-dimensional space defined by the continua of height, backness and lip-rounding (see Appendix 2.4), some degree of discrepancy was considered inevitable. An 'approximate agreement' figure was therefore also calculated. This was the percentage of both exact matches and 'near-matches', the latter occurring when (a) the two different transcriptions represented sounds which were judged (subjectively) to be very similar, such as /ə/ and /ɐ/; or (b) the difference between the two transcriptions was attributable to the length rather than the quality of the vowel involved (as exemplified by the second phoneme in /gwɪn/ and /guɪn/).

Despite having "unique common sense value" (Uebersax, 2010), raw agreement statistics of this kind have been criticized for not taking into account the fact that "a certain amount of agreement [between two ratings] is to be expected by chance" (Cohen, 1960:38). However, in the current study, the likelihood of obtaining an identical transcription of a given grapheme token *by chance* is negligibly small, due to the extremely large set of phonetic symbols from which to choose. For this reason, the calculation of Cohen's kappa was considered inappropriate.

Intra-rater agreement figures were also calculated for the RAT scores. Since scoring was not completed directly from participants' RAT output, but rather on the basis of the transcribed data (i.e. each transcribed realization was mechanically assigned a score of either 1 or 0 according to the stipulations of the mark scheme), there was again no likelihood of agreement arising by chance. It is therefore argued that percentage agreement figures offer an appropriate indication of intra-rater reliability.

Incidentally, it should be noted that the percentage agreement figures for the RAT scores must be at least as high as, though not necessarily identical to, the percentage agreement figures for the transcriptions. This is because, if a given sound is transcribed identically on the two occasions, the score generated on the basis of these transcriptions will necessarily be the same as well; by contrast, in some cases the transcriptions may differ yet the scores remain the same. For example, /e/ and /ɛ/ are both designated as 'correct' realizations of the grapheme <ai>; they would therefore constitute 'disagreement' at the level of transcription but 'agreement' in terms of the scoring.

To calculate inter-rater reliability, a colleague was recruited with many years' experience of working with MFL learners as a teacher, teacher educator and researcher. The sub-sample of RAT recordings used for inter-rating, which took place on two separate occasions, was inevitably much smaller than that used for intra-rating, due to the extremely time-consuming nature of the transcription process; further, considerable time had to be devoted to training the inter-rater in the use of the IPA charts (Appendix 2.4-5). The complete RAT outputs of one participant per school at each time point were therefore used for inter-rating (eight sets of data all told), although one had to be excluded through administrative error. During the first inter-rating session, one

transcription was initially completed collaboratively by the researcher and inter-rater by way of training; two more were then completed by the inter-rater working alone to provide the basis for calculating percentage agreement figures as described above. In the second inter-rating session, for greater speed, the inter-rater completed his transcriptions using the researcher to support him in selecting the IPA symbols which he felt best represented the sounds he was hearing. Percentage agreement figures were again computed for both the transcriptions and the corresponding scores.

### 3.6.4 Self-report interviews

#### 3.6.4.1 Rationale and validity

To investigate learners' strategic reasoning when decoding L2 words, task-concurrent self-report interviews (SRIs) were conducted in quiet, private locations at t1 and t2 with a sub-sample of twenty participants, selected as described in section 3.4.3. Participants saw individual words from the RAT (presented on screen as in the test itself) and were asked to 'think aloud' (Cohen, 1998), that is, to verbalize as much as possible of what was going through their minds as they tried to read each word aloud. However, as discussed below, the approach taken was not one of a 'pure' think aloud; rather, as in Erler (2002:98), "[t]he relationship with each pupil was seen to be that of collaborators working together to find the words to describe the personal experiences of dealing with written French". This is reflected in the use of the term 'self-report interviews' rather than 'think alouds'.

The validity of verbal report data as a source of insight into mental processes has been questioned both generally (Nisbett & Wilson, 1977) and in SLA research specifically (Seliger, 1983). A central argument of both authors is that there is no way of telling what proportion of a given verbal report might genuinely reflect internal processing (if this is possible at all) and what proportion is based simply on the participant's post hoc inferences about what the course of the processing was likely to have been. Indeed, the problem might be worse in the case of young respondents, who might feel pressurized to say something – *anything* – to please the researcher.

A further threat to the validity of self-report data is posed by the potential for reactivity: that is, the possibility that the very act of attending to and verbalizing one's thinking alters its course. At the very least, it is intuitively likely that the one-to-one interview situation will lead participants to pay greater attention to decoding than they usually would. The strategies they report may therefore not reflect (i.e. may surpass) what they usually do when decoding French words at home or in class. However, turning the reactivity argument on its head, Woore (2010) argued that this did not make the findings any less valid; it just means that the SRIs tell us what kinds of strategies learners use, and what knowledge bases they draw upon, when they *really* engage with L2 decoding tasks, applying the kind of conscious reasoning which may form the first step in overcoming automatic L1-based processing (section 2.6.1). Precisely, it is interesting to discover whether, under such circumstances, there are differences in the strategic reasoning of intervention and comparison participants.

Leow and Morgan-Short (2004) and Bowles and Leow (2005) investigate the issue of reactivity specifically in the SLA context, demonstrating that verbalization does not

affect L2 outcome measures. This, they claim, provides "empirical support for the validity of the use of such procedures to gather concurrent data on learners' cognitive processes" (Leow & Morgan-Short, 2004:50). It should be noted, however, that identical *outcomes* do not in themselves prove that there has been no effect on the nature of the processing which generated them, which is in fact the object of interest. The problem of course is that we have no insight into the route that interviewees' processing has taken in the comparison (non-verbalization) condition; the only way to gain insight into this processing is, precisely, to ask them to verbalize it. This circularity besets any attempt to demonstrate the validity of self-reports, although investigations of the neurological correlates of strategy use (e.g. Takeuchi, Ikeda & Mizumoto, 2009) may offer the beginnings of a solution to this dilemma.

Despite the widely-cited critiques mentioned above, self-report methods have gained momentum. To an extent, their widespread use over the last thirty years in various areas of cognitive psychology, including SLA (Bowles and Leow, 2005), itself provides a strong argument in their favour (Kormos, 1998). They have become an established tool for researching L2 learners' strategies (Macaro, 2001), including in the particular sphere of L2 decoding research (Woore, 2010). This last study found SRIs to elicit a rich picture of strategic reasoning from learners of the same population as those in the current study.

However, according to Ericsson and Simon's (1980) model of verbalization, the validity of self-report data differs according to the particular circumstances of its production. They propose that only the current contents of short-term memory ('heeded' information) are directly accessible to self-report; therefore, this procedure becomes less

reliable when informants attempt to describe their thought processes retrospectively (what they *did*) or generally (what they *usually do*). Following these arguments, the current study asked participants to describe their thinking concurrently, whilst attempting to decode individual words. Inevitably, however, some information relevant to participants' decoding processes will be inaccessible. Although strategies – defined by Macaro (2006) as taking place in working memory – may be reported, any subconscious processes cannot be. This means that where participants pronounce words using automatic, L1-based processing, their self-report protocols could be "very sketchy" (Ericsson & Simon, 1980:227); indeed, this may be the origin of some of the apparently uninformative self-reports in Woore (2010), where participants indicated that a pronunciation "just came" to them or that they "just said it".

Ercisson and Simon (1993) also caution against 'introspection', where "subjects are not only requested to verbalise but also to describe or explain their thoughts", or where they are probed for specific information which may not coincide with currently 'heeded' aspects of their processing, because this again goes beyond the current contents of working memory. Despite this advice, in the current study occasional probes were used, in order to invite participants to explain 'why they thought X' or 'how they knew X'. Again, it is acknowledged that the resulting data may not necessarily reflect the way the participants would *usually* think about French decoding; nonetheless, this kind of probing was intended to gain insight into the resources participants were able to draw upon (in the form of their knowledge bases and strategic reasoning) when they really tried hard to decode French words correctly. It is, precisely, interesting to discover whether the programmes of instruction have increased the resources available to participants in such circumstances.

Finally, it should be borne in mind that participants may vary in terms of 'invisible' factors such as their level of awareness of their own processing, their ability to verbalize it and/or their propensity to do so. Thus, differences in the strategies and knowledge bases reported by participants may not reflect true differences in their processing. More generally, an absence of SRI evidence for a particular strategy is not informative: i.e. just because a participant does not mention it does not necessarily mean that it was not used.

#### 3.6.4.2 Administration

In conducting the SRIs, instructions were delivered by the researcher in person, embedded in a general introductory conversation designed to put participants at ease: previous experience (Woore, 2010) showed that establishing and nurturing a rapport with the Y7 interviewees made them more likely to give detailed and uninhibited responses (see Rapley, 2004). Care was taken to tell participants that there were no 'right' or 'wrong' answers, and that whatever they said was interesting because of the light it shed on their thinking.

Once interviews were under way, the trajectory of the discussion on each word was determined by the participants themselves: they were free to pursue their own lines of thinking. However, if they retreated into prolonged silence, prompts to keep thinking aloud were given. To avoid influencing their subsequent reasoning, participants received no feedback concerning the rightness or wrongness of the *contents* of their verbalizations or of the pronunciations they offered for target words; however, general encouragement was given on the process (e.g. "That's really interesting"). Sometimes, follow-up

questions were also used to clarify what the participant meant, or to request confirmation of the researcher's emerging interpretation of what had been reported. It is acknowledged that there are some dangers in this approach, especially given the possible tendency of children in particular to confirm what is put to them (Lewis 2002:113). However, an attempt was made to avoid leading questions, and where (at the analysis stage) leading questions were occasionally identified, any resulting comments were discarded.

SRIs lasted around twenty minutes each, the exact duration depending on (a) the time made available by the teacher, within the constraints of the school day and (b) the participant's loquacity. Only a small number of RAT items were used, to allow in-depth discussion of each one and to keep the data within manageable proportions. Eleven words were selected at each time point (Appendix 3.1), on the basis that they contained graphemes shown to pose particular problems for English learners (Woore, 2006); different words were used at t1 and t2, in order to reduce potential practice effects. Participants worked through the words sequentially, the number covered depending upon the time available and how much they had to say about each one; occasionally, participants also expressed a wish to omit one of the words and move on to the next one, which was allowed. At both time points, the modal number of words covered was six[18].

### 3.6.4.3 Analysis

All SRIs were digitally audio-recorded; a sample transcript appears in Appendix 3.2. One recording was discarded due to idiosyncratic (and therefore uninformative)

[18] At t2, three of the eleven words were not covered by any participants; these are omitted from Appendix 3.1.

comments; four others were also excluded because the participants were in classes which did not contribute to the main analysis (section 4.2.1). To analyse the remaining fifteen SRIs, an approach was piloted which involved transcribing, segmenting and exhaustively coding the data based on the categories used in Woore (2010), modified and expanded as necessary. However, whilst this facilitated between-participant comparisons in terms of the frequency with which various strategies were used, it became clear that a great deal of rich, contextual information was lost. This information concerned, for example, the *way* in which particular strategies were used, how successfully and in what combination.

An alternative approach was therefore adopted (drawing inspiration from the 'pen portraits' of trainee teachers' progress used by Ofsted, 2009) which allowed participants' reasoning to be understood holistically and in context, rather than in a fragmented and atomized way. This comprised several stages. First, each participant's SRI was subdivided into the separate discussions of the individual target words (henceforth referred to as 'word-discussions'). An overall description of the participant's approaches to decoding each word was then produced, working directly from the audio recording rather than from transcripts (which inevitably involve the loss of information such as intonation, loudness and speed of delivery: Walford, 2001). These descriptions were structured around three broad headings, based on the superordinate categories used in Woore (2010): (a) strategies deployed; (b) knowledge bases drawn upon; (c) metacognitive comments on the decoding process, such as reflections on their own self-efficacy, evaluations of their own output and evaluations of their approaches to the decoding tasks. Under each of the three headings, relevant extracts of the word-discussions were transcribed and analysed. A summary was then produced (in note form) of each participant's approaches to decoding at each time point, as demonstrated in

their SRIs as a whole, synthesizing the analyses of the various individual word-discussions. In turn, these summaries were used to create a brief 'pen portrait' (again in note form) of any ways in which that participant's reasoning differed between t1 and t2. Finally, these pen portraits provided the basis for (a) an overview of t1-to-t2 differences in participants' reasoning across the sample as a whole; and (b) the selection of one participant from each group (Programme A, Programme B, Comparison) for more detailed analysis, to be presented in the form of three case studies.

Whilst the analytical approach adopted here offers advantages in terms of a contextualized understanding of participants' reasoning, it clearly involves a high degree of subjective interpretation of the primary data. Further, it was not possible to calculate intercoder reliability in the traditional way, based on a second coder's classification of the data into a set of pre-existing categories. Nonetheless, an independent 'moderator' was recruited who had considerable experience both of (a) qualitative data analysis and (b) of working with Y7 MFL learners as a teacher and teacher educator. The moderator listened to the audio-recordings of twenty randomly-selected word discussions (ten per cent of the total produced by all participants at both time points) and made notes under each of the three headings (strategies, knowledge bases, metacognition), based on his own intuitive interpretation and having had no insight into the researcher's original analysis. The moderator and researcher then compared and discussed their interpretations. There were no instances of the researcher and moderator disagreeing in their interpretation of a word-discussion, although the moderator did make a number of additional comments. However, these almost always related to aspects of the data deliberately excluded from the researcher's analysis, such as inferences about the

processes involved in pronouncing a target word based on the participant's pronunciation of it. In four other cases, small additions were made to the researcher's original analysis.

## 3.7 Fidelity to condition

### 3.7.1 Methods

Whilst it was acknowledged that some differences would inevitably emerge in the implementation of each programme of instruction between the three teachers assigned to it (section 3.2.1), it was nonetheless important that certain aspects of the interventions remained as constant as possible across the various classrooms: for example, the amount of time spent on the decoding segments; the GPCs they covered; the opportunity for learners to attempt to decode unfamiliar words; the additional reference, in Programme A, to the memorized poem as a mechanism for retrieving the pronunciations of target graphemes. The teaching received by all intervention groups was therefore monitored in order to (a) assess fidelity to condition and (b) document any differences which might arise in the decoding instruction as received by the different classes, so that this might be taken into account at the analysis stage. Similarly, the teaching of comparison groups required monitoring in order to assess the extent to which (as envisaged) they did not receive any explicit instruction in French decoding. Finally, the teaching received by all groups (intervention and comparison) should be checked for differences on other theoretically relevant variables such as the amount of time spent on reading and speaking activities. To address these issues, three tools were used:

(1) To promote consistent implementation of key aspects of the decoding programmes, all intervention group teachers received detailed instructions in the form of 'Teacher Packs' (Appendix 1.1), provided both as hard copies and electronically. These included (a) an overview of the relevant programme and brief description of its rationale; (b) a detailed breakdown of its contents and 'scheme of work'; and (c) ready-to-use copies of all necessary resources. Initial briefing meetings were also held with teachers; these additionally served to promote a trusting and mutually supportive relationship with the researcher, which was considered crucial given the considerable levels of commitment required.

(2) Routine monitoring and support meetings ('Programme Healthchecks') were held with participating teachers in each school twice during the intervention period and once after t2. The aims were to (a) find out through discussion with teachers whether the programmes were being implemented as planned; (b) offer guidance and support in case of any problems or questions; (c) obtain feedback on the programmes with a view to making on-going changes if necessary; and (d) help maintain teachers' motivation and commitment to the programmes. The meetings took place in quiet, private locations at mutually convenient times, generally lasting around thirty minutes. Group (rather than individual) discussions were chosen in order to promote teachers' sense of working on the decoding instruction as members of a team and to facilitate the sharing of experiences, questions, problems and solutions. The discussions were allowed to unfold naturally but a framework of questions (Appendix 8.5) ensured coverage of three pre-determined areas: (a) overall progress in terms of covering the scheduled content; (b) possible dimensions of variation between the groups; and (c) any positive or negative experiences of the programmes. Notes were taken by the researcher and subsequently

typed up; teachers' responses to each question were then collated and compared. The 'healthcheck' meetings were also supplemented as necessary by informal channels of support: for example, teachers occasionally sent questions by email and unscheduled 'chats' often occurred when the researcher was in school collecting other strands of data. This on-going presence in school team rooms was considered important in maintaining positive working relationships with participating teachers.

(3) Lessons in both intervention and comparison classes were observed, including the decoding segments in the former. Practical constraints permitted the observation of only two lessons per class (in January and April 2008). This gave twenty-four observations overall, clearly too few to support generalizations about each class's wider teaching during the intervention period. To address this issue to some extent, teachers were asked (and trusted) to nominate a lesson for observation whose plan they considered 'typical' of their teaching with this group (following a procedure used by Vandergrift, as reported in Spada & Fröhlich, 1995). There remained an unavoidable possibility of observer effects, although lesson observations were a routine part of all participating teachers' professional practice through their annual Performance Management cycle[19] and through their work within the university's Initial Teacher Education partnership. Teachers' BIQ responses also provided some basis for triangulating the lesson observation data.

To ensure a consistent and systematic approach to observing lessons, the COLT (*Communicative Orientation of Language Teaching Observation Scheme*: Spada & Fröhlich, 1995) was chosen as an appropriate framework, having been used widely over many years in SLA research and being adaptable to the needs of individual studies

[19] Under the Education (School Teacher Performance Management) (England) Regulations 2006, teachers must undergo an annual Performance Management process which includes regular lesson observations by another qualified teacher.

(Spada & Lyster, 1997). Much of Part B of the COLT was considered either too detailed or not relevant for the current study. A modified version of Part A was therefore used (Appendix 8.1), completed by the researcher in real time. The use of an inter-rater (as recommended by Spada and Fröhlich, 1995) was considered unnecessary, since the lesson observations were subsidiary to the main research questions.

The modified COLT comprised thirty-two categories of observation, recording the following information: (a) the lesson's main activities, the episodes that comprised them, how long was spent on each and whether there was any explicit focus on decoding; (b) the predominant language (L1 or L2) used by teachers and students in each activity; (c) whether the students worked on the activity individually, in pairs, in groups or as a whole class; (d) whether the activity focussed on linguistic forms or on topic content; (e) whether the focus was chosen by the teacher or the students; (f) whether the activity mainly involved listening, speaking, reading or writing, and whether it involved feedback; and (g) the nature of any materials used. Several versions of the modified COLT were piloted with (non-participating) Y7 MFL classes from school A and successively refined.

The COLT data were used to calculate the approximate percentages of lesson time occupied by the various aspects of teaching that were systematically recorded. For each class, these percentages were calculated both (a) for each individual lesson observation and (b) for both lesson observations taken together. The latter figures, which (at least to some extent) look beyond the idiosyncrasies of individual lessons, are summarized in the tables in Appendix 8.3. For intervention groups, separate percentage figures were also calculated for (a) the decoding segments and (b) the remaining (i.e. non-intervention)

teaching. The data for the non-intervention teaching provided the basis for comparisons of the 'usual' teaching received by each class (though it is conceivable that the non-intervention teaching in the intervention classes might have been influenced by the decoding instruction, for example through a raised awareness of decoding issues on teacher's part).

Finally, to facilitate monitoring of the implementation of the programmes, a summary of the key features of each decoding segment (such as the time taken, the sequence of activities and the students' reactions) was drawn up immediately after each observation, drawing on the COLT data, the researcher's recollections of the lesson and any comments made by teachers in post-lesson discussions. A sample summary appears in Appendix 8.4.

### 3.7.2 Findings

This section discusses the findings of (a) the observations of non-intervention teaching in both intervention and comparison groups (section 3.7.2.1), (b) the observations of decoding segments (section 3.7.2.2) and (c) the intervention healthcheck meetings (section 3.7.2.3). Space permits only a brief summary of the most pertinent findings; a full overview of the quantitative COLT data appears in Appendix 8.3.

#### 3.7.2.1 Non-intervention teaching

Few differences were found between the observed lessons in terms of the categories included in the modified COLT; those differences which did emerge were readily

ascribable to idiosyncratic features of the particular lessons observed (for example, one observed lesson happened to be taken by a substitute teacher and involved completing the class's termly listening assessment). In all lessons, whole class, teacher-led instruction generally predominated, with an emphasis on the explicit teaching of vocabulary. All observed lessons tended to focus on linguistic forms (Long, 1991) rather than on function.

However, whilst (outside the intervention segments) most classes included very little or no explicit focus on decoding, this did occupy around a fifth of one of the observed lessons in class D1. It is possible that this was an idiosyncratic feature of this particular lesson, rather than being a systematic feature of teacher D1's teaching. However, cross-reference of the observation data with her BIQ responses (section 3.4.2.2) suggests otherwise: she stated that she "often refer[s] to pronunciation", mentioning in particular the SFC rule and the fact that some groups of letters are "pronounced together". In informal discussion with teacher D1 following the lesson observation, she reported a long-standing emphasis in her teaching on what she called *le secret de la prononciation française* ('the secret of French pronunciation'); indeed this was observed in action in the lesson, when a pupil initially pronounced *je lis* ([ʒəli], 'I read') as [ʒəliz], then produced the correct form (without the final [z]) when told to "remember the secret" of SFCs. The fact that the pupil was able to self-correct suggests that the reminder was familiar to him. There is thus reasonable evidence to suggest that class D1 may have received regular, explicit instruction in at least some aspects of French decoding, a confounding variable which must be borne in mind in the data analysis.

### 3.7.2.2 Decoding segments

The decoding segment summaries revealed a generally encouraging picture. The decoding segments were implemented in all observed lessons in the intervention condition, though the time spent on them varied more widely than had been anticipated (from seven to sixteen minutes), with a difference of three minutes in the mean segment length between schools A and B (thirteen versus ten minutes respectively). However, these observations based on just two segments per class must be cross-referenced with other data, such as teachers' responses in the 'healthcheck' interviews (section 3.7.2.3).

In terms of the sequence of activities outlined in the teacher packs, the segment summaries showed that this was generally followed closely in eight of the twelve observed lessons. There were however four instances of departures from the intended sequences, arising when the teacher (a) asked pupils to work out the pronunciation of the *demonstration* words rather than the list of ten unfamiliar words (class B1 at both t1 and t2); (b) failed to ask pupils to work out the pronunciations of any words for themselves (class A3, t1); and (c) failed to ask pupils both to search the poem for words containing the target grapheme and to work out the pronunciations of the unfamiliar words (class A1, t2, where the lesson was taught by a substitute teacher in the absence of the regular teacher). In post-lesson discussions with the teachers, three of these cases were attributed to 'accidentally' forgetting one of the planned activities, and the fourth to particular time pressures on the day, meaning that more lesson time was needed for other (non-decoding) content. In all four cases, the researcher subsequently discussed with the teachers individually the importance of including all planned activities in the decoding

segments, in order to increase fidelity to condition. The teachers responded positively, although it cannot be guaranteed that this impacted on subsequent teaching.

The other source of insight into the implementation of the programmes of instruction was the COLT observation data. This revealed clear differences between the nature of the teaching in the decoding segments on the one hand and in the non-programme teaching (in both intervention and comparison groups) on the other. In particular, an explicit focus on decoding occupied between 92% and 100% of the segment teaching time in all intervention classes, whereas it featured hardly at all in either the same classes' non-programme teaching or in the teaching received by comparison classes (with the exception of one lesson in class D1, as discussed in section 3.7.2.1).

#### 3.7.2.3 Intervention 'healthcheck' meetings

Teachers' responses during the three intervention healthcheck discussions (hereafter H1, H2 and H3) confirmed that the decoding segments were being implemented every lesson, barring occasional omissions due to teacher absence, class trips, school events, CATS testing and the like. It was encouraging that the project appeared, on this evidence, to have generated a systematic and sustained change in teachers' practice. This suggests that the programmes of instruction were (a) straightforward enough for diverse teachers to use in practice over a period of many weeks and (b) considered to be worth the time investment in pedagogical terms (the teachers' first duty being to their students' MFL learning rather than to the research project).

However, due to intervening factors such as those mentioned above, there was some variation in how far teachers had progressed through the scheduled programmes: at H1, this was true of teachers in school A only; but by H2, all teachers had fallen behind to some extent. Though teachers tried to make up lost ground, in some cases by occasionally conducting two segments per lesson, differences remained in teachers' progress through the scheduled segments at H3 (Table 3.4).

**Table 3.4 Intervention classes' coverage of planned content in the programmes of instruction**

| **Class** | **Number of last segment completed at t2** (out of 38 segments; see teacher packs) | **Prescribed GPC not covered at t2** |
|---|---|---|
| **A1** | 28 | an, in/ine, ain/aine, ien, un, g, eur, eu, au, eau, gn, tion, gu, i$^{+vowel}$ |
| **A2** | 31 | ien, un, g, eur, eu, au, eau, gn, tion, gu, i$^{+vowel}$ |
| **A3** | 31 | ien, un, g, eur, eu, au, eau, gn, tion, gu, i$^{+vowel}$ |
| **B1** | 34 | gn, tion, gu, i$^{+vowel}$ |
| **B2** | 34 | gn, tion, gu, i$^{+vowel}$ |
| **B3** | 30 | g, eur, eu, au, eau, gn, tion, gu, i$^{+vowel}$ |

There was also some variation in the length of time teachers said they usually spent on each decoding segment (Table 3.5). At H1, when roughly seven segments had been completed, this was roughly ten minutes for almost all classes but over fifteen minutes for A3. Following discussions and email exchanges between this teacher and the researcher, she reported reducing the average duration of subsequent segments. At H2 and H3, there was less variation in the time spent on each segment, although on the whole this tended to be slightly longer for Programme A than Programme B (perhaps reflecting the additional demands of memorizing and using the poem), and within Programme B slightly shorter for class B2.

**Table 3.5 Approximate number of minutes reportedly spent on decoding segments in the period preceding each 'healthcheck' interview**

| Class | H1 | H2 | H3 |
|---|---|---|---|
| **A1** | 10 | 10-12 | 10-12 |
| **A2** | 10 | 15 | 12 |
| **A3** | Over 15 | 15 | 12 |
| **B1** | 10 | 10 | 10 |
| **B2** | 10 | 8 | Less than 10 |
| **B3** | 10 | Just over 10 | 10 |

Finally, teachers reported high levels of participant attendance throughout the intervention period. No student missed a substantial number of lessons, defined as more than three lessons' absence in total. This again indicates a secure basis for between-group comparisons. However, the collection of detailed attendance records would have strengthened this claim further.

## 3.8 Ethics

The current study was approved by Oxford's Central University Research Ethics Committee. A copy of the approval letter appears in Appendix 10.1.

Following BERA's (2004) guidelines, the study is based on the principles of voluntary informed consent and the confidential, anonymous treatment of participants' data. Information letters or emails seeking permission to conduct the research were first sent to headteachers, heads of MFL and all participating teachers. For all strands of the project involving data collection, opt-in consent was also sought not only from parents but also from pupils themselves[20]. The information letter to pupils aimed to describe the study in child-friendly language, to ensure that their consent was as genuinely 'informed' as

[20] Sample information letters and consent forms appear in Appendices 10.2-10.4.

possible. The nature of the research, and the fact that participants could opt out at any time without any penalty or prejudice to their progress in school, were also explained to all classes by the researcher in person, giving time to answer any questions that arose. Following this explanation and the distribution of consent letters, students had several weeks in which to consider their response and return their reply slips. Following David, Edwards & Alldred's (2001) problematization of the issue of students' voluntary consent in the context of school-based research, participants were made aware of various possible channels for withdrawing from the study (for example by talking to the researcher, their class teacher or their pastoral tutor).

All data collected in the course of the study was stored on a password-protected laptop or (in the case of hard copies) in a locked university office. Each participant was allocated an ID number (with the key accessible only to the researcher), so that her or his data were not directly associated to her or his name.

A practical problem for the current study is that the opt-in consent applies only to the collection of data, not to the programmes of instruction themselves. These had to be completed by all pupils, since they were delivered on a whole-class basis; it was impracticable for any individual pupil to opt out. However, the activities involved did not differ significantly from those routinely conducted in MFL lessons; therefore, having gained the support of participating teachers, the content of their lessons was considered to be a matter for their own professional judgment. Nonetheless, the design of the programmes of instruction (based on short segments embedded within lessons) was intended to minimize the disruption to participants' coverage of their usual curriculum.

Finally, in the event of a positive evaluation of the programmes of instruction, all materials will be made available to participating teachers in the comparison groups for future use with their classes.

# Chapter 4: Findings I. Reading Aloud Test scores

## 4.1 Introduction

This chapter compares the RAT scores obtained by participants in the different experimental conditions: (a) intervention versus comparison (section 4.3) and (b) Programme A, the 'poem approach' versus Programme B, the 'phonics approach' (section 4.4). It therefore addresses Research Questions 1 and 2:

> *Do the programmes of instruction lead to more accurate decoding of L2 French?*
>
> *Is one of the programmes more effective than the other?*

Converting participants' RAT output into a numerical score follows the method of previous studies using the same instrument (Woore, 2006, 2009). However, the current study adopts a more differentiated approach, generating separate scores for 'correct' and 'acceptable' realizations of each grapheme type.

The chapter also pursues a finer-grained analysis along two other dimensions. First, section 4.5 examines the scores obtained by smaller groupings of participants within each condition, assessing the possibility of school- and class-level effects. Second, section 4.6 examines the scores obtained for individual grapheme types (e.g. <é>, <th>), to investigate whether some are associated with greater improvements in decoding accuracy than others.

As a preliminary to the analysis of RAT scores, however, this chapter begins by considering the extent and distribution of missing (or otherwise problematic) data (section 4.2). Clearly, missing values should be small in number and evenly spread between groups in order to support valid between-group comparisons. The three causes of missing data identified in sections 3.6.3.1 and 3.6.3.3 – inaudibility, omission of items and non-linear decoding (NLD), designated by the codes '997', '998' and '999' respectively – are considered in turn[21].

## 4.2 Preliminaries

### 4.2.1 Missing data

#### 4.2.1.1 Inaudible pronunciations

Within the dataset as a whole, very few grapheme tokens were pronounced inaudibly: on average, 0.3% of all grapheme tokens at pre-test and 0.2% at post-test. When the rates of occurrence of these inaudible tokens were examined at successively smaller groupings of participants, a generally even spread was found, with one exception: consideration of descriptive statistics showed that at both t1 and t2, class B3 appeared to have a considerably higher proportion of inaudible pronunciations than all other classes (Table 4.1). Kruskal-Wallis tests found that the differences between the classes in terms of the distribution of inaudible data were significant at t2 ($H(11)=26.777$, $p=.005$) but not at t1

[21] Data coded as NLD is classified here as 'missing' because – even though the participant did produce a pronunciation of the RAT item in question – it was not possible to record her or his realizations of the individual grapheme tokens making up the word. Since these grapheme tokens cannot be included in the main analysis in this or the following chapter, they are effectively equivalent to 'missing values'.

(H(11)=17.242, *p*=0.101); nor were there significant differences between other levels of participant grouping at either time (e.g. school or experimental condition). However, the lack of statistical significance in the between-class comparisons at t1 may reflect the small numbers in some of the groups, especially class B3.

**Table 4.1 Mean numbers and percentages of grapheme tokens pronounced inaudibly per participant (out of a possible total of 174).**

| **Class** | **A1** N=24 | **A2** N=25 | **A3** N=20 | **B1** N=25 | **B2** N=21 | **B3** N=12 | **C1** N=15 | **C2** N=15 | **C3** N=10 | **D1** N=18 | **D2** N=21 | **D3** N=15 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **t1** mean | 0.4 | 0.2 | 0.3 | 0.4 | 0.1 | 4.3 | 0.2 | 0.4 | 0.0 | 0.2 | 0.3 | 0.6 |
| S.D. | 0.9 | 0.6 | 0.7 | 1.3 | 0.4 | 5.4 | 0.6 | 1.1 | - | 0.5 | 1.3 | 2.3 |
| % | 0.2 | 0.1 | 0.2 | 0.2 | 0.1 | 2.5 | 0.1 | 0.2 | 0.0 | 0.1 | 0.2 | 0.3 |
| **t2** mean | 0.1 | 0.2 | 1.2 | 0.5 | 0.0 | 2.6 | 0.3 | 0.5 | 0.1 | 0.4 | 0.0 | 0.0 |
| S.D. | 0.4 | 0.8 | 3.1 | 1.7 | - | 5.2 | 0.8 | 1.6 | 0.3 | 1.2 | - | - |
| % | 0.1 | 0.1 | 0.7 | 0.3 | 0.0 | 1.5 | 0.2 | 0.3 | 0.1 | 0.2 | 0.0 | 0.0 |

Since the recording conditions in school B were among the most favourable encountered during the data collection process, the higher proportion of inaudible pronunciations generated (at least at t1) by members of B3 appears to reflect inadequacies in their articulation. Informal observations further suggested that this problem may be caused by 'mumbling' or very soft speech (close to subvocalization). It might be speculated that this reflected low self-efficacy on the part of some students as a result of their history of low academic achievement (section 3.3). Even for B3, however, the number of inaudible pronunciations represents only a small proportion of the overall dataset. Inaudible pronunciations were ignored in subsequent analyses.

#### 4.2.1.2 Omission of RAT items

The numbers of missing values resulting from students omitting RAT items were very small (0.4% of all grapheme tokens at t1; 0.3% at t2) and were roughly evenly

distributed across the various groups. Omitted items were therefore ignored in subsequent analyses.

#### 4.2.1.3 Non-Linear decoding

As expected, the amount of Non-Linear Decoding (NLD) in the RAT data was higher than for the two categories of 'genuine' missing data (see footnote 21). Nonetheless, a few participants stood out due to extremely high levels of NLD and were removed from the analysis to avoid skewing subsequent between-group comparisons of RAT scores. This is because a score of zero awarded in the case of NLD has a different meaning than if participants simply decoded a given grapheme incorrectly (because that grapheme may not have been decoded at all in any straightforward sense). An arbitrary threshold was therefore set: any participant for whom over a third of grapheme tokens were coded '999' (to indicate NLD) was removed from the analysis. This applied to five cases at t1 and three at t2, leading to the exclusion of seven participants overall.

Following the exclusion of these outliers, the mean proportions of NLD were relatively low and fairly evenly distributed across most classes (Table 4.2). Histograms (Figure 4.1) also show that, at both t1 and t2, the data are non-Normally distributed, with the modal number of NLD codes per participant being zero and the frequency generally decreasing in inverse proportion to the number (i.e. the higher the number of NLD codes, the fewer participants obtained it). This indicates that most participants did generate pronunciations by decoding graphemes in linear sequence. Non-parametric statistical analyses found no significant differences, either at t1 or t2, in the distributions of NLD codes (a) between the intervention and comparison conditions (Mann-Whitney U=5591,

Z=-0.806, *p*=.420 at t1; U=5461, Z=-1.086, *p*=.277 at t2); (b) between schools (Kruskal-Wallis H(3)=2.246, *p*=.523 at t1; H(3)=3.345, *p*=.341 at t2); or (c) between classes (Kruskal-Wallis H(11)= 15.802, *p*=.149 at t1; H(3)=11.951, *p*=.367 at t2). However, as the descriptive statistics in Table 4.2 show, class B3 again stands out as having higher levels of NLD, equal to double or more than double those of most other classes. The magnitude of the differences suggests that the lack of statistical significance in the between-class comparisons may reflect the small numbers of participants in some classes (especially B3).

**Table 4.2 Mean numbers and percentages of graphemes coded as NLD (Non-Linear Decoding) per participant out of 174 tokens in the RAT.**

| **Class** | **A1 N=24** | **A2 N=25** | **A3 N=20** | **B1 N=25** | **B2 N=21** | **B3 N=12** | **C1 N=15** | **C2 N=15** | **C3 N=10** | **D1 N=18** | **D2 N=21** | **D3 N=15** |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **t1** mean | 10.8 | 18.3 | 14.2 | 12.1 | 13.0 | 28.3 | 16.7 | 14.8 | 17.6 | 14.1 | 11.1 | 10.2 |
| S.D. | 10.4 | 15.1 | 12.8 | 14.4 | 12.6 | 17.8 | 18.3 | 13.9 | 17.0 | 11.9 | 12.5 | 12.3 |
| % | 6.2 | 10.5 | 8.2 | 7.0 | 7.5 | 16.3 | 9.6 | 8.5 | 10.1 | 8.1 | 6.4 | 5.9 |
| **t2** mean | 9.4 | 11.2 | 14.4 | 10.8 | 11.5 | 22.0 | 14.8 | 10.1 | 14.3 | 12.8 | 8.1 | 7.0 |
| S.D. | 11.0 | 11.3 | 16.9 | 10.3 | 12.9 | 12.9 | 15.9 | 11.7 | 16.5 | 15.4 | 6.1 | 8.0 |
| % | 5.4 | 6.4 | 8.3 | 6.2 | 6.6 | 12.6 | 8.5 | 5.8 | 8.2 | 7.4 | 4.7 | 4.0 |
| **Change** | -1.4 | -7.1 | +0.2 | -1.3 | -1.5 | -6.3 | -1.9 | -4.7 | -3.3 | -1.3 | -3 | -3.2 |

**Figure 4.1 Histograms showing numbers of cases of NLD across the whole sample (a) at t1; (b) at t2.**

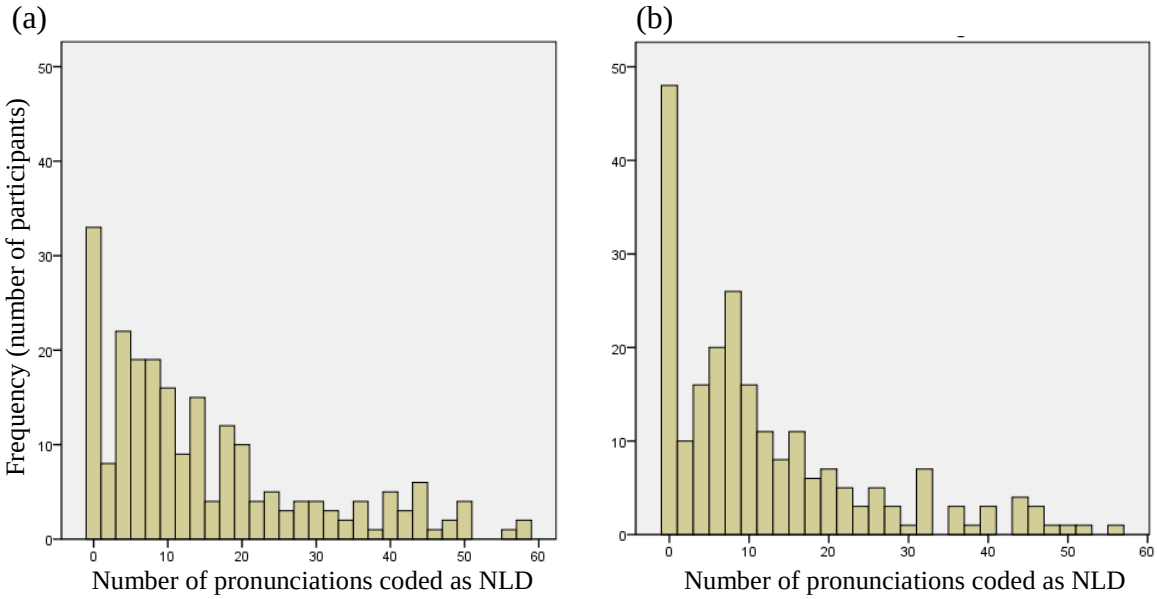



Table 4.2 and Figure 4.1 also show that there was a small but consistent decrease between t1 and t2 in the mean proportion of grapheme tokens coded as NLD in all classes, save for a very small increase in class A3; the largest decrease occurs in class A2, which also had the highest proportion of NLD codes at t1. The proportion of participants obtaining no NLD codes also increases between t1 and t2 from 15% to 22% of the sample.

Returning to the different patterns observed between class B3 and the other participating classes in terms of the proportions of NLD codes – although these differences were not statistically significant – it may again be speculated that this reflects the lower L1 literacy levels within this class (the only participating 'lower stream' group[22]), which might make them less likely to apply the linear mapping principle. To investigate the hypothesis that B3 differed from other participating classes in terms of L1 literacy levels,

[22] This is the school's terminology.

the academic data provided by the sample schools was consulted (section 3.6.2.1). Unfortunately, this incomplete dataset lacks information on the most relevant variables (reading age; verbal CAT scores) for class B3, for which the best available literacy data is the NC level on the reading component of the English SAT tests. Although this variable provides only coarsely gradated data, and though it has some missing values within all participating groups, a clear pattern nonetheless emerges. B3 differs both from other classes in school B and from all classes in school A, the only other school for which this data is available (table 4.3). The highest level (Level 5) was obtained by over half the students in all other classes but by none in B3; conversely, the lowest level (Level 3) was obtained by four out of ten participants in B3 but by very few or none at all in the other classes. These between-class differences in reading level were statistically significant ($\chi^2(10)=37.297$; $p<0.001$). Further, a statistical analysis of the finer-grained verbal CAT scores revealed no significant differences across those nine classes in schools A, B and C (not including B3) for which these data were available (Kruskal-Wallis: $H(8)=11.039$, $p=.199$[23]); in other words, in these classes, the evidence form the verbal CAT scores corroborates that from the reading levels. It therefore seems reasonable to conclude that class B3 does indeed differ from other participating classes in terms of its members' significantly lower L1 literacy profile.

**Table 4.3 Numbers of students in each class obtaining NC levels 3-5 in reading component of English SAT**

| Class | A1 | A2 | A3 | B1 | B2 | B3 |
|---|---|---|---|---|---|---|
| **Level 3** | 3 | 0 | 0 | 0 | 0 | 4 |
| **Level 4** | 7 | 9 | 7 | 9 | 4 | 6 |
| **Level 5** | 13 | 13 | 11 | 15 | 16 | 0 |
| **Missing values** | 1 | 3 | 2 | 1 | 1 | 2 |

[23] A non-parametric test was used due to non-Normal distributions of CAT scores both for the sample as a whole and for various individual classes.

The preceding analysis of missing data has highlighted various differences between class B3 and other participating classes. Most importantly, the descriptive statistics suggest a clear pattern in the proportions of NLD produced in the different classes at both t1 and t2, with B3 having values which appear to be considerably higher than (in many cases, more than double) those of other classes. The fact that these differences were not found to be statistically significant may, it was argued, reflect the small sample sizes in some classes, especially B3 itself (N=12). B3 also appeared to differ from the other classes in terms of the proportion of inaudible data, with these differences being statistically significant at t2; again, larger group sizes may also have resulted in significant differences at t1. Both of these patterns are argued to reflect the L1 literacy levels of participants in class B3, which on the basis of the best (though imperfect) data available were found to be significantly lower than those of other classes; this is as expected, given that class B3 is the only 'lower stream' class in the sample. It was therefore decided to exclude B3 from subsequent statistical analyses at the condition level, because there was a likelihood of creating non-homogeneous groups which would undermine between-group comparisons. (This will have little impact on the power of statistical tests, since the overall sample contained only twelve members from B3). However, B3 is still included in class-level analyses, where it can be considered on its own terms.

### 4.2.2 Fidelity to condition

Section 3.7.2 found that teacher D1, unlike other comparison teachers, included in her teaching some degree of explicit, systematic decoding instruction. For this reason, class D1 cannot be considered truly representative of the 'comparison' condition as

understood in Research Question 1 and is therefore excluded from statistical analyses at the condition level.

### 4.2.3 Intra- and inter-rater reliability

Table 4.4, Column 1 shows the results of the intra-rater reliability checks for the transcription of the data. Despite the reliance on subjective judgments about a wide range of sounds, which in the case of vowels do not naturally fall within categorical boundaries, there was a high proportion of exact matches in the symbols selected to represent a given sound on the two occasions of transcription, indicating that the researcher's own transcription of the audio data was reliable. The proportion of approximate agreements, which included both exact matches and 'near matches', was higher still.

**Table 4.4 Intra-rater reliability figures**

| | 1 Transcription | | 2 Scoring | |
|---|---|---|---|---|
| | Exact matches | Exact and 'near' matches | Stringent mark scheme | Lenient mark scheme |
| Overall percentage agreement | **93.8%** | **94.1%** | **98.5%** | **98.4%** |

In terms of the reliability of the scores generated under both the 'stringent' and 'lenient' mark schemes (assessing the number of 'correct' and 'acceptable' pronunciations respectively), the overall percentage agreement figures were again very high (Column 2, Table 4.4); in fact, as discussed in section 3.6.3.5, these percentage agreement figures must inevitably be at least as high as those for the transcriptions upon which they were

based. The Intraclass Correlation Coefficients also approached the perfect score of 1 and had narrow confidence intervals. The RAT scores, as generated from the researcher's transcriptions of the original spoken output, therefore appear to provide a highly consistent measure of pronunciation accuracy.

The inter-rater reliability check also showed high levels of overall agreement between the two raters in terms of their transcription of the RAT output (Table 4.5, column 1). Although the figures were, unsurprisingly, somewhat lower than in the case of the intra-rater check, the two raters still agreed exactly on their transcription of around 85% of grapheme tokens and there was 'approximate agreement' in almost 90% of cases. There were also very high levels of agreement in terms of the RAT scores derived from these transcriptions (Table 4.5, column 2).

**Table 4.5 Inter-rater reliability figures**

| | 1 Transcription | | 2 Scoring | |
|---|---|---|---|---|
| | Exact matches | Exact and 'near' matches | Stringent mark scheme | Lenient mark scheme |
| Overall percentage agreement | **85.0%** | **89.0%** | **96.8%** | **96.9%** |

### 4.2.4 Summary

Following the exclusion of classes B3 and D1, the dataset has been shown to provide a sufficiently solid basis for valid between-condition comparisons of participants' RAT scores. As shown in Table 4.2, although some data are lost through non-linear decoding, the amounts involved are small (8.0% of grapheme tokens overall at t1; 7.1% at t2) and fairly evenly spread across the classes, schools and experimental conditions (Intervention

A, Intervention B, Comparison). Inter- and intra-rater checks also showed the reliability of the scoring to be very high.

## 4.3 Comparing intervention and comparison groups

### 4.3.1 Introduction

Addressing Research Question 1, this section compares the RAT scores of participants in the intervention group (schools A and B) with those in the comparison group (schools C and D). Two sets of results are presented. First, section 4.3.3 compares the groups' scores under the more stringent of the two mark schemes. As described in section 3.6.3.3, this variable ('score1') tallies the numbers of grapheme tokens judged to have been realized 'correctly' by each participant. Second, section 4.3.4 compares the groups in terms of the number of realizations judged 'acceptable' under the more lenient mark scheme ('score2').

The RAT covers all GPC included in the *planned* programmes of instruction, whereas in reality the programmes were not completed in full (section 3.7.2.3). This means that not all these GPC were in fact covered by all participating classes. To provide a fairer test of the effectiveness of the decoding instruction, the analysis therefore restricts itself to those thirty-three GPC known to have been covered by all participants in the programmes of instruction (Appendix 2.3). This subset of 'target graphemes', comprising 135 of the 174 grapheme tokens in the RAT, excludes the tokens of ten grapheme types and five grapheme groups, unfortunately including a disproportionately high number of nasal

vowels; this limits the conclusions that can be drawn concerning the effectiveness of the instruction for this category of GPC. Research Question 1 can therefore be phrased in the following, more specific way:

*Do the programmes of instruction lead to more accurate decoding of those GPC which they explicitly cover?*[24]

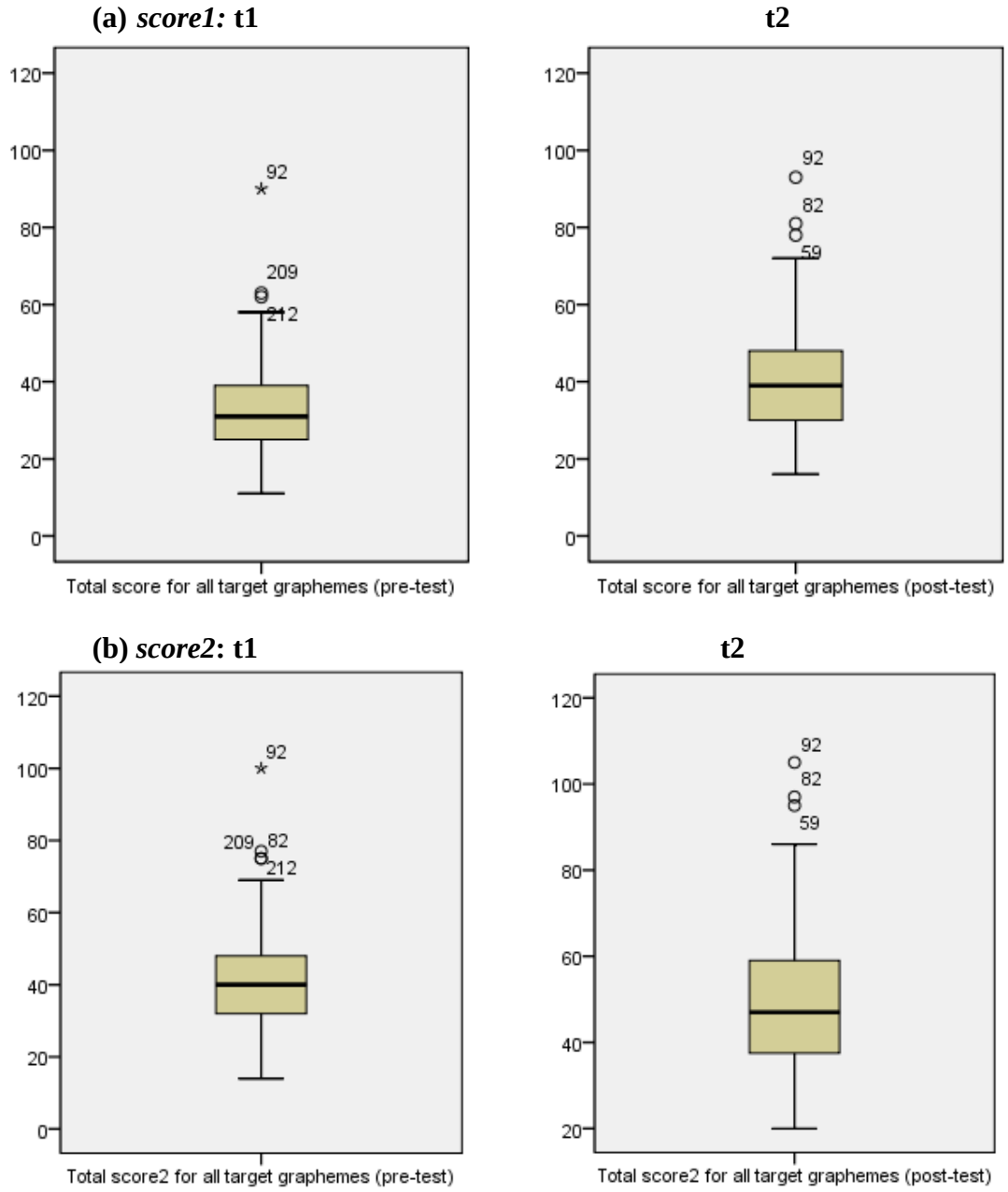
### 4.3.2 Outliers and extreme values

As a preliminary to the analysis, the distributions of *score1* and *score2* for all target graphemes were examined at both t1 and t2 in order to identify outliers and extreme values (Figure 4.2). These are defined by the SPSS statistical software as values falling more than 1.5 times (outliers) or more than three times (extreme values) the inter-quartile range above the third quartile or below the first quartile. To allow any patterns to emerge more clearly, the scores of the participating sample taken as a whole (i.e. all classes except B3 and D1) were used for this purpose. As Figure 4.2 shows, for *score1* there were three outlying or extreme values at each time of testing, generated by five individual participants; for *score2*, there were four such cases at t1 and three at t2, generated by the same five individuals. One participant in particular – case number 92 from class B1 – stands out by dint of obtaining higher scores on both variables at both time points. The five participants associated with outlying or extreme values were excluded from further analysis in this section in order to (a) avoid distorting between-

[24] It would have been desirable to conduct a complementary analysis of the thirty-nine grapheme tokens *not* included in the set of 'target GPC'. However, this analysis was not possible, because each class reached a different point in the planned schedule of decoding segments. In other words, whilst there was a clear subset of 'target GPC' known to have been covered by *all* participants, those GPC *not* covered in the instruction programmes differed from one class to another.

group comparisons of mean scores and (b) increase the validity of statistical tests. Whilst the available background information provided no a priori reason to exclude these participants – for example, their BIQ responses revealed nothing unusual about their previous experiences of learning French – it is possible that some other, undetected confounding variable accounted for their superior French decoding proficiency. The five excluded participants were distributed fairly evenly between the two conditions: three from the intervention group and two from the comparison group.

**Figure 4.2 Box-and-whisker plots showing the t1 and t2 distributions of (a) *score1* and (b) *score2* for the participating sample as a whole.** *Outliers are indicated by circles, extreme values by asterisks.*




### 4.3.3 'Correct' realizations (*score1*)

Prior to statistical analysis, visual inspection of histograms and Q-Q plots suggested that *score1* was non-Normally distributed at both t1 and t2, both for the sample as a whole

and when subdivided by condition. This was confirmed in several cases by significant Shapiro-Wilk test results (whole sample: W(186) =0.968, $p$<0.001 at t1 and W(186)=0.976, $p$=0.003 at t2; intervention group: W(112) =0.962, $p$=0.003 at t1 and W(112)=0.983, $p$=0.183 at t2; comparison group: W(74) =0.960, $p$=0.018 at t1 and W(74)=0.952, $p$=0.007 at t2). Though logarithmic transformation of the scores improved the distributions to some extent, they still showed varying degrees of skewness and kurtosis. The originally-planned statistical analysis of the groups' scores using parametric techniques (section 4.3.3.1) was therefore supplemented by non-parametric tests (section 4.3.3.2) as a point of comparison. Though potentially less sensitive, these are safer in terms of the validity of the analysis.

#### 4.3.3.1 ANOVA

*Score1* values were analysed using a two-way mixed factorial ANOVA with one within-subjects factor (time) and one between-subjects factor (condition). The assumptions of the test were first checked. First, the issue of sphericity was disregarded because the repeated measures variable had only two levels (t1 and t2) (Field, 2009:459). Second, due to the unequal sample sizes in each condition (see Table 4.6, column 1 below), Box's M test of equality of covariance matrices was evaluated at $\alpha$<.001; this retained the null hypothesis that the observed covariance matrices of the dependent variables were equal across groups (Box's M = 5.085, F(3, 1283067.152)=1.674, $p$=.170). Third, Levene's test confirmed that variances were homogenous for both levels of the repeated measures variable: at t1, F(1, 184)=2.301, $p$=.131; at t2, F(1, 184)=0.471, $p$=.493.

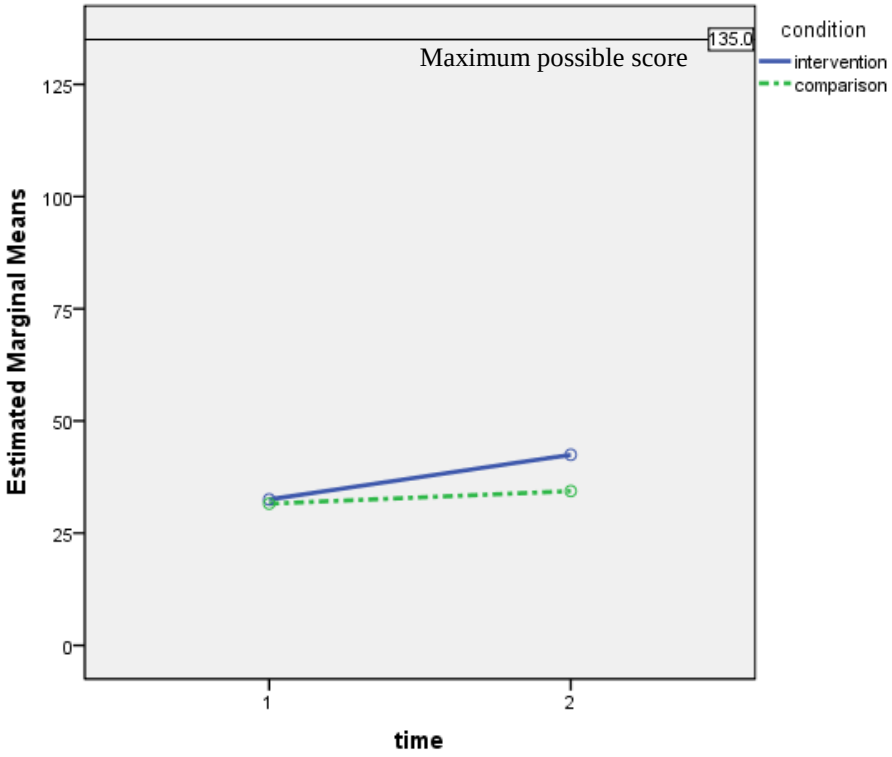
The ANOVA found there to be significant main effects of time, F(1,184)=112.613, $p$<0.001, and condition, F(1,184)=8.889, $p$=.003. There was also a significant interaction effect between the time of testing and the condition, F(1,184)=35.689, $p$<0.001, with a medium effect size ($\eta^2_{partial}$ =0.162): i.e. the time-by-condition interaction accounted for around 16% of the variance in *score1*. This indicates that the intervention and comparison groups differed significantly in the extent to which their *score1* values changed between t1 and t2.

To interpret these results, the interaction graph (Figure 4.3) and descriptive statistics (Table 4.6) were consulted. These indicate that, although participants in both conditions showed an increase in mean *score1* values between the two times of testing, this increase was more pronounced for the intervention group. The two groups' scores did not differ significantly at t1, t(184)=.644, $p$=.520, but did differ significantly at t2, t(184)=4.508, $p$<.001. Indeed, the groups obtained near-identical mean scores at t1 (on average, 24.1% of target grapheme tokens were realized 'correctly' by intervention participants versus 23.4% by comparison participants), but at t2 there is a difference of 6.0 in the mean percentage of target grapheme tokens decoded 'correctly' by the two groups (31.5% for the intervention condition versus 25.5% for the comparison condition). Intervention participants decoded on average just over eight additional tokens 'correctly' compared to comparison participants. This equates to about three whole words on the RAT, since each RAT item contained, on average, just over two and a half target grapheme tokens

**Table 4.6 Results for *score1* (number of target graphemes realized 'correctly') out of 135**

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **median** | **Mean (% of total)** | **S.D.** | **max.** | **range** | **median** | **Mean (% of total)** | **S.D.** | **max.** | **Range** |
| **Intervention (N=112)** | 31 | 32.5 (24.1%) | 9.2 | 58 | 44 | 41 | 42.5 (31.5%) | 12.5 | 72 | 56 |
| **Comparison (N=74)** | 30.5 | 31.6 (23.4%) | 10.2 | 58 | 47 | 31.5 | 34.4 (25.5%) | 11.1 | 64 | 48 |

**Figure 4.3 Estimated marginal means of *score1* for intervention and comparison participants**




***Controlling for L2 proficiency***

To control for possible effects of baseline L2 proficiency, the ANOVA was re-run with the reading comprehension test score (section 3.6.2.3) as a covariate. This reduces the sample size slightly, since the reading comprehension test was not completed by all participants in the RAT sample. Further, as argued in section 3.6.2.3, it must be

acknowledged that the reading comprehension scores provide only a partial measure of French proficiency and one which is not entirely independent of the target variable (decoding proficiency).

The relevant assumptions of the test were first checked. Box's M test retained the null hypothesis that the observed covariance matrices of the dependent variables were equal across groups (Box's M = 5.085, F(3, 1283067.152)=1.674, *p*=.170). Levene's test confirmed that variances were homogenous for both levels of the repeated measures variable: at t1, F(1, 184)=2.301, *p*=.131; at t2, F(1, 184)=0.471, *p*=.493. Finally, the independence of the covariate from the treatment effect was confirmed. First, visual inspection of histograms and non-significant Shapiro-Wilk test results showed that the reading comprehension scores were not non-Normally distributed either for the sample as a whole or when subdivided by condition (Shapiro-Wilk W(176)=0.990, *p*=280 for the whole sample; W(108)=0.988, *p*=.462 for the intervention group; W(68)=0.821, *p*=.413 for the comparison group). A subsequent independent samples t-test found no significant difference between the two groups' reading comprehension scores (t(174) =0.821, *p*=.413).

The ANOVA found a significant interaction effect between the time of testing and the covariate, F(1,173)=5.105, *p*=.025. However, the effect size was small ($\eta^2_{partial}$ =0.029): that is, baseline reading comprehension scores accounted for only around 3% of the variance in *score1*. After controlling for this effect of baseline proficiency in French reading, there remained a significant time-by-condition interaction effect, F(1, 173)=30.724, *p*<.001. The effect size was again medium: $\eta^2_{partial}$ =0.151, i.e. accounting for 15.1% of the variance not attributable to other variables.

***Summary***

The above analysis suggests an affirmative answer to Research Question 1 (*Do the programmes of instruction lead to more accurate decoding of L2 French?*), at least in respect of those GPC explicitly covered in the instruction. Overall, compared to participants in the comparison condition, those who followed one of the programmes of instruction made significantly more progress in French decoding, as measured by the t1-to-t2 change in the number of target grapheme tokens decoded 'correctly' on the RAT. The significant difference between the groups remained even after the effects of baseline L2 (reading) proficiency were partialled out.

However, in the context of the total number of target grapheme tokens (135), the between-condition difference of 6.0 in the percentage of target grapheme types realized 'correctly' (Table 4.6) appears relatively small. So too does the magnitude of the *score1* increase shown by the intervention group (even though this was greater than that shown by the comparison group), as reflected in Figure 4.3 by the shallow slope of the upper line in relation to the full range of possible scores on the test (the reference line indicates the maximum possible score).

#### 4.3.3.2 Non-parametric tests

Due to the asymmetric distributions of *score1* (see above), the differences between the groups' scores were also analysed using non-parametric statistical techniques. First, the differences between each group's *score1* values at t1 and t2 were tested for significance

using Wilcoxon Signed Ranks tests. For both the intervention and comparison groups, the null hypothesis that the distributions of *score1* were the same at t1 and t2 was rejected (intervention group: Z=-8.539, $p$<0.001; comparison group: Z=-3.074, $p$=0.002). Given that both groups' mean scores are higher at t2 than at t1, this suggests that both intervention and comparison participants have made progress in French decoding. Next, the differences between the two groups' scores at each time of testing (t1 and t2) were tested for statistical significance using Mann-Whitney U tests. The null hypothesis that the distribution of *score1* is the same for both groups was retained at t1 (U =4109.5, Z=-.600, $p$=.549) but rejected at t2 (U=2840.5, Z=-4.018, $p$<.001). The Common Language (CL) effect size indicator (McGraw & Wong, 1992), used as an alternative measure of effect size with non-parametric tests (Leech & Onwuegbuzie, 2002), was P=0.67: i.e. if one intervention participant and one comparison participant were selected at random, there would be a 67% chance that the intervention participant produced a higher percentage of 'correct' realizations at t2, despite there being no significant differences between the groups' scores at t1. The non-parametric tests therefore support the findings of the ANOVA in providing an affirmative answer to Research Question 1.

### 4.3.4 'Acceptable' realizations (*score2*)

The development of fully accurate L2 decoding requires both (a) learning to apply L2-specific GPC and (b) developing accurate phonological representations of L2 phonemes. Section 2.6.4 argued that this second component, (b), may be particularly difficult for late-onset learners. Therefore, simply tallying participants' 'correct' pronunciations (*score1*) might underestimate their progress in French decoding. This is because some changes in component (a) – that is, in the extent to which they apply L2-specific GPC –

could go undetected due to inaccuracies in their representations of the relevant L2 phonemes. In the case of at least some graphemes types, participants would have to make progress on both components (a) and (b) in order to produce a 'correct' pronunciation. For example, pronouncing the first syllable of *goinfrez* as [gwan] – i.e. with an anglicised, 'unpacked' realization of the French nasal /wɛ̃/ (Čubrović, 2002) – is arguably 'better' than pronouncing it [gɔɪn] where English rather than French symbol-to-sound mappings appear to have been followed; however, both pronunciations would be judged incorrect under *score1*. This may explain the small difference observed above between intervention and comparison groups in terms of mean *score1* increases.

The more tolerant mark scheme (*score2*), crediting not only 'correct' but also 'acceptable' pronunciations (section 3.6.3.3), was therefore additionally used, to see whether this revealed (a) a greater extent of progress in L2 decoding across all participants and/or (b) a greater difference between the groups in terms of the progress made.

As in the case of *score1* (section 4.3.3), the distributions of *score2* for the subset of target graphemes were inspected as a preliminary to statistical analysis. Visual inspection of histograms and Q-Q plots revealed varying degrees of positive skew in the distributions of both variables at t1 and t2, both for the sample taken as a whole and when subdivided by condition. Shapiro-Wilk tests confirmed that the distributions were non-Normal in many cases (whole sample: $W(186)=0.983$, $p=0.021$ at t1, $W(186)=0.979$, $p=0.006$ at t2; intervention condition: $W(112)=0.980$, $p=0.085$ at t1, $W(112)=0.977$, $p=0.052$ at t2; comparison condition: $W(74)=0.963$, $p=0.031$ at t1, $W(74)=0.953$, $p=0.008$ at t2). Logarithmic transformation of the scores improved the Shapiro-Wilk statistics in some

(but not all) cases, but visual inspection of histograms still revealed varying degrees of skewness. Again, therefore, the outcomes of the planned parametric analysis using a two-way mixed ANOVA procedure (section 4.3.4.1) were triangulated with those of subsequent non-parametric tests (section 4.3.4.2).

#### 4.3.4.1 ANOVA

Following the procedure in section 4.3.3.1, *score2* values for target graphemes were analysed using a two-way mixed factorial ANOVA with one within-subjects factor (time) and one between-subjects factor (condition). The covariate (t1 reading comprehension score) was not included in the initial analysis, in order to maximize the sample size. First, the relevant assumptions of the test were checked. Sphericity was disregarded since the repeated measures variable had only two levels. Box's M test found that observed covariance matrices of the dependent variable did not differ between groups when $\alpha<.001$ (Box's M=5.728, F(3, 1283067.152)=1.885, $p$=.130). Levene's test confirmed homogeneity of variances for *score2* at t2 (F(1,184)=.002, $p$=.961) but not at t1 (F(1,184)=6.097, $p$=.014). Using logarithmically transformed values of *score2* did not improve the homogeneity of variances. This means that the accuracy of the F test for comparisons between the conditions is compromised; results of the ANOVA must therefore be treated cautiously.

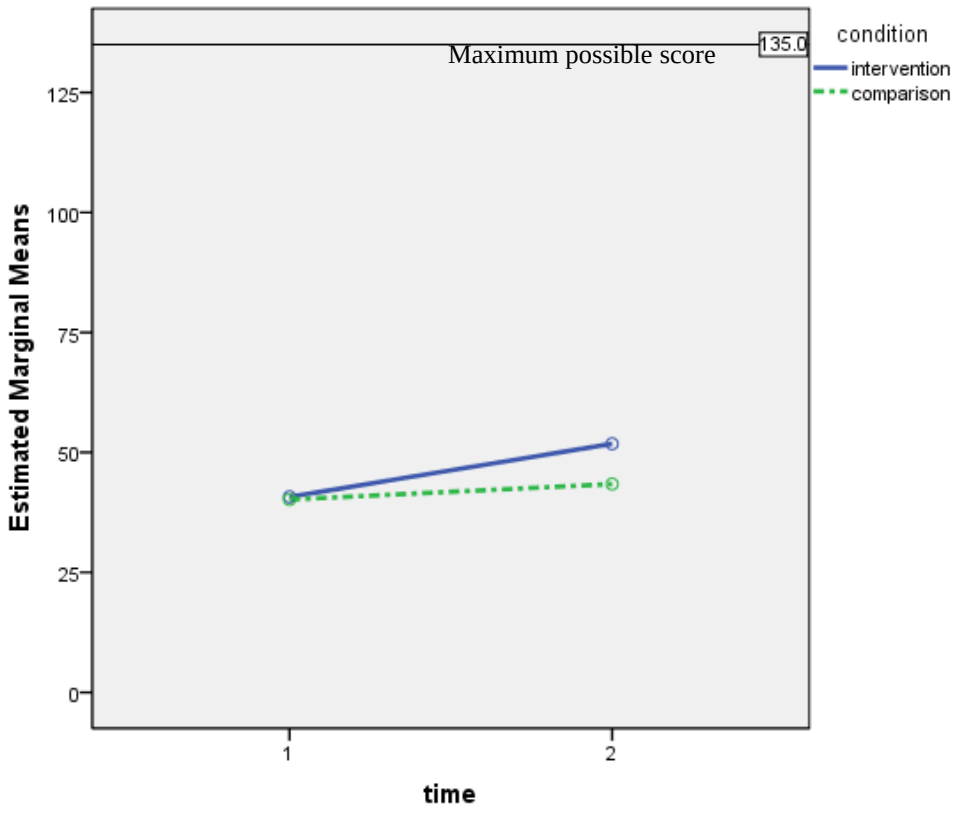
As in the case of *score1*, the ANOVA analysis found significant main effects of time, F(1,184)=117.305, $p$<0.001, and condition, F(1,184)=6.133, $p$=.014. A significant time-by-condition interaction was also found, F(1,184)=35.385, $p$=<.001; the effect size was medium ($\eta^2_{partial}$ =.161), indicating that the interaction accounted for around 16% of the

variance in *score2* not attributable to other variables. In interpreting these results, the interaction graph (Figure 4.4) and descriptive statistics (Table 4.7) again show that, whilst participants in both conditions showed an increase in mean *score2* values between the two times of testing, this increase was more pronounced for the intervention group. The two groups' scores were found to differ significantly at t2, $t(184)=3.896$, $p<.001$, but not at t1, $t(184)=.333$, $p=.740$. Indeed, they again obtained near-identical scores at t1 (on average, 30.1% of realizations were 'acceptable' for intervention participants versus 29.8% for comparison participants), but at t2 the mean proportion of 'acceptable' realizations differs by 6.3 percentage points between the two conditions (38.4% for the intervention group versus 32.1% for the comparison group). On average, intervention participants decoded almost eight and a half additional tokens 'acceptably' compared to comparison participants, the equivalent of about three whole RAT items.

**Table 4.7 Results for *score2* (number of target graphemes realized 'acceptably') out of 135**

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **median** | **Mean (% of total)** | **S.D.** | **max.** | **range** | **median** | **Mean (% of total)** | **S.D.** | **max.** | **Range** |
| **Intervention (N=112)** | 40 | 40.7 30.1%) | 10.4 | 68 | 54 | 50 | 51.8 (38.4%) | 14.7 | 86 | 66 |
| **Comparison (N=74)** | 38.5 | 40.2 (29.8%) | 12.0 | 64 | 47 | 40.5 | 43.4 (32.1%) | 13.8 | 78 | 57 |

**Figure 4.4 Interaction graph for effects of time and condition on *score2*.**




As in the analysis of 'correct' pronunciations, the ANOVA was also re-run with baseline L2 reading comprehension scores as a covariate. Having checked that the assumptions of the test were met (Box's M=4.838, F(3, 871503.334)=1.591, $p$=.189; Levene's F(1,174)=1.156, $p$=.284 at t1 and F(1,174)=0.572, $p$=.450 at t2), the ANOVA found a significant but small interaction effect between the time of testing and the covariate, F(1,173)=13.092, $p$<.001, $\eta^2_{partial}$ =.070, and a significant, medium-sized interaction effect between time and condition, F(1,173)=30.274, $p$<.001, $\eta^2_{partial}$ =.149. Thus, after partialling out the effects of baseline L2 reading proficiency, intervention participants still showed significantly greater increases in 'acceptable' realizations compared to participants in the comparison group. The effect size remained medium, accounting for 14.9% of the variance in *score2* not attributable to other variables.

In summary, the above analysis of *score2* again suggests an affirmative answer to Research Question 1: the programmes of instruction do appear to have led to higher numbers of 'acceptable' realizations of those grapheme tokens explicitly covered in the instruction. However, the magnitude of the score increases is again small in relation to the overall number of target grapheme tokens. Thus, there does not appear to have been a much larger increase in 'acceptable' than 'correct' realizations. This is shown by the very similar gradients of the upper lines in Figures 4.4 and 4.3; indeed, the interaction graph for *score2* appears almost identical to that for *score1* except that all points in Figure 4.4 lie slightly higher on the Y axis, reflecting the greater leniency of the *score2* mark scheme. The relationship between *score1* and *score2* will be investigated further in section 4.3.4.3.

#### 4.3.4.2 Non-parametric tests

Intervention and comparison participants' *score2* values were also subjected to non-parametric statistical analysis. First, a Wilcoxon Signed Ranks test found significant differences between the scores obtained by each group at t1 and t2 (intervention group: Z=-8.525, $p<0.001$; comparison group: Z=-3.412, $p=0.001$), indicating that both groups had made progress as measured by numbers of 'acceptable' realizations of grapheme tokens. Next, at each time of testing (t1 and t2), the differences between the two groups' scores were compared. The null hypothesis that the distribution of *score2* was the same for both groups was retained at t1 (Mann-Whitney U =4166, Z=-0.447, $p=.655$) but rejected at t2 (Mann-Whitney U=3005, Z=-3.575, $p<.001$). The CL effect size statistic indicated a 65% likelihood that any randomly-selected intervention participant produced

a higher number of 'acceptable' realizations at t2 than any randomly-selected comparison participant.

This analysis again supports the findings of the parametric tests. Intervention participants were found to obtain significantly higher *score2* values than comparison participants at t2, despite obtaining almost identical scores at t1. This suggests that participants who followed one of the programmes of instruction made significantly more progress in French decoding than those who did not.

#### 4.3.4.3 Comparing *score1* and *score2*

Following the argument at the start of section 4.3.4, participants who made progress in French decoding proficiency were expected to show a greater increase on the more lenient *score2* than on the more stringent *score1* (which is a subset of *score2*), resulting in a higher proportion of pronunciations judged 'acceptable' but not 'correct' (henceforth 'only acceptable')[25]. To test this hypothesis, the distributions of 'only acceptable' scores (calculated by subtracting *score1* from *score2*) were first checked for Normality. Visual inspection of histograms suggested non-Normal distributions at t1 and t2, both for the sample as a whole and when subdivided by condition; this was confirmed in some cases by significant Shapiro-Wilk test results (whole sample: $W(186)=0.984$, $p=.030$ at t1; $W(186)=0.951$, $p<.001$ at t2; intervention group: $W(112)=0.987$, $p=.364$ at t1; $W(112)=0.933$, $p<.001$ at t2; comparison group: $W(74)=0.972$, $p=.106$ at t1; $W(74)=0.965$, $p=.039$ at t2). Subsequent non-parametric (Wilcoxon Signed Ranks) tests

[25] Note that the 'only acceptable' classification applies only to oral vowels and nasal vowels. For consonants, *score1* and *score2* values are identical, since no distinction is made between 'acceptable' and 'correct' realizations (section 6.3.3.3).

found that the distributions of 'only acceptable' realizations differed significantly between t1 and t2 in the intervention group (Z(112) =-2.593, *p*=.010) but not in the comparison group (Z(74)=-.902, *p*=.367).

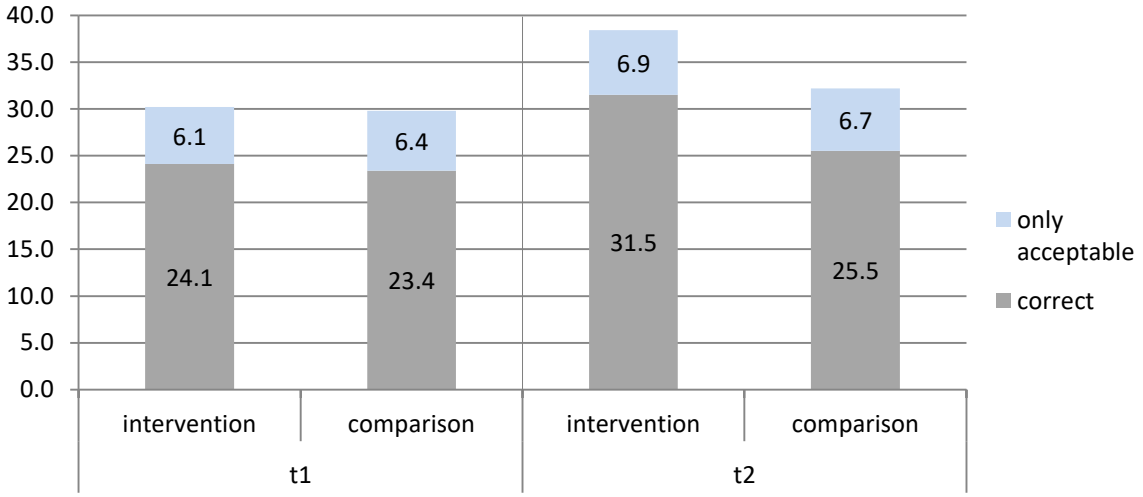
Table 4.8 shows the numbers of 'only acceptable' realizations produced by each group at t1 and t2; Figure 4.5 provides a graphical representation of the relationship between 'correct', 'acceptable' and 'only acceptable' realizations. As predicted, intervention participants show a greater increase in the numbers of tokens realized 'acceptably' compared to those realized 'correctly', resulting in a higher proportion of 'only acceptable' realizations at t2 (6.9%) than at t1 (6.1%). However, the magnitude of the difference is small. The comparison group also shows a very slight, non-significant increase in the percentage of 'only acceptable' realizations (even smaller than that shown by the intervention group).

The very small increases in 'only acceptable' realizations may simply reflect the fact that the t1-to-t2 increases in RAT scores occurred mainly within the category of consonants, for which the 'only acceptable' classification does not apply (see footnote 25), with correspondingly smaller increases in the oral and nasal vowel categories. This was indeed found to be the case: the mean t1-to-t2 increase in the number of 'acceptable' realizations across all participants was 5.0 for consonantal GPC but only 2.4 and 0.3 for oral and nasal vowels respectively.

**Table 4.8 Mean numbers and percentages of target grapheme tokens whose realizations were 'only acceptable' (i.e. *score2-score1*) out of 135**

| | **t1** | **t2** |
|---|---|---|
| **Intervention (N=112)** | 8.2 (6.1%) | 9.3 (6.9%) |
| **Comparison (N=74)** | 8.6 (6.4%) | 9.0 (6.7%) |

**Figure 4.5 Mean percentages of target grapheme tokens whose realizations were (a) 'acceptable' (*score2*: overall column height); (b) 'correct' (*score1*: lower portion of each column); (c) 'only acceptable' (*score2-score1*: upper portion of each column)**




## 4.4 Comparing the two interventions

### 4.4.1 Introduction

As a first step to addressing Research Question 2 (*Is one of the programmes of instruction more effective than the other?*), this section compares the scores obtained by participants in each intervention condition: Programme A (the 'poem approach') versus Programme B (the 'phonics approach'). To provide a clearer overview of the data, results are presented only for the more inclusive variable *score2* ('acceptable' realizations), argued above to offer a more complete view of progress than *score1* ('correct' realizations). In any case, section 4.3.4.3 found that the difference between these two variables showed only a slight increase between the two times of testing, statistically significant for the intervention group only.

Table 4.9 shows the numbers and percentages of 'acceptable' realizations of target grapheme tokens (*score2*) by participants in each condition. Since the non-Normal distributions of this variable at both times of testing have already been demonstrated for the intervention condition as a whole, the same approach to statistical analysis was adopted as in preceding sections.

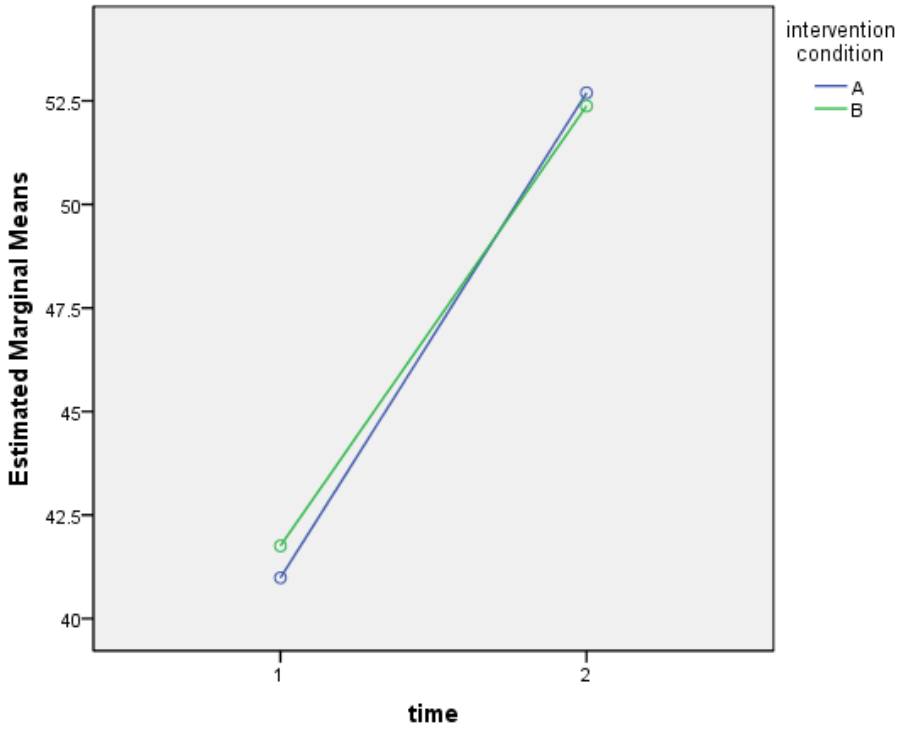
**Table 4.9 Results for *score2* (number of target graphemes realized 'acceptably') out of 135**

| | t1 | | | | | t2 | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | **median** | **Mean (% of total)** | **S.D.** | **max.** | **range** | **median** | **Mean (% of total)** | **S.D.** | **max.** | **Range** |
| **Intervention A (N=68)** | 40 | 41.0 (30.4%) | 11.4 | 69 | 55 | 50 | 52.7 (39.0%) | 15.5 | 95 | 75 |
| **Intervention B (N=44)** | 39 | 41.8 (30.9%) | 10.7 | 75 | 53 | 50 | 52.4 (38.8%) | 16.3 | 97 | 77 |

### 4.4.2 Two-way mixed ANOVA

The assumptions of the test were found to be met. Box's M test indicated equal covariance matrices of the dependent variable across the two groups at the level $\alpha<.001$ (Box's M = 9.559, F(3, 432085.832)=3.121, $p$=.025). Levene's test confirmed that variances were homogenous for both levels of the repeated measures variable: at t1, F(1,112)=.061, $p$=.805; at t2, F(1,112)=0.014, $p$=.906. The ANOVA analysis found a significant and large main effect of time (F(1,112)=158.864, $p$<0.001, $r$=.77) but no significant effect either for condition (F(1,112)=.008, $p$=.927) or for the interaction between time and condition (F(1,112)=0.377, $p$=.540). This is reflected in the very similar gradients of the two lines in the interaction graph (Figure 4.6). Both groups have made similar progress in French decoding between t1 and t2, as measured by the numbers of target grapheme tokens realized 'acceptably' on the RAT.

**Figure 4.6 Interaction graph for effects of time and condition on *score2***




### 4.4.3 Non-parametric tests

The non-parametric tests supported the findings of the ANOVA. First, Wilcoxon Signed Ranks tests found statistically significant differences between the distributions of each group's scores at the two times of testing (t1 and t2), indicating that both groups made progress in French decoding over the period of the intervention (intervention A: Z=-6.535, $p<0.001$; intervention B: Z=-5.529, $p<0.001$). Next, Mann-Whitney U tests found no significant differences between the scores obtained by each group either at t1 or at t2 (t1: U=1534.00, Z=-0.107, $p=0.915$; t2: U=1506.00, Z=-0.270, $p=0.787$).

### 4.4.4 Summary

Both analyses concur in indicating a negative answer to Research Question 2. Participants who followed each of the two programmes of instruction appear to have made similar progress in French decoding, as measured by the numbers of 'acceptable' pronunciations of target graphemes. Neither programme appears to have been more effective than the other in improving decoding outcomes.

## 4.5 Class-level differences

The preceding analysis has compared broad groups of participants at the condition level (Intervention A, Intervention B, Comparison). However, the decoding instruction (or, in the comparison condition, the regular teaching) was implemented at the level of individual classes. This section therefore examines scores at the individual class level in order to gain a more differentiated picture of participants' progress. Did the decoding instruction (or the regular teaching) affect all classes within a given condition evenly in respect of their French decoding, or did some classes show a greater or lesser degree of progress than others in the same condition? Whilst the available background data give no reason to expect strong effects caused by the characteristics of pupils in the classes (as evidenced by BIQ and academic data) or by the nature of the teaching they received (as evidenced by TIQ and lesson observation data), it is nonetheless possible that other, undetected variables may have affected outcomes.

The *score2* values obtained for target graphemes by each individual class were compared at both times of testing (Appendix 9.1). Since each class is here considered on its own

terms, rather than as part of wider between-condition comparisons, results are also presented and analysed for classes B3 and D1 (henceforth referred to as 'excluded classes' as opposed to 'contributing classes', because the former were excluded from the main analysis reported previously). Given that previous sections have shown non-Normal distributions of *score2* for larger groupings of participants, and given the unequal and sometimes small numbers of participants in the various classes (Table 4.10, column 1), non-parametric procedures were employed.

First, in line with the findings in section 4.3.4, Wilcoxon Signed Ranks tests found that all five contributing intervention classes as well as B3 obtained significantly higher scores at t2 than at t1 (class A1: Z=-3.896, $p$=0.000; A2: Z=-4.066, $p$=0.000; A3: Z=-3.249, $p$=0.001; B1: Z=-3.850, $p$=0.000; B2: Z=-3.964, $p$=0.000; B3: Z=-2.674, $p$=0.007). However, in the comparison condition, the scores of only three individual classes (including D1) were significantly higher at t2 than at t1; those of C1, C3 and D2 were not (C1: Z=-1.070, $p$=0.285; C2: Z=-2.366, $p$=0.018; C3: Z=-0.842, $p$=0.400; D1: Z=-2.575, $p$=0.010; D2: Z=-0.654, $p$=0.513; D3: Z=-2.514, $p$=0.012). Therefore, not all classes in the comparison group appear to have made progress in French decoding.

Next, Kruskal-Wallis tests found that, in contrast to the main analysis presented above (sections 4.3.3 and 4.3.4), there were significant differences between the ten contributing classes' scores, not only (as expected) at t2 (H(9)=31.305, $p$<.001) but also at t1 (H(9)=17.823, $p$=.037). This indicates that there may have been baseline differences between individual classes' French decoding proficiency. Subsequent pairwise Mann-Whitney U comparisons (summarized in Appendices 9.2-3) produced both some findings which supported the main analysis and others which did not, including the following: (a)

at t1, there were significant differences between the scores of comparison class D3 and almost all other classes; (b) at t2, there were significant differences within the comparison condition between class D3 and most other classes; (c) at t2, there was also a lack of significant differences between some intervention classes and the comparison class D3. These findings reflect the fact that D3 obtained markedly higher scores than all other classes at both t1 and t2 (Appendix 9.1).

Based on these class-level exceptions to the condition-level patterns found in sections 4.3.3 and 4.3.4, an ANCOVA was planned to check that the main effect of condition and the time-by-condition interaction effect remained significant after controlling for t1 scores. However, this was not possible, because the assumption of homogeneity of regression slopes was not met (indicated by a significant interaction of condition and t1 score, $F(1,186)=4.640$, $p=.033$). Therefore, an alternative analysis was employed where classes were compared on their t1-to-t2 gain scores. This was done by calculating the difference between each participant's *score2* values (i.e. their number of 'acceptable' realizations) at t1 and t2, creating a new variable, *change2* (Table 4.10; Figure 4.7).

Although not all comparison classes obtained significantly higher scores at t2 than at t1 (see above), clear condition-level patterns nonetheless emerge which are consistent with the findings in section 2. When the classes are ranked from highest (1) to lowest (12) according to their mean *change2* values (Table 4.10, columns 7 and 14), the intervention schools A and B jointly contain five of the six classes with the highest mean gain scores. Further, the exception to this pattern is the excluded class B3, whose lower increase in *score2* relative to the other intervention classes is unsurprising given its lower L1 literacy levels (section 4.2.1.3). (Indeed, it could be argued that B3 shows a larger gain score

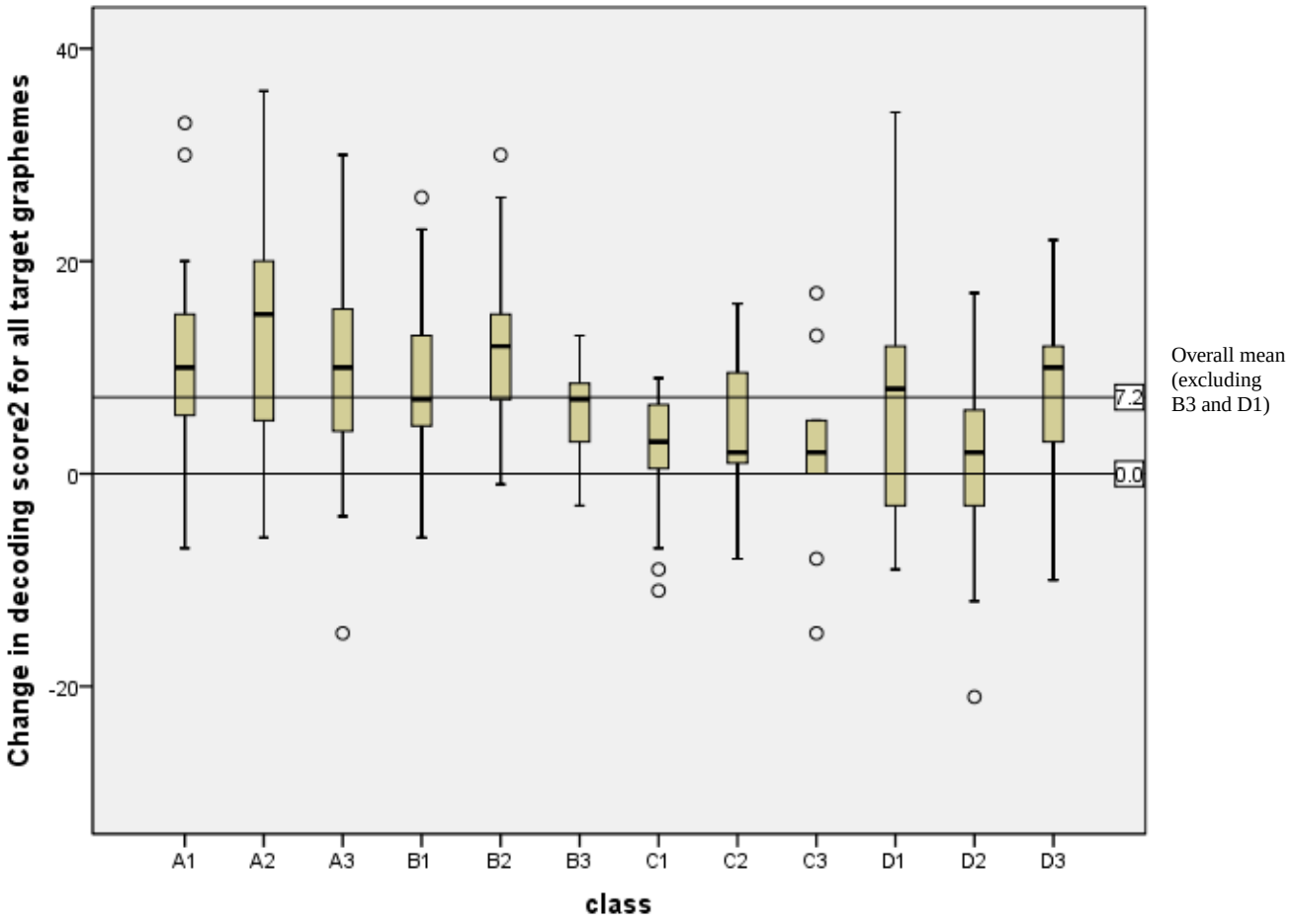
than might be expected based on its prior attainment data, ranking eighth on this measure). Conversely, the comparison schools C and D contain five of the six classes with the lowest mean *change2* values. The exception in this case is D3 (which obtained the highest scores at both t1 and t2); however, its mean gain score, whilst markedly higher than those of the other contributing comparison classes, is also noticeably lower than those of all *contributing* intervention classes. (In the comparison condition, only the excluded class B3 ranked lower). D1 also obtains a gain score almost as high as that of D3, but this can be explained by the separate programme of decoding instruction which this class appears to have followed (section 3.7.2).

These condition-level patterns in the classes' gain scores are also visible in Figure 4.7 in the relationship between the distributions of *change2* on the one hand and the mean value for the ten contributing classes on the other, indicated by the upper of the two reference lines.

**Table 4.10 Results for *change2* (difference in the number of target grapheme tokens pronounced 'acceptably' between t1 and t2) out of 135**

| | intervention condition | | | | | | | comparison condition | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **class** | **median** | **mean (% of total)** | **S.D.** | **max.** | **range** | **rank** | **class** | **median** | **mean (% of total)** | **S.D.** | **max.** | **range** | **rank** |
| **A1 (N=24)** | 10 | 10.5 (7.8%) | 9.2 | 33 | 40 | 4 | **C1 (N=15)** | 3 | 1.7 (1.3%) | 6.2 | 9 | 20 | 11 |
| **A2 (N=25)** | 15 | 13.7 (10.1%) | 9.8 | 36 | 42 | 1 | **C2 (N=15)** | 2 | 4.9 (3.6%) | 6.9 | 16 | 24 | 9 |
| **A3 (N=20)** | 10.5 | 10.7 (7.9%) | 10.8 | 30 | 45 | 3 | **C3 (N=10)** | 2 | 2.1 (1.6%) | 9.2 | 17 | 32 | 10 |
| **B1 (N=25)** | 7 | 9.3 (6.9%) | 8.4 | 26 | 32 | 5 | **D1 (N=18)** | 8 | 6.9 (5.1%) | 10.4 | 34 | 43 | 7 |
| **B2 (N=21)** | 12 | 12.2 (9.0%) | 7.9 | 30 | 31 | 2 | **D2 (N=21)** | 2 | 0.8 (0.6%) | 8.5 | 17 | 38 | 12 |
| **B3 (N=12)** | 7 | 5.5 (4.1%) | 4.9 | 13 | 16 | 8 | **D3 (N=15)** | 6 | 7.1 (5.3%) | 9.0 | 22 | 32 | 6 |

**Figure 4.7 Distributions of *change2* for each class.** *Reference lines indicate (a) zero and (b) the overall sample mean (excluding classes B3 and D1) of 7.2.*




The differences between the classes' *change2* values were tested for statistical significance. In interpreting the results of the statistical comparisons, the unequal and sometimes low numbers of students in the individual classes should again be borne in mind. Despite non-significant results of Shapiro-Wilk tests, visual inspection of histograms and Q-Q plots again indicated non-Normal distributions of *change2* at the individual class level. Non-parametric tests were therefore used. A Kruskal-Wallis test found significant differences between the twelve classes' *change2* values ($H(11)=43.328$, $p<0.001$). Subsequent pairwise Mann-Whitney U comparisons (summarized in Appendix 9.4) produced the following findings at the significance level of $\alpha<.05$. First, no significant differences were found (a) between any classes within the comparison

condition or (b) between any classes within the intervention condition, except that the excluded class B3 (as expected) obtained significantly lower values than A2 and B2. By contrast, of the thirty-six individual pairwise comparisons between intervention classes on the one hand and comparison classes on the other, twenty found statistically significant differences. Further, six of the sixteen non-significant differences involved the excluded class B3, whilst a further three approached significance. However, there were also some exceptions to these patterns, most notably in respect of three comparison classes: (a) the *change2* values of C2 showed no significant differences from those of either A1 or B1; (b) the *change2* values of D3 (and of the excluded class D1) showed no significant differences with those of any other classes except A2. (It is possible that at least some of these few inconsistencies to the overall pattern of significant differences would not have occurred with larger sample sizes).

In conclusion, it appears that the most important factor in determining the classes' mean progress in French decoding (as measured by t1-to-t2 gain scores) is their allocation to either the intervention or comparison condition; it seems unlikely that such consistent condition-level patterns in the classes' scores would have emerged as a result of other, confounding factors[26]. However, it is also clear from Figure 4.7 – and from the standard deviation and range values in Table 4.10 – that many classes in both the intervention and comparison conditions show a wide spread of values on the variable *change2*: i.e. in both conditions, some participants in each class appear to have made considerably more progress in French decoding than others. Further, the relationship of the distributions in Figure 4.7 to the lower of the two reference lines (which indicates zero) shows that some

[26] Considering the available background information on participating teachers (section 3.4.2.2), one potentially confounding variable is their differing levels of teaching experience: teachers in schools A and B had, on average, been teaching for longer than those in schools C and D. This would have been avoided if possible (section 3.3). Against this view, the class of teacher A1 (who had joint least teaching experience) ranked fourth in terms of *change2* values. Nonetheless, this possibility cannot be ruled out.

participants in every class obtained negative *change2* values: i.e. their numbers of 'acceptable' realizations actually *decreased* between the two administrations of the RAT. Indeed, the data reveal that in most classes, at least one participant produced between ten and twenty *fewer* 'acceptable' realizations at t2 than at t1. However, a condition-level pattern again emerges here: the proportion of students obtaining negative *change2* values is generally higher in comparison than in intervention classes (Table 4.11, columns 4-5): ranking the classes from lowest to highest in terms of the proportion of students with negative *change2* values shows that five of the top six (plus the seventh) all belong to the intervention condition. Nonetheless, it must be acknowledged that even in the intervention classes, considerable numbers of participants show decreases in the numbers of grapheme tokens realized 'acceptably'.

**Table 4.11 Numbers of participants obtaining negative *change2* values**

| **Class** | **N** | **Number and percentage of participants obtaining negative *change2* values** | | **Rank (1=lowest percentage of participants obtaining negative values)** |
|---|---|---|---|---|
| **A1** | 24 | 3 | 12.5% | 5 |
| **A2** | 25 | 2 | 8.0% | 2 |
| **A3** | 20 | 2 | 10.0% | 3 |
| **B1** | 25 | 3 | 12.0% | 4 |
| **B2** | 21 | 1 | 4.8% | 1 |
| **B3** | 12 | 2 | 16.7% | 7 |
| **C1** | 15 | 3 | 20.0% | 8= |
| **C2** | 15 | 3 | 20.0% | 8= |
| **C3** | 10 | 2 | 20.0% | 8= |
| **D1** | 18 | 5 | 27.8% | 11 |
| **D2** | 21 | 8 | 38.1% | 12 |
| **D3** | 15 | 2 | 13.3% | 6 |

## 4.6 Analysis by grapheme type

### 4.6.1 Introduction

Sections 4.3 to 4.5 compared groups in terms of their overall RAT scores. The current section pursues a different line of analysis by considering the progress made by participants in each condition (Intervention A, Intervention B, Comparison) on individual grapheme types (e.g. <é>, <th>) rather than on the RAT as a whole. This provides a more nuanced answer to Research Questions 1 and 2: it may be that any t1-to-t2 gains in RAT scores for a given group of participants are not homogeneous across all grapheme tokens in the test. Some grapheme types may show greater improvement than others.

For each grapheme type, the percentages of all its tokens realized 'correctly' (*score1*) and 'acceptably' (*score2*) by participants in a given group were calculated at both t1 and t2. The difference between these figures indicates the change in percentages of 'correct' and 'acceptable' realizations (*change1* and *change2* respectively). These gain scores, which have a theoretical range of -100 to +100, provide an indication of the progress made by participants in respect of each individual grapheme type[27]. Given the scope for human error involved in calculating gain scores for so many individual grapheme types (this was done manually on the basis of the gain scores of the constituent tokens), accuracy was checked by recalculating all figures separately by an alternative method,

[27] Missing values – resulting from inaudible data, omitted items or Non-Linear Decoding (NLD) – are included in this analysis and are counted as incorrect realizations.

comparing the outcomes and resolving any discrepancies[28]. The analysis is again based on *change2* values only, since previous analysis suggested little difference overall between the numbers of 'correct' and 'acceptable' realizations; *change1* values may be analysed in subsequent, follow-up reports. Finally, whilst patterns in the data are highlighted, statistical tests are not systematically applied in this section because of (a) the unmanageable number of individual comparisons required and (b) the limited range of data on which comparisons between many individual grapheme types are based. Much of the analysis in this section must therefore be treated as tentative.

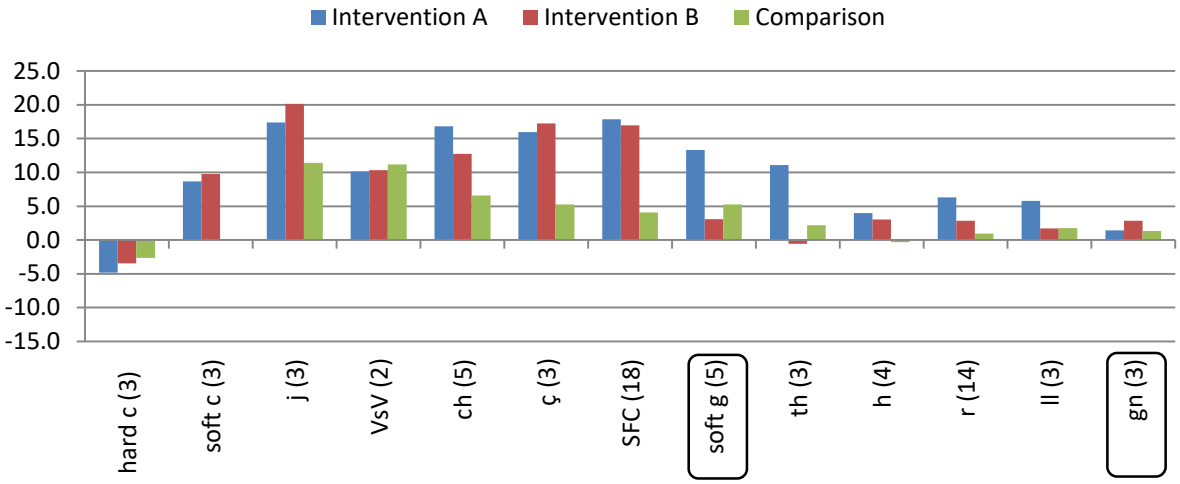
For the participants in each condition, Figure 4.8 provides a graphical representation of *change2* values for individual grapheme types in the categories of (a) consonants, (b) oral vowels and (c) nasal vowels; these broad articulatory categories are used to simplify the presentation of data. Rounded boxes indicate those grapheme types excluded from the subset of target graphemes. Mean changes in raw scores are not shown: percentage figures offer a more reliable basis for comparing grapheme types with different numbers of tokens. However, given the differences between the numbers of tokens exemplifying each grapheme type (as shown in brackets in Figure 4.8), it must be acknowledged that the figures for some grapheme types are more robust than for others: for example, the fourteen instantiations of <r> across a range of different contexts provide a more reliable basis for estimating how accurately participants 'usually' realize this grapheme type than do the three tokens exemplifying <ch>, where there is a stronger possibility that realizations influenced by their particular orthographic context will affect the overall score.

[28] The alternative method involved calculating the mean number of 'correct' and 'acceptable' realizations produced by each group of participants at t1 and t2, averaging these figures across all the tokens within a given grapheme type, and subtracting the t1 values from the t2 values. Comparing the outcomes of the two sets of calculations revealed two discrepancies (due to human error), which were resolved.

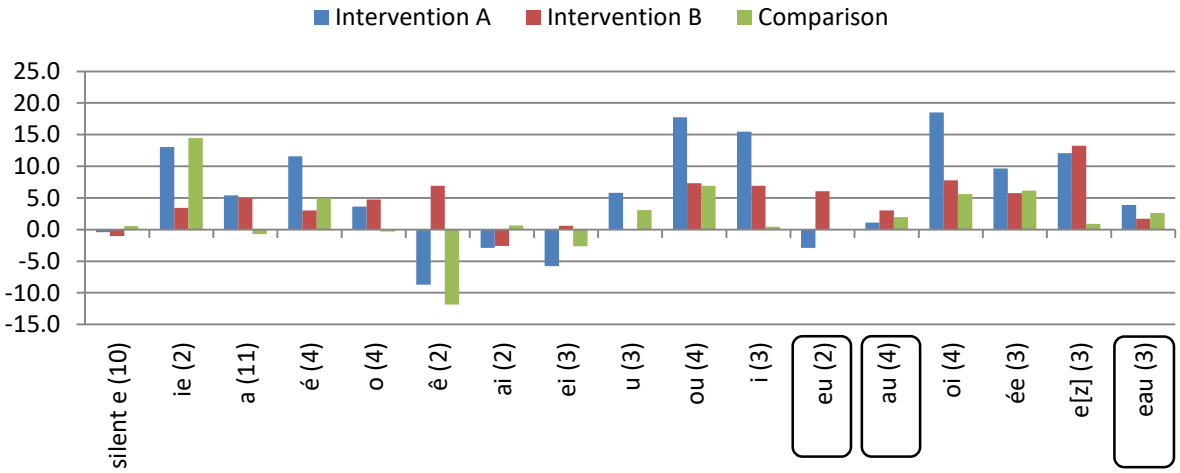
Clearly, when comparing grapheme types in terms of their gain scores, the mean scores at t1 must also be taken into consideration, since this baseline measure determines the maximum possible extent of the score increase at t2: for example, a grapheme type realized 'acceptably' 98% of the time at t1 would show less scope for improvement than one which was realized 'acceptably' only 10% of the time. (Such extreme variation is however unlikely, since the RAT mark scheme includes only those grapheme types which participants in Woore (2006) decoded with less than 75% accuracy: section 3.6.3.3). An additional possibility is that the grapheme types generating the lowest scores at t1 – being the most difficult to decode – will also be those which resist the effects of instruction most stubbornly. To make any such patterns more visible, the grapheme types within each articulatory category were ranked from highest to lowest according to their mean *score2* value across all participants at t1 (Appendix 9.5); the individual charts in Figure 4.8 then follow this sequence from left to right.

**Figure 4.8 Changes in mean percentages of 'acceptable' realizations of grapheme types.** *The number of tokens of each grapheme type appears in brackets. Rounded boxes enclose grapheme types excluded from the subset of target graphemes.*

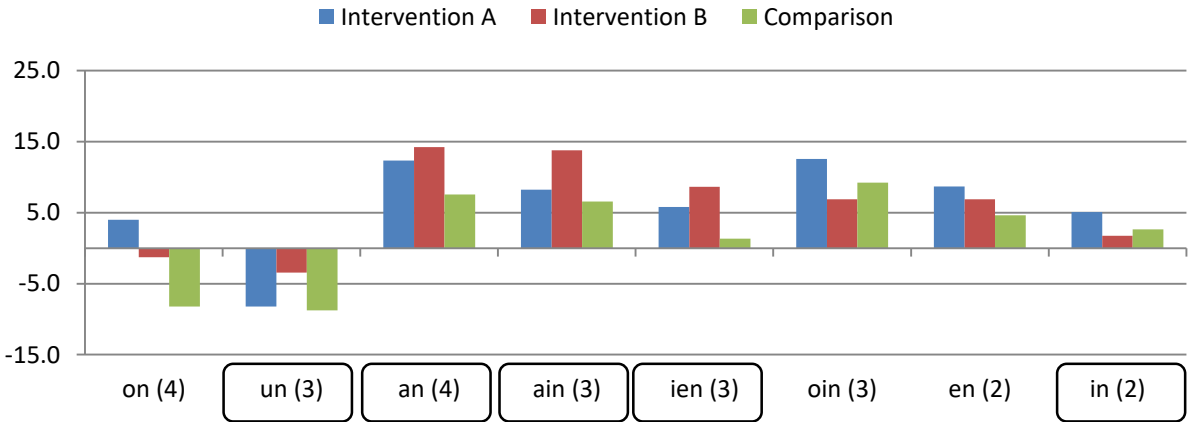
**(a) Consonants**




**(b) Oral vowels**




**(c) Nasal vowels**



### 4.6.2 Consonants

It must again be emphasized that much of the analysis in this section operates at a descriptive level (i.e. in most cases, tests of statistical significance are not conducted) and therefore remains tentative. Nonetheless, a relatively consistent pattern of positive gain scores appears to emerge in respect of consonantal grapheme types (Figure 4.8a). Only in three cases does any group show a negative change in the percentage of 'correct' realizations, as discussed later in this sub-section. There also appear to be fairly consistent relationships between the gain scores of the various conditions. For example, intervention groups A and B both show greater increases than the comparison condition in respect of seven out of thirteen grapheme types. In five of these cases (soft <c>, <j>, <ch>, <ç>, SFC) the difference in gain scores appears to be considerable, whilst for two others (<h> and <r>) it is less pronounced.

The pattern of results for the SFC 'rule' can be seen as particularly robust due to the large number of tokens exemplifying it (N=14), sufficient in this case to permit statistical comparisons. The distributions of *score2* for SFC were first examined. Visual inspection of histograms and significant Shapiro-Wilk test results showed that the distributions differed from Normality both for the sample as a whole ($W(190)=0.983$, $p=0.019$) and when subdivided by condition. Non-parametric tests were therefore used. An independent-samples Kruskal-Wallis test rejected the null hypothesis that the distribution of *change2* was the same across categories of school ($H(2)=19.492$, $p<0.001$). Subsequent pairwise (Mann-Whitney U) comparisons found significant differences between the comparison group on the one hand and both intervention A and intervention B on the other ($U=1776$, $Z=-3.365$, $p=0.001$ and $U=955.5$, $Z=-4.064$,

*p*<0.001 respectively) but not between the intervention groups A and B (U=1421, Z=-0.764, *p*=0.445).

By contrast, there are only two grapheme types for which the comparison group appeared to show a more positive score change than both of the intervention groups (hard <c> and intervocalic <s>) and in each case, the difference in gain scores between the conditions is slight. For a further two grapheme types, the lower of the two intervention groups' *change2* values is roughly equal to that of the comparison condition (<gn>, <ll>), whilst in two other cases (<th> and soft <g>), intervention group A obtains the highest gain scores, with the comparison group marginally outperforming intervention group B. However, two of these grapheme types (<gn> and soft <g>) do not belong to the subset of target graphemes covered by all groups in the programmes of instruction. These descriptively observed patterns tentatively suggest that, overall, explicit decoding instruction may have had a fairly consistent positive effect on participants' decoding of consonantal grapheme types.

As well as varying between conditions, the gain scores associated with individual grapheme types also vary considerably from one grapheme type to another, ranging from +20.1 to -4.8 percentage points' difference in the proportion of tokens pronounced 'acceptably'. Thus, behind the overall score increases in respect of consonants as a whole lies a complex picture, where some grapheme types are associated with much larger improvements in decoding accuracy than others. Four of the grapheme types identified above as possibly showing a consistent advantage for the intervention schools (<j>, <ch>, <ç>, SFC) show increases of between thirteen and twenty percentage points in the proportion of tokens realized 'acceptably' by intervention participants. However,

in other cases the gain scores for the intervention groups are much smaller (e.g. around +10 for soft <c>; between +2 and +6 for <h>, <r> and <ll>) or even negative, as in the case of hard <c>.

Some of these apparent between-grapheme differences may be partly attributable to a broad trend of diminishing score increases with decreasing t1 scores, as shown by a visual appraisal of Figure 4.8a. At the very least, it may be observed that the four grapheme types on the right showing the smallest increases in 'acceptable' realizations for the intervention groups (<h>, <r>, <ll>, <gn>) are also those with very low mean scores at t1 (Appendix 9.5); <th>, which shows higher gain scores for only one of the intervention groups, also has a much lower t1 score than the other grapheme types to its left. Further, the apparent exception to this pattern (soft <g>, which shows a similar pattern of gain scores to <th>, even though its larger t1 score is similar in magnitude to those of its neighbours SFC and <ç>) is not one of the target graphemes covered by all participants in the programmes of instruction.

However, an obvious exception to this general trend is provided by the two contextually-dependent realizations of <c> on the left of Figure 4.8a: 'soft' <c> (pronounced /s/ when followed by <i>, <e> or <y>) shows smaller increases in correct realizations for all groups than most of the grapheme types immediately to its right; 'hard' <c> (pronounced /k/ when followed by <a>, <o> or <u>) actually shows a decrease in all three groups. Further, the intervention groups A and B show much higher increases than the comparison condition in terms of 'acceptable' realizations of soft <c> but also greater decreases in respect of hard <c>.

### 4.6.3 Vowels

In respect of grapheme types in the vowel category, a less coherent picture emerges than for consonants (Figure 4.8b). It is immediately striking that the gains in 'acceptable' pronunciations range more widely (from +23.8 to -8.7 percentage points), with negative values occurring in all groups and in respect of seven of the seventeen grapheme types. In contrast to the consonant category, a visual appraisal of Figure 4.8b also reveals no obvious relationship between the change values of individual grapheme types and their t1 scores. Further, it should be remembered that each bar in Figure 4.8 represents only the *mean* score for each group; additional variation may therefore be expected at the individual class level and, within each class, in the gain scores of individual participants.

Within each vowel grapheme type, there is also much wider and seemingly less consistent variation between the three conditions than in the case of consonants. Nonetheless, a descriptive examination of the data provides preliminary evidence for the emergence of some (limited) condition-level patterns in favour of the intervention groups. For example, the intervention groups A and B both show higher gain scores than the comparison group on six of the fourteen target grapheme types covered by all classes in the instruction programmes (<a>, <o>, <ou>, <i>, <oi>, <ez>) whereas the inverse pattern occurs in only three cases (<ie>, 'silent' <e>, <ai>). For a further seven grapheme types, one of the two intervention groups (A or B) obtains higher gain scores than the comparison group. However, these between-group differences are not tested for statistical significance and, as Figure 4.8b shows, in several cases they are very slight. Finally, the comparison group shows small gain scores of +7 percentage points or less

for all but one grapheme type, whereas the intervention groups obtain larger gain scores of +10 percentage points or more for seven grapheme types.

On the other hand, the gain scores of some vowel grapheme types appear to go against the general advantage for the intervention groups. For example, only in one case (<ez>) do both intervention groups outperform the comparison condition by a sizeable margin. Further, one or both intervention groups actually show *decreases* in the percentages of 'acceptable' realizations of four grapheme types (<eu>, <ê>, <ai>, <ei>), despite the last three of these having been covered in the instruction programmes. Finally, one grapheme type (<ie>) shows an unexpectedly large gain score for the comparison group; further analysis of the data revealed that this was due specifically to participants in school C.

### 4.6.4 Nasal vowels

Grapheme types in this category provide only limited insight into the effectiveness of the instruction programmes, because only three are included in the subset of target graphemes covered by all participants: <en>, <oin> and <on>.

At a descriptive level, Figure 4.8c highlights an apparent contrast between the first two grapheme types – <on> and <un>, which show decreases in the percentage of acceptable realizations in two and three conditions respectively – and the remainder, which all show positive gain scores. It is interesting that, for the two grapheme types which show clear score decreases but for no other nasals, realizations based on English GPC (<on> → /ɒn/; <un> → /ɐn/) would be judged 'acceptable' by the RAT mark scheme. This

probably explains their much higher t1 scores than those of other nasal grapheme types and hence their position on the left of Figure 4.8c: at t1, tokens of <on> and <un> were realized acceptably in 46% and 36% of cases respectively, whereas the third highest-ranked nasal according to t1 scores (<an>) was realized acceptably only 18% of the time (Appendix 9.5).

In terms of between-condition differences, there is again some tentative evidence of an advantage for the intervention participants. Groups A and B both obtained higher gain scores than the comparison condition in respect of five of the eight grapheme types, whereas the inverse pattern does not occur. Again, however, these between-group differences are slight and are untested for statistical significance; further, one of the three grapheme types which did not show an overall advantage for both intervention groups is one of the few nasal vowels which *was* actually covered by all participants in the programmes of instruction. In any case, the magnitude of the increases in 'acceptable' realizations is small across the category of nasal vowels as a whole: for those six grapheme types associated with positive gain scores, the largest increase (+14.7 percentage points) is smaller than the largest increase in the consonant or oral vowel categories. For these reasons, the *change2* values cannot be said to provide convincing evidence that participants in any of the three groups have made more progress than the others in decoding nasal vowels.

## 4.7 Summary

The various strands of analysis in this chapter suggest that participants in the intervention condition made significantly more progress in French decoding than those in the

comparison condition, as measured by increases in the numbers of 'correct' and 'acceptable' realizations of those grapheme types explicitly covered in their programmes of instruction (Research Question 1). However, the magnitude of their advantage over comparison participants was not great, amounting to approximately eight additional grapheme tokens decoded 'correctly' or 'acceptably' at t2. By contrast, no significant differences were found in the progress made by participants in the two intervention groups, suggesting that neither programme of instruction was more effective than the other (Research Question 2). A finer-grained analysis found that the levels of progress made by individual classes (as measured by their t1-to-t2 gain scores) were consistent with these broad condition-level patterns, despite some significant differences within each condition in terms of the classes' scores at t1 and t2. However, there was found to be considerable variation within each class in respect of the gain scores obtained by individual participants. Indeed, there were several participants in each class whose number of 'acceptable' realizations actually *decreased* between t1 and t2; most classes (including those in the intervention condition) had at least one member who pronounced between ten and twenty fewer grapheme tokens 'acceptably' at the second time of testing.

It had also been predicted that an initial dimension of progress in L2 decoding would involve increasing adherence to the principle of linear symbol-to-sound mapping for those participants who did not already observe it. For all groups, percentages of non-linear decoding (NLD) were indeed found to decrease between t1 and t2, though there was no obvious advantage for intervention participants in this respect.

Finally, the extent of progress made by participants in French decoding was found to vary considerably between individual grapheme types (especially within the oral vowel category), suggesting that the instruction may have been more effective in respect of some GPC than others. Within the consonantal category, there appeared to be an underlying trend for gains in the proportions of 'acceptable' realizations to decrease in inverse proportion to the scores generated by that grapheme type at t1.

# Chapter 5: Realizations of grapheme tokens

## 5.1 Introduction

The previous chapter compared the three experimental groups (Intervention A, Intervention B, Comparison) in terms of the numbers of grapheme tokens realized 'correctly' and 'acceptably' at t1 and t2. The current chapter looks beyond this global measure of decoding accuracy, investigating the ways in which individual grapheme tokens were realized by participants at each time point. This addresses Research Question 3:

> *Do the programmes of instruction lead to any other changes in participants' realizations of French graphemes?*

This question was motivated by the assumption that the development of decoding proficiency may follow a longer and more complex trajectory than simply moving from 'incorrect' to 'correct' (or 'acceptable') realizations, thus falling below the threshold of detectability on the RAT: in other words, a participant's realization of a given grapheme may become more target-like in some way, yet still fail to achieve a positive score (section 2.6.3). Further, Woore (2010) found that participants who sought to override their automatic L1-based decoding often focussed simply on making their realizations somehow 'different' from those generated by English GPC, without having a clear view of what the French realization should be. This appeared to result in greater variation in their decoding outcomes: i.e. although their realizations of a given grapheme may

change, this may occur in diverse and unpredictable ways. Such changes would again not be reflected in improved RAT scores.

The analysis in this chapter is based on the phonetic transcription of participants' RAT output (section 3.6.3.3), which provided the basis for the RAT scoring. One possibility would have been to analyse participants' output at the level of grapheme *types*, by amalgamating the data from the various related tokens: for example, participants' realizations of the three RAT items *témoin*, *goinfrez* and *sainfoin* could have provided an overall picture of how the grapheme type <oin> was realized by participants in each group and at each time point. However, this approach was rejected due to the possibility that the participants, accustomed to operating with larger L1 grain sizes beyond the individual grapheme (Ziegler & Goswami, 2005), may pronounce a given token differently according to its orthographic context, as occurs in the L1 (e.g. <a> is realized differently in *mat*, *mate* and *mart*). It was therefore decided to focus more narrowly on the decoding of individual grapheme tokens.

Due to limited space, only a small, illustrative selection of the 174 grapheme tokens available for analysis are presented in this chapter. These are taken from the grapheme types <ç>, <ou> and <oin>: one from each of the categories of consonants, oral vowels and nasal vowels. The rationale for selecting these grapheme types was as follows:

(1) <ç> was chosen as an interesting case because, uniquely amongst French consonantal graphemes, it contains a diacritic. Diacritics are frequently identified in the self-report interviews as a source of difficulty (chapter 6), perhaps because they force participants who notice them to move beyond the confines of English decoding (since they do not

occur in English). On the other hand, learning to decode the grapheme <ç> is theoretically simplified by the fact that the target sound /s/ exists as an English phoneme. The sole task of the beginner decoder is therefore to map the unfamiliar grapheme onto an existing phoneme. Further, this mapping is contextually invariant.

(2) <ou> is theoretically more complex, because the high, back, rounded realization of the target phoneme /u/ appears not to occur in the participants' L1: although a similar sound exists in English Received Pronunciation (e.g. in *food*, [fuːd]), informal observations suggested that participants generally produced a more front vowel [ʉ] here. Thus, in learning to decode French <ou> correctly, they must (a) map the grapheme onto the correct phoneme and (b) develop a more target-like representation of this phoneme in the L2. Following the argument in 2.6.4, it is predicted that participants may instead realize <ou> as the nearest existing L1 sound, [ʉ]. (Segmenting the letter-string correctly – into the digraph <ou> rather than two separate graphemes, <o> and <u> – is also a prerequisite for accurate decoding. However, since <ou> is also a frequent digraph in English, a contrastive analysis would not predict particular segmentation problems here).

(3) Finally, <oin> is predicted to be the most difficult of the three selected graphemes for participants to decode, for several reasons: first, the target sound [wæ̃] is complex, comprising two phonemes. Second, the nasal vowel /æ̃/ does not occur in English, posing similar problems to those described above in respect of <ou>. Third, <oin> is a trigraph. Again, therefore, the letter string must be correctly segmented in order for accurate decoding to occur; however, in this case, <oin> does not exist as a single grapheme in English. This leads to the prediction that participants may have greater problems of segmentation. Two incorrect divisions of the letter-string are possible: (a)

as the vowel digraph <oi> plus the consonant <n>; (b) as two separate vowels <o> and <i> plus the consonant <n>. Both are likely to result in incorrect decoding outcomes.

For each of these grapheme types, one individual token from the RAT is selected and its various realizations by participants in each group at t1 and t2 are analysed. This allows the identification of (a) any within- and between-group patterns in the way the grapheme token is decoded and (b) any changes in these patterns between the two time points.

## 5.2 Realizations of <ç>

### 5.2.1 Realizations produced by the whole sample at t1 and t2

Three RAT items contain <ç>: *caleçon, maçonne, traçant*. Visual inspection of preliminary bar charts suggested that there was broad similarity in the way the sample as a whole pronounced <ç> in these three words. One of the tokens (in *caleçon*) was therefore chosen arbitrarily as the focus of analysis.

Figures 5.1a and 5.1b show all attested realizations of <ç> in *caleçon* at t1 and t2 respectively, listed in decreasing order of number of realizations amongst the sample as a whole at t1: i.e. those produced by the highest number of participants at t1 are at the top. Where two or more realizations are produced by the same number of participants, the order within this group is arbitrary. The same, t1-based ordering is retained for the t2 data to facilitate comparisons between the time points. The bars indicate the percentages of participants in each group who produced each attested realization. Realizations

occurring at one time point but not the other are nonetheless listed in both figures, again to facilitate comparisons; thus, the absence of a bar next to a listed realization indicates that it was not attested at that time point but did occur at the other. Finally, any missing transcriptions resulting from (a) inaudibility (coded 997), (b) the participant's omission of an item (coded 998) or (c) Non-Linear Decoding (NLD: see section 3.6.3.3; coded 999) appear at the foot of the figures.

**Figure 5.1 Attested realizations of <ç> in *caleçon* at t1 and t2 and percentages of participants in each group producing each one.**

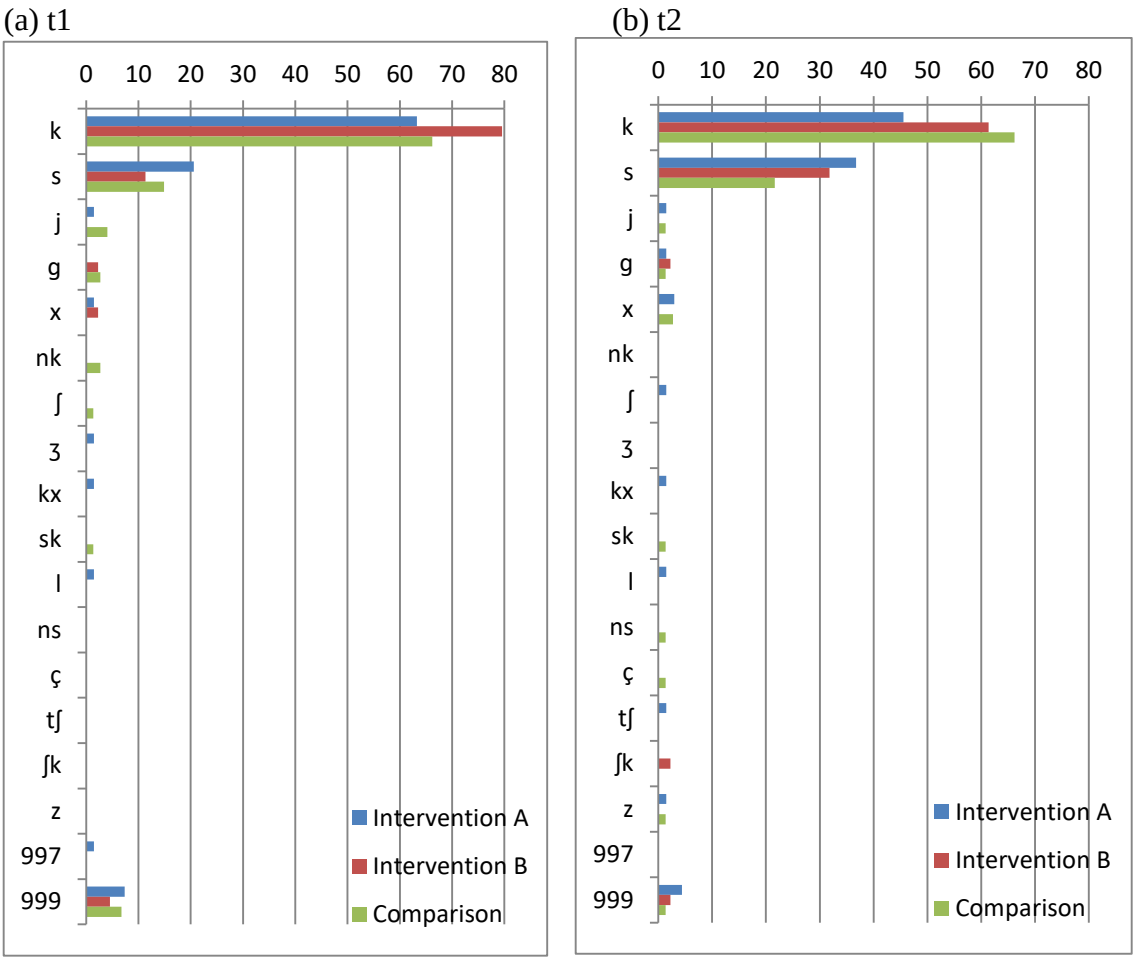



Figure 5.1 shows that the overwhelming majority of participants realized <ç> in one of two ways: as either [s] (the correct French form) or [k]. The form [k] is the more frequent, accounting for 68% of realizations at t1 and 58% at t2; [s] accounts for around 16% of realizations at t1 and 30% at t2. The predominance of these two phonological forms occurs in all three groups and at both time points (although the balance between them varies somewhat both between groups and between time points: see below).

A number of additional realizations (nine at t1; twelve at t2) were also recorded at each time point, but each is produced by only a small number of participants, i.e. by five participants or fewer across the sample as a whole. Indeed five realizations at t1 and

eight at t2 occur only once. The total proportion of participants in each group who produced valid realizations other than [k] or [s] is small: between 4.5% and 13.2% at both time points (Table 5.1).

**Table 5.1 Percentages of participants in each group producing various realizations of <ç> at each time point**

| | Realization | Intervention A | | Intervention B | | Comparison | |
|---|---|---|---|---|---|---|---|
| | | t1 | t2 | t1 | t2 | t1 | t2 |
| **a.** | **[k]** | 63.2% | 45.6% | 79.5% | 61.4% | 66.2% | 66.2% |
| **b.** | **[s]** | 20.6% | 36.8% | 11.4% | 31.8% | 14.9% | 21.6% |
| **c.** | **other** | 7.4% | 13.2% | 4.5% | 4.5% | 12.2% | 10.8% |
| **d.** | **NLD** | 7.4% | 4.4% | 4.5% | 2.3% | 6.8% | 1.4% |

### 5.2.2 Changes over time

First, to assess whether each of the three groups (Intervention A, Intervention B, Comparison) differed in their realizations of <ç> in *caleçon* at the two time points, a series of Pearson's chi square tests were performed. However, problems arose due to the large number of attested realizations, many produced by a small number of participants or by one single participant: in a cross-tabulation of realizations by time points (t1 or t2), the expected count in many cells is very low, violating the assumptions of the chi square test. Therefore, all 'other' valid realizations (i.e. not [k] or [s]) were combined into a single category. These preliminary steps having been taken, no cells had an expected count below 5 for groups A or C. However, in group B, two cells had low expected counts (2.0). Following Field (2009:690), Fisher's Exact Test was therefore used. The tests revealed a significant association between time of testing on the one hand and how <ç> was realized on the other for intervention group A ($\chi(2)=6.124$, $p=.047$) but not for the comparison group ($\chi(2)=0.873$, $p=.646$). For intervention group B, the association approached significance (Fisher's Exact Test,

*p*=.053). However, this statistical test is likely to be conservative because it is based on a partial picture of the overall data, conflating various individual forms into a single category of 'other' realizations.

To assess whether the three groups showed different patterns of change in how <ç> was pronounced between t1 and t2, the realizations they produced at each individual time point were compared statistically, again based on the three categories of /k/, /s/ and 'other'. No significant association was found between group membership on the one hand and how <ç> was pronounced on the other, either at t1 (*p*=.371, Fisher's Exact Test) or at t2 (*p*=.100, Fisher's Exact Test). This suggests that the groups' realizations of <ç> in *caleçon* did not undergo differential patterns of change between t1 and t2.

However, a descriptive analysis of Figure 5.1 and Table 5.1 does allow the tentative identification of some differences in how the groups' decoding develops over time, notwithstanding their different starting points (i.e. the different percentages of participants producing each realization at t1). Given a larger sample size, it is possible (though cannot be known for certain) that these differences may have reached statistical significance. Specifically, the proportion of participants producing the sound [k] remains the same at t2 as at t1 in the comparison group but falls by around 18 percentage points in both intervention groups. Conversely, the proportion of participants producing the correct form, [s], rises by 16 and 20 percentage points in intervention groups A and B respectively but by only around 6 percentage points in the comparison group. There are also differences in the percentage of 'other' realizations produced by the three groups: this rises from 7% to 13% in intervention group A, remains identical in intervention group B but falls slightly in the comparison group. However, these percentage figures

for ‘other’ realizations must be treated cautiously, since the numbers of participants involved are small: for example, the near-doubling of the percentage of ‘other’ realizations in intervention group A is accounted for by an increase of only four participants. Further, the pronunciations classed as ‘other’ differ in some cases between the two time points: there are two which occur at t1 but not t2 and five which occur at t2 but not t1. Finally, although the percentages of Non-Linear Decoding were not included in the statistical analysis, it is noteworthy that these show decreases in all three groups.

### 5.2.3 Summary

The overwhelming majority of participants across the sample as a whole realized the grapheme token <ç> in *caleçon* in one of two ways, as either [k] or [s] (the correct French pronunciation). Nonetheless, various other realizations also occurred at both time points, many of which included the fricative sounds [ʃ], [ʒ], [x] and [ç].

Statistically significant differences were found between the realizations of <ç> produced at each time point by participants in intervention group A, but not by those in the comparison group. The differences for intervention group B approached statistical significance.

No statistically significant association was found at either t1 or t2 between the group (Intervention A, Intervention B, Comparison) and the way <ç> was realized. Nonetheless, a descriptive analysis of the data allowed some differences between the groups to be (tentatively) identified in terms of changes in their realization of <ç> over time. First, there was an increase in the correct form [s] in all groups, but the increase

was around three times larger in the intervention groups than in the comparison group. Second, the proportion of participants producing the form [k] fell considerably in the intervention groups but remained constant in the comparison group. The proportion of participants producing valid realizations other than [k] and [s] also rose in intervention group A, remained the same in intervention group B and fell slightly in the comparison group, although caution is required here as the numbers involved are small. Finally, the percentages of NLD fell in all three groups.

## 5.3 Realizations of <ou>

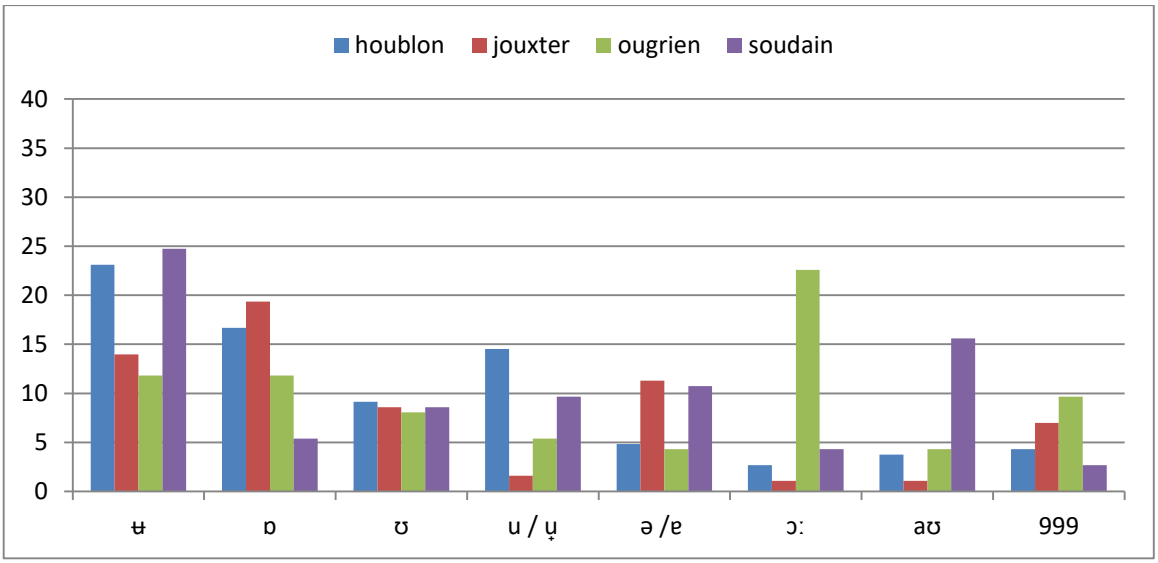
### 5.3.1 Realizations produced by the whole sample at t1 and t2

The second grapheme selected for discussion, <ou>, appears in four RAT items: *jouxter, soudain, houblon, ougrien*. In contrast to <ç>, visual examination of preliminary bar charts suggested that there were differences in the way <ou> was realized in each of these four words both at t1 and t2. To investigate this further, a Pearson's chi square test was used on the baseline (t1) data to check for significant associations between lexical context on the one hand (i.e. the word in which <ou> appeared) and realizations of the grapheme on the other. However, the number of distinct realizations produced by the sample as a whole was extremely high (thirty-nine all told). This again resulted in a large number of cells in the cross-tabulation with an expected count below 5, violating the assumptions of the chi square test. Further, as in the case of <ç>, these realizations were unevenly represented: across all four RAT items, well over half of the attested forms were produced by one participant only in the sample as a whole; thirteen by

between two and six participants each; and seven by between eleven and sixteen participants each. Finally, there was a group of seven realizations – plus instances of NLD – which were produced by over forty participants each. These eight most frequent decoding outcomes were therefore used as the basis for the Pearson's chi square test. A significant association was found between the word in which <ou> appeared and the way the grapheme was realized, $\chi(21)=175.270$, $p<.001$. No cells had an expected count below 5.

Standardized residuals were examined to identify more specifically the locus of significant differences between the groups at the level $p<.05$, as indicated by values exceeding ±1.96. This revealed that all four RAT items were associated with significant differences between the realizations. In *houblon*, significantly more participants than expected realized <ou> as [u] or [ụ], but significantly fewer than expected produced the sound [ɔː]. In *jouxter*, significantly more participants than expected produced (a) [ɒ] and (b) [ə] or [ɐ], but significantly fewer than expected produced [ɔː], [u] / [ụ] and [aʊ]. In *ougrien*, significantly more participants than expected produced [ɔː] and showed NLD, but significantly fewer than expected produced [ʉ]. Finally, in *soudain*, significantly more participants than expected produced [aʊ], but significantly fewer than expected produced [ɒ] and showed NLD. These differences, together with others which did not reach significance, are visually apparent in Figure 5.2, which shows the percentages of all participants who produced each of the most frequently-attested realization of <ou> in each RAT item. In summary, the data clearly show that participants' decoding of the French grapheme <ou> varies according to its orthographic context.

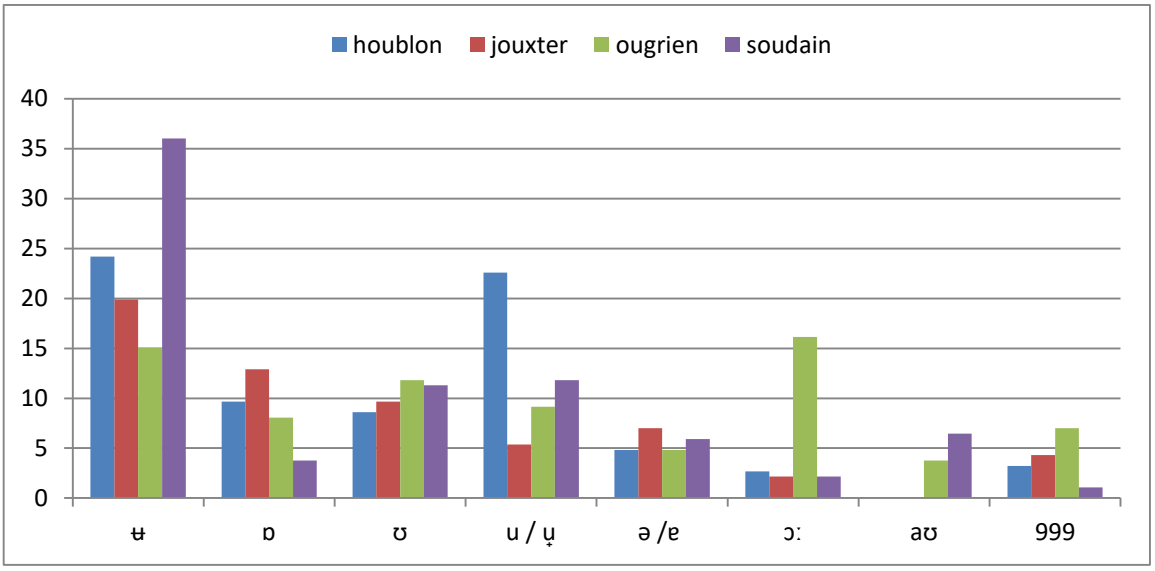
**Figure 5.2 Percentage of all participants producing the eight most frequent realizations of <ou> in the four RAT items at t1.**




The t2 data was analysed using the same procedures as described above, again based on the eight most frequent realizations (which were the same as at t1). A Pearson's chi square test again found a significant association between lexical context and realization of <ou>, $\chi(21)=125.152$, $p<.001$. Three cells in the cross-tabulation had an expected count below 5 (one of 4.0 and two of 4.9), but this was considered an acceptable basis for the test to be conducted (Field, 2009:692).

Standardized residuals and the percentages of participants producing each realization for each RAT item (Figure 5.3) were then examined. All four RAT items were again associated with significant differences between the realizations. The relative frequencies of the different realizations associated with each word were also broadly similar at t2 to those observed at t1, although there appears to be a tendency for the 'acceptable' form [ʉ] and the 'correct' form [u] / [u̟] to occur more frequently in all items, especially *soudain* and *houblon*. Conversely, all other realizations (and NLD) generally appear to decrease in frequency.

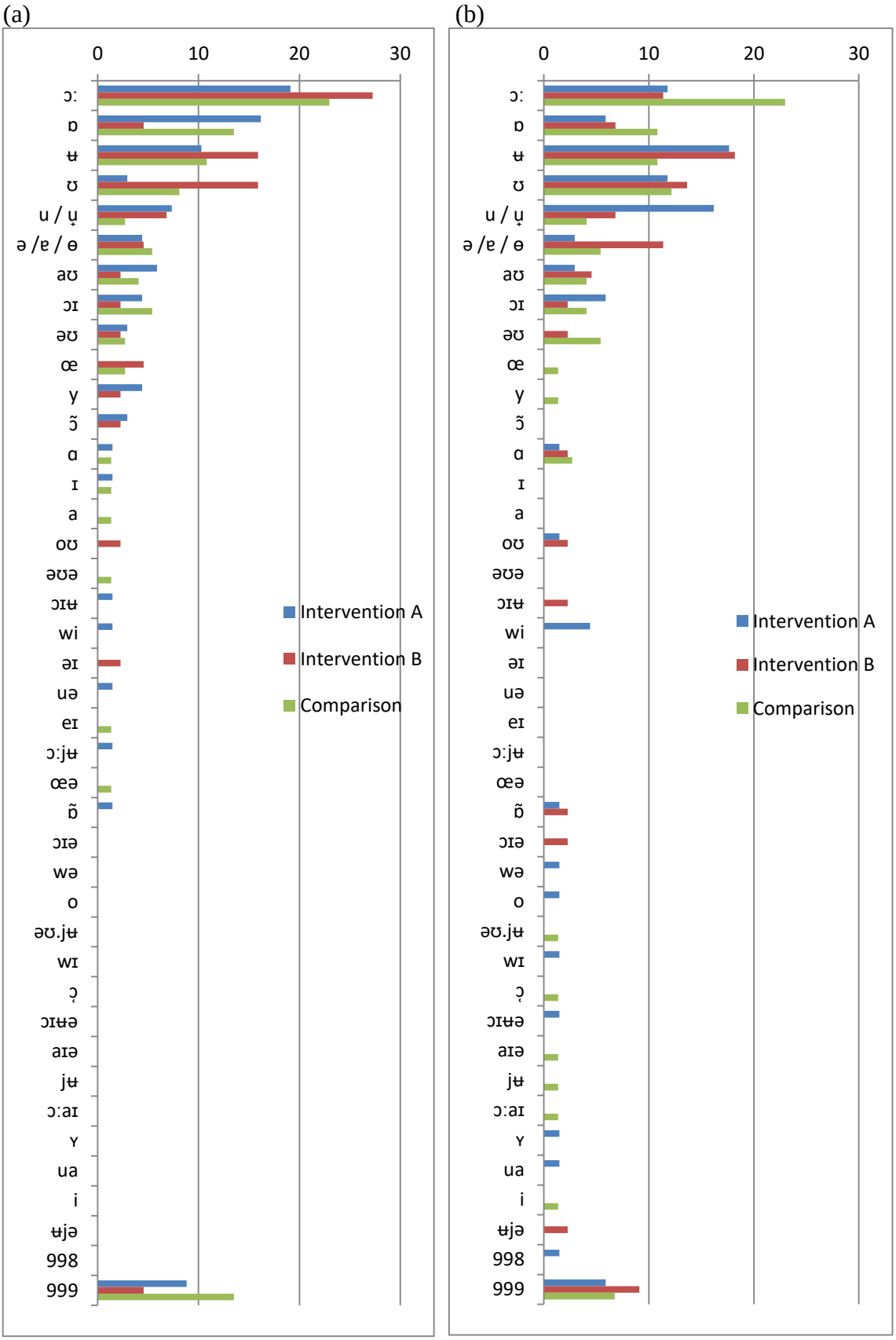
**Figure 5.3 Percentage of all participants producing the eight most frequent realizations of <ou> in the four RAT items at t2.**




### 5.3.2 Analysis of grapheme token <ou> in *ougrien*

In order to investigate differences between the individual groups (Intervention A, Intervention B, Comparison) in terms of their decoding of <ou> at each time point, one particular token (<ou> in *ougrien*) was arbitrarily selected for more detailed analysis. Figures 5.4a-b show all attested realizations of this token at t1 and t2 respectively, displayed according to the same conventions as in Figure 5.1 for <ç> above. Note that some of the originally transcribed realizations were considered minimally different from each other (e.g. [ə], [ɐ] and [ɵ]; [u] and [u̟]) and were therefore grouped together into single categories to facilitate analysis.

**Figure 5.4 Attested realizations of <ou> in *ougrien* at t1 and t2 and percentages of participants in each group producing each one.**



As is immediately apparent in Figure 5.4, <ou> in *ougrien* is realized in many more distinct ways than is <ç> in *caleçon*. A total of thirty-nine realizations were attested across both time points and across all three groups. This statement of course presupposes accurate transcription of the participants' RAT output; the transcription process was associated with high intra- and inter-rater reliability coefficients (section 3.6.3.5).

The number of distinct realizations is broadly similar at both time points and for each of the three groups (Table 5.2): the ratio of attested forms to participants ranges from 1:2.9 in intervention group B at t2 (i.e. on average, there is one distinct realization for around every three participants) to 1:4.6 in the comparison condition at t1. It follows that the majority of attested realizations are produced by only a small number of participants. Indeed, 44% of realizations at t1 and 57% at t2 are produced by only one participant across the whole sample. Further, over half of attested realizations at t1 and two thirds at t2 are produced by three participants or fewer.

**Table 5.2 Numbers of distinct realizations in each group at t1 and t2**

| | | t1 | | t2 | |
|---|---|---|---|---|---|
| | **N** | **Number of realizations** | **Ratio of realizations to participants** | **Number of realizations** | **Ratio of realizations to participants** |
| **Intervention A** | 68 | 18 | 1:3.8 | 18 | 1:3.8 |
| **Intervention B** | 44 | 14 | 1:3.1 | 15 | 1:2.9 |
| **Comparison (C)** | 74 | 16 | 1:4.6 | 18 | 1:4.1 |

Conversely, many participants' realizations of <ou> are concentrated in a small number of the attested phonological forms. The predominance of [ɔː] has already been noted, accounting for between one fifth and one quarter of all participants' realizations at t1 and just under one sixth at t2. Table 5.3 shows all those phonological forms which exceed

the (arbitrary) threshold of being produced by four participants or more across the sample as a whole; these number only eleven out of twenty-five (44%) at t1 and ten out of thirty (one third) at t2. Further, over half of participants at t1 and almost two thirds at t2 produce realizations which fall within the (again arbitrary) group of those produced by at least eight per cent of the whole sample at either time point (i.e. by fifteen or more participants). These forms, which will provide the basis for subsequent statistical analysis, are shaded grey in Table 5.3. Finally, it is notable that eleven of the less frequently-attested realizations (amounting to around 28% of all realizations produced by the whole sample across both time points) are bisyllabic.

**Table 5.3 Realizations of <ou> produced by four or more participants at each time point, ordered by descending frequency**

| t1 | | t2 | |
|---|---|---|---|
| **Attested realization** | **Number and percentage of participants in whole sample producing this realization** | **Attested realization** | **Number and percentage of participants in whole sample producing this realization** |
| ɔː | 42 (22.5%) | ɔː | 30 (16.1%) |
| ɒ | 23 (12.4%) | ʉ | 28 (15.1%) |
| ʉ | 22 (11.8%) | ʊ | 23 (12.4%) |
| ʊ | 15 (8.1%) | u / u̟ | 17 (9.1%) |
| u / u̟ | 10 (5.4%) | ɒ | 15 (8.1%) |
| ə / ɐ / ɵ | 9 (4.8%) | ə / ɐ / ɵ | 11 (5.9%) |
| aʊ | 8 (4.3%) | ɔɪ | 8 (4.3%) |
| ɔɪ | 8 (4.3%) | aʊ | 7 (3.8%) |
| əʊ | 5 (2.7%) | əʊ | 5 (2.7%) |
| œ | 4 (2.2%) | ɑ | 4 (2.2%) |
| y | 4 (2.2%) | | |

In summary, <ou> in *ougrien* is realized by participants in many different ways: almost forty distinct realizations were identified across both time points and across the sample as a whole. On the other hand, most participants' realizations of this grapheme token fall

within a relatively small number of attested phonological forms. In particular, the form [ɔː] predominates at both time points.

### 5.3.3 Changes over time

As in the analysis of <ç>, to assess whether each of the three groups (Intervention A, Intervention B, Comparison) differed in the way they realized <ou> in *ougrien* at t1 and t2, a series of Pearson's chi square tests were performed. To mitigate the problem of low expected cell counts in the cross-tabulation, groups were again compared only in terms of the most frequently-attested realizations, operationalized as those produced by at least 8% of the whole sample at either time point (shaded grey in Table 5.3); the remaining realizations were amalgamated into a single category of 'other' forms. Since Non-Linear Decoding (NLD) was also a frequent outcome for the sample as a whole (accounting for 9.7% of realizations at t1 and 7.0% at t2), it was included in the analysis. The number of cells with an expected count below 5 was zero for intervention group A, very low for the comparison group (N=2) but higher for intervention group B (N=6). Fisher's Exact Test was therefore used for this group. No significant association was found between time of testing on the one hand and how <ou> was realized on the other for either intervention group B (Fisher's Exact Test, $p$=.614) or the comparison group ($\chi(6)=2.889$, $p$=.823); for intervention group A, the association approached significance ($\chi(6)=12.223$, $p$=.057). None of the standardized residuals exceeded the significant threshold of ±1.96. Again, however, these results are based on a partial picture of the data, since a large number of individual realizations are grouped together into the 'other' category. This means that any differential patterns of change within these infrequently-attested realizations are lost to the analysis.

To assess whether the three groups showed differential patterns of change in how <ou> was pronounced between t1 and t2, the realizations they produced at each individual time point were compared statistically, again based on the seven categories of [ɔː], [ʉ], [ɒ], [ʊ], [u] / [ụ], 'other' and NLD. Pearson's chi square tests found no significant association between group and realization of <ou> at either t1, $\chi(12)=15.324$, $p=.224$, or t2, $\chi(12)=12.831$, $p=.381$. However, five cells at t1 and two at t2 had an expected cell count below 5. At t1, the proportion of cells with low expected counts (23.8%) therefore slightly exceeds the threshold of 20% which Field (2009:692) considers acceptable for the chi square test to be valid[29]. This may be associated with a loss of power (i.e. an increased likelihood of Type II error). Two standardized residuals (in the cases of [ʊ] at t1 and [u] / [ụ] at t2) also approached the critical threshold of ±1.96 associated with statistical significance. For these reasons, a descriptive account of the similarities and differences between the groups' realizations at each time point is also presented.

***Time 1***

At t1, as Figure 5.4 and Table 5.4 show, the three groups' decoding of <ou> is similar in that [ɔː] is the most frequent realization in all cases, although the proportion of participants producing this sound is somewhat higher in intervention group B (27%) than in the comparison group (23%), and somewhat lower in intervention group A (19%). Conversely, a descriptive analysis of the data also suggests various (albeit non-statistically significant) differences between the groups at t1. For example, [ʉ] occurs more frequently in intervention group B than in the other two groups. Contrastingly, [ɒ]

[29] Fisher's Exact Test could not be used due to limitations on computational resources.

is well represented in intervention group A and in the comparison group but hardly occurs at all in intervention group B. The correct form [u] is also produced by more participants in intervention groups A and B than in the comparison group. However, these descriptively observed differences do not appear to form any clear pattern at the group level.

**Table 5.4 Percentage of participants in each group producing frequent realizations of <ou> at each time point**

| | Intervention A | | Intervention B | | Comparison | |
|---|---|---|---|---|---|---|
| | t1 | t2 | t1 | t2 | t1 | t2 |
| ɔː | 19% | 12% | 27% | 11% | 23% | 23% |
| ɒ | 16% | 6% | 5% | 7% | 14% | 11% |
| ʉ | 10% | 18% | 16% | 18% | 11% | 11% |
| ʊ | 3% | 12% | 16% | 14% | 8% | 12% |
| u / u̟ | 7% | 16% | 7% | 7% | 3% | 4% |

***Changes at t2***

As revealed by a comparison of the two halves of Table 5.3, there is considerable similarity between t1 and t2 in terms of the most frequent realizations of <ou> across the sample as a whole. Of the eleven forms produced by four or more participants at t1, nine remain in this category at t2 (with one additional realization being added to the list); they also remain in broadly similar order when ranked according to decreasing frequency. As Table 5.2 shows, there is also no evidence of any decrease in the total number of distinct realizations produced by any of the groups. By contrast, there is some variation in the nature of the less frequent realizations between the two time points: across the sample as a whole, nine phonological forms occur at t1 but not at t2; fourteen new ones are then attested at t2 which do not occur at t1. The number of bisyllabic realizations increases

slightly between the two time points (from four at t1 to eight at t2), but the numbers involved are too small to draw any firm conclusions.

At least on a descriptive level, there is also some evidence of differential patterns of change between the groups in terms of how they realize <ou> at the two time points. This can be summarized as follows, based on the five realizations in Table 5.4. First, the proportion of participants producing the form [ɔː] remains constant in the comparison group but shows large decreases of 7 and 16 percentage points in intervention groups A and B respectively. By contrast, the 'acceptable' form [ʉ] becomes more frequent in group A but remains the same or similar in groups B and C. Third, the proportion of participants producing the incorrect form [ɒ] more than halves in group A but shows much lesser change in groups B and C. Fourth, [ʊ] becomes more frequent in group C and slightly less so in intervention group B, but its frequency quadruples in intervention group A. Finally, the correct form [u] more than doubles in frequency in group A but remains constant in intervention group B and shows a very small rise in the comparison group.

### 5.3.4 Summary

No significant differences emerged between (a) the way participants in each group pronounced <ou> in *ougrien* at t1 and t2 or (b) between the groups' realizations at either time point. Nonetheless, there does seem to be some evidence (at least in descriptive terms) of differential patterns of change between the groups. The intervention groups A and B tend to show larger changes in the way <ou> is realized between the two time points, whereas the realizations of the comparison group remain broadly similar. In

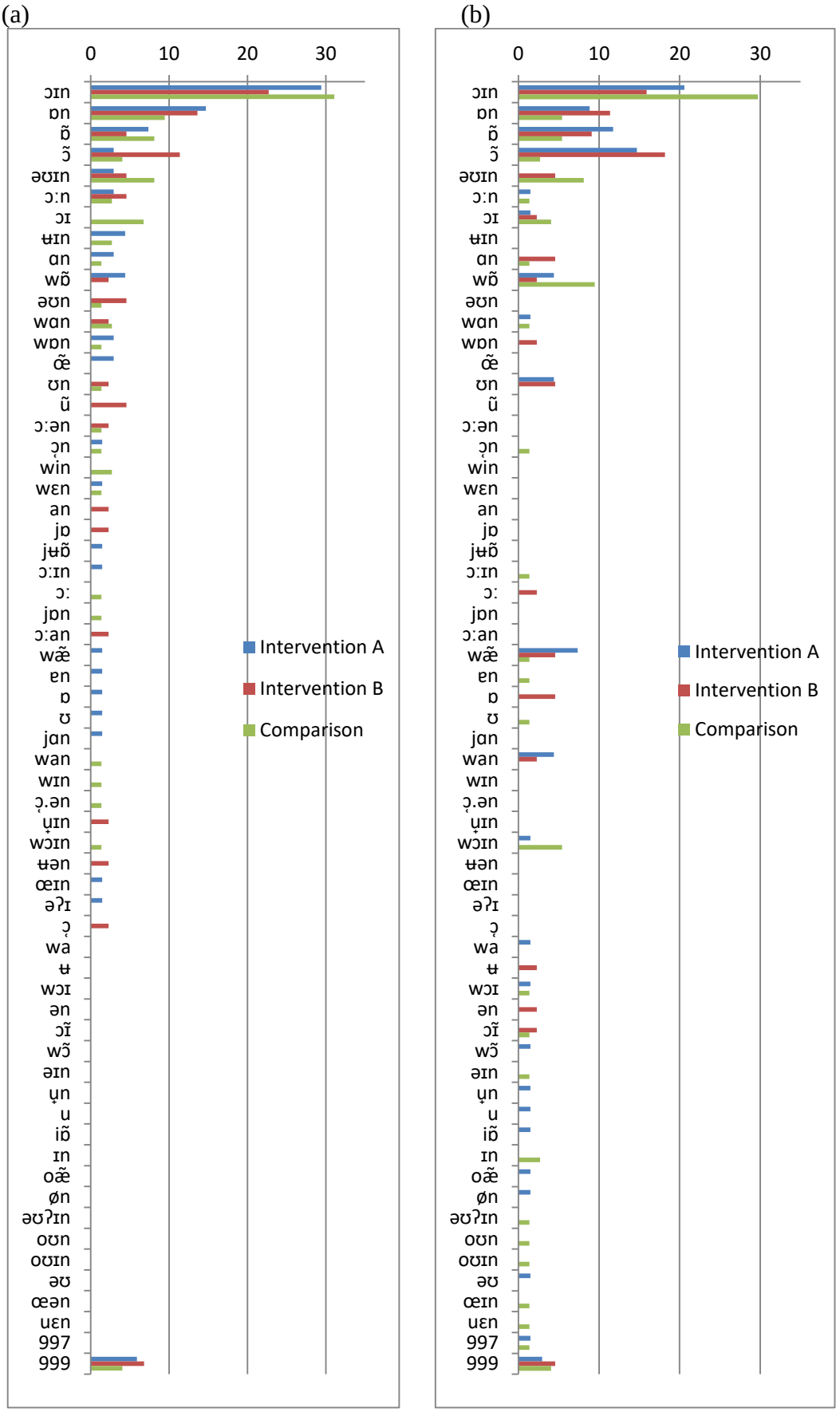
particular, group A shows a large decrease in the number of participants producing the form [ɒ] and large increases in the numbers of participants producing the (a) the 'correct' form [u] / [ụ]; (b) the 'acceptable' form [ʉ]; (c) the form [ʊ]. Group B shows a particularly large decrease in the number of participants producing [ɔː]. The greater overall degree of change observed in intervention group A (compared to intervention group B) is reflected in the fact that the difference between its participants' realizations at t1 and t2 approached significance.

## 5.4 Realizations of <oin>

### 5.4.1 Realizations produced by the whole sample at t1 and t2

The third grapheme type selected for analysis, <oin>, occurs in three RAT items: *témoin*, *goinfrez*, *sainfoin*. Visual inspection of preliminary bar charts suggested no striking differences in the way the three individual tokens were realized by the sample as a whole. One token was therefore chosen arbitrarily as the focus of analysis: <oin> in *goinfrez*. Figures 5.5a-b show how this grapheme token was realized at t1 and t2 respectively, following the same conventions as Figures 5.1 and 5.4.

**Figure 5.5 Attested realizations of <oin> in *goinfrez* at t1 and t2 and percentages of participants in each group producing each one.**



An immediately striking feature of Figure 5.5 is the very large number of distinct phonological forms (sixty-two in total) produced by the sample as a whole across the two time points: even larger than in the case of <ou>. Clearly, some of the attested forms differ only minimally from each other (e.g. [ən] and [ɐn]; [uɪn] and [uˌɪn]); however, the very small numbers of participants producing these forms suggested no benefit to the analysis in combining them into larger categories.

The numbers of distinct attested forms are at similar, high levels at both time points and in each of the three groups (Table 5.5): the ratio of attested forms to participants ranges from 1:3.1 in the comparison condition at t1 to 1:2.3 in intervention group B at t1. As in the case of <ou>, it follows that the majority of the attested realizations are produced by only a small number of participants. Around half of all attested realizations at both time points were produced by only one participant across the sample as a whole; around three quarters of attested realizations at both time points are produced by three participants or fewer.

**Table 5.5 Numbers of distinct realizations of <oin> produced by each group at t1 and t2**

| | | t1 | | t2 | |
|---|---|---|---|---|---|
| | **N** | **Number of realizations** | **Ratio of realizations to participants** | **Number of realizations** | **Ratio of realizations to participants** |
| **Intervention A** | 68 | 23 | 1:3.0 | 23 | 1:3.0 |
| **Intervention B** | 44 | 19 | 1:2.3 | 18 | 1:2.4 |
| **Comparison (C)** | 74 | 24 | 1:3.1 | 27 | 1:2.7 |

Conversely, many participants' realizations of <oin> are concentrated in a small number of the attested phonological forms. Over a quarter of participants at t1 and just under a quarter at t2 realize the grapheme token in the same way, as /ɔɪn/. Table 5.6 shows all

those phonological forms which exceed the (arbitrary) threshold of being produced by four participants or more across the sample as a whole; these number only eight at t1 and ten at t2. Further, well over half of participants at each time point produce realizations which fall within the (again arbitrary) group of those produced by at least five per cent of the whole sample at each time point (i.e. by ten or more participants). These forms, which will contribute to subsequent statistical analysis, are shaded grey in Table 5.6.

**Table 5.6 Attested realizations of <oin> produced by four or more participants at each time point, ordered by descending frequency**

| t1 | | t2 | |
|---|---|---|---|
| **Attested realization** | **Number and percentage of participants in whole sample producing this realization** | **Attested realization** | **Number and percentage of participants in whole sample producing this realization** |
| **ɔɪn** | 53 (28.5%) | **ɔɪn** | 43 (23.1%) |
| **ɒn** | 23 (12.4%) | **ɔ̃** | 20 (10.8%) |
| **ɒ̃** | 13 (7.0%) | **ɒ̃** | 16 (8.6%) |
| **ɔ̃** | 10 (5.4%) | **ɒn** | 15 (8.1%) |
| **əʊɪn** | 10 (5.4%) | **wɒ̃** | 11 (5.9%) |
| **ɔːn** | 6 (3.2%) | **wæ̃** | 8 (4.3%) |
| **ɔɪ** | 5 (2.7%) | **əʊɪn** | 8 (4.3%) |
| **ʉɪn** | 5 (2.7%) | **ɔɪ** | 5 (2.7%) |
| | | **ʊn** | 5 (2.7%) |
| | | **wan** | 4 (2.2%) |

Various patterns can also be discerned amongst the overall set of realizations of <oin> (Figure 5.5). For example, roughly one fifth of the attested realizations begin with the sound [w]. Six others begin with either the glide /j/ or the vowel /i/. Third, twelve of the attested forms, including two of those which were produced most frequently at both time points (/ɒ̃/, /ɔ̃/), include some variety of nasal vowel. Fourth, seventeen attested realizations end in an oral vowel: in other words, there is no distinct phonological counterpart of the <n>, either in the form of nasalization or of a consonantal [n]. Finally,

although most of the attested realizations are monosyllabic, seventeen are bisyllabic, perhaps reflecting the incorrect processing of <o> and <i> as separate vowel graphemes rather than as part of the trigraph <oin>.

To summarize, participants' realizations of <oin> in *goinfrez* are characterized on the one hand by extreme diversity – across both time points, over sixty different phonological forms are produced for this grapheme by the sample as a whole, many by only one participant each – and on the other hand by a high degree of consistency: most participants realize the grapheme using one of only a handful of phonological forms, especially /ɔɪn/. Further, various attested forms (including those which occur only infrequently) can be classified into broad categories defined by shared phonological characteristics, suggesting that there may be at least some degree of systematic variation within the apparent diversity of participants' realizations. This will be explored in more detail below.

### 5.4.2 Similarities and differences between the groups

In order to assess whether the three groups (Intervention A, Intervention B, Comparison) differed in their decoding of <oin> in *goinfrez* between t1 and t2, a cross-tabulation was performed with time of testing on one axis and the most frequent realizations on the other, operationalized as those produced by at least 5% of the whole sample (at least ten participants) at either time point, together with a category containing all 'other' realizations. Six attested realizations ([ɔɪn], [ɒn], [ɔ̃], [əʊɪn], [ə̃], [wɔ̃]) fell above this arbitrary threshold, as indicated by grey shading in Table 5.6. Despite these measures, high numbers of low expected cell counts occurred in all three groups (amounting to

between 29% and 43% of the total number of cells). Fisher's Exact Test was therefore used. This found no significant association between time of testing and realization of <oin> for intervention group A ($p$=.113), intervention group B ($p$=.926) or the comparison group ($p$=.174). Once again, however, it should be noted that these results may be conservative due to the simplified set of data on which they are based.

Did the intervention groups and comparison group show different patterns of change in terms of how <oin> was realized between t1 and t2? Again, to answer this question, the realizations produced by the three groups at each individual time point were first compared. However, a cross-tabulation based on the same set of frequent realizations as used in the t1-t2 comparisons resulted in unacceptably large numbers of cells with low expected counts (around half of all cells). Further, Fisher's Exact Test could not be performed due to computational limitations. Therefore, a smaller set of five frequent realizations was used (defined as those produced by over ten participants at either time point, i.e. excluding [əʊɪn] from the set used previously), together with the category of 'other' realizations. Fisher's Exact Test could now be performed. This found no significant association between group membership on the one hand and realization of <oin> on the other, either at t1 ($p$=.563) or at t2 ($p$=.080). However, it is clear that an analysis based on only six categories of realizations, when over sixty distinct forms were attested amongst the sample as a whole, represents a considerable simplification of the data, which may well have resulted in differences between the groups passing undetected. Further, the test result at t2 approaches significance and the comparison group was found to show a standardized residual of -2.0 (exceeding the threshold of ±1.96 associated with statistical significance) in respect of the realization [ɔ̃]. By contrast, the standardized residuals for [ɔ̃] in the intervention groups A and B were

positive (+0.8 and +1.5 respectively), although they did not reach significance. Once again, therefore, a descriptive account of the similarities and differences between the groups' realizations at each time point is presented.

### 5.4.3 Time 1

Table 5.7 shows the percentages of participants in each group who produced the most frequently-attested realizations of <oin> at t1 and t2, arbitrarily defined as those produced by at least 4% of the whole sample at either time point (see Table 5.6). At t1, in keeping with the finding of the statistical test, Table 5.7 suggests broad similarities between the groups. First, /ɔɪn/ clearly predominates in all three groups, accounting for around 30% of realizations in groups A and C and around 23% in group B. Second, /ɒn/ and /ɒ̃/ occur frequently in all groups: the former accounts for 14-15% of realizations in intervention groups A and B and around 10% in the comparison group; the latter accounts for between 5% and 8% of realizations in all three groups.

On the other hand, there also appear to be some descriptively observable differences between the groups at t1. In several cases, one group stands out from the other two in respect of the number of times a particular phonological form is produced: this is true of /ɒn/ and /ɔɪn/, as noted above, but also /ɔ̃/ (produced by 3-4% of participants in groups A and C but around 11% in group B) and /əʊɪn/ (produced by around 3% of participants in group A, 5% in group B and 8% in group C). However, some of these differences in percentage figures should be viewed cautiously as they are accounted for by very small numbers of individual participants. For example, had /əʊɪn/ occurred just twice more in

group A, the percentage of participants producing this realization would have doubled to 6%.

**Table 5.7 Percentage of participants in each group producing various frequent realizations of <oin> at each time point**

| | Programme A | | Programme B | | Comparison | |
|---|---|---|---|---|---|---|
| | t1 | t2 | t1 | t2 | t1 | t2 |
| **ɔɪn** | 29.4 | 20.6 | 22.7 | 15.9 | 31.1 | 29.7 |
| **ɒn** | 14.7 | 8.8 | 13.6 | 11.4 | 9.5 | 5.4 |
| **ɒ̃** | 7.4 | 11.8 | 4.5 | 9.1 | 8.1 | 5.4 |
| **ɔ̃** | 2.9 | 14.7 | 11.4 | 18.2 | 4.1 | 2.7 |
| **əʊɪn** | 2.9 | 0.0 | 4.5 | 4.5 | 8.1 | 8.1 |
| **wɒ̃** | 4.4 | 4.4 | 2.3 | 2.3 | 0.0 | 9.5 |
| **wæ̃** | 1.5 | 7.4 | 0.0 | 4.5 | 0.0 | 1.4 |

### 5.4.4 Changes at t2

As Table 5.6 and Figure 5.5 show, there is considerable similarity between the two time points in terms of the most frequent realizations of <oin> across the sample as a whole. Of the eight realizations produced by four or more participants at t1, all except [ɔːn] and [ʉɪn] also fall above this threshold at t2; four additional realizations are also added to the list at t2: [wɒ̃], [wæ̃], [ʊn], [wan]. However, [ɔɪn] remains by far the most frequent realization at t2, albeit produced by around 23% of participants rather than around 29% at t1.

A descriptive analysis of Table 5.7 provides some tentative evidence of differential patterns of change in the groups' realizations of <oin> between the two time points. First, although [ɔɪn] remains the most frequent realization amongst the sample as a whole at t2, its frequency has fallen sharply in both intervention groups but has remained fairly constant in the comparison group. The second most frequent form, [ɒn], is associated

with a decrease in realizations of between 2% and 6% in all three groups  Third, both of the single nasal vowels [ɒ̃] and [ɔ̃] show increases for the intervention groups A and B (particularly large in the case of [ɔ̃]) but small decreases for the comparison group. Fourth, the frequency of [əʊɪn] falls slightly in intervention group A (although the numbers involved are very small) but remains constant in intervention group B and the comparison group.  Fifth, the frequency of the 'acceptable' form [wɒ̃] remains the same in both intervention groups but rises sharply in the comparison group.  Finally, the 'correct' form [wæ̃] is associated with increases in frequency in all three groups, although these are larger in the intervention groups A and B (+5.9% and +4.5% respectively) than in the comparison group (+1.4%).

### 5.4.5 Comparing less frequent realizations

The above analysis covers only the most frequently attested realizations of <oin> across the sample as a whole; however, as Figure 5.5 shows, many other realizations at each time point are produced by only one, two or three participants, making between-group comparisons difficult.  Matters are further complicated by the fact that many of these realizations are different at t1 and t2: seventeen occur at t1 but not t2; twenty occur at t2 but not t1.  However, it was noted above that many of the attested realizations fall into broad categories defined by shared phonological characteristics.  Three of these categories can be argued to have particular relevance when assessing participants' progress in decoding <oin>, and were therefore used as the basis for further statistical analysis:

(1) Realizations beginning with the semivowel [w]. These accounted for roughly one fifth of all attested forms, including (a) the correct form [wæ̃]; (b) the 'acceptable' forms [wɒ̃], [wan], [wɑn] and [wɒn]; and (c) various other forms [win], [wɪn], [wɛn], [wa], [wɔɪ], [wɔɪn] and [wɔ̃]. The initial [w] means that all these realizations – even those not judged 'correct' or 'acceptable' – contain at least one target-like component.

(2) Realizations containing a nasal vowel, irrespective of that vowel's other articulatory features. These again amount to around one fifth of all attested forms, including the 'correct' form [wæ̃], the 'acceptable' form [wɒ̃], the frequently-attested single nasal phonemes [ɒ̃] and [ə̃] as well as various other less frequent forms ([œ̃], [ũ], [jʉɒ̃], [ɔ̃ɪ], [wɔ̃], [iɑ̃], [oæ̃]). The inclusion of this category in the analysis again reflects the fact that all these realizations contain at least one component (their nasality) in common with the correct form [wæ̃]. Nasality is also identifiably target-like in that it does not occur in the English phonemic inventory.

(3) Monosyllabic (as opposed to bisyllabic) realizations. Around one quarter of all attested forms are bisyllabic: [əʊɪn], [ʉɪn], [ɔːən], [jʉɒ̃], [ɔːɪn], [ɔːan], [ɔ̣.ən], [ụɪn], [ʉən], [œɪn], [əʔɪ], [iɒ̃], [oæ̃], [əʊʔɪn], [oʊɪn], [œən] and [uɛn]. The inclusion of this category in the analysis reflects the importance of correctly segmenting the letter string as a prerequisite for accurate decoding. Bisyllabic realizations, it is inferred, are likely to arise through the incorrect treatment of <o> and <i> as separate graphemes, which will almost inevitably result in an incorrect realization. Monosyllabic realizations, by contrast, may arise through a correct segmentation of the letter string into the trigraph <oin>. (They may also reflect an incorrect segmentation of the letters into the digraph

<oi> plus the consonant <n>; this problem cannot be detected using the classification into mono- or bisyllables).

Three sets of Pearson's chi square tests were conducted to compare the three groups (Intervention A, Intervention B, Comparison) at each time point according to the proportion of participants whose realizations of <oin> (a) fell into each of these three categories and (b) did not. In addition to the conventional significance level of $\alpha<.05$, the more rigorous threshold of $\alpha<.017$ was also considered, following a Bonferroni correction due to the fact that three analyses were run using the same data. A further difficulty arose because the three categories used as the basis for analysis overlap to some extent (since they are not mutually exclusive): for example, three attested realizations both (a) begin with [w] and (b) contain nasal vowels ([wɒ̃], [wæ̃], [wɔ̃]); two others are both (a) bisyllabic and (b) contain a nasal vowel ([oæ̃], [jʉɒ̃]). To avoid the problem of individual forms contributing to more than one significant result, each of these forms was included in only one analysis and excluded from subsequent ones.

#### 5.4.5.1 Realizations beginning with [w]

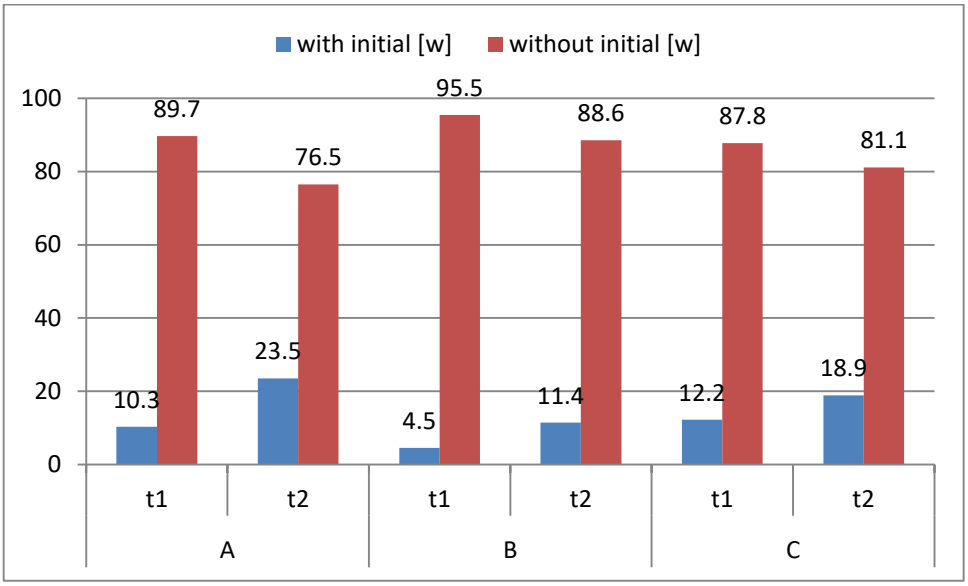
First, chi square tests were performed in order to test whether the three groups showed any change between t1 and t2 in terms of the number of realizations of <oin> beginning with [w]. A significant association was found at the level $\alpha<.05$ (but not at the more rigorous level of $\alpha<.017$) between time of testing (t1 or t2) and whether or not <oin> was realized with an initial [w] in intervention group A ($\chi(1)=4.239$, $p=.040$); no cells had expected counts below 5. However, no significant association was found either for the comparison group ($\chi(1)=1.287$, $p=.257$) or for intervention group B, where Fisher's

Exact Test was used due to the low expected counts in two of the four cells ($p$=.434). This indicates that intervention group A, but not the other groups, differed significantly between t1 and t2 in terms of the number of realizations of <oin> beginning with [w].

To explore whether there were differential patterns of t1-to-t2 change between the groups, the numbers of realizations beginning with [w] produced by the three groups at each individual time point were compared. No significant association was found between group membership on the one hand and whether or not <oin> was realized with an initial [w] on the other, either at t1, $\chi(2)=1.878$, $p$=.391, or at t2, $\chi(2)=2.589$, $p$=.287. One cell at t1 had an expected count below 5; however, a Fisher's Exact Test returned a similar, non-significant result ($p$=.418).

Nonetheless, in all three groups, a descriptive analysis of the data shows that there was an increase between t1 and t2 in the proportion of realizations beginning with [w] (Figure 5.6). This increase was greatest for intervention group A, reflecting the significant result of the between-time chi square test reported above. However, the proportion of realizations beginning with [w] also more than doubles in intervention group B; the increase for the comparison group is smaller.

**Figure 5.6 Percentage of participants in each group realizing <oin> with and without initial [w]**



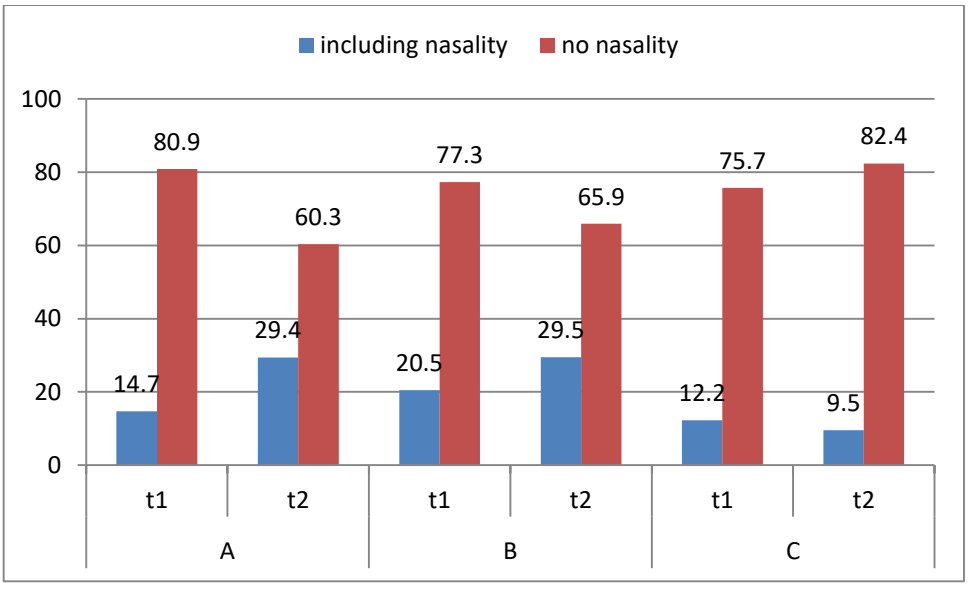

### 5.4.5.2 Realizations including nasality

The same procedures were adopted as described above in relation to realizations beginning with [w]; however, forms which began with [w] were excluded on the basis that they had already contributed to the previous analysis. First, chi square tests revealed a significant association (at the level $\alpha<.05$) between time of testing and the presence or absence of nasality in the realization of <oin> for intervention group A ($\chi(1)=5.559$, $p=.018$); the $p$ value almost reached significance at the more stringent level of $\alpha<.017$. By contrast, no significant association was found for intervention group B ($\chi(1)=1.261$, $p=.261$) or for the comparison group ($\chi(1)=0.083$, $p=.773$). No cells at either time point had expected counts less than 5. The test results suggest that intervention group A, but not the other groups, differed significantly between t1 and t2 in terms of the number of realizations of <oin> containing a nasal vowel.

Second, at t1, no significant association was found between group membership and whether or not the realization of <oin> included a nasal vowel, $\chi(2)=1.007$, $p=.605$. At t2, however, the association between the variables was significant even at the stringent level of $\alpha<.017$ ($\chi(2)=10.854$, $p=.003$). No cells at either time point had expected counts less than 5. An examination of the standardized residuals for nasalized realizations (Intervention A: +1.5; Intervention B: +1.0; Comparison: -2.2) indicated that participants in the comparison group produced significantly fewer nasalized forms than expected, whilst those in the two intervention groups tended to produce more than expected (although these figures did not reach the statistically significant threshold of $\pm 1.96$).

This reflects the descriptive observation that intervention groups A and B showed increases in the proportion of nasalized realizations between t1 and t2, of 14.7 and 9.0 percentage points respectively; however, the comparison group showed a decrease of just under 3 percentage points (Figure 5.7)[30].

**Figure 5.7 Percentage of participants in each group realizing <oin> with and without nasality**




[30] When the statistical tests described above were re-run including the three nasal realizations which also began with [w], the same pattern of significant results was obtained.

#### 5.4.5.3 Monosyllabic versus bisyllabic realizations

Finally, the same analyses were conducted in relation to the proportions of mono- and bisyllabic realizations of <oin>. All realizations beginning with [w] or including nasal vowels were excluded, since they had already contributed to the previous analyses. A significant association was found at the level $p<.05$ between (a) time of testing and (b) whether <oin> was realized with one syllable or two in intervention group A, using Fisher's Exact Test because two of the four cells had expected counts below 5 ($p=.021$); the test result also approached significance at the more rigorous level of $\alpha<.017$. By contrast, no significant association was found either for intervention group B (Fisher's Exact Test, $p=.277$) or for the comparison group ($\chi(1)=0.200$, $p=.655$). This indicates that intervention group A, but not the other groups, differed significantly between t1 and t2 in terms of the number of bisyllabic realizations of <oin>.

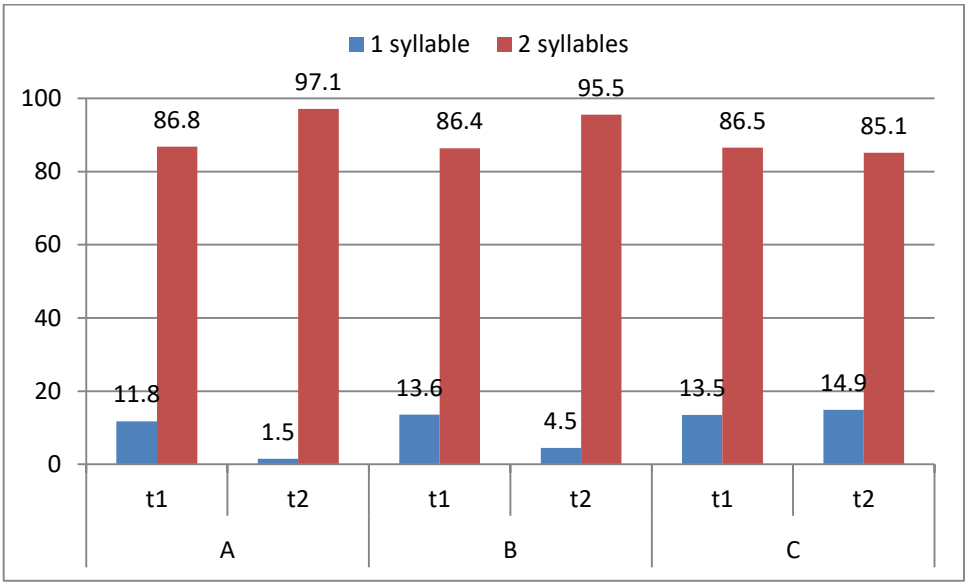
Subsequent analysis also found that there was no significant association at t1 between group membership and whether <oin> was realized with one syllable or two, $\chi(2)=0.118$, $p=.943$. No cells had expected counts less than 5. At t2, however, a significant association even at the more rigorous level of $\alpha<.017$ was found, where Fisher's Exact Test was again used because two out of six cells had low expected counts ($p=.006$)[31].

Finally, an examination of the standardized residuals for bisyllabic realizations (Intervention A: -1.8; Intervention B: -0.7; Comparison: +2.3) indicate that comparison participants produced significantly more of these than expected, whilst intervention

[31] Again, when the statistical tests described above were re-run including the three nasal realizations which also began with [w], the same pattern of significant results was obtained

participants (particularly those in group A) produced fewer than expected, although these values did not reach significance. This reflects the descriptive observation that the proportion of participants producing bisyllabic realizations of <oin> – which can be inferred to arise through incorrectly treating <o> and <i> as separate graphemes – decreased between t1 and t2 in both intervention groups but rose slightly in the comparison group (Figure 5.8). The decrease was of similar magnitude in both intervention groups.

**Figure 5.8 Percentage of participants in each group producing mono- and bisyllabic realizations of <oin>**




### 5.4.6 Summary

Participants in all groups and at both time points realized <oin> in *goinfrez* in a large number of distinct ways. Many of the attested forms were produced by only one or two participants across the sample as a whole. By contrast, a small number of realizations occurred more frequently. The most frequent, [ɔɪn], was produced by over a quarter of all participants at t1. Roughly an eighth of the realizations at t1 have two syllables rather

than one, suggesting that the vowels <o> and <i> may have been wrongly treated as separate graphemes.

At t1, a descriptive comparison of the most frequently attested realizations suggested that the groups were broadly (though not entirely) similar in terms of how they decoded <oin>. Statistical tests also found no significant difference between the groups, although limitations of the sample suggest that this result should be treated cautiously.

At t2, there was broad similarity in terms of which realizations predominated amongst the group as a whole, although the correct realization [wæ̃] and the 'acceptable' [wɔ̃] were more frequent at t2. When the groups were compared statistically at t2 in terms of the most frequently-attested realizations, no significant association was found between group membership on the one hand and how <oin> was realized on the other. However, it was suggested that this result may have underestimated the extent of the differences between the groups, because it involved a considerable simplification of the data (conflating over fifty distinct realizations into a single category of 'other' forms) in order to make the statistical test practicable. A descriptive analysis of the groups' realizations of the most frequent realizations of <oin> was therefore undertaken. This tentatively indicated that the two intervention groups showed different patterns of change compared to the comparison group.

To compare the groups in terms of all attested realizations (rather than simply in terms of a handful of the most frequent ones), three non-exclusive categories were created in order to measure the extent to which participants' realizations (a) included the target-like feature of initial [w]; (b) included the target-like feature of nasality; and (c) were

correctly decoded as monosyllables rather than as bisyllables (bisyllabicity being a likely indicator that the vowels <o> and <i> had been incorrectly processed as separate graphemes). At t1, statistical tests found no significant association between group membership on the one hand and any of these variables on the other. However, at t2, significant associations were found in respect of syllabicity and nasality. Both intervention groups produced fewer (incorrect) bisyllabic realizations than expected, particularly group A, whilst the comparison group produced significantly more than expected. Similarly, intervention group A produced significantly more nasalized realizations than expected; intervention group B also tended to produce more than expected whilst the comparison group produced fewer.

The groups' realizations of <oin> at each individual time point were also compared in terms of these three categories (presence of initial [w]; presence of nasality; syllabicity). Only intervention group A was found to show significant differences between t1 and t2. However, a descriptive analysis of the data appeared to reveal a pattern, whereby (a) all three groups produced more realizations containing nasality and initial [w] at t2 than at t1; (b) both intervention groups produced fewer bisyllabic realizations at t2 than at t1, whilst the comparison group produced slightly more.

## 5.5 Summary

The findings presented in this chapter were primarily intended to address Research Question 3:

*Do the programmes of instruction lead to any other changes in participants' realizations of French graphemes?*

Taken together, the evidence from the three grapheme tokens selected for analysis suggests an affirmative answer to this question. In four of the six statistical analyses conducted, intervention group A is associated with significant differences between the way the grapheme tokens are realized at t1 and t2; in one other case, the difference approaches significance. By contrast, the t1-to-t2 differences are not statistically significant in intervention group B or the comparison group (though they approach significance in one case in intervention group B). However, a descriptive analysis suggests that intervention group B generally tends to pattern with intervention group A and against the comparison group in terms of the changes in its participants' realizations between the two time points. In particular, in the case of all three target grapheme tokens, the intervention groups (A and B) tend to show a large decrease in the number of participants who produce the single, most frequently-attested pronunciation (/k/ for <ç> in *caleçon*; /ɔː/ for <ou> in *ougrien*; /ɔɪn/ for <oin> in *goinfrez*), whereas the comparison group shows little or no change in the number of participants producing this phonological form.

The data analysed in this chapter also produced some incidental findings, not directly related to Research Question 3, which nonetheless provide interesting insights into participants' decoding of the target graphemes. For example, in the case of all three grapheme tokens, there were a small number of dominant realizations produced by the majority of participants, with one phonological form in particular accounting for a large proportion of all realizations (as noted above). However, a wide range of additional

realizations were also produced, usually by one participant or very few participants in each case. The total numbers of attested realizations were high in all cases, but were higher for <ou> (N=39) than for <ç> (N=16), and higher still in the case of <oin> (N=62). An additional finding was that, in the case of <ou> (although not for <ç> or <oin>), the grapheme was realized in significantly different ways by the sample as a whole in the four different words in which it appeared, both at t1 and at t2. In other words, the realization of <ou> depended upon its orthographic context. These incidental findings will be discussed further in chapter 7.

Finally, it must be reiterated that the findings presented in this chapter relate to only three grapheme tokens out of a total number of 174 contained in the RAT. This restriction of the focus was inevitable given the limited space available. One grapheme type was therefore selected from each category of consonants, oral vowels and nasal vowels on theoretical grounds, based on particular difficulties that participants might be expected to encounter according to a contrastive analysis approach. The subsequent selection of the individual tokens on which the analysis was based, from the set of three or four possible tokens for each target grapheme type, was arbitrary. It is therefore possible that a different choice of grapheme types and tokens would have resulted in different findings. It is intended to explore this further in future analysis and reports.

# Chapter 6: Self-report interviews

## 6.1 introduction

This chapter presents the findings of the task-based self-report interviews (SRIs), addressing Research Question 4:

> *Do the programmes of instruction lead to changes in learners' strategic reasoning when decoding French words?*

Findings are presented in three parts. First, section 6.2 provides an overview of the main approaches to decoding which were reported or observed most frequently across all participants' SRIs at t1 and t2. Second, section 6.3 summarizes the extent and nature of any changes in each individual's approaches to decoding between t1 and t2. Finally, section 6.4 presents case studies of three participants (one per condition) whose approaches to decoding at each time point are described in detail.

## 6.2 Overview of approaches to decoding

### 6.2.1 Introduction

The SRIs contained some uninformative responses where participants were unable to explain why they pronounced a word in a given way, as exemplified in (1) to (2). This

may reflect the operation of automatized decoding processes, which in most cases are probably L1-based.

(1) "I don't know really it just sort of pops up in my head the answer … I don't even know how I do it" (SR-C-t1[32])

(2) "that one just sort of came automatically in my mind"; "I just said it … how it's spelt really" (LR-C-t1)

However, as in Woore (2010), such comments were rare, arising in the discussions of less than one word in ten at both time points; they were made by only six participants at t1 and four at t2. By contrast, most word discussions revealed a rich picture of strategic reasoning, as described below under the three broad categories of (1) strategies which participants used to help decode the target words; (2) knowledge bases which they drew upon to support their reasoning; and (3) metacognitive comments on the decoding process, such as reflections on their own self-efficacy, evaluations of their own output and evaluations of their approaches to the decoding tasks. Participants' self-reported reasoning could not always be straightforwardly classified within a given category: there were some cases of overlap. However, this was not considered problematic, since the three categories did not serve directly as the basis for analysis; rather, they were simply used to structure the descriptions of participants' reasoning (section 3.6.4.3). This reflects the decision to adopt a 'holistic' approach to the analysis of the self-report data rather than one based on strict coding and quantification.

[32] In quotations from the SRIs, dots (…) indicate material omitted for clarity, such as false starts or repetitions. Participants are identified, in brackets, by their pseudonym initials, their group (A=Intervention A, B=Intervention B, C=Comparison) and the time point (t1/t2). Where appropriate, the target word under discussion is indicated in angled brackets. Illustrative quotations are generally taken from non-case study participants, since these are analysed in detail later.

### 6.2.2 Decoding strategies

Whilst the SRIs provided evidence for a number of strategies which were used by two or three participants each, three strategies stood out as being used by a large majority of participants at both time points. These were also the three strategies used most frequently by participants in Woore (2010). Each is considered in turn.

#### 6.2.2.1 Dividing words into smaller units

The most frequent approach to decoding evidenced in the SRIs involved dividing the target words into smaller orthographic units. This occurred in around half of all word discussions at t1 and over six in ten of those at t2. In some cases, the subdivision of the word is observable but not commented on explicitly, as in example 3:

(3) "erm the E would be [dɛ] then the D A I N is [dan]" (<dédain>; VP-B-t2[33])

Nevertheless, in over half of cases the participants stated that this was something they did consciously in order to help them work out the word's pronunciation (example 4). Sometimes, participants also commented explicitly on the recombination of the word parts whose pronunciations they have worked out separately (example 5), recalling a synthetic phonics approach.

(4) "well first I kind of split the word in into like sections so work on each section at a time" (<dédain>; EH-B-t2)

[33] Capital letters denote alphabetic letter names, pronounced in English. These are placed in inverted commas where ambiguity would otherwise arise.

(5) "I … like split it up in the sections like [diː] and then I split it up and then I say [dɛɪn] so and then I just put them both together and then it's [diːdɛɪn]" (<dédain>; DC-C-t2)

Occasionally, the strategy of breaking words up into smaller units is explicitly evaluated by participants. For example, DC (comparison group) commented at t1 that splitting *haubert* into three sections "just like makes it a bit easier ... because it's not like a big block". However, this evaluation of the strategy – like all others that occurred – focuses solely on the extent to which it makes the decoding task 'easier' or more manageable, rather than on the accuracy of the outcomes. Clearly, whilst subdividing a word is potentially a helpful strategy, it will not necessarily lead to a correct pronunciation: this will depend on the extent to which French symbol-to-sound correspondences are applied to decode each part of the word.

Further, in around a fifth of cases in which participants divided words up into smaller units, the locus of division did not coincide with grapheme boundaries: for example, digraphs were fractured into two separate letters. This occurred most commonly in the case of <gn> (example 6), which in French but not in English often forms a single grapheme, but also with some vowel digraphs such as <au> (example 7).

(6) "what I'd do is erm split it up … after the G so [pwɔɪŋg] no (3.5) [pɔɪg] or something like that and then I'd say [nɑːd] so [pɔɪgnɑːd]" (<poignard>; DC-C-t2[34])

(7) "erm I split it after the U cos that says Bert ... and that's like an English name ... erm and then I split it after the A ... and it says [ha] ... and then you add the U and then it says [hajʉ]" (<haubert>; DC-C-t1)

[34] Numbers in brackets indicate the lengths of pauses longer than 2.0 seconds. Shorter pauses were not recorded.

In these examples, the strategy of dividing a word into smaller units actually militates against a correct pronunciation, making it less likely (or impossible) that the letters forming the digraph will be realized correctly. Better knowledge of the French grapheme inventory may therefore allow the strategy to operate more effectively in some cases.

#### 6.2.2.2 References to known words

The second most frequently attested approach to decoding in the SRIs involves participants making references to known words to support their discussions about the decoding of target words. This occurs in around one third of word discussions at t1 and over one quarter at t2.

It is not always clear whether a given reference to a known word is part of a conscious strategic action, undertaken proactively when sounding out the word, or whether it is a case of a participant simply noticing orthographic or phonological similarities as they reflect on the word further during the self-report process. For example, DC (comparison group, t2) – having already suggested the pronunciation [hjʉɛlʔɛɪdʒ] for *huilage* – comments that "it sounds like *haulage* in English"; however, this does not seem to lead him to revise his pronunciation of the target word in any way, perhaps suggesting that he might more appropriately have said that 'the spelling reminds me of the English word *haulage*'.

In most cases, however, participants' references to known words do seem to be conscious strategic actions which can be understood in terms of the analogy strategy

chain (section 2.7.4), as described by LR (comparison group, t2) when discussing her decoding of *dédain*: "I sort of see if any of it I've heard in another word … and I put that bit aside and then I say the first bit just how I think it's said and then I add the bit that I've heard before". However, whilst the analogy strategy has the potential to support decoding, several criteria must be met for it to operate effectively. The first is that an appropriate source word must be identified. However, in over half of all cases when participants refer (or attempt to refer) to a known source word at t1 – and in about one quarter of cases at t2 – that source word is English rather than French (examples 8-9). No participant shows any awareness that this may be problematic, even though using English source words will almost inevitably result in incorrect pronunciations.

(8) "I just thought of 'England' but I just changed it [gɪŋgləɴd]" (<cinglant>; JG-C-t1)

(9) "I split it after the U cos that says 'Bert' and that's like an English name" (<haubert>; DC-C-t1)

Further, in about one fifth of cases when participants attempt to use the analogy strategy, their search for a source word is unsuccessful (examples 10-11; note that the intended source word is again English in both cases).

(10) "cos in English the Q's like quick and that but I can't think of nothing in the middle with Q" (<maquille>; JG-C-t1)

(11) "it's like sometimes in French you get words that like hamster which is exactly the same so you know it off like that and there's words that are similar … like [ʃat] like cat … so you can kind of relate them to each other but with this you don't have an English word that is like it" (<jaugez>; EC-B-t1)

French source words are sometimes also used. This occurs in about one quarter of uses of the analogy strategy at t1 over half of cases at t2. About a third of these references to

French source words appear to be effective in supporting a correct (or at least partially correct) realization of the target word, as in examples 12-13:

(12) "and then like I know the A G E from like [bɹɪkɔlɑʒ]" (<huilage>; VP-B-t2)

(13) "I recognize that bit from like [blø]" (discussing the <eu> in <ferreux>; MR-C-t2)

However, even when a French source word is identified, problems arise in around two thirds of these cases. Most often, this is because the participant recalls either its written or spoken form incorrectly. For instance, in example 14, MR wrongly thinks that *français* ([fʁɒ̃se], 'French') is spelt with an acute accent like the target word *dédain*.

(14) "in like [fɹɒnsɛɪ] I think it's got either an E or an A with a thing over it makes it sound different" (<dédain>; MR-C-t2)

Finally, another problem occurs when the letter string of the target word is incorrectly segmented into graphemes, as in example 15. Here, SR recognizes the familiar letter string <moi> within the target word *témoin*, but this reduces the likelihood of the nasal grapheme <oin> being decoded correctly (although in this case it does happen to result in an 'acceptable' realization). Again, this problem may be avoided through better knowledge of the French grapheme inventory.

(15) "well I mean [tɛɪ] is obviously T E then [mwæ] M O I is 'me' in French and then then just add … an N on the end of that and [tɛɪmwan]" (<témoin>; SR-C-t1)

In summary, whilst the analogy strategy is sometimes effective, it may be derailed by various problems. Most often, these arise through participants' inability to locate appropriate and/or accurately stored French source words. Programme A (the 'poem

approach') was of course intended to address some of these issues, by providing a bank of securely-known source words and providing practice in executing the strategy.

#### 6.2.2.3 Trying out and evaluating possible pronunciations

Participants often make evaluative comments about whether their attempted pronunciations 'sound right'. In some cases, however, they make repeated attempts at producing the target word in what appears to be a deliberate strategy of 'trying out' possible pronunciations, as if to facilitate the evaluation process (examples 16-17). This "suggests that they were finding it hard to decide whether their pronunciation was acceptable, reflecting in turn their lack of a secure basis for doing so" (Woore 2010:13). Altogether, the strategy of trying out possible pronunciations appears to occur in around one quarter of word discussions at t1 and around one fifth at t2.

(16) "I think it would be something like erm [hɑːbət] [hɑːbəː] [hoʊbəː] something like that I think it would be \ … \ I think erm [hɑː] [hɑːbɛː] think [hɑːbə] yeah" (<haubert>; LT-A-t1)

(17) "[p] [pɔɪŋg] [p] [pɔɪgnɑːd] [pɔɪ] [pɔɪnɑːd]" (<poignard>; VP-B-t2)

In example 16, LT appears to be trying to decide whether or not to sound the final <t> in *haubert* and how to pronounce the two vowels <au> and <e>. In example 17, it is clearly the medial consonant <gn> which is the focus of the participant's attention, with the remainder of the word's phonological form remaining constant across the different pronunciations.

It is possible that this strategy of trying out alternative pronunciations occurs more frequently than would appear from the available evidence, because participants may

employ it subvocally or using their 'inner voice'. Evidence for this was provided by occasional whispering, murmuring or silent lip movements, which – though impossible to analyse systematically on the basis of the audio recordings – were sometimes commented on by the interviewer. There are also occasional comments such as EC's description of her approach to decoding *témoin*, which she says involved "just in my head thinking through what it would sound like".

Despite its apparent frequency, the strategy of trying out possible pronunciations of a word or grapheme(s) will not in itself lead to correct decoding outcomes, unless participants have a secure basis for evaluating the pronunciations they produce. The evaluations often seem to be based on a rather vague 'feeling' for what 'sounds right' (see section 6.2.3.3), which is often defective or incomplete. For example, whilst the pronunciation of *haubert* which LT finally decides upon in example 16 correctly omits the SFC, both the vowels and initial /h/ are incorrect. Similarly, almost all aspects of the final pronunciation in 17 are incorrect.

### 6.2.3 Knowledge bases

Two main types of knowledge were identifiable which participants drew upon to support their decoding. The first was where participants referred to conscious knowledge either (a) about how a specific grapheme or combination of graphemes should be pronounced or (b) about the nature of French orthography or phonology more generally. The second involved participants appealing to an instinctive 'feeling' for what 'sounds French' (as mentioned above).

#### 6.2.3.1 Knowledge about specific GPC

***a. Diacritics***

In around half of all word discussions at t1 – and around two thirds at t2 – participants referred to knowledge about the pronunciation of individual graphemes or combinations of graphemes. A little under a quarter of these cases at t1 – and a little over one third at t2 – concerned diacritics, even though these occurred in only two of the words covered by most participants at each time point (t1: *témoin, bêcheuse*; t2: *dédain, lancée*). This disproportionate representation of diacritics in participants' comments may reflect the fact that – since they do not exist in English orthography except in occasional loan words – they force those participants who notice them to move beyond a reliance on L1-based decoding and to think specifically about L2 GPC.

Occasionally, participants' comments concerning the effects of a diacritic are correct (allowing for some anglicization in the realization of the L2 phonemes), as in example 18.

(18) "well the E instead of saying [timwan] like with an ordinary 'E' I saw the sign on top and I thought well change it over to [tɛɪ] instead of [ti]" (<témoin>; SR-C-t1)

Much more frequently, however, their beliefs are incorrect, sometimes reflecting apparently idiosyncratic, personal hypotheses about French orthography. For example, TM (Intervention B, t2) thinks that the acute accent in *témoin* "changes [the E] to like the sound of like 'A'", whilst AB (Intervention A, t1) thinks "it might be … just that [the] letter is missed out". At least two participants also seem to believe that the acute accent

relates to a different aspect of phonology entirely, having to do with stress placement, as in example 19. This recalls the participant in Woore (2010) who believed that the cedilla in *caleçon* indicated a rising intonation pattern.

(19) "you sort of like pronounce the E sort of harder and standing out more … like [d**i**dɛɪn] instead of if it wasn't there [did**ɛɪ**n]" (<dédain>; AB-A-t2; **emphasis** indicates stress placement)

More frequently still, participants state that they know the diacritic affects pronunciation, but do not know in what way (examples 20-21). This accounts for just under half of all explicit comments on the pronunciation of graphemes with diacritics at both t1 and t2.

(20) "I know the little thing above the E should make it sound different but I can't remember how" (<lancée>; MR-C-t2)

(21) "that one's got the thing on top and it might do the pronunciation a bit different" (<lancée>; DC-C-t2)

***b. 'Silent' letters***

'Silent' letters (graphemes with no phonological realization) are another recurring feature in participants' explicit comments on symbol-sound correspondences, occurring in around one third of cases at both time points. The SFC 'rule' is particularly well represented (examples 22-23), often correctly leading to the final consonant not being realized phonologically; nonetheless the 'rule' is again often expressed in rather vague and tentative terms (example 23).

(22) "and then miss out the D cos of the last letter you don't sound it" (<poignard>; EC-B-t2)

(23) "erm shouldn't pronounce the S on the end of words I think … well I'm not too sure if it is the S that you shouldn't pronounce is it the vowels that you

shouldn't pronounce no that's wrong though or is it I'm not sure" (<piochais>; TM-B-t2)

There are also some cases where the principle of unsounded final consonants is misunderstood or overgeneralized. This occurs in example 24, where LT fails to take account of the <e> in word-final position. Similarly, in 25, EC seems to conceptualize nasal vowels simply in terms of omitting the final nasal consonant, a misperception also demonstrated by several other participants.

(24) "the S normally you wouldn't hear an S on the end of a word a French word" (<bêcheuse>; LT-A-t1)

(25) "think about how erm you don't always say the last letter and like pronounce it as such \…\ and then I wouldn't say the N" (<témoin>; EC-B-t1)

Unlike in the case of diacritics, the high frequency of participants' comments on 'silent letters' cannot be explained on the basis of any *orthographic* features specific to the L2. It may be instead that the perception of 'silent letters' constitutes a highly salient feature of French orthography for English-speaking learners. The SFC principle is also readily formulated as a simple 'rule of thumb' which may be easy for learners to remember. Conversely, it must also be noted that SFCs appear more frequently than diacritics in the target words, occurring in seven of the thirteen words covered by most participants across both time points.

***c. Other graphemes***

Participants also make explicit comments about a number of other graphemes and grapheme groups. Again, some of these comments are correct (examples 26-27) but many are not, being based on English GPC (28-29). Example 28 is particularly

interesting in that MR appears to apply her knowledge of the English split digraph <o_e>, again with no apparent awareness that transferring this GPC from the L1 to the L2 may be problematic.

(26) "the O and the I I'd say [pwa]" (<poignard>; LT-A-t2)

(27) "normally in English I'd say [ɛɪdʒ] at the end but I said it as [aʒ] because it's French" (<huilage>; RS-C-t2)

(28) "[əʊni] \...\ because the E turns the O to the [əʊ]" (<maçonne>; MR-C-t2)

(29) "you know what E and Z makes yeah which is [ɛz] and G and E is [gɛ]" (<jaugez>; MR-C-t1)

***d. General comments on French decoding***

Finally, in a little over one in ten word discussions at both t1 and t2, participants comment on French decoding more generally, often to support or justify their statements about the sounds of specific graphemes (described above). Most cases can be divided into two broad categories. Around half at both time points concern patterns in French orthography or phonology (examples 30-31).

(30) "like A U cos you see it quite a few times in English it's like quite common but then in French it's not" (<jaugez>; DC-C-t1)

(31) "cos I think there's loads of letters that are silent but you don't actually say them confuses me" (<témoin>; AB-A-t1)

The remainder are vague observations that French decoding is somehow different from English, although the nature of this difference remains unspecified (examples 32-33). These comments recall Woore's (2010:15) conclusion that participants' attempts to pronounce target words were often "explicitly driven by knowledge that French decoding is different from English decoding. However, beyond this basic awareness of difference,

participants lacked specific knowledge about what the French pronunciations of words should be".

(32) "because … it's like from a different country I would ... think they would pronounce it different" (<bêcheuse>; RS-C-t1)

(33) "English language is like think it's different to French" (<jaugez>; JG-C-t1)

### 6.2.3.2 Origins of knowledge

As in Woore (2010), the origins of participants' knowledge about French decoding are usually unclear. However, in some cases participants claim that it derives from their own observations of the language (examples 34-35).

(34) "it's [ɑʒ] … because in other words I've heard it like that" (<huilage>; LR-C-t2)

(35) "well just erm from looking at French words … they hardly ever pronounce the X" (<ferreux>; SR-C-t2)

In other cases, participants say their knowledge results from explicit instruction. When this knowledge is correct (as in example 36, allowing for the anglicized pronunciation), this suggests that explicit instruction can successfully impact upon learners' strategic reasoning. However, in many other cases, the knowledge said to result from explicit instruction is recalled only vaguely or imperfectly (example 37).

(36) "we like have kind of learnt that the the E with the little accent above it is an [ɛɪ] sound so that's how I did it" (<dédain>; EC-B-t2)

(37)
**Participant**: cos French sometimes … I think you miss out a letter or something
**Researcher**: how do you know that
**Participant**: erm I think the teacher told us
**Researcher**: and which letters do you miss out
**Participant**: I don't know I can't remember (<témoin>); JG-C-t1)

It therefore seems that we cannot take for granted that explicit instruction in French GPC will produce secure foundations of conscious knowledge to support decoding, which was one of the principles underlying the programmes of instruction in the current study.

#### 6.2.3.3 'Feeling for French'

As mentioned in section 2.1, in around one fifth of all word discussions at each time point, participants appeal to a vague 'feeling' for what 'sounds French' to help them evaluate their attempted pronunciations. In exceptional cases, this instinctive sense of what 'sounds right' leads to correct judgments, as in example 38; however, LR's RAT scores were far above the mean at both time points, suggesting that she may have been unusually proficient in French decoding. She may therefore have been able to draw upon a more secure bank of implicit knowledge than other participants.

(38) "[fɛɹʉks] it just didn't really sound right … so I took it off and then I did [fɛɹʉː] and then it sounded much better" (<ferreux>; LR-B-t2)

Despite the frequency with which participants invoke this 'feeling for French', however, in most cases it leads only to a vague sense that the pronunciation they have produced might be somehow incorrect; it does not help them identify a more accurate pronunciation (examples 39-40).

(39) "I'm just not sure I'm just not feeling it sounds right and stuff" (<poignard>; EC-B-t2)

(40) "it just doesn't sound right … it just sounds like it should be something else … it just doesn't really feel like that's how it should be pronounced" (<dédain>; SR-C-t2)

### 6.2.4 Metacognition and affective reactions

The SRIs at both time points are permeated by an overwhelming sense of tentativeness and hesitancy. Indeed, most participants explicitly mention doubts about their pronunciations on at least one occasion (examples 41-42).

(41) "I dunno how to like say it properly" (<dédain>; JG-C-t2)

(42) "just cos there's lots of like groups of letters that I don't know how to pronounce" (<piochais>; EC-B-t2)

Conversely, most participants also express confidence in their pronunciations at least once. Very occasionally, the pronunciation in question is indeed correct; however, in the overwhelming majority of cases, their confidence is misguided, as the pronunciations reflect English GPC or derive from analogies to English words (examples 43-44).

(43) "I suppose it's quite easy cos you just split it up like [pəʊ] and then [ɪg] and then [nɑd] [pəʊɪgnɑd]" (<poignard>; JG-C-t2)

(44) "I'm happy with that first bit ... and I'm confident in that bit cos it's like basically like 'chairs' … except it's just an I and instead of R" (<piochais>; DC-C-t2)

Occasionally, participants express a sense of alienation when faced with target words, which appear 'odd' or 'weird' in comparison to the familiar spellings of English words (examples 45-47). There is a sense here of English orthography as the norm, French as deviant and chaotic. This is ironic given that French orthography is actually shallower (more phonologically transparent) than English (section 2.4.3).

(45) "like really weird letters that go tog[ether] in like a word ... it's like there's no actual word in English or anything like it in English that looks like it" (<jaugez>; EC-B-t1)

(46) "there's more just letters all dotted about … it just looks like a load of gobbledygook" (<poignard>; AB-A-t2)

(47) "this one is another sort of odd word really ... the first four letters look … kind of random really I think they're just picked out letters from somewhere" (<quignon>; VP-B-t1)

## 6.3 Summary of changes in participants' approaches to decoding

A detailed examination of the data showed that there were some small variations in each participant's approaches to decoding at the two time points. This was to be expected given that (a) the strategies or knowledge bases used may depend to some extent upon the particular words being decoded, which differed between the two time points; and (b) contextual factors may affect participants' motivation both for the decoding task itself and for the act of self-reporting. However, for almost all participants, there was little or no evidence of any major differences in strategic reasoning at t1 and t2, in any of the three categories of analysis. This was despite the fact that both programmes of instruction aimed to strengthen participants' knowledge of French GPC, whilst the 'poem approach' additionally aimed to develop their use of the analogy strategy.

Table 6.1 summarizes any t1-to-t2 changes observed in each participant's approaches to decoding, cross-referenced with their RAT scores at each time point.

**Table 6.1 Summary of differences in participants' strategic reasoning between t1 and t2.**

| Partici-pant (school; gender) | Attain-ment level in French* | RAT score (t1) | RAT score (t2) | RAT score change (t1 to t2) | Differences in approaches to decoding observed between t1 and t2 |
|---|---|---|---|---|---|
| *Comparison mean* | | 40.2 | 43.4 | 3.2 | |
| *Intervention mean* | | 40.7 | 51.8 | 11.1 | |
| AB (A;♀) | Low | 29 | 34 | 5 | There is no evidence of any significant changes in AB's reasoning. The predominant strategy at both time points is one of breaking words up into smaller units; this seems to be linked to attempts to identify English words embedded within the French ones. |
| LT (A;♀) | High | 68 | 53 | -15 | There is no evidence of any significant changes in LT's reasoning, although at t2 she perhaps expresses herself less tentatively than at t1; however, it is possible that this simply reflects a greater ease in the interview situation at t2. Note that the marked decrease in LT's RAT scores may reflect the fact that her test was disrupted at t2. |
| LV (A;♀) | High | 47 | 73 | 26 | There appear to be some differences in LV's approaches to decoding at the two time points. The main difference is that the trying out of different possible pronunciations, which permeates her discussions at t1, is observed much less frequently at t2. There is also an apparent strengthening of her knowledge of French GPC, which is generally (though not always) more accurate at t2 than at t1. Conversely, some notable misconceptions identifiable at t1 persist at t2, despite the explicit instruction. |
| VP (B;♂) | Low | 38 | 44 | 6 | There is no evidence of any significant shift in VP's reasoning, although there is an apparent tendency for him to express himself more tentatively at t2 than at t1. |
| TM (B;♂) | Low | 38 | 50 | 12 | There is no evidence of any significant shift in TM's reasoning. He is not very loquacious but his predominant strategy at both time points is one of breaking words up into smaller units. |
| EC (B;♀) | High | 54 | 59 | 5 | There is little evidence of any changes in EC's reasoning, except that (a) she makes less use of the analogy strategy at t2 than at t1 and (b) there is perhaps a slight increase in the extent of her knowledge of specific GPC, which is her predominant resource at both time points. |
| EH (B;♀) | High | 41 | 45 | 4 | There is little evidence of any changes in EH's reasoning; indeed some misconceptions which hinder her L2 decoding at t1 persist at t2, despite the decoding instruction. |

* as nominated by the teacher.

*Continued* →

| | | | | | |
|---|---|---|---|---|---|
| DC (C;♂) | Low | 32 | 33 | 1 | There is no evidence of any significant changes in DC's reasoning, nor even in his apparent lack of awareness that the English and French decoding systems differ. |
| JG (C;♂) | Low | 23 | 34 | 11 | There is some evidence of some slight changes in JG's reasoning, but these seem to have been negative rather than positive: strategies and knowledge bases which are successfully deployed at t1 do not appear at t2. There is also more evidence at t2 of feelings of negative self-efficacy in relation to L2 decoding. Note that, although JG shows a considerably larger RAT score increase than most others in the comparison condition, he starts from a very low baseline at t1, when his score is well below the mean values. |
| LH (C;♀) | High | 50 | 53 | 3 | At both time points, this participant expressed herself hesitantly, interpreted by the interviewer as the result of anxiety; at t2, she elected to stop the interview after only four words, limiting the scope of any between-time comparisons. However, there appears to be no evidence of any significant changes in her reasoning; indeed, there is no conclusive evidence that she deploys any strategies at either t1 or t2. At both times, she shows a vague awareness that the French and English decoding may differ, but cannot say how. |
| OK (C;♂) | High | 44 | 61 | 17 | Although OK appears to be a highly reflective learner who adopts a critical stance to the relationship between the English and French orthographies, there is little apparent change in his reasoning. |
| MR (D;♀) | Low | 29 | 38 | 9 | There is no evidence of any significant changes in MR's reasoning, although she perhaps shows a greater degree of uncertainty in her decoding outcomes at t2; however, this could be due to contextual factors. |
| RS (D;♂) | Low | 44 | 47 | 3 | There is little evidence of any changes in RS's reasoning. At both time points, there is a predominance of comments on the fact his pronunciations do not 'sound right' but little evidence for the deployment of strategies or declarative knowledge to help him improve them. |
| LR (D;♀) | High | 63 | 78 | 15 | LR expresses herself very hesitantly at t1, interpreted by the interviewer as the result of anxiety; at t2, she appears more confident, which may affect between-time comparisons. Nonetheless, there is little evidence of change in her reasoning. At both time points, she appears to rely mainly on an instinctive 'feeling' for how the target words should be pronounced; at t2, this instinct appears more likely to lead to correct decoding outcomes. |
| SR (D;♂) | High | 59 | 66 | 7 | There is no evidence of any significant changes in SR's reasoning. At both time points, he appears to rely heavily on his 'feeling' for how the target words should be pronounced, even though this instinctive judgment is almost always incorrect. |

## 6.4 Case studies

This section analyses in detail the reasoning of three participants (one from each condition), as revealed in their SRIs at each time point. Due to the small overall sample size, it is not possible to claim that these individuals exemplify the reasoning of participants in their respective conditions; rather, they are selected as being interesting cases who provide examples of some of the ways in which beginner-learners' approaches to L2 decoding may develop over several months of learning French under different types of instruction. All case studies are drawn from the group of teacher-nominated 'high attainers' in French, in order to reduce some of the scope for variation which might be expected on the basis of different proficiency levels.

The three participants selected for detailed analysis were:

(1) OK (Comparison), because the apparent lack of change in his approaches to decoding contrasts with a relatively large RAT score increase. He also appears to be a highly reflective learner; his critical stance towards the relationship between English and French decoding might have been expected to lead to greater changes in his strategic reasoning as he gained experience of the language;

(2) EH (Intervention B), because, although she draws upon a range of strategies and knowledge bases to support her decoding at both time points, these are often ineffective or only partially effective in promoting correct outcomes. There is also little apparent development in her reasoning between the two time points despite following the programme of instruction. Further, her RAT score increase is below the mean value for the intervention group.

(3) LV (Intervention A), because she does demonstrate some differences in her approaches to decoding at the two time points: her conscious knowledge of French GPC seems to be stronger. This may contribute to her very large RAT score increase. However, some misconceptions which

hinder her decoding at t1 seem to persist at t2, despite having followed the programme of instruction.

### 6.4.1 Case study 1: OK (comparison group)

#### 6.4.1.1 Background

OK (class C3) reported in his BIQ that he studied French at primary school for about one hour per week for six months; he also learnt a little German and Spanish there. He reported having no French-speaking relatives or friends, no extracurricular contact with French and no history of visiting a French-speaking country. His verbal quotient in the CAT tests was 131, considerably above the sample mean of 105.5 (range: 78-141; N=142 valid cases). His average KS2 point score of 34.33 was similarly high (sample mean: 29.3; range: 19-35; N=129 valid cases).

OK's SRIs covered eight target words at t1 (*témoin, quignon, jaugez, bêcheuse, cinglant, haubert, obtient, maquille*) and six at t2 (*dédain, poignard, huilage, ferreux, piochais, lancée*).

#### 6.4.1.2 Approaches to decoding at t1

***Strategies***

At t1, at least three possible strategies to support decoding can be identified in OK's SRI, reflecting those used most frequently by the sample as a whole: (1) breaking words up into

smaller units; (2) referring to known words with similar spellings; (3) trying out and evaluating possible pronunciations by saying them aloud.

***a. Dividing words into smaller units***

When discussing three of the eight target words, OK reports that he "split it up into two parts", probably into syllables (as he perceives them): *quig/non*; *cin/glant*; *hau/bert*. Whilst this may make the decoding task more manageable, the resulting word segments usually appears to be decoded using English GPC. This is exemplified in OK's description of how he builds up his pronunciation of the first 'syllable' of *quignon*: "the Q and the U make it go [kwɐ] and the I gives it like a [kwɪ] and there it's G which makes it go [gə]". Further, his subdivision of this word incorrectly splits <gn> into two separate graphemes, making it much less likely (if not impossible) to generate the correct phonological form /ɲ/.

***b. Drawing on known words as sources of evidence***

There are four instances (occurring in three word discussions) where OK supports his reasoning about how the target word should be pronounced through reference to a known word which either looks or sounds similar to it. In three of these cases, the source word is English. For example, when discussing *obtient*, he comments that "it looks a bit like 'obedient' ... so I used that … except missing out … the E D bit". His pronunciation of the target word ([ɒbtiənt]) then reflects various aspects of the English source word.

In this example, OK's remark that the target word "looks a bit like 'obedient'" suggests that the source word may come to mind simply through his noticing of orthographic similarities.

Contrastingly, in the case of the target word *maquille*, he appears (though this cannot be known for certain) to engage in a deliberate search for appropriate source words in order to inform his decision as to whether <qu> (or <q> as he perceives it) should be pronounced [k] or [kw]. Again, however, the source word is English:

> in the English language sometimes things like that happen but it could be [məkwɪl] because it's got a U at the end of it ... because the only times in English when … it's not got a U so it sounds different is like the country 'Qatar'

OK explains his reliance on English here by his limited knowledge of French orthography: "I'm not sure about what happens in French if it's Q U if there's any French sayings … that sound like [kə]". This suggests that he *is* aware of the possible limitations of using English source words to inform his reasoning. OK is in fact highly likely to have encountered various French words containing <qu> (e.g. *quel âge as-tu?*, [kɛlɑʒaty], 'How old are you?'; *quinze*, [kæ̃z], 'fifteen'; *qu'est-ce que c'est?*, [kɛskəse], 'what's that?'). However, these forms were apparently not accessible to him as sources of analogy.

There is also one example at t1 where OK refers to a French source word to support his reasoning. Having suggested that the grapheme <au> in *haubert* might be pronounced "[hɔː] … like O R", this sound then seems to recall the known French word *hors d'œuvre* ([ɔʁdœvʁ], 'starter'). However, OK's stored phonological representation of this word is heavily anglicized ([hɔːdəːvz]), perhaps because he knows it as the French-into-English loanword rather than directly from French. This would clearly reduce its potential effectiveness as part of the analogy strategy. In any case, however, his use of this source word to inform his realization of <au> is based on a mistaken belief about its spelling: "I'm not sure is [hɔː] erm [hɔːdəvz] erm is that spelt H A U".

***c. Trying out and evaluating possible pronunciations***

In two word discussions, OK appears to use the strategy of 'trying out' possible pronunciations. For example, his uncertainty (mentioned above) about how to realize the <qu> in *maquille* leads him to try out two different pronunciations of the word, one of which is finally decided upon as sounding more 'French':

> I'm not so sure … it's the [kj] erm [kjʉ] bit it's Q and U because it could be [məkwılə] or [məkılə] … but [makwıl] doesn't sound as French as ... [makılə]

Similarly, when discussing *cinglant*, he becomes aware that his realization of the second vowel has changed from [a] to [ɑ], leading him to try out and evaluate the two possible pronunciations:

> [sıŋglant] the erm it and [sıŋglɑnt] [sıŋglɑnt] seems more yeah I think ... [glɑnt] [sıŋglɑnt] with the 'A' I think

Interestingly, in both these cases, OK's instinctive feeling for what 'sounds right' in French rejects the realization which would typically be generated by English GPC and correctly identifies the 'better' pronunciation, at least as reflected in the lenient RAT mark scheme. However, it appears not to be sensitive to other non-target-like aspects of the words' pronunciations: for example, in *cinglant*, he does not comment on his anglicized realization of <in> as [in] or of the final <t>.

***Knowledge bases***

In three word discussions, OK comments explicitly on the pronunciation of specific graphemes or combinations of graphemes. First, discussing *témoin*, he says that "the 'O' and the 'I' make it sound like [ɔɪ]". Second (as quoted above), he describes how he builds up the pronunciation of the first syllable of *quignon* grapheme by grapheme: "the Q and the U make it go [kwɐ] and the I gives it like a [kwɪ] and there it's G which makes it go [gə]". In both these cases, however, all the GPC he describes are consistent with English orthography but incorrect for French. In the third case, when attempting to explain why the second part of *cinglant* should be pronounced [lant] in French but [lɑnt] in English, he describes (using non-standard metalanguage) a feature of French phonology which he says he has noticed:

> **OK** I think it's because I've heard … things and the A sometimes … sound like that
> **RW** can you think of examples
> **OK** well [ʒuma] [ʒumapɛl]* has erm a higher A as well
> **RW** OK what do you mean by higher that's interesting
> **OK** well it just it seems to have like a lift at the end which makes it sound a bit different
>
> * *je m'appelle*, /ʒəmapɛl/, 'my name is'

OK may be thinking here of the contrast between the lengthened, back /ɑ/ which occurs in certain orthographic contexts in his variety of Southern English (e.g. *class,* pronounced [klɑːs]) and the front /a/ which would occur in many of these contexts in supralocal French (e.g. *classe*, 'class', pronounced [klas]). However, the <a> in *cinglant* does not follow this pattern in French because it forms part of the nasal grapheme <an> (/ɒ̃/); OK mistakenly treats it as a separate grapheme in its own right. Again, therefore, his reasoning is undermined by inadequate knowledge of the French grapheme inventory. This is despite the fact that he has almost certainly encountered words containing <an> early in his MFL studies (e.g. *français*, [fʁɒ̃se], 'French'; grand, [gʁɒ̃], 'big').

Finally, in four cases (in three word discussions), OK reports drawing on a vague 'feeling' for what 'sounds French', linked in two cases (described above) to the strategy of trying out and evaluating possible pronunciations. In two other cases, the evaluation process appears to take place more quickly, presumably reflecting greater certainty about the pronunciations concerned. One example occurs when he explains why he pronounced *cinglant* with an initial [s] rather than [k]: "[kɪŋglant] didn't seem erm didn't sound … as good as [sɪŋglant] so … I just did what I thought would … sound right". The other example relates to the initial <j> in *jaugez*: "well the J it doesn't seem right to s(ay) [dʒ] [dʒɑːgɛz] so they'd go [ʒɑː] sound".

In both these cases, OK's instinct for what 'sounds right' identifies the correct pronunciation of the target grapheme; in the case of *jaugez*, this is despite the fact that the French and English GPC differ. This shows that, at least in respect of certain graphemes, OK has developed accurate implicit knowledge of the French decoding system, distinct from his knowledge of the English system. Again, however, he fails to detect the incorrect realizations of various other graphemes in these words. One way of interpreting this is that his implicit knowledge of the French decoding system is incomplete. Alternatively, it may result from a conscious focus on particular graphemes perceived as 'problematic', so that other graphemes are ignored.

***Self-efficacy and feelings about decoding***

In common with many other participants, OK's SRI is characterized by tentative language and an explicit acknowledgment of uncertainty. Two of the target words prompt comments on the 'strangeness' of the French language, implicitly comparing it to the familiar sounds and spellings of English: *quignon* "sounds really strange"; *bêcheuse* is "quite different to every spelling" he knows.

Five individual graphemes are singled out as generating particular uncertainty, two containing diacritics. OK uses his own metalanguage to refer to these, commenting that *bêcheuse* has "got the E thingy". In *témoin*, his lack of knowledge of how to deal with the acute accent is used to explain his reliance on English decoding. His comment suggests that he knows the diacritic may affect the pronunciation, though he does not know in what way:

> I'm not sure about the E thing … cos it's got this special little mark above it so and I'm not so sure about how to pronounce that so … I just used those in English E

The other graphemes about which OK reports uncertainty are the <qu> in *quignon* (discussed above), the <au> in *jaugez* and the <ch> in *bêcheuse*. In this last case, he explicitly considers both the standard English and French realizations as 'candidate' pronunciations: "I wasn't sure if the C and the H would produce a [tʃɐ] or a [ʃɐ]".

OK's careful deliberations here and elsewhere – whereby he reflects on his own lack of knowledge, considers different possible pronunciations, monitors and evaluates his output – recall Erler's (2002) conclusion that more accurate decoders took longer to pronounce French words because they were trying consciously to override automatic, English-based decoding.

However, as we have seen, OK's questioning approach does not seen to extend to all graphemes, but only to some which he identifies as problematic. Indeed, there are four target words about which he is more confident: *obtient* "was pretty easy"; he thinks his pronunciation of *cinglant* is "good"; he is happy with his pronunciation of *témoin* except for the acute accent, which "might not be as good"; and he finds *quignon* easy to pronounce because "there's nothing like any little … marks [i.e. diacritics] in it and so I just read it how it looked". However, OK's pronunciations of all four of these words are entirely consistent with English GPC. Thus, his confidence in the accuracy of his pronunciations is misplaced. In these examples, he seems to resemble the other kind of learner in Erler's (2002) study, who simply read the words as English without being aware of the problems this entailed. The comments about diacritics in the examples quoted above perhaps offer an explanation for this apparent contradiction: it may be only when OK encounters orthographic features unfamiliar from English (be these diacritics or "strange" letter combinations) that he is prompted to think consciously about how those features should be pronounced. In other words, he may read French words as English until a problem is encountered which makes this difficult or impossible to do.

#### 6.4.1.3 Approaches to decoding at t2

***Strategies***

At t2, OK's SRI shows somewhat less evidence of strategic behaviour. Nonetheless, the same three types of strategies are identifiable as at t1.

***a. Dividing words into smaller units***

There are three word discussions in which OK says that, to decode the target word, he "split it up", again following what he probably perceives to be syllable boundaries: *dé/dain; lan/cée; poig/nard*. Although *dédain* and *lancée* are segmented in accordance with French grapheme boundaries, the <gn> in *poignard* is incorrectly subdivided into separate letters – exactly as with *quignon* at t1 – making a correct realization of the grapheme much less likely, if not impossible.

In the case of *dédain*, there is possible evidence for a further strategic narrowing of the focus onto the <ai> due to the perceived difficulty of this letter group:

> basically I break them into two parts and basically look at … the vowels in the words … because … some of them put together make a different sound

OK evaluates the strategy of breaking words up into smaller units positively: "it makes it easier cos … you can just work out one piece at a time". He here demonstrates the kind of metacognitive strategy evaluation suggested to be characteristic of effective learners (Macaro, 2006); however, his evaluation is based purely on the ease of completing the task rather than on the correctness of the outcomes. In fact, the three words which he mentions splitting up into smaller units are all pronounced in ways entirely consistent with English GPC.

***b. Drawing on known words as sources of evidence***

At t2, there is only one instance where OK refers to a known word to support his reasoning. This occurs in the discussion of *piochais*, where it supports a correct decoding outcome: he says he knows that <ch> is pronounced [ʃ] from the name of a French wine, although he cannot recall the name specifically.

***c. Trying out possible pronunciations***

There is one instance (when discussing *piochais*) which could be interpreted as OK consciously trying out and evaluating possible pronunciations of the target word, though it is unclear to what extent the behaviour here is consciously strategic. Again, OK appears to focus on the realization of a specific grapheme (in this case <o>), with the remainder of the word being pronounced identically.

> [pjɔːʃɛz] erm I'm not yeah [pjəʊʃɛz] think I'm not sure about that

***Knowledge bases***

In two word discussions, OK consciously contrasts his pronunciation of certain graphemes with the way he would pronounce them in English. For example, he comments that the <i> in *piochais* would be realized as the vowel /i/ in English but as the semivowel /j/ in French (which is correct); and that he would pronounce the second syllable as [tʃɛɪs] in English rather than as [ʃɛz] in French (correct apart from the SFC). Similarly, he pronounces the <ui> in *huilage* near-correctly as [wi], but says that in English he would pronounce it [ɔɪ]. These

comments suggest an awareness of French orthography as a system distinct from the English one, even though he recognizes that his knowledge of the French system is incomplete. This is reflected in this comment about the pronunciation of the <ui> in *huilage*:

> I'm not sure if that's the right way to say it in French but it definitely would sound different in … English I think it would sound like [hɔɪ] erm [hɔɪlɪdʒ]

As at t1, OK also makes some explicit comments (in three word discussions) about how individual graphemes should be realized in French. For example, discussing *piochais,* he correctly states that "the C and the H in French is erm [ʃɒ] which is a [ʃə] sound" (supported by his reference to a French wine, as described above). However, in other cases his 'knowledge' is incomplete or inaccurate. In *dédain*, he incorrectly believes that "the A and the I … make the [ɛɪ] sound", apparently influenced by English GPC; this statement also reflects an incorrect subdivision of the trigraph <ain>. He is also unsure what effect the acute accent has in *dédain*, although he knows that it "changes … the pronunciation of some words". Finally, discussing *ferreux,* OK thinks that the SFC <x> should be pronounced as the velar fricative [x]:

> it's the E U bit because [fɛɹiəːx] erm it makes the X turn into a different word erm different letter turning into G H it would sound like that **…** I think the X is the erm [fɛɹiəːx] sound at the end

Not only is this hypothesis about the pronunciation of <x> inconsistent with both L1 and L2 GPC, but in fact the sound [x] does not even exist in the phonemic inventories of either French or most varieties of English (including, apparently, his own). The source of the belief is unclear, although OK claims to have "learnt that myself". Hypothetically, his knowledge of the *sound* may derive from his brief study of Spanish or German at primary school, although neither language represents it in writing with <x>. The reference to <gh> as a

representation of /x/ is puzzling; perhaps OK has experience of another orthography which contains this GPC, such as Irish Gaelic (but this is speculation). This example therefore reveals the extent to which – left to their own devices – learners may develop unpredictable and idiosyncratic hypotheses about L2 decoding, recalling the participant in Woore (2010) who believed that the cedilla in *caleçon* indicated a rising intonation pattern.

Finally – as at t1 – there are three instances when OK reports drawing upon implicit knowledge about what 'sounds right'. This instinctive 'feeling' is sometimes correct – for example, he thinks that pronouncing the <ch> in piochais as [ʃ] as opposed to [tʃ] "just sounds more French" – but it is not always a reliable guide. For example, he says that his idiosyncratic pronunciation of *ferreux* (discussed above) "sounds right in my head", despite including a sound which does not even exist in French.

***Self-efficacy and feelings about decoding***

Generally, as at t1, OK's discussions are characterized by tentative language and an acknowledgment of uncertainty. An interesting case occurs in the discussion of *dédain*, which he says he is unsure how to pronounce because "I don't know what the word means … I've never heard of it before basically". This could be interpreted as suggesting that OK does not expect to be able to pronounce French words correctly unless he already knows them orally.

Unlike at t1, there are no explicit comments at t2 about French orthography or pronunciation being 'strange', although his unfamiliarity with the French system is implicitly highlighted when he says that the <x> in *ferreux* "doesn't really seem like it's part of the word it just

seems like it's been added on there". Several other graphemes are also highlighted as particularly problematic. As at t1, the "little thingy" (acute accent) in *dédain* causes uncertainty, as do the vowel digraphs <oi> in *poignard* and <ui> in *huilage*.

As at t1, there are also three instances where OK expresses confidence in his decoding of the French words. In one case, this confidence is reflected in a correct pronunciation: he feels "pretty sure" that the <ch> in *piochais* should be pronounced [ʃ]. In both other cases, however, OK's confidence again seems to reflect a misguided belief that the words can be read more or less as English: thus *poignard*, which he thinks he has correctly pronounced as [pɔɪgənɑːd], "looks more English than some of the words", whilst the first part of *lancée*, "you just pronounce in English [lan]". Again, then, OK's critical approach to L2 decoding apparently does not extend to all aspects of the orthography.

#### 6.4.1.4 Change between t1 and t2

At both t1 and t2, OK reveals himself to be a highly articulate learner (as might have been expected given his high verbal CAT score) who adopts a critical approach to decoding the target French words: he is aware that the L1 and L2 orthographic systems differ in some respects, is aware of specific graphemes which he considers problematic (especially diacritics and certain vowel digraphs), monitors his own output and consciously tries out alternative pronunciations in order to judge which sounds more 'French'. His discussions are characterized by a tentative language. To the extent that conscious reasoning may be a way of overcoming automatized, L1-based decoding processes, this is exactly the kind of learner who might be expected to do this successfully.

However, OK's critical approach does not seem to extend to all graphemes. At both t1 and t2, his pronunciations of some words are entirely consistent with English rather than French GPC. In other cases, although one or more graphemes may be pronounced in identifiably French (or at least non-English) ways, the remainder of the pronunciations still follow English decoding conventions. Ironically, it is precisely those words which are pronounced in the most 'English' manner whose pronunciations he feels most confident about; in some cases, he explicitly comments that the reason he found them easy to pronounce was that he felt he could simply read them as if they were English.

This uncritical relationship towards some aspects of his L2 decoding is also seen when OK explicitly draws upon a feeling for what 'sounds French' to evaluate his realizations of particular 'problematic' graphemes. Whilst his 'feeling' is sometimes successful in identifying the correct realization of the target grapheme(s) under evaluation, it appears blind to the English-based realizations of the remaining graphemes in the word. Overall, therefore, there is some evidence to suggest that OK is happy to pronounce words according to English GPC unless he encounters particular orthographic features which force him to do otherwise, such as diacritics or unfamiliar letter combinations which do not occur in English.

In terms of the strategies OK draws upon to support his L2 decoding, these are broadly similar at t1 and t2. First, at both time points, he reports breaking several words up into smaller units, which he says makes the decoding task more manageable. However, this strategy can also militate against a successful outcome, such as when the digraph <gn> is incorrectly split into two separate graphemes in *quignon* (t1) and *poignard* (t2). This shows that knowledge of the L2 grapheme inventory is a necessary foundation for the successful application of this strategy.

Second, OK refers to known words to support his reasoning about how to pronounce some target words. This occurs more frequently at t1, when the source words are predominantly English. French words are also referred to: *hors d'œuvre* (t1) and an unidentified name seen on a wine bottle (t2). However, his encounters with both these French source words may have been extracurricular and mediated by his L1, which has implications for the accuracy of their stored pronunciations. OK appears not to draw upon any French words learned in MFL lessons, a potentially valuable resource to support his L2 decoding.

It is interesting that OK's RAT scores show a relatively large improvement between t1 and t2 (Table 6.1). Since there seems to have been little change in his reported reasoning, this may be largely due to improvements in his implicit knowledge of the French decoding system.

### 6.4.2 Case study 2: EH (Intervention B)

#### 6.4.2.1 Background

According to her BIQ responses, EH (class B1) studied French (but no other languages) at primary school for about one hour per week for two years. She reportedly has no French-speaking relatives or friends and no routine extracurricular contact with French, although she has visited a French-speaking country 'between two and four times'. Since CAT scores and KS2 points were unavailable for school C, the only relevant attainment data for EH is that she obtained a Level 5 in her English SAT at the end of KS2. In the sample as a whole, around 41% of participants obtained a Level 5, 50% a Level 4 and 9% a Level 3.

#### 6.4.2.2 Approaches to decoding at t1

***Strategies***

At t1, at least two distinct strategies can be identified in EH's SRI: (1) breaking words up into smaller units; (2) referring to known 'source' words to support her reasoning. In some cases these two strategies appear to be closely linked.

Although it is not always commented on explicitly, some of the pronunciations which EH produces strongly suggest that the target words are being processed in smaller units. This is true of *jaugez* where there is a pause between the realization of the first syllable (attempted twice) and the second: "[ʒɔː] no [ʒɔːʒ] [ɛz]". It also clear in various pronunciations of *cinglant* (e.g. "[siŋ] erm oh [l] [lɪ] [ənt]", "[sɪk] erm [sɪŋkəʔɪʔʊnt]"), where pauses or glottal stops seem to suggest a division of the word either into syllables or into three orthographic units (*cing/l/ant*).

In three cases when discussing these two words, EH also explicitly mentions dividing them into smaller units. The first example occurs when, discussing *jaugez*, she reports "just thinking about erm the 'U' again I think and the 'G'"; this is linked to a specific belief about the decoding of the <u> (see below). The other two examples are closely linked to what appear to be strategic references to known source words. First, following her initial mention of the <ug> in *jaugez*, EH next focuses on the <ja>, apparently influenced by a known phrase which she thinks starts with the same two letters:

the J and the A at the beginning erm in other words it's … kind of pronounced like [ʒɛɪ] instead of like [dʒaʊ] or something … because like in [ʒəmapɛl]* it's spelt like 'J' I think it's spelt 'J A' no I can't remember erm it I think that's how it's spelt

* *Je m'appelle*, /ʒəmapɛl/, 'I am called'

This apparent attempt to use the analogy strategy draws on a familiar French phrase – almost certainly covered in MFL lessons at primary and/or secondary school – and leads to the correct realization of <j> as [ʒ] rather than as English [dʒ]. However, the strategy is also undermined by problems. First, there seems to be confusion over the target letter string: whilst EH states it to be <ja>, which seems to involve the incorrect segmentation of the grapheme <au>, her suggested alternative pronunciation [dʒaʊ] appears to be an English realization of the first *three* letters (<jau>), in which the vowel digraph is intact. Further, there is a mismatch between the vowel [ə] in the source phrase [ʒəmapɛl] and the vowel in the pronunciation [ʒɛɪ] which she appears to derive from it by analogizing. This suggests that it may in fact only be the initial consonant which is the focus of her attention. (There is also a possible influence of alphabetic letter names here: see below). Second, whilst the phonological form of the source phrase is recalled correctly, its orthographic form is not (although this misspelling need not affect the use of the analogy strategy to determine the realization of the initial <j>).

A similar picture emerges in the discussion of *cinglant*. Here, EH explicitly comments that she divides the word into two parts (*cing/lant*), focussing initially on the first syllable. The point of segmentation is again influenced by her reference to a known source word, *cinq* ([sæ̃k], 'five').

**EH*** I might actually pronounce the G differently cos I'm kind of splitting in two now I'm looking at one
**RW** where are you splitting it
**EH** erm like down the middle … after the G cos like [saŋk] is the G's kind of more pronounced I'm not very good at spelling so I think it ends in a G [saŋk] and like

> it's produced more as a the English word ((i.e. letter)) Q actually I think it might be a Q so like might pronounce it differently to like [s] like [siŋ]
>
> * EH confuses the terms 'word' and 'letter' throughout her SRIs; explanatory notes in ((double brackets)) indicate where this appears to occur.

It is interesting here that EH uses as a source word one of the basic numbers: as argued in Woore (2006), these may be problematic due to irregularities in their GPC, such as the presence of several non-silent final consonants. Nonetheless, in this case the reference to *cinq* is partially successful, leading to an 'acceptable' (albeit anglicized) realization of the nasal vowel. On the other hand, it also seems to lead to the <g> being pronounced as the devoiced /k/, apparently through a mistaken belief that the spelling of the source word is <cing> rather than <cinq>, i.e. an exact match of the first syllable of the target word. Consequently, EH seems to transfer the phonological form of the whole source word into her realization of the target word. However, at this point, she begins to doubt whether she has correctly recalled the spelling of *cinq*, deciding that it in fact ends in <q> rather than <g>. Rather than simply replacing her incorrect realization of the final consonant /k/ with a voiced /g/, she also rejects her 'acceptable' realization of the nasal ([an]), reverting to the English-based realization of <ing> as /iŋ/. It therefore seems that, whilst the analogy strategy operates successfully when EH perceives there to be a direct orthographic match between the target letter group and the source word as a whole, she is not able to use it to derive the pronunciation of the nasal vowel <in> in isolation.

***Knowledge bases***

**a. Beliefs about specific GPC**

There are several instances where EH comments explicitly on the pronunciation of specific graphemes or combinations of graphemes. Diacritics prompt comments on both occasions they are encountered (in *témoin* and *bêcheuse*). However, although she notices the acute accent in *témoin* and knows that it has some significance for the pronunciation of the word, she does not know what this significance is; nor is she able (or chooses not) to refer to the diacritic using standard metalanguage, describing it instead in vague terms:

> well we had the E with the something over the top I think that that's pronounced differently than it would like be in English so it's got something different

Similar comments are generated by the circumflex in *bêcheuse*: again, EH has a vague sense that it may make some kind of unspecified 'difference' to the pronunciation or that it may indicate the omission of <ê> altogether.

> **EH** that E might be pronounced differently cos it's got … the erm can't remember what it's called … on top …
> **RW** OK and what does it do …
> **EH** erm I think sometimes it means to pronounce it differently and sometimes it means to miss it out …
> **RW** how do you know that
> **EH** think I remember looking at like a book and it had the all the things at the beginning saying what they meant

Her claim that this incorrect belief about the L2 decoding system derives from 'looking at a book' may relate to the systematic lists of GPC often provided at the start of 'teach yourself' language books (e.g. Carpenter 2003; Fournier 2003; Moys 1996; Knight 1994; Overy &

Lecanuet 2003). However, whatever the source of this information, it has clearly not led to secure knowledge capable of supporting accurate French decoding.

EH also comments explicitly on the decoding of other graphemes or combinations of graphemes besides those containing diacritics. First (as described above), she initially focuses on the <u> in *jaugez*, stating that "sometimes U is missed out". This paves the way for the use of the analogy strategy to decode the target letter string <ja>, which clearly reduces the likelihood of a correct realization of the digraph <au>. A second explicit comment concerns the <se> in *bêcheuse*. She seems to conceptualize this as a digraph, in which the <e> is "part of the S"; this is reflected in correct realizations both of the <s> (as [z]) and of the 'silent' final <e>.

> cos you might not have to pronounce that E cos so you might have to go like [bɪʒʉːz] [ʒʉz] and then not like an [ɛ] at the end

Third, following her division of *cinglant* into parts (see above), EH is asked about her pronunciation of the second syllable <lant>. She pronounces the <a> as schwa based on a belief that [ənt] is "kind of what an ending would sound like" in French, whereas in English it would be pronounced [ant]. The origin of this belief is unclear, although it appears to reflect a personal hypothesis about the nature of French phonology. However, it is clearly antithetical to accurate decoding: ironically, it is in English that unstressed vowels tend to be reduced to schwa. Further, EH shows no awareness that the <a> in fact forms part of the nasal digraph <an>.

**b. Beliefs about general principles of French decoding**

A general principle of French decoding on which EH comments is her belief that a given grapheme should be "pronounced the same" in different orthographic contexts. This arises in relation to what she perceives as a threefold occurrence of <e> in *bêcheuse*. French orthography does indeed show less contextual variation in its GPC than English (section

2.4.3), but in this instance EH's comment fails to take account of the fact that the first <ê> bears a diacritic, the second is part of the digraph <eu> and the third is 'silent' due to its position at the end of the word. (In fact, EH later revises her view when, "looking at it again", she notices the circumflex. As described above, she also subsequently comments on the final <e>, correctly noting that it is not realized phonologically).

**c. Knowledge of the French alphabet**

In two word discussions, EH explicitly appeals to her knowledge of the letter names in the French alphabet. The first instance occurs when discussing *témoin*:

> **EH** I'm trying to think how you pronounce like the alphabet in French and then kind of translating the words and then putting it together
> **RW** right what do you mean by the alphabet then so can you give me an example …
> **EH** yeah like erm say erm A you'd say it like more kind of like [ɛ] or erm like I think it's W something like [igɹɛk]
> **RW** OK … but how does that apply to this word because there's no letter … W
> **EH** I remembered that M was kind of the same erm and like erm 'I' was more sounding like an 'E' I think

Informal observations within the university's ITE partnership suggest that learning the French alphabet is often an early requirement for beginner MFL learners. This may be linked to communicative aims, such as being able to spell one's name when booking a hotel room by telephone. However, EH's knowledge of the alphabet is inaccurate: for example, she confuses the names for 'W' ([dubləve]) and 'Y' ([igʁɛk]). More importantly, however, she clearly has a misconception that knowledge of the L2 alphabet is directly relevant to decoding: her reference to "thinking how you pronounce like the alphabet and then kind of … putting it together" suggests that she is using the names of letters, rather than their phonemic values, to 'sound out' the target word. This misconception coincidentally leads to a correct outcome in the case of the <m>, whose letter names in French and English are "kind of the same", as are its phonemic values. In many cases, however, it seems likely to lead to confusion and error. Further, the use of alphabetic letter names for decoding purposes

presupposes the treatment of letters as individual graphemes rather than potentially as parts of di- or trigraphs. This occurs in the case of the <i> in the current example: EH's comment that "'I' was sounding more like an 'E'" – probably deriving from the fact that the letter name for English 'E' is similar to that of the French 'I' (/i/) – does not take account of the fact it is part of the nasal grapheme <oin>. Exactly the same problem of segmentation occurs subsequently when EH appeals to the French alphabetic name for the letter G to help her decode *quignon*, when in fact the <g> is part of the digraph <gn>.

**d. 'Feeling for French'**

EH makes at least one explicit appeal to an instinctive 'feeling' for what 'sounds French'. One possible example occurs in the discussion of *cinglant* quoted above, where she feels that the second syllable should be pronounced [lənt] because it "just kind of fits … I think it's kind of what an ending would sound like". A clearer example occurs in the discussion of *quignon*, where she decides that the first <n> should not be sounded because it "sounds better without the N in". When asked what she means by 'sounds better', she replies:

> erm cos like if you went [kwɪʒɒn] it sounds better than going like [kwɪʒnəʔɒn] … [kwɪʒʔɒn] sounds more kind of fluent … cos you kind of have to stop to put the N in

EH's instinctive 'feeling' here is explicitly presented as something she has recourse to in the absence of more secure knowledge ("I'm not quite sure how this word would actually sound in French so…"). In fact, better acquaintance with the French grapheme inventory might have avoided her misguided focus on <n> in the first place, allowing her to treat it correctly as part of the digraph <gn>.

### *Self-efficacy and feelings about decoding*

EH makes several explicit comments about her lack of secure knowledge of French decoding or about her lack of confidence in her pronunciations. One example (quoted above) occurs in her discussion of *quignon*, when she remarks that she is "not quite sure how this word would actually sound in French". When discussing *bêcheuse*, she explains that she feels puzzled "probably cos it's just quite a long word". However, all RAT words are between six and eight letters long. This perception may therefore reflect the presence of other orthographic features unfamiliar from English, such as the high number of vowels or the diacritic. Finally, the clearest expression of EH's negative self-efficacy in terms of French decoding occurs in her discussion of *témoin*. Here, she contrasts her own pronunciations with those of a more competent French speaker, which she perceives would be more accurate and more fluent than her own letter-by-letter decoding.

> yeah like when I usually see words in French I they just don't really sound right as like erm someone else who's like really good at French would say them … like say you listen to like a French tape or something and someone's saying like that word… they'd like say it really fast and like they get all the pronunciations right whereas when I do it I kind of once I say it I kind of go along with the each word ((i.e. letter)) so I kind of say it a bit more kind of slowly

This comment echoes those made by some of Erler's (2002) participants, who complained of a mismatch between their own decoding of familiar L2 words and the pronunciations they heard in class.

### 6.4.2.3 Approaches to decoding at t2

***Strategies***

**a. Dividing words into smaller units**

At t2, there are several instances where EH reports focussing on a particular letter-string within the target word. First, she explains that she divides *dédain* into two sections (which appear to be coterminous with its perceived syllables) in order to "work on each section at a time". EH may also focus on smaller sub-lexical units, perhaps those she considers particularly difficult: for example, when discussing *lancée*, she initially reports that she "was thinking again of the E with the erm accent on". Third, she focuses on the <ux> in *ferreux*, incorrectly subdividing the digraph <eu>; this is reflected in her trisyllabic pronunciation [fɛɹɛʔuks], where the <e> and <u> appear to have been decoded as two separate vowels.

**b. Referring to known words**

In the final example above, EH's division of the target word into smaller units is linked to an attempted analogy to a familiar word with similar spelling. Again, the source word is a basic number, *deux* ([dø], 'two'), recalling her similar use of *cinq* ([sæ̃k], 'five') at t1.

> I think the U and the X at the end was like a number I think like but then I'm not sure if that's right now I'm thinking of that (6.6) like can't remember which number it is but it's like spelt D E U X or something

The source word here is appropriate and its spelling accurately recalled, but EH makes an analogy only to the last two letters <ux> rather than to the whole rime <eux> or to the

digraph <eu>, reflecting her incorrect segmentation of the word. Incidentally, it is noticeable how insecure EH's knowledge is even of an elementary word such as *deux*, which is almost certain to have been covered early in the Y7 MFL curriculum as well as at primary school and to have been encountered in spoken or written form many times since. This perhaps puts into perspective the aim, inherent in the programmes of decoding instruction, of helping participants develop secure knowledge of a large body of French GPC.

There are two other occasions when EH draws upon known words to support her decoding. In both cases, the sources are again French words likely to have been covered as part of the MFL curriculum. First, she makes an analogy to the word *âge* (/ɑʒ/, 'age') when discussing *huilage*: "I remember that … A G E I remember saying [ɛɪdʒ] as like [ɑʒ]". Here, she either ignores or is unaware that the source word, unlike the target word, contains a circumflex; nonetheless, her pronunciation of the target word [hɥliɑʒ] contains an 'acceptable' realization of <age> (except for the extraneous /i/, perhaps resulting from a misperception of the position of <i> in the letter string). In the second example, EH draws an analogy to the source word *là* (/la/, 'there') to support her decoding of <la> in *lancée*:

> that's like [la] cos I think cos like think it's 'there' or something is spelt L A and that's how you say it but I'm not sure

EH is unsure in her knowledge of the source word *là*; this may explain the lack of reference to the grave accent, not present in the target word (although in this case the difference in spelling would not prevent the analogy strategy operating effectively). However, as in previous examples, the segmentation of the target letter string is problematic. The nasal digraph <an> is incorrectly subdivided, making it impossible for the analogy strategy to generate a correct outcome.

**c. Trying out possible pronunciations**

In two word discussions, EH explicitly considers two or more options for a given grapheme or combination of graphemes. In the case of *lancée*, she wonders whether the <ée> might be pronounced "[ɪ] … or [ɛ]". However, the clearest example occurs in her discussion of how to pronounce <pio> in *piochais*:

> and then just thinking P whether you say it like [pɪ] or like [pə] or [pʊ] just erm pronounce it with the 'I' whether it changes … whether it's … is it [pɪ] or is it whether it's like still I dunno [pa] or whatever I'm not sure

The large number of possible options she considers here, some with no obvious connection to the written form, suggests a sense of uncertainty: in the absence of any secure knowledge of the relevant GPC, it is as if she is drawing vowel sounds at random from her L1 phoneme inventory in order to see if they happen to 'sound right'.

There is some evidence that EH also tries out possible pronunciations using her 'inner voice'. For example, when she first sees *piochais*, there is a pause during which, she later says, she is "just like going through it in my head". Similarly, asked why she stops just before attempting to pronounce *poignard*, she explains, "I was about to say something and then I was like no that's not it". EH therefore appears to adopt a questioning approach to French decoding, frequently monitoring and evaluating her own pronunciations, not only those she says aloud but also pre-vocalization.

***Knowledge bases***

**a. Beliefs about the decoding of specific graphemes**

As at t1, EH makes several explicit comments about the way individual graphemes or groups of graphemes are pronounced. This might be expected to occur more frequently or with greater accuracy at t2, since one of the aims of the decoding instruction was to develop participants' knowledge of French GPC. Correspondingly, there is at least one word discussion in which EH comments correctly on the realization of a specific grapheme and feels confident in this knowledge: in *piochais*, she says that "the 'C' and the 'H' I think it's like spelt like [ʃ] like so [piɒʃ] [az] so I think that's right". However, in all other cases (occurring in three word discussions), EH's knowledge of French decoding still appears insecure. The first example occurs when discussing *dédain*:

> **EH** so like I would also think whether the word ((i.e. letter)) changes whether it's at the beginning or the end or has like a can't remember what they're called
> **RW** this little thing on the E
> **EH** yeah

Her knowledge of the diacritic in <é> is also vague, despite having covered this explicitly in the programme of decoding instruction: "I can't remember actually I know it does something". Asked specifically whether she remembers covering <é> in class, she says she cannot. It is of course possible that she was absent during the lesson in which this GPC was taught. However, explicit instruction appears to have been equally unsuccessful in the case of <ain> in the same target word, about which she is also asked. She does recall learning about this grapheme (one of the most recent ones covered by her class), but again cannot recall its pronunciation.

The second word discussion in which EH comments explicitly on the decoding of specific graphemes concerns *lancée*, where the focus is again the <é>. As at t1, therefore, diacritics generate explicit comments on both occasions that they appear in the target word list. Once again, she is aware that the diacritic may "change just saying the E normally" but does not know what its specific effect might be:

> **RW** what's the difference there between normally and with an accent
> **EH** I wish I'd know that

Similar sketchiness is apparent in EH's knowledge of the SFC 'rule', again despite the fact that (as she mentions) she was "doing it in class" as part of the decoding instruction. She dimly recalls the principle that some letters in word-final position are not realized phonologically, but is unsure which letters this applies to and seems to show confusion with the fact that word-initial <h> is also not pronounced. However, on this occasion her final pronunciation [pəʊʔɪgnɑː] does correctly omit the final [d], which had appeared in her initial attempt at pronouncing the word. Therefore, conscious processing appears to have positively affected her decoding here.

**b. Knowledge of the French alphabet**

As at t1, EH appears to support her decoding by appealing to her knowledge of the French alphabet. This can be inferred to occur in her discussion of *dédain*:

> I'm not really sure how to pronounce each letter … I can remember like some that we did in class like recently … like erm a 'B' is like [bɛɪ] or something like that 'A' was like [ɛː] erm quite a lot of the words like together like 'E' is like like sounds more like an 'I' in English erm (4.2) oh I have such poor memory

This comment could be interpreted as referring to the decoding instruction; however, her reference to <b> as [bɛɪ] (an anglicized version of the correct French name for 'B') suggests that she is thinking at least partly of alphabetic letter names. Further, part of her comment sounds like an attempt to express the similarity between the English alphabetic name for 'E' and the French name for 'I' (/i/) (see above).

The same issue of the letters E and I sounding similar occurs in the discussion of *poignard*:

> **EH** the 'I' I was thinking … it was like an 'E' or something but then I was thinking no because you pronounce an 'E' like an 'I' so kind of thing so I was wondering whether you keep the 'I' the same or do you say it differently
> **RW** so what was your conclusion what did you come up with
> **EH** ((laughs)) I wasn't sure actually

Again, this is probably (though not certainly) a reference to alphabetic letter names rather than to phonemic values. Whether or not this is so, this example again reveals EH to be a reflective decoder, attempting to use her conscious knowledge as a resource to evaluate and revise her pronunciation of the target word. Unfortunately, however, the knowledge she draws upon is unclear and may (if it is knowledge of the alphabet) be inappropriate for the task; further, the focus on the single letter <i> fails to treat it correctly as part of the digraph <oi>. The end result is simply a state of confusion.

### *Self-efficacy and feelings about decoding*

As revealed in many previous examples, EH's SRI at t2 is pervaded by a sense of uncertainty; she frequently states that she has "poor memory" and is unable to recall material covered in lessons. Phrases such as "I can't really remember" and "I'm not sure" occur in

almost all word discussions; most also contain at least one pause several seconds in length. The extract from the discussion of *poignard*, quoted above, seems to epitomize her uncertainty when attempting to decode unknown French words:

> and then just thinking P whether you say it like [pɪ] or like [pə] or [pʊ] just erm pronounce it with the whether it changes ... erm whether it's s is it [pɪ] or is it whether it's like still I dunno [pa] or whatever I'm not sure

#### 6.4.2.4 Change between t1 and t2

At both t1 and t2, EH talks articulately about L2 decoding. Like OK, she demonstrates a consistently critical stance towards the task, monitoring and evaluating her pronunciations and questioning the reliability of the knowledge she draws upon. However, despite the fact that she has followed a programme of instruction in French decoding, her reasoning does not appear to show any significant development between the two time points. This is in line with her RAT scores, which also show little increase between t1 and t2.

In terms of the strategies she uses, these are broadly similar at t1 and t2. There is clear evidence that she strategically divides words into smaller units to make the decoding task more manageable: sometimes into syllables and sometimes into individual graphemes or grapheme groups which receive detailed attention. However, as with OK, a persistent problem at both time points is that the segmentation of the words often violates grapheme boundaries, with the constituent letters of both vowel and consonantal digraphs (e.g. <gn>, <oi> and <eu>) being treated as separate graphemes. This is then reflected in the fact that some of EH's attempted pronunciations contain too many syllables due to the vowels being realized separately (e.g. [pœʔɪgnɑː], [fɛɪɛʔuks]). Although one of the aims of the decoding

instruction was to develop participants' knowledge of L2 graphemes, there is no evidence that EH is any more likely to segment letter strings more accurately at t2 than at t1.

EH also deploys a form of the analogy strategy at both time points, using the pronunciation of known words to support the decoding of similarly spelt (parts of) target words. Encouragingly, the source words are not English, as in the case of many participants, but French in all cases and almost certainly learned as part of her MFL studies. Further, EH shows some success in isolating the sound corresponding to the target grapheme(s) and transferring this into the target word. However, her use of the analogy strategy is undermined by (a) her uncertain and/or inaccurate knowledge of the spelling of the source words; and (b) the fact that the strategy is often applied to incorrect letter strings, resulting from the faulty segmentation of the target words. There is no indication of improvement in these respects between t1 and t2.

In terms of the knowledge bases she draws upon, EH comments explicitly on the decoding of individual graphemes or combinations of graphemes at both t1 and t2. However, her knowledge is often incomplete and inaccurate. For example, at both time points she focuses on diacritics whenever she encounters them, but her knowledge does not extend beyond a vague notion that they affect the pronunciation in some unspecified way. There does not seem to be any development in EH's conscious knowledge of French GPC between t1 and t2, even though this was one of the explicit aims of the decoding instruction.

At both time points, EH also appeals (or appears to appeal) to her knowledge of the French alphabet. This seems to reveal a misconception – which the decoding instruction apparently fails to dislodge – that alphabetic letter names are directly relevant to the task of decoding,

when in fact they are distinct from the letters' phonemic values (though clearly they do overlap to some extent). Further, the focus on individual letters is counterproductive, since it may once again make them less likely to be recognized as the constituent elements of multi-letter graphemes.

Finally, EH does not appear to feel any more secure in her L2 decoding in her second SRI than in her first.

### 6.4.3 Case study 3: LV (Intervention A)

#### 6.4.3.1 Background

LV (class A2) reported in her BIQ that she studied French at primary school for about thirty minutes per week for three years; she also learnt a little Spanish there. Her family appear to be supportive of her French learning: she visits France 'once per year or more' and, between the ages of four and seven, used a commercial, multimedia-based French course for young children (www.early-advantage.co.uk, accessed 18/05/2011). However, she reportedly has no French-speaking friends or relatives. Her verbal quotient in the CAT tests was 114, above the sample mean of 105.5 (range: 78-141; N=142 valid cases). Her average KS2 point score of 33 was similarly high (sample mean: 29.3; range: 19-35; N=129 valid cases).

### 6.4.3.2 Approaches to decoding at t1

***Strategies***

**a. Trying out possible pronunciations**

All but one of LV's word discussions at t1 are permeated by the trying out of different possible pronunciations of the target word. In some cases, multiple realizations are produced, some of which are also repeated several times either consecutively or at different points in the discussion. This is exemplified in the following extract from her discussion of *quignon*:

> [kj] [kjʉnɒn] [kjʉnəʊ] [kwɪgnɒn] [kwʉ] [kwʉ] erm I'm not sure what whether or not to pronounce the 'U' … [kwinɒn] don't know oh erm looking at it as an English word it would look like [kwɪgnɒn] and erm I'm not sure how they'd say it in French I'm not sure whether they'd pronounce the 'U' or not erm but if they didn't then it'd be like … I still don't think I've pronounce[d] the 'N' very well … [kjʉnɒ] [kjʉgnɒn] [kwɪ] [kwignɒn] [kwɪgnɒn] it's the 'U' and then the 'I' I'm not sure about

This systematic vocalization of alternative pronunciations suggests that LV is deliberately trying out alternative pronunciations in order to facilitate the process of evaluating them. Indeed, when discussing *jaugez*, she comments on this explicitly:

> [I] kind of always think of loads of options and then try and sort of … say right doesn't look like that doesn't look like that … if I'm really stuck on them I'll say it out loud but normally I'd do it in my head sort of thing

The discussion of *quignon* quoted above also exemplifies the fact that LV almost always pronounces part of the word in a constant way, providing a sort of phonological framework

within which alternative realizations of one or more graphemes can be tried out and evaluated. For example, in her first two pronunciations of *quignon*, the first part [kjʉ] is kept the same whilst two different realizations of the ending are attempted ([nɒn] and [nəʊ]); in the third pronunciation of the whole word, LV reverts to the initial ending [nɒn], but this time tries out a different form for the start of the word ([kwɪg]); the fourth pronunciation ([kwɪnɒn]) then maintains all aspects of the previous one ([kwɪgnɒn]) except that the [g] is removed. Various permutations of these options are then further tried out in the remainder of the discussion.

LV's trying out of different pronunciations seems to respond to specific aspects of the decoding about which she is unsure. For example, her comment that she is unsure "whether they'd pronounce the U or not" seems to be reflected in the fact that <qui> is a principal locus of variation in the extract quoted above. By contrast, the realization of <non> as [nɒn] is constant across most of the pronunciations; this is the part of the word which she says "just looks slightly easier for some reason". (On the other hand, she does also say that she may not have "pronounce[d] the N very well"; it is unclear which <n> this refers to, although both are subject to some more limited variation). However, LV's confidence in her pronunciation of <non> is misplaced, since it is consistent with English rather than French GPC. Indeed, this reflects a more general pattern whereby those word-parts which LV realizes in a fairly stable way – suggesting that she is more confident of how to pronounce them – tend to be read as if they were English. This suggests that, whilst aware that some graphemes are realized differently in French and English, she underestimates the extent of the differences between the two orthographies.

LV's repeated attempts at pronouncing target words also highlight her lack of any secure basis for evaluating the pronunciations she produces. No amount of repetition will provide a definitive answer. Indeed, in some cases, none of the options under evaluation is correct, as in the case of [kwɪ] and [kjʉ] as possible realizations of <qui>. Further, in the case of two words (*bêcheuse*, *haubert*), LV's deliberations come full circle: the pronunciation she finally settles on is identical to her very first attempt. In this sense, her strategy of trying out possible pronunciations is counterproductive, absorbing time and concentration which could be better spent on other forms of strategic reasoning.

**b. Referring to known source words**

There are at least four instances at t1 when LV refers (or attempts to refer) to known words to support her reasoning. In her discussion of *bêcheuse*, she provides an explicit description of the analogy strategy, which she says she would attempt to use in order to help her choose between three possible options for the realization of <eu>. This begins with a conscious mental search for source words containing the target grapheme:

> **LV** I try and just think of as many French words as I can and then if one of them which it probably won't just springs to mind and it helps then I kind of if there's an example of it sounding like a U then I'll just kind of go with U and if
> **RW** OK can you just clarify what you mean by you're looking for other French words which might help
> **LV** well I'll just I'll just run through erm all the French words I know how to spell like in all the tests I've done or anything done and I'll try and I'll try and remember how they pronounce it and so if there's an E U that sounds more like an E then I'll probably go with that

However, in this case, the strategy fails because she cannot identify an appropriate source word. This problem also occurs when trying to pronounce <ant> in *cinglant* and <e> in *jaugez*.

A different problem arises in another attempt to use the analogy strategy when discussing *bêcheuse*. This time, attempting to decide whether <ch> should be pronounced [tʃ] or [ʃ], LV refers to the known word *dimanche* (/dimɔ̃ʃ/, 'Sunday'):

**LV** [bitʃʉz] [biʃʉz] I'm not sure whether to say S H or C H cos like with things like [dɪmɔ̃ʒ] or erm they say sort of G so [bidʒʉz]
**RW** when do they say like a G then
**LV** erm [dɪmɔ̃ʒ] it's I think it's spelt C H E but I'm not sure

Here, she expresses uncertainty as to whether the proposed source word really does contain the target letter string. In fact, she has recalled the spelling correctly but gets the phonological form wrong, making the final fricative voiced [ʒ] rather than voiceless [ʃ]. This then leads her to transfer the sound [ʒ] – or an affricated version of it – into the target word (although she does later revert to the correct sound [ʃ]). In this example, then, the mechanics of the strategy operate smoothly but it is undermined by a defective phonological representation of the source word.

LV also refers to this same source word to support her reasoning about the <oin> in *témoin*, when she is asked to explain why she has produced a series of pronunciations with no final /n/ ("[timɔɪ] [timɔː] [timɔː] or [timɔɪ]"). Although she appears to pronounce the source word *dimanche* with a nasal vowel – albeit an incorrect one, [ɔ̃] – she seems to conceptualize this sound in terms of 'not pronouncing' the <n>. This is then reflected in her pronunciations of the target word, which end in oral vowels. Here, then, the problem seems to lie not in an incorrect phonological representation of the source word, but in LV's conscious analysis of the sounds it contains. This may be interpreted as a problem of phonological awareness in

the L2, in contrast to Durğunoglu, Nagy and Hancin-Bhatt's (1993) conclusion that this skill readily transfers from the L1.

**c. Dividing words into smaller units**

Finally in terms of LV's strategic behaviour, there are a few instances where she divides words into smaller units to support her decoding. In *témoin,* which she divides into two syllables, she comments explicitly that she "sort of break[s] it up … [tiː] and [mɔɪ] sort of thing". In other cases, she focuses on specific graphemes or combinations of graphemes which pose particular difficulties. For example, in *quignon* she comments, "it's the 'U' and then the 'I' I'm not sure about"; this suggests that she may have segmented the letter string incorrectly, since this <u> is in fact part of the digraph <qu>. In *bêcheuse,* she focuses on the <ch>, feeling unsure "whether to say S H or C H". It is interesting here that the English spellings <sh> and <ch> are used as a way of expressing the two sounds /ʃ/ and /tʃ/ respectively, again evoking the idea of English orthography as the 'norm' to which the French system is compared.

***Knowledge bases***

**a. Beliefs about the decoding of specific graphemes**

LV makes several explicit comments about the pronunciation of individual graphemes or groups of graphemes. In some cases these appear to be retrospective reflections on the pronunciations she has produced. On the other hand, she does draw upon explicit knowledge of the effects of the diacritic when attempting to pronounce *témoin*:

**LV** I know sometimes they don't pronounce things the way we do
**RW** right what do you mean by that exactly
**LV** well they have like … things like on some of the Es and that makes it … sound different … I don't know what sound it makes but I know it's different

LV's knowledge of the diacritic again fails to extend beyond a vague sense that it makes the <é> sound somehow 'different'. The other target word in which a diacritic appears is *bêcheuse*; in this case, she notices that "it's got that little thing again", but makes no comments about its possible effects.

**b. Knowledge of the French alphabet**

On one occasion, LV refers to the French alphabet to support her reasoning. The following extract (from the discussion of *jaugez*) exemplifies the way in which the use of alphabetic letter names can impede decoding:

**LV** don't they say J like G and G like J or is that just me I'm sure they did [dʒi] erm [dʒidʒɛɪz] erm [dʒigɛɪz] [giz] [dʒɛɪgiz] [dʒigiz] ((laughs))
**RW** OK that was very interesting what you said about the J and the G can you expand on that a bit more …
**LV** well … when we were learning about the alphabet I'm sure that's what she said that she got mixed up about the J and the G so I just sort of remembered that …
**RW** … by learning the alphabet what do you mean
**LV** well we were saying erm the letters and we'd say how it sounded and stuff [ɛɪ] [bɛɪ] [dɛɪ] sort of thing and we've written them down like [dɛɪ] under D and
**RW** OK … and so that's changed that ((the <j>)) but what's it done to this ((the <g>)) …
**LV** well it's kind of made that [dʒi] and the other one [dʒɛɪ] sort of thing … [dʒidʒɛɪz] [dʒidʒɛz] [dʒidʒiz] ((laughs))

In this example, LV suddenly recalls a 'rule of thumb' relating to the pronunciation of the letter names for J and G. There is no doubt that it is the names of the letters rather than their

phonemic values to which she is referring, because she mentions the alphabet explicitly and recites the first few letters as examples – albeit mispronouncing the first letter A as English [ɛɪ] rather than French [a]. The letter names for J and G are then used to inform the decoding of <j> and <g> in the target word. However, two problems occur. First, the letter names are incorrectly pronounced with an initial [dʒ], as in English, instead of the French [ʒ]. Second, rather than simply using these (incorrect) initial phonemes in processing the target word, the letter names are incorporated intact, including their respective vowel sounds [i] and [ɛɪ], eclipsing any other pronunciations which might have been considered for the vowel graphemes <au> and <ez>. Coincidentally, [ɛɪ] is an 'acceptable' realization of <ez> according to the RAT mark scheme; however, the realization of <au> as [i] would have appeared anomalous without this detailed insight into the mechanisms lying behind it.

It is also interesting that the initial result of LV's appeal to alphabetic letter names – the pronunciation [dʒidʒɛɪz] – is subjected to her usual process of evaluation and reformulation: clearly, she is unhappy with it, immediately trying out alternatives. However, she ascribes this to her uncertainty about the letter names in French; had this knowledge been more secure, she says, she would have been less hesitant about the resulting pronunciation.

**c. 'Feeling for French'**

It was argued above that, when trying out possible pronunciations, LV seems to evaluate them based on a 'feeling' for what 'sounds right'. This can perhaps be seen in action when she explicitly rules out [jʉ] as a possible realization of <eu> in *bêcheuse* on the basis that "I think that's kind of urgh" – "urgh" being a hyperbolic expression of revulsion.

On one occasion, when discussing *cinglant*, LV also refers explicitly to her instinctive sense of what 'sounds French'. Little or no conscious thinking seems to have been involved in pronouncing the word; indeed she cannot explain why she pronounced it the way she did. Interestingly, this is also the only word for which she does not repeatedly try out alternative pronunciations.

> [sɪŋg] [sɪŋglɒn] that's kind of what happens if I just sort of said it I'd say [sɪŋglɒn] cos I know that that's kind of has a slight French sound to it erm I don't know why I pronounced the A as an O though that just [glɒn] just kind of sounded French

Here, LV's instinctive feeling that her pronunciation of <glant> "just sounded French" is correct: the final <t> is correctly 'silent' and the nasal /ɒ̃/ is realized 'acceptably' as [ɒn]. Both these pronunciations contrast with the way the graphemes would be pronounced in English. This suggests that at least some aspects of the French decoding system have become part of LV's implicit knowledge. By contrast, she shows no awareness that her realization of the first syllable is incorrect, with the <in> being pronounced according to English GPC rather than as the French nasal /æ̃/. Her implicit knowledge is therefore only partial.

### *Self-efficacy and feelings about decoding*

As noted above, LV's multiple and repeated pronunciations of almost all target words seems to embody a sense of uncertainty. This is also reflected in an explicit comment made at the very start of her SRI, when trying to pronounce *témoin*:

> I would say like I've no idea … I kind of I don't know cos I'd never seen any of the words before

This comment suggests that LV does not *expect* to be able to pronounce unfamiliar L2 words.

#### 6.4.3.3 Approaches to decoding at t2

***Strategies***

**a. Trying out possible pronunciations**

At t2, LV appears to use the strategy of trying out possible pronunciations in a similar way to at t1, to help decide between alternative realizations of specific graphemes. However, she does so much less frequently, in only two of five word discussions.

In *ferreux*, LV tries out two alternative realizations of <eu>: [ʉ] and [ɛɪ] (the realization of <r> also appears to change along with the vowel sound, apparently without her awareness). By contrast, the initial [fɛɪ] or [fɛʁ] and the SFC form a constant phonological 'framework' within which the pronunciation of <eu> is systematically varied. Her lesser degree of uncertainty about these two components of the framework is reflected in her comments that (a) she 'would not change' the pronunciation of the <ferr>, probably meaning that it is the same in French as in English; and (b) she has not "heard any … words in French that end in X". However, as observed at t1, LV lacks a secure basis for deciding between the two realizations of <eu>; indeed, both are incorrect. Her conscious reasoning (thus far) therefore seems to succeed only in problematizing her decoding, without offering any solutions: "if I kind of without too much thought I would have said it like … [fɛʁʉ] but with more thought thought I wouldn't really know how".

The other word associated with trying out different pronunciations is *piochais*:

> hmm I think I'd probably say it without going into too much depth of all the individual little sections it would probably be like erm [pʉʃɛɪ] or [piʃɛɪ] [pʉʃɛɪ] I can't tell … I've got a horrible habit of like missing out little bits like now it's the 'I' 'S' I just don't know how you'd pronounce it with the 'I' 'S' really so I thought [pʉʃɛɪz] [pʉʃɛɪz] [pʉʃɛɪz] [pʉʃɛɪ] sort of like a 'Y' sort of like a very faint 'Y'

In this example, LV initially tries out two realizations for <io> ([ʉ], [i]) whilst the pronunciation of remainder of the word ([p__ʃɛɪ]) provides an invariant framework. Again, however, she lacks the basis for deciding between these two realizations of the vowel, which are in any case both incorrect.

LV then questions whether her pronunciation of the final part of the word is correct, apparently concluding that her non-realization of the final <s> – which is correct due to the SFC 'rule' – may result from an oversight on her part. Her strategy of trying out possible pronunciations is therefore again deployed to evaluate a new pronunciation in which the <s> is realized as [z]. This leads her to reject the new pronunciation and revert to [pʉʃɛɪ], which is identical to the very first pronunciation she produced at the start of the discussion.

**b. Referring to known words**

In two word discussions, LV refers (or attempts to refer) to known words to support her decoding. This strategy is used effectively to determine the pronunciation of <eu> in *ferreux*. Towards the end of her discussion, having repeatedly tried out the sounds [ʉ] and [ɛɪ], she draws an analogy to the French word *deux* ([dø], 'two') in order to decide between these two options. This word appears in the poem used in Programme A, although it had almost certainly been covered previously within the MFL curriculum.

**LV** well two is sort of [œ] [œks] sort of
**RW** two what do you mean
**LV** two like [də] the number two is … it ends E U X but it's sort of [dʉ] or [d] or [də] so [fɛʁ] [fɛɹə] it'd probably be more like that cos that's the only E U X I can recall that's probably how I'd do it

Here, LV correctly recalls the source word's written form and, although she expresses some uncertainty over its pronunciation, at least one of her realizations ([də]) is 'acceptable'. The phoneme [ə] is then imported into the target word, resulting in a pronunciation in which all elements are realized 'acceptably'. This example clearly shows the potential value of the analogy strategy (and of conscious reasoning more broadly): it leads to the rejection of the previously considered options for the vowel <eu>, both incorrect, and replaces them with a new sound which is consistent with French GPC, albeit in anglicized form.

LV's other attempted use of the analogy strategy, when trying to pronounce <poi> in *poignard*, is less successful. As at t1, it is impeded by a failure to identify appropriate source words.

isn't *poison* or something … it's sort of it's like P O I'm sure I've seen it … like in the poem or something and it's like P O U S S something … I think it was about snake or something I can't quite remember

Here, various dim recollections point her towards French words spelt similarly to the target word, but she is unable to recall them more precisely. She explicitly refers to the poem learnt as part of Programme A, which was designed to provide a bank of securely-known source words for just this purpose; however, her knowledge of it is partial and inaccurate. Her reference to a word beginning with <pouss> and perhaps related to snakes seems to reflect a confusion of the words *poulain* (/pulæ̃/, 'chick'), *poisson* (/pwasõ/, 'fish') and *serpent*

(/sɛʁpɔ̃/, ‘snake’), all of which appear in the poem; *poisson* may also be the source of her reference to the similarly-spelt English word ‘poison’.

**c. Dividing words into smaller units**

LV explicitly mentions dividing words into smaller units on three occasions. In *poignard*, it is described as a way to make the decoding task more manageable, since the word is “longer so there’s more to think about”. She divides the word correctly into syllables, keeping grapheme boundaries intact (*poi/gnard*). In *piochais*, by contrast, the digraph <ai> is incorrectly fragmented as part of a three-way segmentation (*pio/cha/is*), although this appears not to be reflected in LV’s pronunciations of the word.

***Knowledge bases***

**a. Beliefs about specific GPC**

In four word discussions, LV comments explicitly on the pronunciation of specific graphemes, with an apparent influence on her subsequent decoding. This occurs twice when discussing *dédain*, first with regard to the diacritic, which she identifies using correct metalanguage:

> the erm what’s it called the accent on the E makes it sound different so I probably wouldn't pronounce it just erm how you normally would so like [i] … I’m sure it would be sort of more [ɛ] … normally without it I’d probably say erm [di] but it’s got the accent so I’d probably say [dɛ] [dɛdɔ̃]

However, although LV expresses confidence in her belief that <é> should be pronounced [ɛ] rather than [i] ("I'm sure…"), it is in fact incorrect. This is despite having received explicit instruction in this grapheme as part of the decoding programme (though there is of course a slim possibility that LV was absent on the day <é> was covered).

In the same word discussion, LV also refers to explicit knowledge about the realization of <ain>. This knowledge, explicitly identified as deriving from the decoding instruction, correctly recognizes that the vowel is nasalized; however, the specific nasal phoneme used is incorrect (/ɔ̃/ instead of /æ̃/).

> we've been studying the erm the poem which I think helped with like breaking up the little sounds so I'd probably say sort of [dɛdɔ̃] because I think we've looked at … that ending and I'm sure it sounded like that

By contrast, when discussing *huilage*, LV makes two correct comments on the decoding of specific graphemes, both again said to derive from explicit instruction and both subsequently reflected in the pronunciation of the target word. The first refers to <age>, the second to <h>:

> [ʉlɑːʒ] erm because … we've looked at A G E and I'm sure it's [ɑːʒ] … we were talking about erm way way before when we first looked at the poem like oh I'm not sure if it is in the poem we were looking at H and like not not pronouncing it

Later in the discussion, she reiterates her view about the 'silent' <h>: "I don't think you pronounce the H much if at all".

In the discussion of *ferreux*, LV also correctly identifies that the final <x> is 'silent', this time (as quoted above) apparently based on her own observations of the language. Although the SFC rule is included in the instruction programme, and although the poem contains several words ending in 'silent' <x> (*deux, vieux, yeux, agneaux*), this is not mentioned.

**b. Knowledge of the French alphabet**

At t2, there is one instance (when discussing the <u> in *huilage*) where LV appeals to knowledge of the French alphabet, or at least can be inferred to do so:

> when we first came in and we did the letters we wrote down how they sound erm U was sort of [ʉ] I think erm [ʉ] so erm I just assume that it hadn't changed but it's really guesswork really

It is possible that LV's recollection of "when we did the letters" refers to the decoding instruction, although the fact that this was "when we first came in" – perhaps referring to the start of Y7 – suggests that she is thinking of alphabetic letter names, as observed at t1. Coincidentally, in the case of <u>, the name and phonemic value of the letter are identical (/y/, here anglicized to [ʉ]), although in most cases they differ. In this case, however, the <u> actually represents the semivowel /ɥ/.

**c. Feeling for French**

On one occasion, LV explicitly appeals to an instinctive feeling for how the target word should be pronounced. This occurs when she explains that she did not sound the <g> in *poignard* "cos I couldn't quite think of it with the G it didn't sound right". Whilst this may be true, her instinct is apparently unable to generate a correct realization of <gn>.

Implicit knowledge also seems to play a role in generating LV's initial pronunciation of *ferreux*, which (as noted above) she produces "without sort of too much thought". Again,

however, the resulting pronunciation is incorrect and LV immediately suggests a second possibility (also incorrect).

***Self-efficacy and feelings about decoding***

Perhaps due to the fact that there is less reliance on repeatedly trying out different pronunciations, LV's SRI at t2 gives less of a sense of overriding uncertainty. Nonetheless, she makes some explicit statements that she is unsure of her pronunciations, such as when she says that her realization of <uil> in *huilage* "is guesswork really" or when she notes that she is not sure "if I should pronounce the 'I'" in *poignard*.

However, LV does express confidence in her decoding knowledge on at least one occasion. This is in relation to her belief about the 'silent' <h> in *huilage*: "I'm sure about that". At least in respect of this particular GPC, explicit instruction appears to have successfully led to secure declarative knowledge which she then draws upon with confidence to support her decoding.

**6.4.3.4 Change between t1 and t2**

The most immediately striking difference between LV's two SRIs is her much greater reliance, at t1, on the strategy of trying out different pronunciations of the target words in order to evaluate alternative realizations of specific graphemes or grapheme groups. This permeates almost all her discussions at t1, whereas it occurs in only two word discussions at t2. Although this strategy may reflect a determination not to pronounce the target words (or at least parts of them) simply as if they were English, it also seems indicative of a sense of

uncertainty about how they should sound. Further, the fact that some of the alternative realizations are repeated many times without LV coming to a conclusion about which one sounds 'right' reflects her lack of a secure basis for evaluating the pronunciations she produces. To this extent, the strategy can be seen as counterproductive, since it absorbs time and mental resources which could be used for more effective types of reasoning.

Conversely, there seems to be a strengthening of LV's conscious knowledge of French GPC between the two SRIs. At t1, she makes several explicit comments about how particular graphemes should be pronounced, but these tend to be vague and/or defective. At t2, by contrast, most of her comments are at least partially accurate and are often reflected in correct or 'acceptable' realizations of the graphemes concerned. This may be causally related to her lesser reliance on trying out alternative pronunciations: better knowledge would be expected to lead to less uncertainty about how particular graphemes should be pronounced. In several cases at t2, LV also states that her GPC knowledge derives from the programme of decoding instruction. This is encouraging in that developing participants' knowledge of French GPC was one of the stated aims of the programmes.

Despite this apparent improvement in LV's GPC knowledge, there is some evidence to suggest that she continues to appeal to alphabetic letter names to support her decoding. Learning the French alphabet was apparently covered early in LV's MFL lessons and was reinforced by letter name posters (complete with anglicized transcriptions), seen by the researcher in her classroom. Although the purpose of learning the alphabet may be linked to communicative aims such as being able to spell one's name when making a telephone booking, for LV (just as for EH in school B) it seems to have led to a misconception that letter names are directly useful for decoding. In fact, as is clearly exemplified at t1,

appealing to letter names can seriously impede accurate decoding. If (as appears to be the case) this misconception persists at t2, the decoding instruction has not managed to dislodge it.

Another aim of Programme A was to develop participants' use of the analogy strategy by providing them with (a) a bank of securely-known source words and (b) practice in using the strategy. However, there is no evidence of any increase in the frequency or effectiveness with which LV uses the analogy strategy. At t1, before the programme of instruction, she demonstrates several attempts to use it, although these are undermined by inaccuracies in the stored orthographic or phonological forms of the source words and by an inability to identify appropriate source words in the first place. At t2, she attempts to use the strategy less often (though this may of course be influenced by differences in the particular target words covered) and when she does, similar problems are observed in identifying appropriate source words. On one occasion, LV does attempt to refer to a source word from the decoding poem, but fails to do so accurately.

Finally, LV shows a large increase in RAT scores between t1 and t2, more than double the mean value for the intervention group (Table 6.1). It is plausible to hypothesize that this may have resulted (at least partly) from her improved conscious knowledge of French GPC. However, it is impossible to know to what extent it is caused by other factors as well (or instead), such as developments in her implicit knowledge.

# Chapter 7: Discussion and limitations

## 7.1 Research Questions 1 and 2

*Do the programmes of instruction lead to more accurate decoding of L2 French?*

*Is one of the programmes more effective than the other?*

These questions were addressed using the Reading Aloud Test (RAT) used by Woore (2006, 2009), argued to provide a valid measure of French decoding proficiency as defined in this thesis (section 3.6.3.4). Participants' pronunciations of target words were scored at the individual grapheme level; intra- and inter-rater reliability checks indicated that this scoring was highly reliable.

### 7.1.1 Non-Linear Decoding

Around 8% of grapheme tokens analysed at t1 could not be scored in a straightforward way, because the pronunciation of the RAT item could not be unambiguously mapped onto the linear sequence of its graphemes. Such cases, coded as 'Non-Linear Decoding' (NLD), provided the basis for an additional strand of analysis, beyond that based on the accuracy of individual grapheme-to-phoneme mappings. In section 3.6.3.3, it was argued that the proportion of graphemes coded NLD might be expected to decrease as L2 decoding proficiency increased, as learners became more proficient in applying the basic principle of linear grapheme-to-phoneme mapping; this in turn provides the basis for subsequent improvements in specific GPC. A small but consistent decrease was indeed found in the

mean percentage of NLD codes generated by all but one participating classes. Further, the proportion of all participants generating no NLD codes (i.e. those whose pronunciations of all target words showed a one-to-one mapping between graphemes and phonemes) rose from 15% at t1 to 22% at t2. However, there was no significant difference between (a) the intervention and comparison groups or (b) between intervention groups A and B in terms of the proportion of NLD codes at either time point.

Incidental observations suggest that these NLD codes often involved pronunciations which could be characterized in terms of the omission, addition, displacement or duplication of phonemes (though this is not to say that they were generated by such processes). These errors recall the 'deviations from the grapheme sequence' described by Frith, Wimmer and Landerl (1998) amongst young beginner-readers of L1 English; similar problems were also reported by Hickey (2005) in her analysis of L2 Gaelic word recognition errors made by young (third grade) L1 English readers. The reduction in cases of NLD amongst the sample as a whole could therefore be interpreted as a reflecting developments in some participants' basic literacy skills (underpinning both L1 and L2 decoding). This recalls developmental models of L1 literacy such as Ehri's (1995, 2005), where young beginner readers gradually progress from a 'partial alphabetic' phase, in which they recognize words based on their global shape or a few salient letters, to a 'full alphabetic' phase, characterized by more systematic intra-word analysis. However, the origins of NLD codes were not systematically analysed in the current study, because participants' pronunciations of RAT items were examined only at the level of individual GPC, not as whole words. This represents a potentially rich source of further inquiry.

### 7.1.2 Main analysis

The main analysis of RAT data involved comparing groups in terms of the numbers of grapheme tokens at each time point which they realized (a) 'correctly', that is, in a native-like or near-native-like way and (b) 'acceptably', that is, in a manner consistent with French GPC but allowing for non-native realizations of L2 phonemes. These two categories of outcome were measured by the variables *score1* and *score2* respectively.

An initial finding was that, on both variables, both intervention and comparison participants showed statistically significant mean increases between t1 and t2, even though those in the comparison condition had received no explicit decoding instruction. However, the increases for the comparison group were extremely small: around three additional grapheme tokens out of 274 decoded 'correctly' or 'acceptably' after six months' additional MFL lessons. This matches previous findings based on cross-sectional surveys of KS3 learners from the same population (Woore, 2006; Erler, 2008).

By contrast, previous *longitudinal* data gathered using the same RAT as in the current study (Woore, 2009) found that a sample of eighty-five MFL learners made no discernible progress in French decoding at all between the end of Y7 and the end of Y8. Correspondingly, when the values of *score1* and *score2* in the current study were examined at the individual class level (rather than treating the intervention and comparison conditions as homogeneous wholes), no significant differences were found between the mean t1 and t2 scores for three of the five participating comparison classes. Overall, therefore, the balance of evidence suggests that, in the absence of explicit instruction, most beginner learners in MFL classrooms make little or no discernible progress in French decoding over the course of one

or two years' language learning, as measured by the number of graphemes decoded correctly. This does not of course mean that it is impossible for L1 English learners to acquire L2 French decoding proficiency implicitly, but rather that, on average, this has not occurred under the prevailing instructional approaches within the limited curriculum time available.

Another preliminary finding relating to the sample as a whole concerns the distributions of *score1* and *score2*. At t1, when participants had had little experience of reading French, there was a clear grouping of scores in the range from 20 to 40 (out of 274). This therefore seems likely to reflect the accuracy with which unknown French words are decoded when (predominantly) English symbol-to-sound mappings are applied; this score is not zero due to the overlap between French and English GPC. The distributions of *score1* and *score2* also show a marked positive skew at both time points. The diminishing rightward 'tails' may result from those participants who *do* correctly follow French rather than English decoding conventions to a greater or lesser extent (with more doing so to a lesser extent than to a greater extent). At t2, the positive skew remains but the centre of gravity of the sample as a whole shifts rightwards and there is a more even slope beyond the score of about 40, suggesting that more participants are now applying (at least some) French symbol-to-sound mappings.

### 7.1.3 Comparing intervention and comparison groups

Although both intervention and comparison groups showed mean increases in the proportions of grapheme tokens realized 'correctly' and 'acceptably' between t1 and t2, the increases were significantly larger for the intervention group. Research Question 1 can therefore be answered affirmatively. This finding appears to be robust in that it is based on a relatively

large sample of participants (N=186) drawn from ten different classes in four different schools. Though these schools cannot claim to be representative of English secondary schools – notably in respect of their generally below average numbers of learners with Special Educational Needs, English as an Additional Language, minority ethnic backgrounds and low socio-economic status – they are arguably not untypical of many mixed-sex comprehensive schools in non-inner city areas.

Whilst the research design means that school effects cannot be ruled out as a confounding variable, it appears unlikely that both intervention schools (A and B) would show larger increases in RAT scores than both comparison schools (C and D) as a result of factors other than the explicit instruction they received. Further, based on published performance indicators, participants in school B might have been expected to perform less well than those in schools C and D, not better (section 3.4.1). It is also notable that participants in schools A and B made more progress than those in school D despite receiving between thirty and forty minutes less MFL teaching per week.

Teacher effects also represent a potentially confounding variable. In particular, a pattern emerged – which would have been avoided if possible – whereby the intervention teachers tended to have more years' teaching experience than those in the comparison condition (in several cases, considerably more). On the other hand, one intervention teacher had joint least teaching experience, yet her class performed similarly to other intervention classes and better than all comparison classes in terms of their progress in L2 decoding. Nonetheless, in order to gauge the possible extent of any teacher effects, it would have been helpful to measure participants' progress in other aspects of L2 learning, providing a benchmark against which to compare their progress in decoding more specifically.

The decoding instruction took place entirely within timetabled MFL lessons, requiring no additional teaching time. Nonetheless, in the regular discussions with intervention group teachers, none expressed concern that their classes had fallen behind parallel classes in the same school, in terms of either (a) their attainment in the four skills of listening, speaking, reading or writing French or (b) their coverage of prescribed Schemes of Work. Thus, there was no evidence that the intervention groups' additional progress in decoding came at the expense of other aspects of MFL learning. However, a limitation of the current study is that this was not tested systematically. This was because (a) existing measures, such as National Curriculum levels, were considered unreliable; and (b) it was impracticable to administer additional tests due to the already large testing burden.

There is also no evidence concerning the longer-term effectiveness of the instruction programmes, due to the lack of delayed post-test data (section 3.3). On the other hand, due to the extended nature of the intervention, learners' knowledge of some individual GPC was tested several months after these GPC had been explicitly covered in the programmes of instruction: in other words, there was already a considerable delay between the teaching and the testing (though these same graphemes may also have occurred incidentally in the word lists used in subsequent segments). Further analysis of the current dataset could explore the possible relationship between the gain scores associated with individual GPC and the recency with which they were explicitly covered in the programmes of instruction.

### 7.1.4 Comparing Programmes A and B

To evaluate the relative effectiveness of the two programmes of instruction, the score increases of participants in intervention groups A and B were compared. No statistically significant differences were found at either t1 or t2 in terms of either 'correct' or 'acceptable' realizations. The answer to Research Question 2 is therefore negative, at least in relation to RAT scores. Encouraging participants to refer to source words in the poem as a mechanism for retrieving the pronunciation of target graphemes appears to have conferred no advantage over and above the phonics approach adopted in Programme B.

However, as argued in section 3.5.2, this may reflect limitations in the design of Programme A. In particular, it may be that a more systematic, 'staged' approach to strategy instruction (as proposed by Macaro, 2001 and adopted in studies such as Macaro and Erler, 2008a) would have had a greater effect on participants' strategic reasoning and hence on decoding outcomes. Such an approach would include (a) awareness raising and strategy modelling, (b) scaffolded practice in using the strategy and (c) the progressive removal of scaffolding to develop participants' independence in using the strategy.

In their current form, the two programmes were therefore broadly similar: both involved (a) explicitly presenting one or two (or occasionally more) individual GPC each lesson, followed by (b) opportunities for learners to work out the pronunciations of unfamiliar words containing these (and other) graphemes based on a detailed, intraword analysis of their orthographic forms and (c) corrective feedback from the teacher. This may explain the very similar levels of progress in French decoding made by participants in the two groups. Therefore, whilst a valid (negative) answer has been provided to Research Question 2, the

*underlying* question – whether a programme of L2 decoding instruction based on the analogy strategy approach is more or less effective than a phonics-based approach – requires further investigation.

Finally, it is also possible that – despite the lack of staged strategy instruction – participants in Programme A did increase in their attempts to use the analogy strategy, but that these attempts were undermined by inadequate knowledge of the 'source words' contained in the poem: in other words, that they had not memorized the poem securely. The self-report interviews provide some evidence to support this possibility, albeit on a very limited scale. A measure of all participants' knowledge of the words contained in the poem (in both written and spoken forms), administered at t2, would have allowed a more systematic investigation of the relationship between participants' knowledge of the poem on the one hand and their decoding outcomes on the other.

### 7.1.5 Interpreting the effects of the instruction

The magnitude of the significant difference between the scores of intervention and comparison participants is crucial in judging the effectiveness of the programmes. At t2, there was a significant difference of 6.0 in the mean percentage of target grapheme tokens (i.e. those tokens covered by all participants in the programmes of instruction) which were realized 'correctly' by the two groups, whereas at t1 the two groups' scores did not differ significantly. This indicates that, on average, intervention participants decoded just over eight additional tokens correctly compared to comparison participants: the equivalent of about three whole words on the RAT, since each RAT item contained, on average, just over two and a half target grapheme tokens. This amounted to a medium effect size ($\eta^2_{partial}$

=0.15) after controlling for t1 reading comprehension scores, which provided an (albeit partial) measure of baseline L2 proficiency.

However, it might have been expected that groups would register little improvement in terms of the number of 'correct' realizations (i.e. *score1*), since to some extent this required them not only to acquire L2-specific GPC, but also to develop more accurate, L2-specific phonemic representations – widely assumed to be a difficult task for late onset learners (Scovel, 1969; Birdsong, 1999). This was the motivation for implementing an additional, less stringent mark scheme, recording the numbers of grapheme tokens realized 'acceptably' by participants (i.e. *score2*). Intervention participants did indeed show a statistically significant t1-to-t2 increase in the number of 'acceptable' but not 'correct' realizations (referred to as 'only acceptable' realizations), whereas comparison participants did not; however, the magnitude of this increase for the intervention group was very small, amounting to around one single grapheme token realized 'only acceptably'. For both intervention and comparison groups, therefore, the overall increases in 'acceptable' realizations were very similar to the increases in 'correct' realizations. However, this may simply reflect the fact that the RAT score increases related mainly to consonantal GPC, for which *score1* and *score2* were identical (section 3.6.3.3) and the 'only acceptable' classification consequently did not apply.

What is the *significance* of decoding around eight additional grapheme tokens 'correctly' or 'acceptably'? From one perspective, this seems a small return on the considerable teaching time invested in the instruction programmes (around thirty ten- to fifteen-minute segments, totalling between 5 and 7.5 hours' instruction). On the other hand, there are several reasons for evaluating these score increases more positively. First, the programmes of instruction in

the current study are relatively short compared to other phonics-based interventions evaluated in the literature, both in L1 (e.g. Gaskins et al., 1988: around twenty minutes per day for a whole academic year) and in L2 (e.g. Yen, 2004: twenty hours in total; Liaw, 2003: forty minutes per week for one semester).

Second, although in total between five and seven and a half hours were dedicated to decoding instruction in the current study, the design of the programmes meant that each individual GPC was the *explicit* focus of only one segment (although the same GPC may also have occurred incidentally in the lists of unfamiliar words used in previous or subsequent segments). Thus, where the intervention group outperforms the comparison group in respect of a given grapheme type, this can be seen as the product of only about ten to fifteen minutes' instruction targeted specifically on this GPC.

Ultimately, however, the improved decoding outcomes resulting from the programmes of instruction can only be interpreted in the light of additional, contextualizing information. Further research is therefore needed to investigate the effects (if any) of a given change in L2 decoding proficiency on other theoretically relevant language-learning outcomes (section 2.3), such as reading comprehension or vocabulary learning (as suggested by Hamada and Koda, 2008). Data gathered as part of the current study, though not reported in this thesis for space reasons, may allow an initial exploration of this issue.

It is also important to note that the degree of progress in L2 decoding made by participants within each group (Intervention A, Intervention B, Comparison) varied considerably: some showed much larger or smaller RAT score increases than the mean figures for their group. Further research is needed to investigate the origins of these individual differences in

decoding progress. One possibility is that they may reflect differences in participants' phonological awareness, which Durğunoglu, Nagy and Hancin-Bhatt (1993) argued to be non-language-specific: that is, phonological awareness developed in the L1 should be available for transfer when acquiring L2 literacy. This possibility could have been explored through a phonological awareness test administered as part of the current study, not conducted due to the already large testing burden.

Further, for some participants in every class, *score2* values (that is, the number of grapheme tokens realized 'acceptably') actually *decreased* between t1 and t2; indeed, in most classes, at least one participant produced between ten and twenty fewer 'acceptable' realizations. A condition-level pattern emerged here, with the proportion of participants obtaining negative *change2* values tending to be higher in comparison than intervention classes. Nonetheless, even in intervention classes, considerable numbers of participants showed *score2* decreases: at face value, a puzzling finding. It cannot be explained as the result of L1-entrenched processing leading to 'negative transfer', since there is no theoretical basis for expecting such transfer effects to *increase* as learners gain more experience of processing the L2. One tentative explanation lies in the conscious reasoning which Erler (2003) suggested some of her participants deployed in order to overcome their automatic, L1-based processing. It is possible that, during the course of Year 7, some participants in both conditions (intervention and comparison) became more aware that L2 GPC are different from those of the L1 in at least some cases. When they tried hard to pronounce words in a 'French' way (encouraged by the 'formal' Reading Aloud Test), at least some learners may have consciously sought to make their realizations different from English-based ones, as observed amongst beginner decoders in Woore (2010). However, some participants may have overgeneralized this principle of 'difference from the L1', applying it even in cases where L1-based decoding

would, paradoxically, have led to 'acceptable' realizations. On this view, their decrease in 'acceptable' pronunciations could be seen in terms of the initial decline in performance within a U-shaped curve model of L2 learning (McLaughlin, 1990). The greater tendency for comparison than intervention participants to show *score2* decreases may then reflect the fact that the former were more likely to override their L1-based decoding without knowing the correct realizations to put in its place; by contrast, more of those intervention participants who rejected L1-based realizations may have done so in favour of what they knew to be the correct French forms.

One possible implication of the U-shaped curve model, applied to the current study, is that the number of 'acceptable' pronunciations may eventually increase, given sufficient time and practice in L2 decoding. A longer-term intervention may therefore be necessary in order for the benefits to be seen in terms of accurate decoding outcomes. In the current study, delayed post-test data may also have revealed whether any longer-term improvements emerged even in the absence of additional explicit instruction.

### 7.1.6 Analysis of individual GPC

Participants' performance was also analysed at the level of individual GPC, in order to assess whether some were associated with greater degrees of improvement than others. 'Gain scores' were calculated for each grapheme type (for example, <ç>), measuring the extent of any increase (or decrease) in the percentage of 'acceptable' realizations of its various tokens (in this example, the <ç> in *caleçon*, *maçonne* and *traçant*). Considerable variation was found at the level of individual grapheme types, which for convenience were grouped into three categories: consonants, oral vowels and nasal vowels. However, the analysis was

mainly descriptive rather than being tested statistically due to (a) the large numbers of individual comparisons involved and (b) the small token counts of some grapheme types.

In respect of the thirteen consonantal GPC, a relatively consistent pattern of gain scores emerged. Almost all were associated with mean increases in the percentage of 'acceptable' realizations by participants in all three groups (Intervention A, Intervention B, Comparison). Further, the intervention groups showed larger increases than the comparison group in respect of seven of the eleven GPC covered by all participants in the programmes of instruction; and in five of these cases (soft <c>, <j>, <ch>, <ç>, Silent Final Consonants or 'SFC'), the difference in gain scores was considerable. In the case of SFC, which had a larger number of individual tokens than other GPC, statistical tests confirmed that the difference in gain scores between the intervention and comparison groups was significant. By contrast, the inverse pattern – where the comparison group achieved a larger gain score than both intervention groups – arose in respect of only two consonantal GPC. Thus, it may be tentatively suggested that the programmes of instruction had a fairly consistent positive effect on participants' decoding of consonants.

However, some consonantal GPC were associated with much larger improvements in decoding accuracy than others. It was hypothesized that grapheme types associated with higher baseline (t1) accuracy scores might tend to show smaller t1-to-t2 gain scores, because they offered less room for improvement. In fact, however, the opposite pattern emerged: gain scores tended to be lower for grapheme types whose t1 scores were also lower. This tentatively suggests that the most difficult consonantal graphemes for participants to decode (as evidenced by low t1 accuracy scores) were also those most resistant to change through explicit instruction (as evidenced by low t1-to-t2 gain scores). In turn, it had previously been

found that the t1 accuracy scores associated with individual GPC broadly matched the predictions of a contrastive analysis of the English and French writing systems (Woore, 2006). Together, these findings can be interpreted within a cognitive processing framework in terms of the triggering of L1-entrenched symbol-sound associations by L2 input; problems then arise when letters or letter combinations look the same in the two languages, but map onto different phonemes.

An exception to this general trend, whereby gain scores diminished as t1 accuracy scores decreased, was provided by the two contextually-dependent realizations of <c>. A contrastive analysis would predict little difficulty in respect of these GPC: in French, as is often the case in English, <c> is 'soft' (i.e. pronounced as /s/) when followed by <i>, <e> or <y> and 'hard' (pronounced as /k/) elsewhere. In fact, however, soft <c> was associated with small gain scores, whilst hard <c> actually showed a decrease in 'acceptable' realizations in all three groups. Further, the two intervention groups (A and B) showed larger increases in 'acceptable' realizations of 'soft' <c> than the comparison group, but also slightly larger *decreases* in respect of 'hard' <c>. Following the argument above, a tentative explanation for these findings might be that the decoding instruction led to a greater tendency among intervention participants to question their L1-based grapheme-phoneme mappings, using conscious reasoning to 'override' them. Intervention participants may have become particularly sensitized to the fact that French <c> is pronounced /s/ in some contexts (perhaps influenced by the non-L1 grapheme <ç>, also pronounced /s/); they may then have overgeneralized this 'rule', applying it inappropriately in contexts where, paradoxically, simply following the English pattern (i.e. pronouncing <c> as /k/) would have generated correct pronunciations. In other words, their L1-based processing may have been 'destabilized', but secure L2-based processing may be yet to develop. If so, this again

suggests that acquiring proficiency in at least some aspects of L2 decoding may be a long-term process, even with the help of explicit instruction.

In respect of vowel graphemes, much greater variation was found between the gain scores of individual grapheme types than in the consonant category. There was also no obvious relationship between individual GPCs' gain scores and t1 accuracy scores. This diverse picture may be explained by the fact that individual vowel GPC show greater variation in terms of their levels of difficulty for beginner decoders, since they vary according to both (a) whether they require different grapheme-to-phoneme mappings in French and English and (b) the degree of discrepancy between the phonological realizations of the target phonemes in the two languages. However, without a more detailed picture of the instruction actually received by each class, it is impossible to know whether the variation in gain scores is in fact due to properties of the grapheme types themselves (e.g. some might be more amenable to explicit instruction or implicit acquisition than others) or whether it simply reflects variations in the teaching experienced by participants within each condition (e.g. explicit discussion of particular GPC may have arisen incidentally in comparison classes; certain segments within the programmes of instruction may have been covered more hurriedly than others or with lower levels of student engagement).

Compared to consonantal graphemes, there is also less evidence that explicit instruction had a consistently positive effect on the decoding of vowel graphemes. *Decreases* in the percentage of 'acceptable' realizations occurred in all three groups (Intervention A, Intervention B, Comparison) and in respect of seven of the seventeen vowel GPC. This perhaps reflects the greater difficulty generally associated with decoding vowels than consonants, because (following a contrastive analysis approach) it is in respect of vowels that

French and English GPC differ most markedly (Sprenger-Charolles, 2004). A descriptive analysis of the mean gain scores associated with individual GPC did reveal some evidence of a general advantage for intervention over comparison participants; however, the size of the differences between the groups' mean gain scores varied widely and was in many cases slight. Further, the intervention groups actually showed decreases in the number of 'acceptable' realizations of three vowel GPC covered by all participants in the programmes of instruction (<ê>, <ai>, <ei>). This could again (tentatively) be interpreted as reflecting a 'destabilization' of intervention participants' L1-based processing. Decoding all three of these graphemes according to frequent L1 GPC would result in 'acceptable' pronunciations: <ai> and <ei>→/ɛɪ/; <ê>→/ɛ/, ignoring the circumflex. Therefore, participants may have consciously avoided automatic, L1-based realizations even in cases where these would have been 'acceptable'.

Finally, in respect of nasal vowel GPC, limited conclusions can be drawn because only three of these (<en>, <oin>, <on>) were covered by all participants in the programmes of instruction. However, in terms of 'acceptable' realizations, a descriptive analysis highlighted an interesting contrast whereby two nasal grapheme types (<on> and <un>) showed negative gain scores (i.e. were associated with decreased decoding accuracy), whilst the remaining six grapheme types showed positive gain scores. As in the case of the vowels <ê>, <ei> and <ai> discussed above, a possible explanation for this is that decoding both <on> and <un> according to English GPC would result in 'acceptable' realizations (/ɒn/ and /ɐn/ respectively), whereas this is not true of any other nasal grapheme types. Again, therefore, it may be that some participants have moved away from L1-based realizations in the absence of a better alternative.

Support for this argument can be found by comparing the different groups' gain scores in respect of (a) <on>, which was covered by all participants in the instruction programmes and (b) <un>, which was not. The latter grapheme type shows a decrease in all three groups, whereas <on> shows a larger decrease in mean score for the comparison condition but only a very small decrease for intervention group B and an increase for intervention group A. Again, therefore, it is possible that comparison participants were more likely to reject L1-based decoding of <on> yet not know the correct realization, whereas intervention participants who rejected the L1-based realization may have been more likely to replace it with the correct French alternative. However, it must be acknowledged that this argument is speculative, being based on limited evidence.

## 7.2 Research Question 3

*Do the programmes of instruction lead to any other changes in participants' realizations of French graphemes?*

To address these questions, chapter 5 compared participants' realizations of grapheme tokens (a) between the three groups, Intervention A, Intervention B and Comparison and (b) between the two time points, t1 and t2. The analysis was based on the phonetic transcriptions of participants' RAT output at the level of individual GPC (section 3.6.3.3). Although it is acknowledged that the transcription process involved subjective judgments, especially concerning the identification of vowel phonemes, both intra- and inter-rater agreement coefficients were high. This suggests that the representations of pronunciations are acceptably valid and reliable.

For space reasons, the analysis was restricted to a small subset of participants' total RAT output. One grapheme type from each of the three categories of consonants, oral vowels and nasal vowels was therefore selected: <ç>, <ou> and <oin>. Based on a contrastive analysis, these were predicted to pose increasing levels of difficulty for participants. One individual grapheme token (from the ten tokens in total exemplifying these three grapheme types) was then arbitrarily chosen for analysis. These appeared in the words *caleçon* (<ç>), *ougrien* (<ou>) and *goinfrez* (<oin>). It is acknowledged that different findings may have been obtained had different grapheme types and/or tokens been selected; further analysis is planned to explore this possibility.

The evidence from the three grapheme tokens selected for analysis generally indicated an affirmative answer to research question 3. Statistical tests were used to compare the three groups' output at each time point in terms of the most frequently-attested realizations of <ç>, <ou> and <oin>. Further tests examined realizations of <oin> in terms of whether or not they exhibited each of three target-like features: (a) pre-vocalic /w/; (b) nasality; (c) monosyllabicity. In four of these six statistical analyses, significant t1-t2 differences in the realizations of the grapheme tokens were found for intervention group A, but not for the other two groups; however, in one case the difference approached significance for intervention group B. This suggests that the 'poem' approach led to greater changes in the way the target graphemes were decoded than either the 'phonics' approach or an absence of instruction.

However, it was argued that these statistical tests were likely to provide a conservative picture of changes in decoding outcomes, since they conflated a large number of distinct pronunciations of each token into a single category of 'other' realizations. Additional, descriptive analyses were therefore undertaken. For all three tokens, these revealed broadly

similar patterns of t1-t2 differences in the two intervention groups (even though these generally failed to reach significance in intervention group B), whereas comparison participants showed either few differences or contrasting patterns of difference. This (tentatively) suggests that the two programmes of instruction led to similar changes in the way the target grapheme tokens were decoded. This interpretation of the findings is consistent with the analysis of RAT scores (see above), which found an advantage for the two intervention groups over the comparison group.

A key additional question concerns the nature of the changes in decoding between t1 and t2. Specifically, is there evidence that participants' realizations of <ç>, <ou> and <oin> became more target-like in some way, beyond any changes in the numbers of 'correct' and 'acceptable' realizations, as reflected in RAT scores? Alternatively, if a principal task for beginner learners is to overcome automatic, L1-based processing, then progress in decoding may be identifiable simply in departures from L1-like realizations of L2 letter-strings in those cases where the two languages' GPC differ. These issues are explored below in relation to each of the three grapheme tokens <ç>, <ou> and <oin>.

It should be noted at this point that participants' RAT output records only the *product*, not the *process* of their decoding. Therefore, interpreting a given pronunciation as the likely product of English GPC involves making inferences about the processing involved. Such inferences are not unusual in studies of L2 decoding (e.g. Durğunoglu, Nagy & Hancin-Bhatt, 1993; Hickey, 2005); however, a particular difficulty for the current study lies in the fact that participants' pronunciations of RAT items are systematically recorded only at the level of individual GPC, rather than holistically (as whole words). This reduces the scope for inferring the origins of some incorrect forms.

### 7.2.1 Realizations of <ç>

Across the sample as a whole, the overwhelmingly predominant realization of <ç> in *caleçon* was /k/, produced by 68% of all participants at t1 and 58% at t2. This realization plausibly results from L1-based decoding: in this orthographic context, /k/ is the canonical realization of <c> (without the diacritic) in English, as in *recondition* (and equally in French, as in *balcon*, /balkɔ̃/, 'balcony'). The second most frequent realization of <ç> was /s/, the correct French form, accounting for 16% of all realizations at t1 and 30% at t2. However, /s/ is also a frequent realization of English <c> in other orthographic contexts. According to Berndt, Reggia and Mitchum (1987), the two most frequent realizations of English <c> across all orthographic contexts are /k/ (in 76% of cases) and /s/ (in 23% of cases).

Various other realizations of <ç> in *caleçon* were also attested amongst the sample as a whole: nine at t1 and twelve at t2, each produced by very few participants (often only one). This pattern of one or two dominant realizations together with a wide range of additional ones, each produced by very few participants, is consistent with the positive skew in RAT score distributions described above: most participants may have decoded <ç> using automatic, L1-based processing, as reflected in the predominant realization /k/; however, some learners may have sought consciously to move beyond English GPC to make their pronunciation sound more 'French' – without, however, necessarily knowing what the correct French pronunciation should be. This could result in a variety of different forms, worked out consciously by participants as solutions to the decoding 'puzzle'. This recalls Tarone's (1982) argument that L2 interlanguage is likely to display higher levels of variation when greater attention is being paid. Of course, some of the realizations may also have arisen

through other causes, such as misperceptions of the letter string; however, there is no space here to speculate further about the origins of all the attested realizations.

Statistical analyses, based on the proportions of participants realizing <ç> as /k/, /s/ or some 'other' form, found significant t1-to-t2 differences for intervention group A but not for the other two groups; however, the results for intervention group B approached significance. A descriptive analysis tentatively identified further between-group differences in the way participants' decoding of <ç> developed. The most obvious finding was that the proportion of participants producing /k/ fell in both intervention groups but remained constant in the comparison group. Conversely, /s/ was associated with much larger increases in both intervention groups than in the comparison group. Third, the percentage of 'other' realizations rose in intervention group A, remained the same in intervention group B and fell slightly in the comparison group. Finally, all groups showed a decrease in the proportion of realizations coded as NLD. However, caution must be exercised in respect of these last two changes, since the actual numbers of participants involved are small.

Given these descriptively observed patterns, a tentative but plausible hypothesis is that some participants in the two intervention groups, but not in the comparison group, may have moved away from their initial reliance on English-based decoding of <ç> (ignoring the cedilla, which does not occur in English). At t2, some of these participants then produced the correct realization, /s/, whilst others (principally in group A) have adopted 'other', non-English realizations which are nonetheless still incorrect. By contrast, although there was a rise in correct realizations in the comparison group, this may have been driven principally by participants moving away from NLD or from 'other' realizations, rather than from the L1-based form /k/. Of course, these hypotheses are speculative, since the data is analysed only at

the group level: it does not track the realizations produced by individual participants. Nor can it be said with certainty that /k/ does in fact result from the application of English GPC in all cases.

Notably, of the three dimensions of potentially differential development identified here between the groups – (a) the increase in frequency of the 'correct' form /s/; (b) the decrease in frequency of the (putatively) L1-based form /k/; (c) the change in frequency of 'other' realizations – only the first is captured by the RAT mark scheme, which simply tallies all 'correct' or 'acceptable' realizations of consonantal graphemes. In other words, some positive changes appear to have occurred in participants' decoding of <ç>, especially in the intervention groups, which are not reflected in their RAT scores. This is again consistent with the view that participants' progress in French decoding may follow a more complex trajectory than simply moving from 'incorrect' to 'correct' pronunciations.

### 7.2.2 Overall realizations of <ou>

A striking initial finding was that, both at t1 and t2, the distributions of the seven most frequent realizations of <ou> produced by the sample as a whole differed significantly between the four RAT items in which this grapheme appeared (*houblon*, *jouxter*, *ougrien*, *soudain*). More information would be needed in order to explore the processes lying behind these different decoding outcomes in more detail. However, it is plausible to suggest that this contextual variation in the decoding of <ou> results from the L1-to-L2 transfer of processing mechanisms which, due to the inconsistency of English GPC, operate at a larger orthographic 'grain size' than individual graphemes (Ziegler & Goswami, 2005); English <ou> is also realized variably according to its orthographic context (Berndt, Reggia & Mitchum, 1987).

By contrast, in the phonologically more transparent French orthography, individual GPC provide a more consistent basis for decoding. Operating at larger 'grain sizes' is therefore likely to generate incorrect outcomes. According to this argument, one indication of improved French decoding proficiency amongst L1-English learners would be reduced variability in the pronunciation of a given grapheme across different lexical contexts. This represents a possible avenue for further research.

### 7.2.3 Realizations of <ou> in ougrien

As in the case of <ç>, participants' decoding of <ou> in *ougrien* was dominated by a small number of realizations at both time points, with one form in particular ([ɔː][35]) occurring much more frequently than all others, at least at t1. There were then many additional realizations (almost forty across the sample as a whole across both time points), each produced very infrequently, often by a single participant.

Many of the more frequent realizations of <ou>, defined as those produced by at least four participants, are again (speculatively) interpretable as the product of L1-based decoding. For example, the predominant form [ɔː] reflects the English pronunciation of this vowel in a similar orthographic context in *ought*, whilst [aʊ] and [ə] also occur as realizations of English <ou> in other orthographic contexts (Berndt, Reggia & Mitchum, 1987). Further, [ɒ] and [ʊ] – whilst not readily interpretable as English realizations of <ou> – are possible English realizations of the individual vowel graphemes <o> and <u> respectively (as in *pot* and *put*).

[35] Square brackets are used here, since the sounds produced by participants could not always be unambiguously attributed to an underlying phoneme.

On the other hand, the correct French form ([u]) and the 'acceptable' [ʉ] also featured amongst the most frequent realizations at both time points. The form [ʊ], mentioned above, could also be interpreted as an approximation of the correct realization [u], in that the two pronunciations share various articulatory features (both are high, back, rounded vowels).

Finally, some frequent realizations were consistent with neither English nor French GPC; indeed, two were distinctively French sounds, not present in most varieties of British English ([œ], [y]). This may again be hypothesized to reflect a conscious attempt to produce realizations which are somehow 'different from English' or which 'sound French'.

In terms of the less frequently-attested realizations of <ou>, many of these are also difficult to account for on the basis of either French or English GPC. There is not space to speculate about the origin of each one individually. However, it is notable that around one quarter of all realizations produced by the whole sample across both time points are bisyllabic, perhaps reflecting participants' incorrect segmentation of the letter-string into two separate vowel graphemes (<o> and <u>).

Statistical tests based on the five most frequently-attested realizations amongst the sample as a whole found no significant differences (a) between the t1 and t2 realizations for any group, although the results for intervention group A approached significance; or (b) between the three groups' realizations at either t1 or t2. However, a descriptive analysis did provide (tentative) evidence of differential patterns of t1-to-t2 change between the groups. First, the frequency of [ɔː], argued to be a likely product of L1-based decoding, decreased considerably in the two intervention groups but remained constant in the comparison group. The frequency of [ɒ], interpretable as an English-based realization of incorrectly segmented <o>,

also more than halves in intervention group A but shows much smaller decreases in the other two groups. Conversely, the frequencies of the correct and acceptable forms ([u], [ʉ]) increase considerably in intervention group A (more than doubling in the case of [u]) but show much smaller increases or remain constant in the other groups. Finally, the frequency of [ʊ] decreases in intervention group B, increases in the comparison group but shows a much larger (fourfold) increase in intervention group A.

This evidence therefore tentatively suggests that intervention group A may have made more progress than the other two groups in decoding <ou> in *ougrien*. This is reflected in an increase in 'correct' and 'acceptable' forms (as reflected in RAT scores), but also in other changes not reflected in RAT scores: (a) a decrease in (likely) English-based realizations; (b) a decrease in the incorrect form [ɒ]; and perhaps (c) an increase in [ʊ], which could be interpreted as a 'near miss' for the target form [u] (though it could also arise through the English-based realization of the incorrectly-segmented <u>). By contrast, intervention group B and the comparison group show less consistently 'positive' patterns of change in respect of these realizations. On the other hand, intervention group B does pattern with intervention group A in showing a large decrease in the frequency of the (likely) L1-based realization [ɔː].

### 7.2.4 Realizations of <oin>

In contrast to <ou>, preliminary observations suggested no differences in the realization of <oin> in the three RAT items in which it appeared. This may reflect the fact that <oi> is associated with much less variation in English, having only one realization, /ɔɪ/, in most orthographic contexts (Berndt, Reggia & Mitchum, 1987).

Participants' realizations of <oin> in *goinfrez* followed a similar pattern to the previous two grapheme tokens: only five forms were produced by more than ten participants each, but these accounted for well over half of all realizations at each time point. Further, one form in particular clearly predominated ([ɔɪn]). There were then a large number of additional realizations (sixty-two in total across both time points), each produced by very few participants, in many cases only one.

Again, a number of realizations may (speculatively) be interpreted as resulting from English-based decoding. This is true of the most frequent realization, [ɔɪn], where <oi> appears to have functioned as a digraph (as in English *ointment*). The fourth most frequently attested realization, [əʊɪn], is also consistent with English GPC, although in this case the vowels function as separate graphemes: <o> as in *go* and <i> as in *in*; several other attested forms can be interpreted along similar lines (e.g. [əʊʔɪn], [əʊən], [oʊɪn], [oʊən]).

By contrast, some realizations are consistent with French GPC. This is most clearly true in the case of the correct realization [wæ̃], but perhaps also of various 'acceptable' forms as well (e.g. the nasal vowels [wɒ̃] and [wɑ̃]; various 'unpacked' realizations of the nasal vowels such as [wan], [wɑn] and [wɒn]).

However, the overall range of sixty-two phonological forms produced for <oin> clearly surpasses what can readily be explained on the basis of either French or English GPC. This may again reflect participants' conscious efforts to pronounce the grapheme in a way which 'sounds French' – or at least, does not 'sound English'. Nonetheless, various patterns were identified, some of which could be interpreted as containing at least one target-like component. For example, roughly one fifth of attested forms began with [w], sharing the

initial phoneme of the correct form /wæ̃/. Second, twelve attested forms, including two of those produced most frequently at both time points ([ɒ̃], [ɔ̃]), correctly included a nasal vowel, even though the vowels themselves were usually incorrect. Third, although most of the attested realizations were monosyllabic, seventeen were bisyllabic; as argued above, this may reflect the incorrect processing of <o> and <i> as separate vowel graphemes rather than as part of the trigraph <oin>.

Statistical analyses based on the most frequently attested realizations of <oin> in *goinfrez* found no significant differences (a) within each group, between the realizations produced at t1 and t2 or (b) within each time point, between the three groups. However, some of the standardized residuals exceeded the threshold of ±1.96 associated with statistical significance. A descriptive analysis subsequently found tentative evidence for a greater degree of positive development in participants' decoding of <oin> in the two intervention groups compared to the comparison group, in that they showed (a) greater movement away from English-based realizations ([ɔɪn], [əʊɪn]); (b) increased frequency of the correct form [wæ̃]; and (c) greater movement towards the nasal forms [ɔ̃] and [ɒ̃]. These last two sounds, whilst not associated with positive scores on the RAT, are arguably target-like at least with regard to their nasality, a feature of the French phonemic inventory absent in English. By contrast, the 'acceptable' realization [wɒ̃] formed an exception to this general pattern, since its frequency increased much more sharply in the comparison group than in the intervention groups.

To allow a statistical analysis of changes amongst the whole set of realizations of <oin> (rather than simply amongst the most frequently attested ones), the groups' output at each time point was also analysed in terms of the three broad categories identified above, each

defined by the presence or absence of one component shared with the 'correct' form /wæ̃/: (a) initial [w]; (b) nasality; (c) monosyllabicity. (Any given pronunciation was only included in one of these analyses to avoid distorting the results, although in fact this was not found to affect the outcomes of the tests). Significant increases were found in intervention group A, but not the other two groups, in terms of the number of realizations which included nasal vowels and which had only one, rather than two syllables; again, these 'positive' changes would not have been reflected in RAT scores, except insofar as they related to increases in the specific 'correct' or 'acceptable' forms. Intervention group B was again descriptively observed to pattern with intervention group A in terms of these changes.

### 7.2.4 Summary

For all three grapheme tokens subjected to detailed analysis, a general trend appeared to be observed whereby intervention participants showed greater progress away from what could plausibly be interpreted as L1-based realizations and towards forms which were arguably more target-like in some way. In many cases, the observed changes in decoding outcomes would not have been reflected in RAT scores, suggesting that the discussion of research Question 1 may have underestimated the extent of the intervention groups' progress in decoding French. This in turn suggests that learners' progress in French decoding may follow a complex trajectory, recalling McLaughlin's (1990) U-shaped model of L2 proficiency.

## 7.3 Research Question 4

*Do the programmes of instruction lead to changes in learners' strategic reasoning when decoding French words?*

This question was addressed through task-concurrent self-report interviews (SRIs) conducted with fifteen participants, comprising roughly equal numbers of teacher-nominated high- and low-attainers in French. SRIs – though unable to access participants' subconscious mental processes – were nonetheless argued to offer valid (if incomplete) insight into their conscious reasoning. This conscious reasoning was, in turn, argued to offer participants a way of overcoming automatic, L1-based processing, which might otherwise become further entrenched even through exposure to L2 input (Ellis, 2008). Findings were analysed under three broad headings, derived from Woore (2010): (a) the strategies participants used; (b) the knowledge bases they drew upon; and (c) metacognitive comments, such as evaluations of their own decoding processes or outcomes.

It is acknowledged that the approaches taken by participants to decoding may have been determined at least partly by the characteristics of the particular words they were decoding (for example, by the extent to which these words happened to trigger associations with familiar L1 or L2 words). Since the words used in the SRIs differed between t1 and t2 (in order to avoid a practice effect), this must be borne in mind when comparing participants' reasoning between time points. Further, the small number of SRI participants means that findings cannot necessarily be considered representative of the wider sample.

The programmes of instruction were intended to strengthen participants' knowledge of L2-specific GPC, in turn providing a more secure basis for conscious reasoning. Programme A (the 'poem approach') was further intended to develop participants' use of the analogy strategy cluster (Woore, 2007, 2010), by (a) helping them to build a bank of securely-known 'source' words to which unfamiliar words could be compared and (b) providing practice in using the strategy. (Limitations in the model of strategy instruction used have already been highlighted: section 7.1.4). However, changes in participants' strategic reasoning would not necessarily be expected to translate immediately into improved RAT scores: although there was no formal time limit for the RAT, participants nonetheless tended to complete it relatively quickly (taking, on average, about three and a half minutes to read fifty-three items at both time points); this may have provided insufficient time for conscious reasoning to operate, at least as extensively as it did in the SRIs. Therefore, it was anticipated that Research Question 4 might detect changes in participants' strategic reasoning which would only later be reflected in improved RAT scores, once participants had had sufficient practice for any improved decoding mechanisms to become more rapid.

However, the answer to Research Question 4 was broadly negative. With one exception (discussed below), little difference was found between t1 and t2 in the strategies, knowledge bases or metacognitive comments reported by participants, either in the comparison or intervention conditions. At both time points, most participants' reasoning was very similar to that found in Woore (2010), as discussed below.

### 7.3.1 Strategies

The most frequent strategies used by participants were (a) breaking words up into smaller orthographic units, to make decoding more manageable; (b) making analogies to known words; and (c) trying out possible pronunciations aloud in order to evaluate them. However, although these strategies all have the potential to support participants' decoding of unknown words, various problems undermined their successful operation.

First, dividing words into smaller units can lead to correct outcomes only insofar as the resulting units are then decoded according to French GPC. In fact, however, their pronunciations were usually consistent with English GPC. Further, this strategy sometimes militated against accurate decoding by incorrectly segmenting the letter string according to L1 rather than L2 norms (again interpretable as the result of automatic, L1-entrenched processing). Thus, pairs of letters which often form single digraphs in French but not in English (e.g. <gn>) were often fragmented into two separate graphemes, virtually guaranteeing an incorrect pronunciation. Some vowel digraphs were also incorrectly segmented even though they are also digraphs in the L1 (e.g. <au>), recalling the finding from chapter 5 that <ou> and <oin> were sometimes incorrectly pronounced bisyllabically. Improved knowledge of the French grapheme inventory might have been expected to lessen this problem. Whilst the segmentation of letters into graphemes was implicitly covered in the programmes of instruction (since it is impossible to teach French GPC without first defining what the graphemes are), it may have been beneficial to place more emphasis on raising participants' awareness of differences between the L1 and L2 grapheme inventories.

Second, the analogy strategy (which was applied in particular to orthographic units which participants perceived as problematic) was occasionally used successfully, echoing findings from early literacy acquisition in L1 English (Goswami, 1995; Moustafa, 1995; Gaskins et al., 1988). However, in the current study, participants often referred to source words which were English rather than French, inevitably leading to incorrect pronunciations, apparently without awareness that this might be problematic. There were also several occasions when participants (a) tried to identify French source words, but were unable to do so; or (b) having identified an appropriate French source, incorrectly recalled its spelling and/or pronunciation.

Further, there was no evidence that the 'poem' approach had led participants to apply the analogy strategy more often, or that it had helped them develop a more secure bank of memorized source words to which they could refer. Although the case study participant from intervention group A (LV) did try to refer to source words from the poem, her recollection of them was confused and inadequate. Thus, despite the additional time devoted to memorizing the poem in the current study, the same problem arose as in Woore (2007), where participants' use of the analogy strategy was undermined by poor knowledge of the sounds and spellings of the poems. Perhaps a shorter poem would have been more easily memorized, although this would have had implications for the number of GPC it was able to exemplify. As argued in section 7.1.4, a more systematic, 'staged' approach to strategy instruction may also have been more effective in changing participants' strategic behaviour.

Third, the trying out of possible pronunciations was almost always undermined by the lack of a secure basis for evaluating them. There was an overwhelming reliance on an instinctive 'feeling' for what 'sounded right'. Although this occasionally succeeded in identifying the correct pronunciation, it more often led to the judgment that all the various pronunciations

under consideration sounded somehow ‘wrong’, without being able to indicate what the correct form should be. Indeed, the attempted pronunciations sometimes came full circle, with participants finally settling on the same pronunciation as the one they had originally produced and rejected. This strategy can therefore be counter-productive in that it prevents participants using other, potentially more effective approaches to decoding.

Often, participants’ evaluations of the various pronunciations they produced were also restricted to specific graphemes which they perceived as problematic. The remaining graphemes were generally ignored, even though they were often realized according to English GPC.

Given the prevalence of the ‘trying out pronunciations’ strategy both in the current study and in Woore (2010), it may be helpful for teachers to alert students to the likely limitations of relying on an instinctive feeling for what ‘sounds French’.

### 7.3.2 Knowledge bases

In terms of the knowledge bases participants drew upon, there were frequent appeals to conscious knowledge of how particular graphemes (or larger orthographic units) should be pronounced. The origins of this knowledge are often unstated; however, the SRIs contain at least some examples where (a) explicit instruction appears to have successfully improved participants’ conscious knowledge of French GPC and (b) this has in turn led to improved decoding outcomes. Conversely, there are also instances where participants’ recollection of explicit instruction is vague and/or defective.

In many cases, participants' comments about specific GPC involve 'silent' letters, perhaps a particularly salient feature of French orthography for English learners due to the contrast with their L1. Participants' comments about SFC were frequently correct, perhaps because this principle of French orthography can be readily expressed (and recalled) as a simple 'rule of thumb'. Participants' knowledge of the SFC principle also led to correct pronunciations on several occasions, including in some cases where this involved modifying their original pronunciations. This demonstrates that conscious reasoning can successfully override automatic L1-based processing to produce more accurate L2 outcomes (Ellis, 2008).

Diacritics (usually referred to by participants using pseudo-metalanguage such as "the little thingy above the E") were the other recurring feature in participants' explicit comments about French GPC, perhaps because – once they have been noticed – they demand conscious attention due to their unfamiliarity from the L1. However, although participants' knowledge of the effects of diacritics was occasionally correct, it usually was not. Participants often thought that diacritics affected the pronunciation of graphemes in some way, although they were not sure how. This could in fact be seen as the result of a logical deduction rather than of pre-existing knowledge: the "little thingy" above the letter must do *something*; otherwise, why would it be there?

Finally, two participants (both in intervention groups) appealed to their knowledge of French alphabetic letter names, which they had previously learnt in class. Even assuming this knowledge is secure and accurate (which it was not always), problems are likely to arise when letter names are used in lieu of phonemic values. Clearly, these two things do overlap (Treiman, 2008): for example, the French name of the letter V (/ve/) contains the phoneme /v/ that it symbolizes. However, participants were observed to incorporate into the target word

not only the initial consonant of the letter name, but also its vowel, usurping the place of the vowel represented in the spelling. Appealing to letter names also seems likely to exacerbate the tendency to treat letters individually, even when they should be part of di- or trigraphs. If the evidence here were replicated more widely, teachers wishing to teach alphabetic letter-names to beginner learners might be advised to make clear the purposes and limitations of this exercise. It was disappointing that the decoding instruction did not appear to dislodge the two participants' pre-existing misconception about the role of alphabetic letter names in decoding.

### 7.3.3 Metacognition

There was very little evidence of any changes in participants' metacognitive or affective comments. At both time points, the SRIs were permeated by an overwhelming sense of uncertainty and tentativeness; doubts were frequently expressed concerning the correctness of the pronunciations produced. These doubts often centred on particular graphemes which were perceived as problematic, usually those containing diacritics or letter combinations unfamiliar from the L1. Occasionally these were explicitly described as 'odd' or 'weird' in comparison to an implicit L1 norm. Whilst this may be ironic given that English orthography is actually less phonologically transparent than French (and therefore 'odder' in these terms), it accurately reflects participants' history of processing L1 text, which has led them implicitly to 'expect' L1-based patterns (Ellis, 2008).

However, besides these particular 'problem' graphemes, other graphemes within the same words are usually realized according to English GPC, without any apparent awareness on the participants' part that this might be incorrect. Further, although most participants made at

least one positive comment expressing confidence in their L2 decoding outcomes, this confidence was almost always misplaced: the pronunciations concerned again tended to be based on English GPC or on analogies to English words.

Taken together with the preceding discussion, these findings are consistent with the view that (a) L2 input triggers automatic, L1-based processing, especially where the L2 orthographic patterns are similar to those of the L1; (b) participants may be unaware that this results in incorrect L2 pronunciations; (c) when participants encounter 'unexpected' orthographic features – those which are unfamiliar from the L1 and of which they have little previous experience of processing – automatic L1-based processing is more likely to be disrupted, bringing these features into conscious attention where they can be treated as a problem to be solved through reasoning. It is therefore possible that learners would benefit from a greater degree of explicit awareness-raising that even L1-resembling letter strings are often pronounced differently in the L2. This may widen the sphere of influence of their conscious reasoning.

### 7.3.4 RAT scores of SRI participants

Just as there was little or no evidence of positive development in most participants' reasoning between t1 and t2, there was also little change in their RAT scores: most showed increases around or below the mean increases for their group. The only exception in the intervention condition was LV, who realized 26 additional grapheme tokens 'acceptably' at t2 than at t1. This compared to a mean increase of 11 for intervention group A as a whole. LV was also the only participant to show considerable positive changes in her reasoning. This *may* lie behind her increased RAT score, although other factors may also be involved, such as

developments in her implicit knowledge. Conversely, three comparison participants in the SRI sub-sample showed RAT score increases three to five times higher than the mean increase for their group (which was three additional 'acceptable' realizations), although in one case this may be due to the participant's exceptionally low t1 score. Since there is no evidence for any corresponding improvement in these participants' reasoning, their score gains seem more likely to result from improved implicit knowledge. To investigate further the relationship between conscious reasoning and decoding outcomes, it may be helpful for future research to analyse the length of time taken by participants to complete the RAT in relation to the scores they achieve on it.

# Chapter 8: Conclusions

## 8.1 The problem

There are strong grounds to believe that decoding proficiency – the ability to convert unfamiliar written forms into accurate phonological representations of them – is a foundation skill which facilitates many other aspects of L2 learning. Besides playing an important role in reading comprehension (the traditional home of decoding research), a strong case can be made for its contribution to vocabulary learning. Generating accurate pronunciations for words which are not already part of the learner's 'sight vocabulary' is crucial in developing independence from the teacher (or other spoken models).

At the same time, evidence suggests that beginner learners in English MFL classrooms generally have poor decoding proficiency and make poor progress in this aspect of their learning. This may be partly attributable to what Pachler (2002:6) calls the "highly challenging systemic pressures" faced by MFL learners, such as the low status of L2 learning in wider society and the limited curriculum time devoted to it. Moreover, reading (and decoding more specifically) appears to have been relatively neglected in MFL classrooms in recent years.

Over the last decade, official guidance has advocated greater explicit focus on developing learners' knowledge of L2 GPC, perhaps influenced by a renewed focus on synthetic phonics in the teaching of L1 English literacy. However, the assessment criteria – 'Attainment Targets' – prescribed by the *National Curriculum* have remained largely unchanged from

earlier versions, placing little emphasis on decoding and offering no clear pathway for its development. Given the importance of measuring pupils' performance in the current educational climate, it seems reasonable to ask whether the needs of the assessment criteria have overshadowed the recommendations concerning the teaching of decoding. Interesting in this respect is the very existence of a viable comparison group in the current study, apparently receiving no explicit instruction in French GPC despite official guidance.

This thesis has argued, however, that learners' difficulties with French decoding cannot be wholly ascribed to 'external' factors such as these: rather, they may also derive from the nature of the L2 writing system and its relationship to that of the L1. Writing systems such as English and French have often been seen as "convenient" pairings for L2 learners, because they have much in common: for example, both are alphabetic systems using Roman script and sharing various GPC. These commonalities may indeed provide many opportunities for 'positive transfer', interpretable as the triggering of L1-based processing mechanisms by L2 input and resulting in correct (or at least, 'acceptable') outcomes.

However, there are also many cases in which L1-based processing of L2 input generates incorrect pronunciations. This is evident in several features of the data gathered in the current study. First, by far the most frequent realizations of French graphemes were consistent with English GPC, even where these differed from the French ones. This was particularly true of vowels, the area where French and English GPC differ most markedly. Second, participants often segmented French words according to English norms, resulting in French graphemes being incorrectly subdivided; this almost inevitably led to incorrect pronunciations. Third, there is some initial indication that participants' pronunciations of a given vowel grapheme varied according to its orthographic context, as occurs very frequently

in English but less so in French. This is consistent with the view that participants have developed processing mechanisms at larger grain sizes than individual graphemes in order to cope with the considerable inconsistency of English GPC. However, when this large grain size of processing is transferred to the 'shallower' French orthography – in which more graphemes are pronounced identically irrespective of their context – incorrect pronunciations arise.

In terms of L2 decoding, then, it is an oversimplification to view English and French as a "convenient" pairing. The considerable overlap between the two writing systems is facilitative in some respects, but may also encourage the incorrect triggering of L1-based processing in cases of 'misleading similarity': that is, where letters or letter strings are visually similar to those of the L1, yet have different phonological realizations. Indeed, the presence of many cognates in French and English – helpful from the point of view of lexical access – may be particularly disruptive in this respect. Beginner-decoders of French, already literate in L1 English, therefore face different challenges compared to L1 readers of typologically more distant writing systems such as Chinese. However, the scale of their learning problem may be underestimated – by both learners and teachers – due to the apparent familiarity of the orthography. This may explain the relative neglect of decoding which has been observed in English MFL classrooms, in which the L2 is usually European and uses the Roman alphabet.

## 8.2 An attempted solution

Despite the emphasis on decoding in recent official guidance, there is a lack of convincing evidence concerning the effectiveness of explicit L2 decoding instruction both in MFL

classrooms and more widely. Responding to this issue, the current study evaluated two programmes of instruction designed to strengthen participants' knowledge of French GPC: Programme A (the 'poem' approach) and Programme B (the 'phonics' approach). The programmes aimed to provide a secure basis for participants' conscious reasoning, which may offer a way of interrupting and overriding automatized, L1-based processing. Of course, the ultimate goal must be for learners to develop automaticity in L2 decoding; however, accuracy is a necessary precursor to automatization and conscious reasoning may therefore provide a first step along this road.

The two programmes were delivered by class teachers in approximately thirty ten- to fifteen-minute segments, embedded in lessons spanning around nineteen weeks. In both programmes, each segment included (a) the presentation of one or two (or occasionally more) GPC, focussing on those graphemes whose realizations differ in French and English; (b) opportunities for learners to practise decoding unfamiliar French words containing these (and other) graphemes; (c) form-focussed feedback from the teacher. Both programmes therefore contrast with the model of progression enshrined in the *National Curriculum* Level Descriptors (QCA, 2007b), where beginner-learners' reading is initially expected to be restricted to familiar words. Programme A additionally aimed to develop participants' use of an analogy strategy by (a) helping them build a bank of securely-known 'source words' to which unfamiliar words could be compared and (b) providing practice in using the strategy. The strategy involves using the 'source words' to derive the pronunciations of unfamiliar words with similar spellings. However, there were limitations in the design of the strategy instruction component of Programme A, meaning that the two programmes of instruction were arguably more similar than different.

***Did the programmes of instruction lead to more accurate decoding of L2 French?***

Participants who followed the programmes of instruction were found to make significantly more progress in French decoding (at least in the short term) than those in the comparison group, as measured by increases in the number of graphemes pronounced 'correctly' and 'acceptably' on a decoding test. As in previous studies of the same population, participants receiving no explicit decoding instruction made little or no such progress. These findings may be considered relatively robust, because the achieved sample was large and was drawn from ten different classes in four different schools (although the possibile influence of school and teacher effects cannot be ruled out). The sample schools were arguably not untypical of the wider population of English secondary comprehensive schools in non-inner city areas.

However, the magnitude of the significant difference between the intervention and comparison groups (around eight grapheme tokens realized 'acceptably') may appear small in relation to the amount of time invested in the instruction. On the other hand, it is difficult to know how much more could be achieved in the very limited curriculum time available. Further, the total duration of the decoding instruction was considerably less than in some other studies evaluating similar programmes, both in L1 and L2. Ultimately, further research is needed to investigate the impact (if any) of a given increase in decoding proficiency on other L2 learning outcomes, such as vocabulary learning. Data gathered as part of the current study (but not written up here for space reasons) may provide an initial basis for investigating the relationship between these variables, as called for by Hamada and Koda (2008) in what they herald as "a new line of inquiry" in L2 writing system research.

Within the mean gain scores obtained by the different groups, there was also considerable variation, both between individual participants and between the individual grapheme types involved. First, the increases in 'acceptable' realizations shown by intervention participants appeared more consistent in respect of consonants than vowels. There was also tentative evidence that those consonantal graphemes which were more difficult for learners to decode in the first place (because their realizations were more dissimilar to English) were also those which more stubbornly resisted the effects of explicit instruction. These findings suggest that explicit decoding instruction may be more effective (or at least, more rapid) in respect of some graphemes than others.

Second, some individuals within each group showed much larger overall score increases than others. Further research should explore the origins of these individual differences in the effectiveness of the decoding instruction. Further, some participants in each class (even in the intervention group) actually showed a *decrease* in their mean score: in other words, they pronounced *fewer* graphemes 'acceptably' at t2 than at t1. This may at least partly reflect a destabilization of their decoding mechanisms, leading to the conscious avoidance of L1-based pronunciations, in some cases without knowing what the correct L2 pronunciation should be. The resulting decrease in decoding accuracy is consistent with U-shaped models of developing L2 proficiency.

***Did the programmes of instruction lead to any other changes in participants' realizations of French graphemes?***

Analysis of participants' actual realizations of individual grapheme tokens on the decoding test suggested that both of the intervention groups tended to show movement (a) away from

(putatively) L1-based pronunciations and (b) towards pronunciations which were interpretable as more target-like in some ways. By contrast, comparison participants tended to show little change, or different patterns of change. However, many of these changes in intervention participants' realizations were not reflected in increased decoding test scores, since they did not result in 'acceptable' realizations. This underlines the fact that acquiring L2 decoding proficiency may follow a longer and more complex trajectory than simply progressing directly from 'incorrect' to 'correct' pronunciations (again as in the U-shaped curve model). Contrastingly, official guidance for MFL teachers apparently expects learners to know 'regular' GPC within a year and 'exceptions' within two.

Of course, learners' rates of progress in L2 decoding may also depend on the specific L1 and L2 writing systems involved. Various other Roman alphabetic languages, having more consistent GPC than French, may permit more rapid development of L2 decoding proficiency, as is true of beginner *L1* readers of these languages. At the same time, given the increasingly multilingual populations of primary and secondary classrooms in many parts of the UK (NALDIC, 2011a, 2011b), it cannot be assumed that MFL learners will necessarily share the same L1, as they did in the current study. Learners may be literate in additional writing systems with more or less facilitative relationships to the L2. Clearly, it is a challenge for classroom teachers to take practical account of their learners' possibly diverse prior knowledge, helping them build bridges between their L1 and L2. However, the pupil-pupil pairwork in the current study, where learners discussed with each other how to pronounce unfamiliar L2 words before receiving feedback from their teacher, seems to offer a possible pedagogical approach to this issue.

In terms of the nature of participants' errors, analysis of their output on the decoding tests revealed that – whilst the predominant realizations of each grapheme were consistent with English GPC at both time points – a diverse range of other pronunciations were also attested, surpassing what could be explained on the basis of L1-to-L2 transfer. It was hypothesized that, in at least some cases, they may have resulted from conscious reasoning. For example, if (as indicated in previous research) participants sought consciously to make their pronunciations 'different from English', without however knowing what the correct French forms should be, a wide range of possible realizations may result. Further research is now needed to explore the origins of learners' decoding errors in more detail. In particular, the current study analysed participants' output only in an 'atomized' way, that is, at the level of individual GPC. Examining their pronunciations of words as wholes may therefore provide further insight into the processing mechanisms involved, particularly in the case of deviations from the grapheme sequence and word substitution errors. These cases were excluded from the analysis in the current study.

***Did the programmes of instruction lead to changes in learners' strategic reasoning when decoding French words?***

The self-report interviews revealed little or no evidence of improved strategic reasoning amongst most participants, either in the intervention or comparison groups, although it should be acknowledged that the few participants involved may not be representative of the wider sample (still less of KS3 learners generally). Overall, at both time points, participants reported a similar range of error-prone approaches to those identified in previous exploratory research with the same population. It is clear that conscious reasoning *can* help learners overcome automatic, L1-based decoding mechanisms, at least when they really focus on the

task; however, this reasoning will result in correct pronunciations only insofar as it is based upon secure knowledge of L2 GPC or of the language more broadly. For most participants in the self-report sub-sample, the programmes of instruction did not appear to generate this kind of knowledge.

There was one exception to this pattern. One participant who had followed the 'poem' programme did show improved strategic reasoning, based on strengthened knowledge of French GPC. She also showed a well above-average increase in the number of graphemes pronounced correctly in the decoding test. However, it is not possible to claim definitively that these two outcomes were causally related. The findings from the self-report interviews are therefore inconclusive; further investigation is required to identify how learners deploy strategies in order to decode French words and how they might be helped to do so more effectively.

***Was one of the programmes of instruction more effective than the other?***

None of the three strands of analysis suggested convincingly that one of the two programmes of decoding instruction had been more effective than the other. First, no significant differences were found between the increases in numbers of graphemes realized 'correctly' or 'acceptably' by participants following one or the other programme. Second, although the statistical analysis of participants' grapheme realizations possibly indicated a greater degree of t1-to-t2 change amongst 'poem' than 'phonics' participants, a subsequent descriptive analysis suggested that the two intervention groups in fact showed broadly similar patterns of change. Finally, as noted above, the self-report interviews provided no convincing evidence of any positive developments in participants' strategic reasoning, save in the case of one

participant from the 'poem' group. However, even she showed no improvement in her use of the analogy strategy, which was undermined by poor knowledge of the source words she attempted to use, despite this being an explicit aim of the programme.

These findings may reflect the fact that the two programmes could be considered broadly similar: as noted above, both involved the presentation to learners of individual GPC, followed by opportunities for them to decode unfamiliar French words and to receive feedback on this task. The additional, strategy-based component of the Programme A may have received insufficient emphasis to make a difference. A more systematic programme of strategy instruction – adopting a staged approach with progressive removal of scaffolding, as proposed by Macaro (2001) – may have been more effective. A shorter poem may also have been easier for learners to memorize, although this would have had implications for the number of graphemes it could exemplify.

## 8.3 Future directions

The current study sought to produce findings which teachers would perceive as directly relevant to their own practice in MFL classrooms – a context which has been underrepresented in the SLA literature, perhaps contributing to a sense of disjuncture between research and practice. The programmes of instruction were therefore implemented in 'real' classrooms as part of timetabled MFL provision. Whilst this entailed some compromises in research design, it nonetheless demonstrates that systematic decoding instruction can be successfully accommodated in the MFL curriculum, as recommended in official guidance. Further, teachers reported no detriment to participants' progress in other

aspects of their L2 learning, although it is acknowledged that this could have been investigated more systematically.

The significant but small advantage for participants who received explicit decoding instruction relates only to the specific programmes of instruction used in the current study, not to decoding instruction more generally: in other words, it is possible that alternative programmes may be more effective (or indeed less so). This requires further exploration. However, given the findings of the current study, especially concerning the possibly complex trajectory of learners' progress in L2 decoding, it may be speculated that attempting to provide systematic coverage of 'regular' French GPC within one academic year is an ambitious goal within existing levels of lesson provision, despite the stated expectations of the *Key Stage 3 Framework*.

Additional evidence presented in this thesis may be useful in informing the design of future programmes of instruction. First, it may be helpful to pay particular attention to raising learners' awareness of differences between the L1 and L2 grapheme inventories, since segmenting the letter-string correctly into L2 graphemes is a prerequisite for accurate decoding. Second, learners need to be aware that it is not only those graphemes which are visually unfamiliar from the L1 which require their conscious attention: rather, even L2 letter strings which resemble those in L1 words may be pronounced differently. Third, the teaching of alphabetic letter names in the L2 – often a staple component of beginner MFL curricula – is potentially confusing for learners, who may form a misconception that the letter names (rather than their phonemic values) form the basis for sounding out L2 words. Therefore, if the L2 alphabet is to be taught, it is recommended that teachers make clear to learners the purpose and limitations of this exercise.

Given the emphasis placed on GPC knowledge in the *Key Stage 2 Framework*, and in view of the recent expansion of MFL teaching in primary schools, the effectiveness of explicit L2 decoding instruction should also be investigated with this younger age group. From the point of view of learners' progress in L2 learning at secondary school, it would clearly be advantageous for them to reach Y7 already proficient in L2 decoding. However, there appears to be some way to travel before this ambition becomes reality.

Finally, in terms of developing practical tools for assessing pupils' progress in decoding, it is suggested that the Reading Aloud Test – used in the current study as a research instrument – could also be adapted for pedagogical purposes, since it provides the basis for formative feedback in terms of (a) the extent to which participants are decoding words on a linear, grapheme-by-grapheme basis, thus indicating whether they need help with this basic principle of decoding; and (b) the accuracy with which they have pronounced individual graphemes, highlighting those GPC they should focus on improving. By contrast, existing mark schemes for L2 pronunciation used in the MFL context tend to rely on holistic judgments and are therefore less helpful in formative terms. In future, the development of appropriate software may also allow the computerized assessment of learners' decoding outcomes, again with personalized feedback being provided at the individual grapheme level.

# References


Adair, J. G. (2004) 'Hawthorne Effect', in M. S. Lewis-Beck, A. Bryman and T. Futing-Liao (eds.) 2004, *The SAGE Encyclopedia of Social Science Research Methods*. Thousand Oaks, CA: SAGE Publications.

Adair, J. G., Sharpe, D. and Huynh, C. (1989) 'Hawthorne Control Procedures in Educational Experiments: A Reconsideration of their Use and Effectiveness', in *Review of Educational Research*, vol. 59, no. 2, pp. 215-228.

Akamatsu, N. (1999) 'The Effects of First Language Orthographic Features on Word Recognition Processing in English as a Second Language', in *Reading and Writing: An Interdisciplinary Journal*, vol. 11, pp. 381–403.

Akamatsu, N. (2002) 'A Similarity in Word-Recognition Procedures Among Second-Language Readers with Different First Language Backgrounds', in *Applied Psycholinguistics*, vol. 23, pp. 17-133.

Akamatsu, N. (2005) 'Effects of Second Language Reading Proficiency and First Language Orthography on Second Language Word Recognition', in V. Cook & B. Bassetti (eds.) 2005, *Second Language Writing Systems*. Clevedon: Multilingual Matters, pp. 238-259.

Alderson, J. C. (1984) 'Reading in a foreign language: a reading problem or a language problem?', in J. C. Alderson and A. H. Urquhart (eds.) 1984, *Reading in a Foreign Language*. London: Longman, pp. 1-24.

Allen, P., Fröhlich, M. and Spada, N. (1984) 'The Communicative Orientation of Language Teaching: An Observation Scheme', in J. Handscombe, R. Orem and B. Taylor (eds.) (1984) *On TESOL '83. The question of control*. Washington, DC: TESOL, pp. 231-252.

AQA (Assessment and Qualifications Alliance) (2008) *GCSE Specification: French*. [Online]. Available from <http://www.aqa.org.uk> [Accessed 23.10.2011].

Arteaga, D., Herschensohn, J., Gess, R. (2003) 'Focusing on Phonology to Teach Morphological Form in French', in *The Modern Language Journal*, vol. 87, no. 1, pp. 58-70.

Atkins, B. T., Duval, A., Milne, R. C., Cousin, P.-H., Lewis, H. M. A., Sinclair, L. A., Birks, R. O., Lamy, M.-N. (1998) *Collins-Robert French-English, English-French Dictionary: Unabridged*. Fifth edition. Glasgow: HarperCollins.

Atkinson, E. (2000). 'In Defence of Ideas, Or Why 'What Works' Is Not Enough', in *British Journal of Sociology of Education*, vol. 21, no. 3, pp. 307-330.

Baddeley, A. D., Eldridge, M. and Lewis, V. (1981) 'The role of subvocalisation in reading', in *The Quarterly Journal of Experimental Psychology*, Section A, vol. 33, no. 4, pp. 439-454.

Baddeley, A. D., Gathercole, S. and Papagno, C. (1998) 'The Phonological Loop as a Language Learning Device', in *Psychological Review*, vol. 105, no. 1, pp. 158-173.

Baddeley, A. D. and Hitch, G. J. (1974) 'Working memory', in G.A. Bower (ed.) 1974, *Recent Advances in Learning and Motivation*, vol. 8. New York: Academic Press, pp. 47-89.

Baddeley, A. D. and Logie, R. H. (1999) 'Working Memory: The Multiple Component Model', in A. Miyake and P. Shah (eds.) 1999, *Models of Working Memory: Mechanisms of Active Maintenance and Executive Control*. Cambridge: Cambridge University Press, pp. 28-61.

Baluch, B. and Besner, D. (1991) 'Visual Word Recognition: Evidence for Strategic Control of Lexical and Nonlexical Routines in Oral Reading', in *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 17, no. 4, pp. 644-652.

Barbéry, S. (2006) *Dico des Rimes*. [Online]. Available from <https://www.barbery.net/lebarbery/index.htm> [Accessed 30.05.2006].

Bee, H. (2000) *The Developing Child.* London: Allyn and Bacon.

BERA (*British Educational Research Association)* (2004). *Revised Ethical Guidelines for Educational Research*.

Berndt, R. S., Reggia, J. A. and Mitchum, C. C. (1987) 'Empirically derived probabilities for grapheme-to-phoneme correspondences in English', in *Behavior, Research Methods, Instruments and Computers*, vol. 19, no. 1, pp. 1-9.

Bernhardt, E. B. (2000) 'Second-Language Reading as a Case Study of Reading Scholarship in the 20th Century', in M. Kamil, P. B. Mosenthal, P. D. Pearson and R. Barr (eds.) 2000, *Handbook of Reading Research*, vol. 3. London: Lawrence Erlbaum Associates, pp. 791-811.

Birdsong, D. (1999) ‘Introduction: Whys and Why Nots of the Critical Period Hypothesis for Second Language Acquisition’, in D. Birdsong (ed.) 1999, *Second Language Acquisition and the Critical Period Hypothesis*. Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 1-22.

Black, P. J. and Jones, J. (2006) ‘Formative Assessment and the Learning and Teaching of MFL: Sharing the Language Learning Road Map with the Learners’, in *Language Learning Journal*, vol. 34, no. 1, pp. 4-9.

Black, P. J. and Wiliam, D. (1998) ‘Inside the Black Box: Raising Standards Through Classroom Assessment’, in *PhiDelta Kappan*, vol. 80, no. 2, pp. 139-148. [Online]. Available from <https://www.pdkintl.org/kappan/kbla9810.htm> [Accessed 16.08.2010].

Bongaerts, T. (1999) ‘Ultimate attainment in L2 pronunciation: The Case of Very Advanced Late L2 Learners’, in D. Birdsong (ed.) 1999, *Second Language Acquisition and the Critical Period Hypothesis*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 133-159.

Bowles, M. A. and Leow, R. P. (2005) ‘Reactivity and Type of Verbal Report in SLA Research Methodology: Expanding the Scope of Investigation’, in *Studies in Second Language Acquisition*, vol. 27, pp. 415-440.

Brook, D. (1999) *Approaches to Teaching Phonics: A Resources Booklet for Students in Initial Teacher Education*. Derby: University of Derby.

Brown, G. D. A. and Deavers, R. P. (1999) ‘Units of Analysis in Nonword Reading: Evidence from Children and Adults’, in *Journal of Experimental Child Psychology*, vol. 73, pp. 208–242.

Brown, T. L. and Haynes, M. (1985) ‘Literacy Background and Reading Development in a Second Language’, in T. H. Carr (ed.) 1985, *The Development of Reading Skills: New Directions for Child Development*, no. 27. San Francisco: Jossey Bass.

Bruck, M., Genesee, F. and Caravolas, M. (1997) ‘A Cross-Linguistic Study of Early Literacy Acquisition’, in B. A. Blachman (ed.) 1997, *Foundations of Reading Acquisition and Dyslexia: Implications for Early Intervention*. Mahwah, New Jersey: Lawrence Erlbaum Associates, pp. 145-162.

Bryant, P. (1993) ‘Phonological Aspects of Learning to Read’, in R. Beard (ed.) 1993, *Teaching Literacy, Balancing Perspectives*. London : Hodder & Stoughton, pp. 83-94.

Butler, R, (1988) ‘Enhancing and Undermining Intrinsic Motivations: The Effects of Task-Involving and Ego-Involving Evaluation on Interest and Performance’, in *British Journal of Educational Psychology*, vol. 58, pp. 1-14.

Byrne, B. and Fielding-Barnsley, R. (1991) ‘Evaluation of a Program to Teach Phonemic Awareness to Young Children’, in *Journal of Educational Psychology*, vol. 83, no. 4, pp. 451-455.

Cable, C., Driscoll, P., Mitchell, R., Sing, S., Cremin, T., Earl, J., Eyres, I., Holmes, B., Martin, C. and Heins, B. (2010). *Languages Learning at Key Stage 2: A Longitudinal Study. Final Report*. Research commissioned by the Department for Children, Schools and Families. [Online]. Available from <www.education.gov.uk/publications> [Accessed 05.01.2012].

Campbell, D. & Stanley, J. (1963) *Experimental and Quasi-Experimental Designs for Research*. Chicago, IL: Rand-McNally.

Carpenter, C. (2003) *Teach Yourself Beginner’s French*. London: Hodder Education.

Carrell, P. L. (1991) ‘Second Language Reading: Reading Ability or Language Proficiency?’, in *Applied Linguistics*, vol. 12, no. 2, pp. 159-179.

Chambers, G. (1993) ‘Taking the ‘De’ out of Demotivation’, in *Language Learning Journal*, vol. 7, pp. 13-16.

Chamot, A. U. and El-Dinary, P. B. (1999) ‘Children's Learning Strategies in Language Immersion Classrooms’, in *Modern Language Journal*, vol. 83, no. 3, pp. 319-338.

CILT (Centre for Information on Language Teaching) (2010). *Language Trends 2010 Secondary*. [Online]. Accessed 29.11.2006 at www.cilt.org.uk/home/research_and_statistics/language_trends_surveys.

Coady, J. (1997) ‘L2 Vocabulary Acquisition Through Extensive Reading’, in James Coady and Thomas Huckin (eds.) 1997, *Second Language Vocabulary Acquisition*. Cambridge: Cambridge University Press, pp. 225-237.

Cohen, A. D. (1998) *Strategies in learning and using a second language*. London: Longman.

Cohen, J. (1960) ‘A Coefficient of Agreement for Nominal Scales’, in *Educational and Psychological Measurement*, vol. 20, no. 1, pp. 37–46.

Coleman, J. A. (2009). 'Why the British Do Not Learn Languages: Myths and Motivation in the United Kingdom', in *Language Learning Journal*, vol. 37, no. 1, pp. 111-127.

Coleman, J. A., Galaczi, Á. and Astruc, L. (2007) 'Motivation of UK school pupils towards foreign languages: a large-scale survey at key Stage 3', in language Learning Journal, vol. 35, no. 2, pp. 245-281.

Coltheart, M. (1978) 'Lexical Access in Simple Reading Tasks', in G. Underwood (ed.) 1978, *Strategies of Information Processing*. London: Academic Press, pp. 151-216.

Coltheart, C. (2005) 'Modeling Reading: The Dual-Route Approach', in M. J. Snowling and C. Hulme (eds.) 2005, *The Science of Reading. A Handbook*. Oxford: Blackwell, pp. 6-23.

Coltheart, M., Curtis, B., Atkins, P. and Haller, M. (1993) 'Models of reading aloud: dual route and parallel-distributed-processing approaches', in *Psychological Review*, vol. 100, pp. 589-608.

Concise Oxford English Dictionary (2002). Tenth Edition. Oxford: Oxford University Press.

Cook, V. and Bassetti, B. (eds.) (2005a) *Second Language Writing systems*. Clevedon: Multilingual Matters.

Cook, V. and Bassetti, B. (2005b) 'An introduction to researching Second Language Writing systems', in V. Cook & B. Bassetti (eds.) (2005) *Second Language Writing systems*. Clevedon: Multilingual Matters, pp. 1-70.

Cook, V., Vaid, J. and Bassetti, B. (2009) 'Writing Systems Research: A New Journal for a Developing Field', in *Writing System Research* vol. 1, no. 1, pp. 1-3.

Coulmas, F. (1999) *The Blackwell Encyclopedia of Writing Systems*. Oxford: Blackwell.

Courtney, L., Pignot-Shahov, V., Porter, A. and Mitchell, R. (2011), *Call for Contributions: Research Perspectives on Modern Foreign Languages in UK schools*. University of Southampton. [Online]. Accessed 15.04.11 at <http://www.soton.ac.uk> (No longer available at this address).

Coveney, A. (2001) *The Sounds of Contemporary French*. Exeter: Elm Bank Publications.

Čubrović, B. (2002) 'Assimilation of Recent French Loanwords in English: Transphonemisation of Nasal Vowels'. Paper presented at the 10th ELSNET Summer School, University of Southern Denmark, 14th-27th July, 2002. [Online]. Available from <http://www.summerschool2002.nis.sdu.dk/studentsession/cubrovic.doc> [Accessed 19.12.2009].

Daneman, M. and Carpenter, P. A. (1980) 'Individual Differences in Working Memory and Reading', in *Journal of Verbal Learning and Verbal Behavior*, vol. 19, pp. 450-466.

Daneman, M. and Carpenter, P. A. (1983) 'Individual Differences in Integrating Information Between and Within Sentences', in *Journal of Experimental Psychology: Learning, Memory, and Cognition*, vol. 9, no. 4, pp. 561-584.

D'Angiulli, A., Siegel, L. S. and Serra, E. (2001) 'The Development of Reading in English and Italian in Bilingual Children', in *Applied Psycholinguistics*, vol. 22, pp. 479-507.

David, M., Edwards, R. and Alldred, P. (2001) 'Children and School-Based Research: "Informed Consent" or "Educated Consent"?' in *British Educational Research Journal*, vol. 27, no. 3, pp. 347-365.

Day, R. R. and Bamford, J. (1998) *Extensive Reading in the Second Language Classroom*. Cambridge: Cambridge University Press.

DCC (Dorset County Council) (2009) *Secondary School Provision in Purbeck*. [Online]. Available from <http://www.dorsetforyou.com/media.jsp?mediaid=139681&filetype=pdf> [Accessed 26.01.2011].

DCSF (Department for Children, Schools and Fanilies) (2007) *Performance Tables*. [Online]. Available from <http:// www.dcsf.gov.uk/rsgateway/DB/SFR/s000900/index.shtml> [Accessed 18.03.2007].

DCSF (Department for Children, Schools and Fanilies) (2009) *Key Stage 3 Framework for Languages*. [Online]. Accessed 03.11.2010 at <http://nationalstrategies.standards.dcsf.gov.uk/mfl> (No longer available at this address).

Deary, I. J., Strand, S., Smith, P. and Fernandes, C. (2007) 'Intelligence and Educational Achievement', in *Intelligence*, vol. 35, Issue 1, pp. 13-21.

Defior, S. (2004) 'Phonological awareness and learning to read: a cross-linguistic perspective', in T. Nunes and P. Bryant (eds.) 2004, *Handbook of Children's Literacy*. London: Kluwer Academic Publishers, pp. 631-650.

DFE (Department for Education) (2010) *The Importance of Teaching: The Schools White Paper*. London: DfE.

DfEE (Department for Education and Employment) and Welsh Office Education Department (1995) *Modern Foreign Languages in the National Curriculum*. London: HMSO.

DfEE (Department for Education and Employment) (1998) *The National Literacy Strategy*. London: HMSO.

DfEE (Department for Education and Employment) (1999) *The Induction Period for Newly Qualified Teachers*. Circular 5/99. London: DfEE.

DfES (Department for Education and Skills) (2002) *Languages for all: Languages for Life*. [Online]. Accessed 16.08.2006 at <http://www.dfes.gov.uk/languages> (No longer available at this address).

DfES (Department for Education and Skills) (2003) *Key Stage 3 National Strategy. Framework for Teaching Modern Foreign Languages: Years 7, 8 and 9*. [Online]. Accessed 16.08.2006 at <http://www.standards.dcsf.gov.uk/secondary/keystage3/respub/mflframework/mflfwkdl/> (No longer available at this address).

DfES (Department for Education and Skills) (2005) *Key Stage 2 Framework for Languages*. [Online]. Available from <http://www.primarylanguages.org.uk/> [Accessed 06.08.2006].

DfES (Department for Education and Skills) (2006a) *National Statistics First Release: Schools and Pupils in England.* [Online]. Available from <http://www.education.gov.uk/rsgateway/DB/SFR/s000682/SFR38-2006.pdf> [Accessed 26.01.2011].

DfES (Department for Education and Skills) (2006b) *Primary National Strategy: The Primary Framework for Literacy and Mathematics. Core Position Papers Underpinning the Renewal of Guidance for Teaching Literacy and Mathematics: The Renewed Primary Framework.* London: DfES.

DfES (Department for Education and Skills) (2006c) *Value Added Measures*. [Online]. Accessed 24.01.2007 at <http://www.dfes.gov.uk/performancetables/schools_05/sec4.shtml> (No longer available at this address).

De Jong, P. F. (2006) 'Units and Routes of Reading in Dutch', in *Developmental Science*, vol. 9, no. 5, pp. 441-442.

Diack, H. (1965) *In Spite of the Alphabet: A Study of the Teaching of Reading*. London: Chatto and Windus.

Dobson, A. (1997) *MFL Inspected: Reflections on Inspection Findings 1996/7*. London: CILT.

Dulay, H. C. and Burt, M. K. (1974) 'Natural Sequences in Child Second Language Acquisition', in *Language Learning*, vol. 24, no. 1, pp. 37-53.

Durgunoğlu, A. Y., Nagy, W. E. and Hancin-Bhatt, B. J. (1993) 'Cross-Language Transfer of Phonological Awareness', in *Journal of Educational Psychology*, vol. 85, no. 3, pp. 453-465.

Ehri, L. C. (1995) 'Phases of Development in Learning to Read Words by Sight', in *Journal of Research in Reading*, vol. 18, no. 2, pp. 116-125.

Ehri, L. C. (2005) 'Development of Sight Word Reading: Phases and Findings', in M. J. Snowling and C. Hulme (eds.) 2005, *The Science of Reading. A Handbook*. Oxford: Blackwell, pp. 135-154.

Ehri, L. C. and McCormick, S. (1998) 'Phases of Word Learning: Implications for Instruction with Delayed and Disabled Readers', in *Reading and Writing Quarterly*, vol. 14, no. 2, pp. 135-163.

Ellis, N. C. (2008) 'Usage-based and form-focused language acquisition: the associative learning of constructions, learned attention and the limited L2 endstate', in P. Robinson and N. C. Ellis (eds.) 2008, *Handbook of Cognitive Linguistics and Second Language Acquisition*. Abingdon: Routledge, pp. 372-405.

Ericsson, K. A. and Simon, H. A. (1980) 'Verbal Reports as Data', in *Psychological Review*, vol. 87, no. 3, pp. 215-251.

Ericsson, K. A., and Simon, H. A. (1993) *Protocol analysis: Verbal reports as data*. Revised Edition. Cambridge, MA: MIT Press.

Erlam, R. (2003) 'The Effects of Deductive and Inductive Instruction on the Acquisition of Direct Object Pronouns in French as a Second Language', in *Modern Language Journal*, vol. 87, no. 2, pp. 242-260.

Erlam, R. (2005) 'Language Aptitude and its Relationship to Instructional Effectiveness in Second Language Acquisition', in *Language Teaching Research*, vol. 9, no. 2, pp. 147-171.

Erler, L. (2002) *Reading in a Foreign Language – Near-Beginner Adolescents' Experiences of French in English Secondary Schools*. DPhil Thesis, Oxford University.

Erler, L. (2004) 'Near-Beginner Learners of French Are Reading at a Disability Level', in *Francophonie*, vol. 30, pp. 9-15.

Erler, L. (2006) 'Getting started on the successful acquisition of a foreign language', in *Gifted and Talented*, vol. 10, no.1, pp. 8-10.

Erler, L. (2007). 'Using rhymes to assess students' strategies for deciding how written French words should sound', in *Francophonie*, vol. 35, pp. 22-29.

Erler, L. (2008) 'Rhyme identification at Key Stage 3 – the results', in *Francophonie*, vol. 37, pp. 3-7.

Evans, M. J. (2007) 'The Languages Review in England: foreign language planning in the absence of an overarching policy', in *Language Learning Journal*, vol. 35, no. 2, pp. 301-303.

Field, A. (2010) *Discovering Statistics Using SPSS* (Third Edition). London: Sage Publications.

Fielding, M. (1999) 'Target Setting, Policy Pathology and Student Perspectives: Learning to Labour in New Times', in *Cambridge Journal of Education*, vol. 29, no. 2, pp. 277-287.

Fisher, L. (2001). 'Modern foreign languages recruitment post-16: The pupils' perspective', in *Language Learning Journal*, vol. 23, pp. 33–40.

Flege, J. E. (1999) 'Age of Learning and Second Language Speech', in D. Birdsong (ed.) 1999, *Second Language Acquisition and the Critical Period Hypothesis*. Mahwah, NJ: Lawrence Erlbaum Associates, pp. 101-131.

Fournier, I. (2003) *Talk French*. Revised Edition. Harlow: BBC Active.

Frith, U., Wimmer, H. and Landerl, K. (1998) 'Differences in Phonological Recoding in German- and English-Speaking Children', in *Scientific Studies of Reading*, vol. 2, no. 1, pp. 31-54.

Frost, R (2005) 'Orthographic Systems and Skilled Word Recognition Processes in Reading', in Margaret J. Snowling and Charles Hulme (eds.) 2005, *The Science of Reading. A Handbook*. Oxford: Blackwell, pp. 272-295.

Frost, R. (2006) 'Becoming Literate in Hebrew: the Grain Size Hypothesis and Semitic Orthographic Systems', in *Developmental Science*, vol. 9, no. 5, pp. 439-440.

Frost, R., Katz, L. and Bentin, S. (1987) 'Strategies for Visual Word Recognition and Orthographical Depth: A Multilingual Comparison', in *Journal of Experimental Psychology: Human Perception and Performance*, vol. 13, no. 1, pp. 104-115.

Fukkink, R. G., Hulstijn, J. and Simis, A. (2005) 'Does Training in Second-Language Word Recognition Skills Affect Reading Comprehension? An Experimental Study', in *Modern Language Journal*, vol. 89, no. 1, pp. 54-75.

Gaskins, I. W., Downer, M. A., Anderson, R. C., Cunningham, P. M., Gaskins, R. W. and Schommer, M. (1988) 'A Metacognitive Approach to Phonics: Using What You Know to Decode What You Don't Know', in *Remedial and Special Education Journal*, vol. 2, no. 1, pp. 36-41.

Gillborn, D., and Youdell, D. (2000) Rationing Education. Policy, Practice, Reform and Equity. Buckingham and Philadelphia: Open University Press.

Glushko, R. J. (1979) 'The Organization and Activation of Orthographic Knowledge in Reading Aloud', in Human Perception and Performance, vol. 5, no. 4, pp. 674-691.

Goodman, K. S. (1976) 'Reading: a psycholinguistic guessing-game', in H. Singer and R. Ruddell (eds.) 1976, *Theoretical Models and processes of reading.* Newark, German: International Reading Association, pp. 497-508.

Gorard, S. (2001) *Quantitative Methods in Educational Research: The Role of Numbers Made Easy*. London: Continuum.

Gorard, S. (2005) *Value-Added is of Little Value*. Paper presented at the British Educational Research Association Annual Conference, University of Glamorgan, 14th-17th September 2005. [Online]. Available from <http://www.leeds.ac.uk/educol/documents/143649.htm> [Accessed 29.06.2010].

Goswami, U. (1995) 'Phonological Development and Reading by Analogy: What Is Analogy, and What Is Not?' in *Journal of Research in Reading*, vol.18, no. 2, pp. 139-145.

Goswami, U., Gombert, J. E. and De Barrera, L. F. (1998) 'Children's Orthographic Representations and Linguistic Transparency: Nonsense Word Reading in English, French and Spanish', in *Applied Psycholinguistics*, vol. 19, pp. 19-52.

Grabe, W. and Stoller, F. L. (2002) *Teaching and Researching Reading*. Longman: Harlow.

Graham, S. (2010). *Language learner strategies: the link with motivation*. Paper presented at UKPOLLS (U.K. Project on Language Learner Strategies) Conference, 'Improving Language Learning Through the Strategic Classroom: Findings and Applications of Research', 21.04.2010, British Academy, London.

Graham, S. and Macaro, E. (2008) 'Strategy Instruction in Listening for Lower-Intermediate Learners of French', in *Language Learning*, vol. 58, no. 4, pp. 747–783.

Graves, M. F. and Dykstra, R. (1997) 'Contextualizing the First-Grade Studies: What is the Best Way to Teach Children to Read?' in *Reading Research Quarterly*, vol.32, no.4, pp.342-344.

Grenfell, M. (1992) 'Reading and Communication in the Modern Languages Classroom. Perspectives on Reading'. CLE Working Papers, no. 2, pp. 64–77.

Grenfell, M. (ed.) (1995) *Reflections on Reading: From GCSE to 'A' level*. London: CILT.

Grevisse, M. (1988) *Le Bon Usage: Grammaire Française*. Twelfth edition: revised by André Goosse. Paris: Duculot.

Hagger, H. and McIntyre, D (2000) 'What Can Research Tell us about Teacher Education?' in *Oxford Review of Education*, vol. 23, pp. 483-494.

Hamada, M. and Koda, K. (2008) 'Influence of First Language Orthographic Experience on Second Language Decoding and Word Learning', in *Language Learning*, vol. 58, no. 1, pp.1-31.

Harlen, W. and Malcolm, H. (1999) *Setting and Streaming*. Second edition. The Scottish Council for Research in Education: Glasgow.

Hawkins, P. (1984) *Introducing Phonology*. London: Routledge.

Hawkins, E. (2005) 'Out of this nettle, drop-out, we pluck this flower, opportunity: re-thinking the school FL apprenticeship', in *Language Learning Journal*, vol.32, pp. 4-17.

Hickey, T. (2005) 'Second language Writing Systems: Minority Languages and Reluctant Learners', in V. Cook and B. Bassetti (eds.) 2005, *Second Language Writing systems*. Clevedon: Multilingual Matters, pp. 398-423.

Hill, P. W. and Rowe, K. J. (1996) 'Multilevel Modelling in School Effectiveness Research', in *School Effectiveness and School Improvement*, vol. 7, no. 1, pp. 1-34.

Hobbs, G. and Vignoles, A. (2010) 'Is Children's Free School Meal 'Eligibility' a Good Proxy for Family Income?', in *British Educational Research Journal*, vol. 36, no. 4, pp. 673-690.

Hoover, W. A. and Gough, P. B. (1990) 'The Simple View of Reading', in *Reading and Writing: An Interdisciplinary Journal*, vol. 2, pp. 127-160.

Horwitz, E. K., Horwitz, M. B. and Cope, J. (1986) 'Foreign Language Classroom Anxiety', in *Modern Language Journal*, vol. 70, no. 2, pp. 125-132.

Hurry, J. (2004) 'Comparative Studies of Instructional Methods', in T. Nunes and P. Bryant (eds.) 2004, *Handbook of Children's Literacy*. Dordrecht: Kluwer Academic Publishers, pp. 557-574.

Hurry, J., Sylva, K. and Riley, J. (1999) 'Evaluation of a Focused Literacy Teaching Programme in Year 1 Classes: Child Outcomes', in *British Educational Research Journal*, vol. 25, no. 5, pp. 637-649.

Institut Français de Vienne (2008) *La Francophonie dans le monde à l'occasion de la Journée internationale de la Francophonie*. [Online]. Available from <www.ambafrance-at.org/IMG/pdf/Dossier_francophonie.pdf> [Accessed 08.08.2011].

IPA (International Phonetic Association) (2005) *Handbook of the International Phonetic Association: A guide to the use of the International Phonetic Alphabet*. Cambridge: Cambridge University Press.

Johnston, R. and Watson, J. (2005) 'Effects of Synthetic Phonics Teaching on Reading and Spelling Attainment: A Seven Year Longitudinal Study'. Edinburgh: Scottish Executive. [Online]. Available from <http://www.scotland.gov.uk/Resource/Doc/36496/0023582.pdf> [Accessed 09.10.2007].

Jolly Learning (2011) *Introduction to Jolly Phonics*. [Online]. Available from <http://jollylearning.co.uk/overview-about-jolly-phonics/> [Accessed 12.08.2011].

Jorm, A. F. and Share, D. L. (1983) 'Phonological Recoding and Reading Acquisition', in *Applied Psycholinguistics*, vol. 4, pp. 103-147.

Just, M. A. and Carpenter, P. A. (1980) 'A theory of reading: From eye fixations to comprehension', in *Psychological Review*, vol. 87, no. 4, pp. 329-354.

Katamba, F. (1989) *An Introduction to Phonology*. London:Longman.

Katz, L. and Feldman, L. B. (1983) 'Relation Between Pronunciation and Recognition of Printed Words in Deep and Shallow Orthographies', in *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 9, no. 1, pp. 157-166.

Katz, L. and Frost, R. (1992) 'The Reading Process is Different for Different Orthographies: The Orthographic Depth Hypothesis', in L. Katz and R Frost (eds.) 1992, *Orthography, Phonology, Morphology and Meaning*. Amsterdam: Elsevier Science Publishers, pp. 67-84.

Kenstowicz, M. (1994) *Phonology in Generative Grammar*. Oxford:Blackwell.

Kintsch, W. and Rawson, K. A. (2005) 'Comprehension', in Margaret J. Snowling and Charles Hulme (eds.), *The Science of Reading. A Handbook*. Oxford: Blackwell, pp. 209-226.

Knight, T. W. (1994) *Living French*. Fifth Edition. London: Hodder Arnold.

Koda, K. (1989) 'Effects of L1 Orthographic Representation on L2 Phonological Coding Strategies', in *Journal of Psycholinguistic Research*, vol. 18, no. 2, pp. 201-222.

Koda, K. (1990) 'The use of L1 reading strategies in l2 reading: Effects of LI Orthographic Structures on L2 Phonological Recoding Strategies', in *Studies in Second Language Acquisition*, vol. 12, pp. 393-410.

Koda, K. (1996) ‘L2 Word Recognition Research: A Critical Review’, in *Modern Language Journal*, vol. 80, no. 4, pp. 450-460.

Koda, K. (1998) ‘The role of phonemic awareness in second language reading’, in *Second Language Research*, vol. 14, no. 2, pp. 194-215.

Koda, K. (1999) ‘Development of L2 Intraword Orthographic Sensitivity and Decoding Skills’, in *The Modern Language Journal*, vol. 83, no. 1, pp. 51-64.

Koda, K. (2004) *Insights into Second Language Reading*. Cambridge: Cambridge University Press.

Koda, K. (2007) ‘Reading and Language Learning: Crosslinguistic Constraints on Second Language Reading Development’, in *Language Learning*, vol. 57, supplement 1, pp. 1-44.

Koda, K. and Reddy, P. (2008) ‘Research Timeline: Cross-linguistic transfer in second language reading’, in *Language Teaching*, vol. 41, no. 4, pp. 497–508.

Kormos, J. (1998) ‘Verbal Reports in L2 Speech Production Research’, in *TESOL Quarterly*, vol. 32, no. 2, pp. 353-358.

Kuhl, P. K. (1994) ‘Learning and Representation in Speech and Language’, in *Current Opinion in Neurobiology*, vol. 4, pp. 812-822.

Kyriacou, M. (2008) *The Development of Narrative Writing in Primary School Children: Designing and Evaluating an Experimental Intervention*. DPhil thesis, Oxford University.

Kyriakides, L., Creemers, B., Antoniou, P. and Demetriou, D. (2010) ‘A Synthesis of Studies Searching for School Factors: Implications for Theory and Research’, in *British Educational Research Journal*, vol. 36, no. 5, pp. 807-830.

LaBerge, D. and Samuels, S. J. (1974) ‘Toward a Theory of Automatic Information Processing in Reading’’, in *Cognitive Psychology*, vol. 6, pp. 293-323.

Ladefoged, P. (1993) *A Course in Phonetics*. Third Edition. New York: Harcourt Brace.

Lado, R. (1957) *Linguistics across cultures: applied linguistics for language teachers*. Ann Arbor: University of Michigan Press.

Lange, M. and Content, A. (1999) ‘The grapho-phonological system of written French: statistical analysis and empirical validation’. [Online]. Available from < http://homepages.widged.com/mlange/> [Accessed 29.11.2006].

Lange, M. and Content, A. (in preparation) 'The reading aloud of one- and two-syllable words. Different problems that require different solutions? Insights from a quantitative analysis of the print-to-sound relations'. [Online]. Accessed 29.11.2006 at http://homepages.widged.com/mlange/.

Lee, J. and Schallert, D. L. (1997) 'The Relative Contribution of L2 Language Proficiency and L1 Reading Ability to L2 Reading Performance: A test of the Threshold Hypothesis in an EFL Context', in *TESOL Quarterly*, vol. 31, no. 4, pp. 713-739.

Leech, N. L. & Onwuegbuzie, A. J. (2002) *A Call for Greater Use of Non-Parametric Statistics*. Paper presented at the annual meeting of the Mid-South Educational Research Association, Chattanooga, TN, 6th-8th November 2002. [Online]. Available from <http://www.eric.ed.gov/PDFS/ED471346.pdf> [Accessed 12.08.2010].

Leow, R. P. and Morgan-Short, K. (2004) 'To Think Aloud or Not To Think Aloud: The Issue of Reactivity in SLA Research Methodology', in *Studies in Second Language Acquisition*, vol. 26, pp. 35-37.

Levy, E. S. and Law, F. F. (2010) 'Production of French Vowels by American-English Learners of French: Language Experience, Consonantal Context and the Perception-Production Relationship', in Journal of the Acoustical Society of America, vol. 128, no. 3, pp. 1290-1305.

Lewis, A. (2002) 'Accessing, Through Research Interviews, the Views of Children with Difficultis in Learning', in *Support for Learning*, vol. 17, no. 3, pp. 110-116.

Liaw, M. (2003) 'Integrating Phonics Instruction and Whole Language Principles in an Elementary School EFL Classroom', in *English Teaching and Learning*, vol. 27, no. 3, pp. 15-34.

Littlemore, J. (2009) *Applying Cognitive Linguistics to Second Language Learning and Teaching*. Basingstoke : Palgrave Macmillan.

Long, M. H. (1990) 'Maturational Constraints on Language development', in *Studies in Second Language Acquisition*, vol. 12, pp. 251-285.

Long, M. H. (1991) 'Focus on Form: A Design Feature in Language Teaching Methodology', in K. De Bot, R. Ginsberg and C. Kramsch (eds.) 1991, *Foreign language Research in Cross-Cultural Perspective*. Amsterdam: John Benjamins, pp. 39-52.

Lupker, S. J. (2005) 'Visual Word Recognition: Theories and Findings', in M. J. Snowling and C. Hulme (eds.) 2005, *The Science of Reading. A Handbook*. Oxford: Blackwell, pp. 39-60.

Luyten, H. (2003) 'The Size of School Effects Compared to Teacher Effects: An Overview of the Research Literature', in *School Effectiveness and School Improvement*, vol. 14, no. 1, pp. 31-51.

Macaro, E. (2001) *Learning Strategies in Foreign and Second Language Classrooms*. London: Continuum.

Macaro, E. (2003) *Teaching and Learning a Second Language: a Review of Recent Research*. London: Continuum.

Macaro, E. (2006) Strategies for language learning and for language use: revising the theoretical framework *Modern Language Journal*, vol. 90, no. 3, pp. 320-337.

Macaro, E. (2008). 'The decline in language learning in England: getting the facts right and getting real', in *Language Learning Journal*, vol. 36, no. 1, pp. 101-108.

Macaro, E, and Erler, L (2008a) 'Raising the Achievement of Young-Beginner Readers of French Through Strategy Instruction', in *Applied Linguistics*, vol. 29, no. 1, pp. 90-119.

Macaro, E. and Erler, L. (2008b) 'Basic Literacy Skills in L2 French and Language Learning Motivation', paper presented at AILA (Association Internationale de Linguistique Appliquée) 15th World Congress of Applied Linguistics, 'Multilingualism: Challenges and Opportunities', 24th-29th August 2008. Essen, Germany.

Macaro, E. and Mutton, T. (2009) 'Developing Reading Achievement in Primary Learners of French: Inferencing Strategies Versus Exposure to "Graded Readers"', in *Language Learning Journal*, vol. 37, no. 2, pp. 165–182.

Massaro, D. W. and Cohen, M. M. (1990) 'Perception of Synthesized Audible and Visible Speech', in *Psychological Science*, vol. 1, pp. 55-63.

McGraw, K. O. and Wong, S. P. (1992), 'A Common Language Effect Size statistic', in *Psychological Bulletin*, vol. 111, no. 2, pp. 361-365.

McGurk, H. and MacDonald, J. (1976) 'Hearing Lips and Seeing Voices', in *Nature*, vol. 264, pp. 746–748.

McLaughlin, B. (1990) 'Restructuring', in *Applied Linguistics*, vol. 11, no. 2, pp. 113-128.

Mitchell, I. (2002) 'Developing Reading Skills', in A. Swarbrick (ed.) 2002, *Aspects of Teaching Secondary Modern Foreign Languages*: *Perspectives on Practice*. London: Routledge, pp. 107-122.

Mitchell, R. (2003) 'Rethinking the Concept of Progression in the National Curriculum for Modern Foreign Languages: A Research Perspective', in *Language Learning Journal*, no.27, pp. 15-23.

Miyake, A. and Shah, P. (1999) *Models of working memory: Mechanisms of Active Maintenance and Executive Control.* Cambridge: Cambridge University Press.

Moore, L., Graham, A & Diamond, I. (2003) 'On the Feasibility of Conducting Randomized Trials in Education: Case Study of a Sex Education Intervention', in *British Educational Research Journal*, vol. 29, no. 5, pp. 673-689.

Morais, J., Cary, L., Alegria, J. and Bertelson, P. (1979) 'Does Awareness of Speech as a Sequence of Phones Arise Spontaneously?' in *Cognition*, vol. 7, pp. 323-331.

Moustafa, M. (1995) 'Children's productive phonological recoding', in *Reading Research Quarterly*, vol. 30, no. 3, pp. 464-476.

Moys, A. (1996) *Colloquial French: The Complete Course for Beginners*. London: Routledge.

Muljani, D., Koda, K. and Moates, D. R. (1998) 'The development of word recognition in a second language', in *Applied Psycholinguistics*, vol. 19, pp. 99-113.

Nagy, W. E., Anderson, R. C. and Herman, P. A. (1987) 'Learning Word Meanings from Context during Normal Reading', in *American Educational Research Journal*, vol. 24, no. 2, pp. 237-270

Nagy, W. E., Herman, P. A. and Anderson, R. C. (1985) 'Learning Words From Context', in *Reading Research Quarterly*, vol. 20, no. 2, pp. 233-253.

NALDIC (National Association for Language Development in the Curriculum) (2011a) *EAL and Ethnicity in Schools Nationally and by LA, January 2011*. [Online]. Available from < http://www.naldic.org.uk/docs/resources/KeyDocs.cfm>. [Accessed 04.09.2011].

NALDIC (National Association for Language Development in the Curriculum) (2011b) *EAL Pupils in Primary and Secondary Schools 1997-2011*. [Online]. Available from < http://www.naldic.org.uk/docs/resources/KeyDocs.cfm>. [Accessed 04.09.2011].

Nisbett, R. E. and Wilson, T. D. (1977) 'Telling More Than We Can Know: Verbal Reports on Mental Processes', in *Psychological Review*, vol. 84, no. 3, pp. 231-259.

Nuffield Languages Inquiry (2000). *Languages: The Next Generation*. London: The Nuffield Foundation.

Nunes, T. (2004) 'Introduction' to section 'Looking Across Languages', in T. Nunes and P. Bryant (eds.) 2004, *Handbook of Children's Literacy*. London: Kluwer Academic Publishers, pp. 625-630.

Nunes, T. and Bryant, P. (eds.) (2004) *Handbook of Children's Literacy*. London: Kluwer Academic Publishers.

Odlin, T. (1989) *Language Transfer: Cross-linguistic influence in language-learning*. Cambridge: Cambridge University Press.

Ofsted (Office for Standards in Education, Children's Services and Skills) (2009) *The inspection of Initial Teacher Education 2008–2011: A Guide for Inspectors on the Management and Organisation of Initial Teacher Education Inspections.* Manchester: Ofsted.

Ofsted (Office for Standards in Education, Children's Services and Skills) (2011). *Modern languages: achievement and challenge 2007–2010.* [Online]. Available from <http://www.ofsted.gov.uk/Ofsted-home/Publications-and-research> [Accessed 12/01/2011].

OUFFML (Oxford University Faculty of Medieval and Modern Languages) (2010) *Policy Statement on Disabilities and Compliance with SENDA.* [Online]. Available from <http://www.mod-langs.ox.ac.uk/accessibility> [Accessed 16.07.2010].

Overy, R. and Lecanuet, J. (2003) *French in Three Months*. Third Edition. London: Dorling Kindersley.

Pachler, N. (2002) 'Foreign Language Learning in England in the 21st Century', in *Language Learning Journal*, vol. 25, pp. 4-7.

Papagno, C., Valentine, T. and Baddeley, A. D. (1991) 'Phonological short-term memory and foreign-language vocabulary learning', in *Journal of Memory and Language*, vol. 30, no. 3, pp. 331-347.

Paribakht, T. S. and Wesche, M. (1996) 'Enhancing Vocabulary Acquisition Through Reading: A Hierarchy of Text-Related Exercise Types', in *The Canadian Modern Language Review*, vol. 52, no. 2, pp. 155-178.

Paulesu, E. (2006) 'On the Advantage of 'Shallow' Orthographies: Number and Grain Size of the Orthographic Units or Consistency Per Se?' in *Developmental Science*, vol. 9, no. 5, pp. 443-444.

Peereman, R. and Content, A. (1998) 'Quantitative analyses of orthography to phonology mapping in English and French'. [Online]. Available from <http://student.vub.ac.be/~acontent/OPMapping.html> [Accessed 25.11.2006].

Perfetti, C. A. (1985) *Reading Ability*. New York: Oxford University Press.

Perfetti, C. A. and Zhang, S. (1991) 'Phonological Processes in Reading Chinese Characters', in *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 17, no. 4, pp. 633-643.

Perfetti, C. A. and Zhang, S. (1995) 'Very Early Phonological Activation in Chinese Reading', in *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 21, no. 1, pp. 24-33.

Phillips, D. and Filmer-Sankey, C. (1993) *Diversification in Modern Language Teaching: Choice and the National Curriculum*. London: Routledge.

Pinker, S. (1994) *The Language Instinct: The New Science of Language and Mind*. London: Allen Lane.

Plaut, D. C., McClelland, J. L., Seidenberg, M. S. and Patterson, K. E. (1996) 'Understanding normal and impaired word reading: computational principles in quasi-regular domains', in *Psychological Review*, vol. 103, pp. 56-115.

Pollard, A. and James, M. (eds.) (2004) *Personalised Learning: A Commentary by the Teaching and Learning Research Programme*. London: ESRC Teaching and Learning Research Programme. [Online]. Available from <http://www.tlrp.org/documents/personalised_learning.pdf> [Accessed 12.05.2010].

Pugh, K. (2006) , 'A Neurocognitive Overview of Reading Acquisition and Dyslexia Across Languages', in *Developmental Science*, vol. 9, no. 5, pp. 448-450.

QCA (Qualification and Curriculum Authority) (2007a) *Key Stage 2 Scheme of Work for languages: Overview of French Units 1–12*. [Online]. Available from <http://webarchive.nationalarchives.gov.uk> [Accessed 05.01.2011].

QCA (Qualification and Curriculum Authority) (2007b) *The National Curriculum for England and Wales*. [Online]. Available from <http://curriculum.qcda.gov.uk/> [Accessed 05.05.2009].

Rapley, T. (2004) 'Interviews', in C. Seale, G. Gobo, J. Gubrium, D. Silverman (eds.) (2004), *Qualitative Research Practice*. London: SAGE Publications, pp. 15-33.

Robinson, P. and Ellis, N. C. (eds.) (2008), *Handbook of Cognitive Linguistics and Second Language Acquisition*. Abingdon: Routledge.

Robson, C. (2002) *Real World Research.* Second Edition. Oxford: Blackwell.

Rochet, B. (1995) 'Perception and Production of L2 Speech Sounds by Adults', in W. Strange (ed) 1995, *Speech Perception and Linguistic Experience: Theoretical and Methodological Issues*. Timonium MD: York Press, pp. 379-410.

Rose, J. (2006) *Independent review of the Teaching of Early Reading: Final Report*. London: DfES.

Saito, Y., Horwitz, E. K. and Garza, T. J. (1999) 'Foreign Language Reading Anxiety', in *The Modern Language Journal*, vol. 83, no. 2, pp. 202-218.

Sammons, P. (2007) *School Effectiveness and Equity: Making Connections. A Review of School Effectiveness and Improvement Research - Its Implications for Practitioners and Policy Makers*. Reading: CfBT Education Trust.

Scheerens, J. and Bosker, R. J. (1997) *The Foundations of Educational Effectiveness*. New York: Pergamon.

Scholfield, P. and Chwo, G. S. (2005) 'Are the L1 and L2 Word Reading Processes Affected More by Writing System or Instruction?', in V. Cook & B. Bassetti (eds.) (2005) *Second Language Writing systems*. Clevedon: Multilingual Matters, pp. 215-237.

Scovel, T. (1969) 'Foreign Accents, Language Acquisition, and Cerebral Dominance', in *Language Learning*, vol. 19, pp. 245–253.

Scovel, T. (2000) 'A Critical Review of the Critical Period Research', in *Annual Review of Applied Linguistics*, vol. 20, pp. 213-223.

Segalowitz, N. S. and Segalowitz, S. J. (1993) 'Skilled Performance, Practice and the Differentiation of Speed-up from Automatization Effects: Evidence from Second Language Word Recognition', in *Applied Psycholinguistics*, vol. 14, pp. 369-385.

Seidenberg, M. S. (1985) 'The Time Course of Phonological Code Activation in Two Writing Systems', in *Cognition*, vol. 19, pp. 1-30.

Seidenberg, M. S. and McClelland, J. L. (1989) 'A distributed developmental model of word recognition and naming', in *Psychological Review*, vol. 96, pp. 523-568.

Seliger, H. W. (1983) 'The Language Learner as Linguist: Of Metaphors and Realities', in *Applied Linguistics*, vol. 4, no.3, pp.179-191.

Selinker, L. (1972) 'Interlanguage', in *International Review of Applied Linguistics*, vol. 10, no. 3, pp. 209-31.

Seymour, P. H. K. (2005) 'Early Reading Development in European Orthographies', in M. J. Snowling and C. Hulme (eds.) 2005, *The Science of Reading. A Handbook*. Oxford: Blackwell, pp. 296-315.

Seymour, P. H. K., Aro, M. and Erskine, J. M. (2003) 'Foundation Literacy Acquisition in European Orthographies', in *British Journal of Psychology*, vol. 94, pp. 143–174.

Shallice, T., Warrington, E. K. and Mccarthy, R. (1983) 'Reading Without Semantics', in *The Quarterly Journal of Experimental Psychology Section A*, vol. 35, no. 1, pp. 111-138.

Share, D. L. (1995) 'Phonological recoding and self-teaching: *sine qua non* of reading acquisition', in *Cognition*, vol. 55, pp. 151-208.

Sharpe, K. (2001) *Modern Foreign Languages in the Primary School: the what, why and how of early MFL teaching*. London: Kogan Page.

Siddons, L. (2001) 'Practical Reflections on the Sound-Spelling Link', in *Francophonie*, vol. 23, pp. 10-14.

Smith, F. (1971) *Understanding Reading*. New York: Holt, Rinehart and Winston.

Spada, N. and Fröhlich, M. (1995) *Communicative Orientation of Language Teaching Observation Scheme: Coding Conventions and Application*. Sydney: National Centre for English language Teaching and Research, Macquarie University.

Spada, N. and Lightbown, P. M. (1999) 'Instruction, L1 Influence and Developmental Readiness in Second Language Acquisition', in *Modern Language Journal*, vol. 83, pp. 1-22.

Spada, N. and Lightbown, P. M. (2008) 'Form-Focused Instruction: Isolated or Integrated?' in *TESOL Quarterly*, vol. 42, no. 2, pp. 181-207.

Spada, N., Lightbown, P. M. and White, J. L. (2005) 'The Importance of Form-Meaning Mappings in Explicit Form-Focussed Instruction', in A. Housen and M. Pierrard (eds.) 2005, *Investigations in Instructed Second Language Acquisition*. Berlin: Walter de Gruyter, pp. 199-234.

Spada, N. and Lyster, R. (1997) 'Macroscopic and Microscopic Views of L2 Classrooms', in *TESOL Quarterly*, vol. 31, no. 4, pp. 787-795.

Sprenger-Charolles, L. (2004) "Linguistic processes in reading and spelling: the case of alphabetic writing systems: English, French, German and Spanish", in Nunes and Bryant (eds.) 2004, *Handbook of Children's Literacy*, pp. 43-66.

Sprenger-Charolles, L. and Casalis, S. (1995) "Reading and spelling acquisition in French first-graders: longitudinal evidence", in *Reading and Writing: An Interdisciplinary Journal*, vol. 7, pp. 39-63.

Sprenger-Charolles, L., Siegel, L. S. and Bonnet, P. (1998) 'Reading and Spelling Acquisition in French: The Role of Phonological Mediation and Orthographic Factors', in *Journal of Experimental Child Psychology*, vol. 68, pp. 134-165.

Stables, A. and Wikeley, F. (1999) 'From bad to worse? Pupils' Attitudes to Modern Foreign Languages at Ages 14 and 15', in *Language Learning Journal*, vol. 20, pp. 27-31.

Stahl, S. A., Duffy-Hester, A. M. and Stahl, K. A. D. (1998) 'Everything You Ever Wanted to Know about Phonics (But Were Afraid to Ask)', in *Reading Research Quarterly*, vol. 33, no. 3, pp. 338-355.

Stanovich, K. E. (1980) 'Toward an interactive-compensatory model of individual differences in the development of reading fluency', in *Reading Research Quarterly*, vol. 16, no. 1, pp. 32-71.

Stanovich, K. E. (1986) 'Matthew Effects in Reading: Some Consequences of Individual Differences in the Acquisition of Literacy', in *Reading Research Quarterly*, vol. 21, no. 4, pp. 360-407.

Strand, S. (2004) 'Consistency in Reasoning Test Scores Over Time', in *British Journal of Educational Psychology*, vol. 74, pp. 617–631.

Stuart, M. (1995) 'Through printed words to meaning: issues of transparency', in *Journal of Research in Reading*, vol. 18, no. 2, pp. 126-131.

Stuart, M. (1999) 'Getting Ready For Reading: Early Phoneme Awareness and Phonics Teaching Improves Reading and Spelling in Inner-City Second Language Learners', in *British Journal of Educational Psychology*, vol. 69, pp. 587-605.

Sylva, K. and Hurry, J. (1996) 'Early Intervention in Children with Reading Difficulties: An Evaluation of Reading Recovery and a Phonological Training', in *Literacy, Teaching and Learning: An International Journal of Early Literacy*, vol. 2, no. 2, pp. 49-68.

Sylva, K., Hurry, J. Mirelman, H., Burrell, A. and Riley, J. (1999) 'Evaluation of a Focused Literacy Teaching Programme in Reception and Year 1 Classes: Classroom Observations', in *British Educational Research Journal*, vol. 25, no. 5, pp. 617-635.

Stroop, J. R. (1935) 'Studies of interference in serial verbal reactions', in *Journal of Experimental Psychology*, vol. 18, pp. 643–662.

Takeuchi, O., Ikeda, M. and Mizumoto, A. (2009) 'Establishing the cerebral basis for language learner strategies: A NIRS study comparing L1 and L2 strategy use'. Paper presented at *First and Second Languages: Exploring the Relationship in Pedagogy-related Contexts*, 28th March 2009, Oxford.

Tarone, E. E. (1982) 'Systematicity and Attention in Interlanguage', in *Language Learning*, vol. 32, pp. 69-84.

Torgerson, C. J. (2001) 'The Need for Randomized Controlled Trials in Educational Research', in *British Journal of Educational Studies*, vol. 49, no. 3, pp. 316-328.

Treiman, R. (2008) *Learning about Letters as a Foundation for Learning to Read and Spell*. Paper presented on 14.01.2008 at Oxford University Department of Educational Studies, Oxford.

Treiman, R., Mullenix, J., Bijeljac-Babic, R. and Richmond-Welty, E. D. (1995), 'The Special Role of Rimes in the Description, Use, and Acquisition of English Orthography', in *Journal of Experimental Psychology: General*, vol. 124, no. 2, pp. 107-136.

Uebersax, J. (2010) 'Statistical Methods for Rater and Diagnostic Agreement' [Online]. Available from <http://www.john-uebersax.com/stat/agree.htm> [Accessed 14.07.10].

Underwood, G. and Batt, V. (1996) *Reading and Understanding*. Oxford : Blackwell.

Van berkel, A. (2005) 'The Role of the Phonological Strategy in Learning to Spell in English as a Second Language' , in V. Cook and B. Bassetti (eds.) 2005, *Second Language Writing systems*. Clevedon: Multilingual Matters, pp. 97-121.

Van Heuven, W. J. B. (2005) 'Bilingual Interactive Activation Models of Word Recognition in a Second Language', in V. Cook and B. Bassetti (eds.) 2005, *Second Language Writing systems*. Clevedon: Multilingual Matters, pp. 260-288.

Van Orden, G. C. and Kloos, H. (2005) 'The Question of Phonology and Reading', in M. J. Snowling and C. Hulme (eds.) 2005, *The Science of Reading. A Handbook*. Oxford: Blackwell, pp. 61-78.

Wade, P., Marshall, H. and O'Donnell, S. (2009). *Primary Modern Foreign Languages: Longitudinal Survey of Implementation of National Entitlement to Language Learning at Key Stage 2. Final Report*. Research commissioned by the Department for Children, Schools and Families. [Online]. Available from <http://www.dcsf.gov.uk/research> [Accessed 01.03.10].

Walford, G. (2001) *Doing Qualitative Educational Research: A Personal Guide to the Research Process*. London: Continuum.

Walter, C. (2004) ‘Transfer of Reading Comprehension Skills to L2 is Linked to Mental Representations of Text and to L2 Working Memory’, in *Applied Linguistics*, vol. 25, no. 3, pp. 315-339.

Walter, C. (2007) ‘First- to second-language reading comprehension: not transfer, but access’, in *International Journal of Applied Linguistics*, vol.17, no.1, pp.14-37.

Walter, C. (2008) ‘Phonology in Second Language Reading: Not an Optional Extra’, in *TESOL Quarterly*, vol. 42, no. 3, pp. 455-474.

Wang, M. and Koda, K. (2005) ‘Commonalities and Differences in Word Identification Skills Among Learners of English as a Second Language’, in *Language Learning*, vol. 55, no. 1, pp. 71-98.

Watts, C.J. (2003). *Decline in the Take-up of Languages at Degree Level*. London: Anglo-German Foundation

Wigfield, A. and Eccles, J. S. (2000). ‘Expectancy–Value Theory of Achievement Motivation’, in *Contemporary Educational Psychology*, vol. 25, pp. 68–81.

Williams, M., Burden, R. and Lanvers, U. (2002). “'French is the language of lover and stuff' – student perceptions of issues related to motivation in learning a foreign language”, in *British Educational Research Journal*, vol. 28, no. 4, pp. 503-528.

Wimmer, H. (2006) ‘Don’t Neglect Reading Fluency!’, in *Developmental Science*, vol. 9, no. 5, pp. 447-448.

Wimmer, H. and Goswami, U. (1994) ‘The Influence of Orthographic Consistency on Reading Development: Word Recognition in English and German Children’, in *Cognition*, vol. 51, pp. 91-103.

Wimmer, H. and Hummer, P. (1990) ‘How German-speaking First Graders Read and Spell: Doubts on the Importance of the Logographic Stage’, in *Applied Psycholinguistics*, vol. 11, pp. 349-368.

Woore, R. (2003) *Developing learners’ L2 French decoding ability*. *Best Practice Research Scholarship Level 1 Report*. [Online]. Accessed 20.08.2004 at <http://www.teachernet.gov.uk/professionaldevelopment/resourcesandresearch/bprs/search/index.cfm?report=1054> (No longer available at this address).

Woore, R. (2004) *Developing and applying knowledge of L2 grapheme – phoneme correspondences amongst beginner learners of German*. MSc dissertation, Oxford University.

Woore, R. (2006) *Investigating Beginner Learners' Decoding of Second Language French: Methodological Issues and Some Preliminary Substantive Findings*. MSc dissertation, Oxford University.

Woore, R. (2007) '"Weisse Maus in Meinem Haus": Using Poems and Learner Strategies to Help Learners Decode the Sounds of the L2″,in *Language Learning Journal*, vol. 35, no. 2, pp. 175-188.

Woore, R. (2009) 'Beginners' progress in decoding L2 French: some longitudinal evidence from English Modern Foreign Languages classrooms', in *Language Learning Journal*, vol. 37, no. 1, pp. 3-18.

Woore, R. (2010) 'Thinking aloud about L2 decoding: an exploration into the strategies used by beginner learners when pronouncing unfamiliar French words', in *Language Learning Journal*, vol. 38, no. 1, pp. 3-17.

Yen, H. Y. (2004) *An Examination of the Effect of Explicit Phonics Instruction and Authentic Readings on EFL Elementary Pupils*. Masters Dissertation, National Cheng Kung University, Taiwan. [Online]. Available from <http://etdncku.lib.ncku.edu.tw/ETD-db/ETD-search/view_etd?URN=etd-0625104-143036> [Accessed 08.07.2001].

Ziegler, J. C., Jacobs A. M., Stone, G. O. (1996) 'Statistical Analysis of the Bidirectional Inconsistency of Spelling and Sound in French', in *Behaviour Research Methods, Instruments and Computers*, vol. 28, no. 4, pp. 504-515.

Ziegler, J. and Goswami, U. (2005) 'Reading Acquisition, Developmental Dyslexia and Skilled Reading Across Languages: A Psycholinguistic Grain Size Theory', in *Psychological Bulletin*, vol. 131, no. 1, pp. 3-29.

Ziegler, J. and Goswami, U. (2006) 'Becoming literate in different languages: similar problems, different solutions', in *Developmental Science*, vol. 9, no. 5, pp. 429-436.