

Gonzaga University

## The Repository of Gonzaga University

---

TESOL Faculty Scholarship

Teaching English to Speakers of Other  
Languages

---

2012

### **A Multi-Method Investigation of the Effectiveness and Utility of Delayed Corrective Feedback in Second-Language Oral Production**

James Hunter

Follow this and additional works at: <https://repository.gonzaga.edu/tesolschol>



Part of the [Bilingual, Multilingual, and Multicultural Education Commons](#)

---

A MULTI-METHOD INVESTIGATION  
OF THE EFFECTIVENESS AND UTILITY  
OF DELAYED CORRECTIVE FEEDBACK  
IN SECOND-LANGUAGE ORAL PRODUCTION

By

JAMES DUNCAN HUNTER, MA

Module 3 (thesis) submitted to the  
School of Humanities  
of the University of Birmingham  
in partial fulfilment of the degree of  
DOCTOR OF PHILOSOPHY  
in  
Applied Linguistics

Centre for English Language Studies  
Department of English  
University of Birmingham  
Edgbaston  
Birmingham B15 2TT  
UK

NOVEMBER 2011

UNIVERSITY OF  
BIRMINGHAM

**University of Birmingham Research Archive**

**e-theses repository**

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

## DEDICATION

I would like to dedicate this work to my parents, Ann Hunter and Keith Hunter  
and to Bridget, Ceilan, Zeke, and Fionn, the loves of my life.

## ACKNOWLEDGEMENTS

I wish to acknowledge a huge debt of gratitude to several people whose guidance and support have been indispensable throughout my doctoral studies, in particular my parents. I would like to thank my supervisor, Charles Owen, and to apologise for taking so long. I would also like to thank my colleagues and students in Spokane and Abu Dhabi for participating in the research and for giving me valuable feedback and suggestions.

For those who proof-read drafts and gave suggestions, Ann Hunter, Bridget Green, and Charles Owen, I am greatly indebted and beg your forgiveness for any remaining inaccuracies, which are all my own.

Many thanks to Dr. Clodagh Brook for giving me a home in Birmingham.

Finally, a special thank you to Dr. Ron Harris for the inspiration.

## TABLE OF CONTENTS

CHAPTER 1: INTRODUCTION.....	1
1.1 Statement of the problem.....	1
1.2 Review of Module I and II .....	3
1.3 Overview of research.....	6
1.4 Significance of the study .....	7
CHAPTER 2: THEORETICAL AND METHODOLOGICAL BACKGROUND .....	10
2.1 Introduction .....	10
2.2 Theoretical foundations .....	10
2.2.1 Interlanguage and the origin of error .....	10
2.2.2 Developmental sequences .....	14
2.2.3 L1–TL interface .....	15
2.2.4 Sociocultural perspectives.....	16
2.3 Fluency .....	17
2.4 The connection between complexity, accuracy, and fluency .....	21
2.5 Methodological issues .....	23
2.5.1 Grammaticality judgements .....	23
2.5.2 Reaction time .....	27
2.5.3 The ecological validity of teacher research.....	29
CHAPTER 3: PILOT STUDY OF A TIMED GRAMMATICALITY JUDGMENT TEST...	32
3.1 Introduction .....	32
3.2 Research questions .....	32
3.3 Participants .....	34
3.4 Design and procedures.....	34
3.4.1 Timed judgement tests .....	34
3.4.2 Selection of test items .....	35
3.4.3 Linguistic profile of the pilot test items .....	37
3.4.4 Recording of the test items.....	42
3.4.5 Design of the testing platform.....	42
3.4.6 Elimination of unreliable items.....	46
3.4.7 Summary of stage one.....	48
3.5 Stage two data collection and findings .....	49
3.5.1 Anomalous responses by NS group .....	49
3.5.2 Response bias .....	51
3.5.3 The relationship between item length and judgement data.....	52
3.5.4 The relationship between judgement accuracy and reaction time .....	53
3.5.5 Judgements as ungrammatical and grammatical.....	54
3.5.6 Group mean differences .....	57
3.6 Differential performance by two NNS participants.....	61
3.7 Implicational hierarchies .....	64
3.8 Discussion.....	65
3.9 Summary of findings .....	66
CHAPTER 4: ELICITED IMITATION AND CORRECTION TESTS .....	68
4.1 Introduction .....	68
4.2 Research questions .....	68
4.3 Description of ‘Running Lists’ .....	69

4.4	The role of memory in elicited imitation.....	71
4.5	The Running List Test (RLT).....	75
4.6	Participants .....	75
4.7	Methodology.....	77
4.7.1	Pedagogical use of the RLT .....	77
4.7.2	Operationalization of accuracy and fluency.....	79
4.7.3	Automatic calculations of words per minute .....	81
4.7.4	Correlation of fluency measures .....	83
4.7.5	Fluency baselines for individual participants.....	86
4.7.6	Accuracy – grammaticality and acceptability .....	88
4.7.7	Diagnostic use of accuracy and fluency measures .....	90
4.8	Findings .....	91
4.8.1	Overall differences between RLT 1 & 2 .....	91
4.8.2	Individual differences between RLT 1 & 2.....	94
4.8.3	Relationship between accuracy and fluency .....	99
4.8.4	Comparison of performance on own and ‘All Do’ errors.....	103
4.8.5	Item learning and rule learning .....	105
4.9	Discussion.....	106
	CHAPTER 5: TIMED GRAMMATICALITY JUDGEMENT TESTS.....	110
5.1	Introduction .....	110
5.2	Research questions .....	114
5.3	Participants .....	115
5.4	Methodology.....	115
5.4.1	Administration of the TGJT .....	115
5.4.2	Scoring the TGJT items .....	116
5.4.3	Data analysis .....	116
5.4.4	Response bias.....	117
5.4.5	Comparison of production and recognition performance .....	119
5.5	Findings .....	119
5.5.1	Relationship between measures of accuracy.....	119
5.5.2	Relationship between measures of fluency.....	121
5.5.3	Response bias .....	125
5.5.4	Differential reaction time when judging items as grammatical and ungrammatical.....	127
5.5.5	Errors and mistakes .....	128
5.5.6	Comparison of judgements of own and ‘All Do’ errors .....	133
5.5.7	Comparison of TGJT performance by both groups .....	134
5.5.8	TGJT influence on post-test .....	136
5.6	Discussion.....	138
	CHAPTER 6: INVESTIGATION OF THE SMALL TALK DATABASE.....	142
6.1	Introduction .....	142
6.2	Research questions .....	142
6.3	Rationale for a learner error database.....	143
6.4	Error taxonomies and analytical frameworks.....	150
6.5	Description of the error database.....	153
6.6	Annotation procedures.....	155
6.7	An example of an error analysis: conditional clauses .....	157
6.7.1	Comparison of <i>if</i> conditionals in native speaker data and Small Talk data .....	158

6.7.2	Target and non-target uses of conditionals .....	160
6.7.3	Conjunction errors in conditional clauses .....	169
6.8	Pedagogical implications .....	173
6.9	Discussion.....	178
CHAPTER 7: CONCLUSION – IMPLICATIONS AND RECOMMENDATIONS .....		182
7.1	Overview of the research.....	182
7.2	Summary of principal empirical findings.....	182
7.3	Theoretical implications .....	185
7.4	Technological developments .....	186
7.5	Recommendations for future research.....	187



## List of figures

Figure 1: A model for an ecological approach to corrective feedback and interlanguage research.....	5
Figure 2: Example of standardized scores from timed grammaticality judgement test .....	28
Figure 3: Instructions for the TGJT .....	44
Figure 4: Feedback on the TGJT items .....	45
Figure 5: 3-D scatter plot of mean score, mean RT, and SD of RT .....	58
Figure 6: All TGJT participants, ranked by combined accuracy, RT, and SD of RT .....	60
Figure 7: Judgement data from ‘Gloria’ (NNS) .....	62
Figure 8: Judgement data from ‘Falshehri’ (NNS).....	63
Figure 9: Standardized z-scores for RLT 1 items for one participant .....	87
Figure 10: Comparison of RLT 1 & 2 for ‘Falshehri’ .....	95
Figure 11: Comparison of RLT 1 & 2 for ‘Ralhassan’ .....	97
Figure 12: Comparison of RLT 1 & 2 for ‘Riko’ .....	98
Figure 13: Comparison of RLT and TGJT data for ‘Falshehri’ .....	122
Figure 14: RLT-TGJT comparison for ‘Ayako’, ungrammatical RLT items only .....	130
Figure 15: Three ungrammatical items from ‘Ayako’, RLT 1 and 2 .....	131
Figure 16: An example of syntactic error tagging.....	150
Figure 17: Small Talk database Worksheet Entry form .....	156
Figure 18: Error analysis and tagging form for Small Talk database.....	156
Figure 19: Distribution of conditional sentences in Small Talk database, by level .....	160
Figure 20: Distribution of accurate and inaccurate conditional structures, by level .....	162
Figure 21: Distribution of all conditional clauses (accurate and inaccurate) by level, .....	163
Figure 22: Accurate use of conditional clauses by level, compared with NS production.....	165

Figure 23: Conditional clause errors, by level.....	166
Figure 24: Conditional clause errors, by L1 background .....	168
Figure 25: Subordinating conjunction errors in conditional clauses, by level .....	169
Figure 26: The ‘Focused Worksheet Maker’ form in the Small Talk database .....	174

## List of tables

Table 1: Items for pilot TGJT .....	36
Table 2: Analysis of 30 typical errors from intermediate L1 Arabic speakers.....	37
Table 3: Examples of equivalent errors to the test items found from speakers of L1s other than Arabic in the Small Talk database .....	40
Table 4: Divergent responses by NS group .....	50
Table 5: Summary of incorrect judgements by NS and NNS groups.....	52
Table 6: Correlations of length of audio recording of time with Accuracy and RT of responses .....	53
Table 7: Mean reaction times for incorrect and correct judgements, NNS participants .....	53
Table 8: Mean times for incorrect and correct judgements, NS participants .....	54
Table 9: Mean reaction times for judgements as ungrammatical and grammatical, NS participants .....	54
Table 10: The five items with highest mean reaction times for NS group .....	55
Table 11: Implication hierarchy of NNS judgements of items 18, 15, and 13.....	64
Table 12: Summary of differences between ELC and IEP groups.....	76
Table 13: RLT grade calculation spreadsheet .....	78
Table 14: ELC student grades on practice and RLT 1 (in percentages).....	78
Table 15: Comparison of items and word counts between worksheet item, teacher's reformulation, and students RLT recording .....	82
Table 16: Correlation matrix for word counts on worksheet, teachers' reformulations, and students' RLT recordings .....	83
Table 17: Correlation matrix for measures of fluency .....	84
Table 18: Descriptive statistics for RLT test data .....	91
Table 19: T-test comparison of RLT 1 and RLT 2, ELC group.....	92
Table 20: Paired-sample t-test of mean differences between RLT 1 and RLT 2, ELC group .....	92
Table 21: T-test comparison of RLT 1 and RLT 2, IEP group .....	93
Table 22: Paired-sample t-test of mean differences between RLT 1 and RLT 2, IEP group .....	94
Table 23: T-test of performance measures for own and others' errors, both RLTs, ELC group .....	104
Table 24: T-test of performance measures for own and others' errors, both RLTs, IEP group .....	104

Table 25: Correlation matrix for measures of fluency and accuracy on RLT and TGJT, ELC group.....	120
Table 26: Correlation matrix for measures of fluency and accuracy on RLT and TGJT, IEP group.....	121
Table 27: T-test of TGJT measures for own and ‘All Do’ errors, ELC group.....	134
Table 28: T-test of TGJT measures for own and ‘All Do’ errors, IEP group .....	134
Table 29: T-test of TGJT measures for ELC and IEP groups .....	135
Table 30: T-test comparison of means on RLT 1 and RLT 2 according to TGJT participation .....	137
Table 31: Breakdown of utterances in error database by L1 .....	154
Table 32: Breakdown of utterances in error database by proficiency .....	154
Table 33: Random sample of ten utterances from Small Talk database .....	146
Table 34: Distribution of if-clauses in the Corpus of Contemporary American English .....	159
Table 35: Mean utterance length, by level .....	172
Table 36: Conditional clause sentences spoken by ‘faltoaimy’, by level.....	175

## **List of appendices**

(All Appendices are included on the accompanying CD-Rom.)

- Appendix 1. Automatic calculation of speech rate using Praat
- Appendix 2. Audio recordings by Chinese L1 and Arabic L1 students of item: *If I had hit that barrier, I would have died.*
- Appendix 3. RLT data analysis, ELC and IEP groups (Microsoft Excel document)
- Appendix 4. Running List Test 1 items for all students (pdf document)
- Appendix 5. Running List Test 1 'All Do' items (pdf document)
- Appendix 6. RLT - TGJT comparison and data analysis, ELC and IEP groups (Microsoft Excel document)
- Appendix 7. Example worksheet for second conditional
- Appendix 8. Error taxonomy from the Small Talk database

## **List of abbreviations**

CF .....	Corrective Feedback
EAP .....	English for Academic Purposes
ELC .....	English Language Center (multilingual group – Ch. 4)
IEP .....	Intensive English Program (monolingual group – Ch. 4)
IL .....	Interlanguage
NNS .....	Non-Native Speaker (of English)
NS .....	Native Speaker (of English)
RLT .....	Running List Test
RT .....	Reaction Time
TGJT .....	Timed Grammaticality Judgement Test
TL .....	Target Language

---

## CHAPTER 1: INTRODUCTION

---

### 1.1 Statement of the problem

What is the state of the art in language instruction today? What advances has a half-century of Applied Linguistics research brought to the teaching profession? As the ELT profession matures and evolves, teacher training at the certificate or postgraduate level is becoming increasingly the norm, but what does theory-based teacher education tell us is cutting-edge pedagogy? How has Second Language acquisition (SLA) research contributed to this knowledge base? Surprisingly and perhaps disconcertingly for novice teachers, the field is full of controversies and contradictory findings. In the literature on corrective feedback (CF), which is the central concern of this investigation, there is a serious and growing claim that CF has no beneficial effects and is in fact harmful (Krashen 2002; Truscott 1996; 1999; 2004; 2007). While these charges have given rise to a renewed research focus on CF procedures, approaches, and effects, they have also created considerable uncertainty among practitioners as to what if anything to do about learner error. The dominant oral CF methodologies at present are recasts and elicitation, perhaps because these have been the most researched to date, but both have been challenged as having indeterminate status in both teacher intention and student perception (Hauser 2005; Mackey et al. 2007). One of the few findings related to CF that is uncontroversial is the fact that the majority of students seem to want it (Chun et al. 1982; Chenoweth et al. 1983; Margolis 2010) – although on this point Truscott asserts: ‘The obligation that teachers have to students is not to use whatever form of instruction the students think is best, but rather to help them learn’ (Truscott 1996: 359), pointing out that

‘students entering language or writing classes do not always know what is best for them’ (1996: 362 n.5).

If we are to take seriously claims that CF is ineffective and harmful, there is a definite need for more research on CF methodologies and results, a need which has begun to be addressed over the past decade (e.g. Nassaji and Swain 2000; Lochman 2002; Basturkmen et al. 2004; Loewen 2004; Panova and Lyster 2004; R. Ellis et al. 2006; Rolin-Ianziti 2006; R. Ellis et al. 2009; Sheen 2010). But even if such research indicated that current CF methods are ineffective (which it does not), this would not validate the conclusion that CF in general should be abandoned: it would simply point to a need for refinement of existing practices or invention of new ones.

One reason for Truscott’s insistence on abandonment arguably lies in his espousal of the view that linguistic input is the necessary and sufficient condition for second language acquisition (Krashen 1981; 1982; 1985; Schwartz 1986; Zobl 1995). For instance, Truscott explains: ‘Probably accuracy is improved through extensive experience with the target language – experience in reading and writing’ (Truscott 1996: 360). In this view, teachers simply have to provide input, and lots of it, and the acquisition will take place – a position which has led one cynical observer to conclude that ‘the fundamental message of Krashen’s theory is that you do not have to know very much to be a good language teacher’ (Gregg 1986: 121).

Whether or not one accepts Krashen’s theories, language instruction presents a fundamental problem: learners do not seem able to acquire what they are not ready for, and there are currently no systematic practical approaches to this problem, with the possible exception of work by Pienemann (e.g. Meisel et al. 1981; Pienemann 1989, 1992). Theories abound, the most famous being the ‘natural order’ of acquisition (see Goldschneider and



DeKeyser 2001 for an overview), but the practical application of these to the instruction of individual learners has not been forthcoming. Two recent comments exemplify the resignation with which this state of affairs is generally met:

Instruction needs to take into account the learner's built-in syllabus [in order to] ensure that learners are developmentally ready to acquire a specific target feature. However, this is probably impractical as teachers have no easy way of determining what individual students know. It would necessitate a highly individualized approach to cater to differences in developmental level among the students. (R. Ellis 2008b: 3)

Another problem with focus on form instruction is practical; that is, it involves class size. The views expressed by Long (1991) and Long and Robinson (1998) seem optimally suited to classrooms that are small enough to enable teachers to verbally address their students' problematic forms. In many settings, however, classes are large and individual attention and student–student interaction is not possible. (El-dali 2010: 67)

Thus one challenge facing language pedagogy is to track individual language development systematically in teaching contexts which make it very difficult. This research is motivated by this challenge, and will approach the problem from a number of directions that draw theoretical justification from the fields of SLA, psycholinguistics, and insights from corpus linguistics. Before laying out the research agenda, however, the principal directions and findings from Modules I and II will be reviewed in order to establish the context for the study.

## **1.2 Review of Modules I and II**

The main focus of Module I of this research was a review of the available literature on error analysis and corrective feedback (CF) as it pertains to the oral production of adult second language learners in instructed SLA. That paper also set out the research agenda, the investigation of a methodology of delayed corrective feedback using an oral communication approach called Small Talk (Harris 1998). The CF methodology, which will be investigated here, takes a systematic and ecological (see Section 2.5.3) approach to the collection and

communication of CF, and is summarized in Figure 1. As can be seen, this approach very intentionally makes CF a part of the linguistic environment, in the sense that it encourages teachers to monitor and observe student interaction and spoken production very carefully, leading one commentator to call it a ‘teacherless form of pedagogy’ (Charles Owen, personal communication). In addition, the linguistic content of the CF is intended to inform and guide the instructional focus on form by providing both individual and aggregate data on the types of language forms that the participants are using, whether successfully or not.

Module II investigated the validity and reliability of the process in which teachers observe student interaction and collect errors in the CF cycle, represented in the bottom half of Figure 1. Using video recordings of Small Talk sessions, the investigation established the degree of consensus between different teacher participants in their provision of CF, in other words, the quantity and nature of the items that they would have provided as ‘worksheets’ to the students. The principal findings were first that while teachers do not focus on identical errors (the consensus between all participants averaged about 45%), their provision of CF was highly attuned to the degree of accuracy and fluency that the student participants demonstrated; second, Small Talk is very successful in terms of getting learners to engage in authentic, fluent communication, whether fluency is defined following Brumfit (1979: 115) as the learner’s ‘truly internalized grammar’, or following more traditional definitions such as ‘the capacity to produce speech at normal rate and without interruption’ (Skehan 2009: 510). This expanded study builds on those findings, investigating whether Small Talk and the CF methodology are successful in promoting greater accuracy and complexity (Skehan 1998; Larsen-Freeman 2009) as well as fluency.

Figure 1: A model for an ecological approach to corrective feedback and interlanguage research

### 1.3 Overview of research

This investigation into the effectiveness and utility of delayed CF addresses the following research questions:

- 1) How effective is delayed CF in pushing learners towards greater accuracy and complexity in their oral production?
- 2) To what extent can learners recognize the accuracy of their own production?
- 3) What information about learner language development can be provided by a database of learner errors?

The investigation was approached from three methodological perspectives, each with its own research focus:

- 1) The capacity of learners to correct (reformulate) the language errors produced by themselves and peers during conversational interaction (Small Talk) was measured by means of an elicited imitation and correction task, and analysed using quantitative statistical procedures as well as qualitative analysis of language production. The design did not attempt to compare experimental and control groups, but rather to establish baseline measures of accuracy and fluency for individual participants against which to compare subsequent performance. This stage is reported in Chapter 4.
- 2) The capacity of learners to recognize the accuracy of their own reformulations, produced in the stage described above, was measured using a computerized timed grammaticality judgement task, which measured accuracy and reaction time. These results were also analysed using quantitative statistical procedures in order to determine the relationship between accuracy and fluency in production and reception. This stage is reported in Chapter 5. Since this stage of the research departs from customary psycholinguistic research methods in several important ways, a pilot study

was conducted in order to establish the validity and reliability of the testing platform. This is reported in Chapter 3, prior to the main research phases. The pilot study also addresses the inter-rater reliability of grammaticality judgements by proficient English speakers as well as learners.

- 3) The language errors which have been collected as part of the CF methodology (some of which are discussed in the course of this research) together form an error database. The annotation and analysis of this database, which draw on methodological approaches and assumptions from corpus linguistics, are discussed in Chapter 6, and ways in which such a tool can be applied to problems in interlanguage (IL) analysis and to syllabus design are proposed.

This research thus focuses on two central components of the ecological model depicted in Figure 1, namely the investigation of learner behaviour (reformulation and recognition of accuracy) and the investigation of the error database.

#### **1.4 Significance of the study**

Underlying each of these perspectives is the assumption that CF is not only an essential part of language instruction, but also a valuable research tool, particularly in teacher development. This research is thus an attempt to show that CF, and especially systematic delayed CF, can contribute to language learning, teacher development, and SLA research in ways that have not yet been fully explored by the field. However, to judge from recent publications in the field, CF in the second decade of the 21<sup>st</sup> century has been relegated to the sidelines of both instructional and research practices, becoming one, rather problematic, type of ‘incidental’ focus on form (Loewen 2004). This represents a sea-change in SLA, given that founding work in this field was focused precisely on CF (Corder 1967; 1981; see e.g. Seidlhofer 2003: 169 for a standard view of the origins of SLA).

The principal reason for this change, as mentioned in Section 1.1, is the allure of the ‘input-only’ position. This, as Long suggests, and as experience in L2A and instruction bear out, ought to be more controversial than it generally is:

The impossibility of learning some L2 items from positive evidence alone means that a theory that holds that nativelike mastery of a SL can result simply from exposure to comprehensible samples of that language is inadequate. (Long 1990: 660)

Essentially, the ‘input hypothesis’ derives from the idea that first and second language acquisition are fundamentally the same (Dulay et al. 1982) or are constrained by the same principles (White 1989). These views have proved to be tenacious, in spite of vigorous assaults (e.g. Schachter 1988; Bley-Vroman 1990), and in relation to CF have had two major consequences. First, the role of negative evidence, particularly in the form of CF, has been undermined (e.g. Carroll 1995: 356, who argues that since it is essentially metalinguistic, CF is always an interruption of communication, and therefore removes itself as a candidate for intake). Second, building on findings from L1 research (Brown 1973), SLA scholars claimed a ‘natural order’ of acquisition of syntactic forms (Bailey et al. 1974; Dulay and Burt 1974a) and argued that since the order in which specific features of language are acquired is immutable, CF could not be effective unless it happened to coincide with the relevant stage, in which case it would be unnecessary anyway. These views fail to distinguish important differences between types of L2A (e.g. adult vs. child, instructed vs. naturalistic – see Foster-Cohen 2001 for an overview) and also assume a monotonic continuum of syntactic (or lexical, or phonological, or morphological) development that may fail to account for multiple parallel forms (Larsen-Freeman 2006) and variability in competence (Tarone 1983; R. Ellis 1985).

This research therefore seeks to add to the body of work that suggests that CF can contribute in important ways to instructed language acquisition. The following chapter will

outline the theoretical foundations for the instructional methodologies under investigation and the research procedures employed.

---

## CHAPTER 2: THEORETICAL AND METHODOLOGICAL BACKGROUND

---

### **2.1 Introduction**

This chapter reviews the principal methodological and theoretical bases for the investigations in this research. First, the key theoretical issues will be outlined with reference to published research in this area. Second, methodological concerns and approaches which span multiple areas of the research will briefly be introduced.

### **2.2 Theoretical foundations**

This section surveys theories of SLA as they pertain to corrective feedback (CF), beginning with the concept of interlanguage and major theoretical explanations for L2 error. Some of the limitations of these are briefly discussed before the survey moves on to an exploration of fluency in speech production and the relationship between this and the concepts of accuracy and complexity, and the central role played by formulaic language.

#### **2.2.1 Interlanguage and the origin of error**

According to Lakshmanan and Selinker (2001: 359), the construct of interlanguage (IL), a system in its own right at least partially different from the L1 and the target L2, is one of the few points agreed on by all SLA researchers (but see below). Interlanguage has its origins in the work of Corder (1967), who suggested that errors provide information not only about how much a learner has learned, but also about how the language was learned. In the context of the behaviourist view of language learning as stimulus and response, imitation and reinforcement, and positive and negative transfer from L1 in which errors were seen as ‘dangerous thoughts’ to be eradicated with all the thoroughness and promptitude of which the watchful



teacher is capable' (French 1949: 98), Corder pointed out that errors of the kind regularly produced by second language learners are too creative and systematic to be solely the result of incorrect imitation or first language interference.

Even more significantly, Corder (1967) suggested that far from being undesirable, errors are a crucial part of the learning process. They illustrate that the learner is forming and testing hypotheses about the structure and functions of the language, and that they are therefore 'a device the learner uses in order to learn' (p. 167). Errors, he argued, thus shed light on the '*built-in syllabus*' (p. 166, italics in original) of the learner and constitute evidence of the acquisition process. Corder (1967) also famously distinguished between *errors* of competence and *mistakes* of performance (or 'slips'), a distinction which will be questioned in this research. Corder, it must be remembered, was attempting to apply a Chomskyan framework to build a second language acquisition theory. Chomsky's original formulation of the competence/performance dichotomy was:

Linguistic theory is primarily concerned with the ideal speaker–hearer, in a completely homogeneous speech community who knows its language perfectly and is unaffected by such grammatically irrelevant conditions as memory limitations, distractions, shifts of attention and interest, and errors (random or characteristic) in applying his knowledge of the language in actual performance. (Chomsky 1965: 3-4)

There is simply no room in this definition for a second language learner, nor, as many have pointed out (e.g. Sampson 1997), for a real human being of any description. Corder wanted to establish that learners have some kind of systematic mental representation of the L2 and that this, and not, say, poor memory or a failure of operant conditioning, was responsible for their 'errors'. In doing so, he was applying theoretical terminology to real speakers in very heterogeneous speech communities, which is why he coined the term 'transitional competence' to refer to the learner's 'knowledge of the language to date' (p. 167). This was picked up by Nemser (1971: 117), who introduced the concept of a learning 'plateau' in

learners' 'approximative systems' and claimed that 'effective language teaching implies preventing, or postponing as long as possible, the formation of permanent intermediate systems and subsystems (deviant phonological and grammatical structures)'; and by Selinker (1972), who introduced the term 'interlanguage', and 'fossilization' to refer to this plateau, a stabilized state after which the learner finds it difficult to produce a target form with consistency or at all, regardless of further instruction or input. The most salient example of this is foreign accent in the L2 speech of learners who have achieved near-native competence in all other areas.

A number of possible internal and external causes for fossilization have been suggested including age, lack of desire to acculturate, communicative pressure, and, critically germane to the subject of error treatment, lack of learning opportunity and excessive positive feedback (Lyster and Ranta 1997: 41; Boshier 1990: 92). Briefly, these imply that without sufficient feedback and practice in L2, non-target forms that nevertheless achieve a communicative goal will stabilise; and that positive feedback which signals comprehension of incorrect forms will lead to the stabilisation of these forms (Boshier 1990; for an overview of related research, see R. Ellis 2008b: 28–31).

Corder's use of the term *performance* (and *mistakes*) is problematic if we equate it with Chomsky's. L2 performance clearly is affected by conditions which are both grammatically relevant and systematic. One such condition is L1 knowledge, as discussed above: a learner may have the underlying knowledge (competence) of an L2 form but produce an error or an L1 equivalent during performance. Bialystok and Sharwood Smith (1985: 101) thus present an alternative view of this distinction, offering *knowledge* (linguistic and pragmatic competence) and *control* (the processing system used to control this knowledge during actual performance) as the two elements at play in production. In their view, interlanguage is thus

not what learners speak, ‘as though it had an independent existence’, but rather:

the outcome of mental functioning which attributes to the learner specific limitations in two aspects of mental processing. The result is a linguistic system which is unlike that used by the native speaker, but one which is none the less systematic in the structural sense. IL, in these terms, could be defined as the systematic language performance (in production and recognition of utterances) by second-language learners who have not achieved sufficient levels of analysis of linguistic knowledge or control of processing to be identified completely with native speakers. (1985: 116)

This view goes some way toward explaining the variability in learner performance of elements that are within their competence, and the existence of very proficient students who have disproportionate levels of accuracy and fluency. R. Ellis (1994: 395) suggests that learners might opt for greater knowledge or greater control, but that one would be developed at the expense of the other.

The question of why error occurs is an extremely complex one, and any theory of SLA that seeks explanatory adequacy has to consider L1A processes, age and conditions of target language (TL) and previous L2 (or L<sub>n</sub>) acquisition, L1–TL (and L<sub>n</sub>) interface, psychological factors in the individual and group, performance issues at the time of production, current TL proficiency, and exposure to the TL outside of the classroom. Even if it were possible to control for all of these factors and thereby to produce a truly homogeneous group of students, one could not expect that the language acquisition of such a group would be uniform. Some people are simply better language learners than others, for reasons of genetic endowment, intelligence, experience, personality, motivation, and so on. Despite the claims of some (e.g. Hammerly 1991), it is unreasonable to assume that every learner experiences and processes the TL in the same way, let alone that every learner would produce the same – or no – errors during instruction. For this reason, the quest for universals and common developmental sequences should not be permitted to obscure the need for individualized instruction.

### 2.2.2 Developmental sequences

Corder was certainly not the first to draw a comparison between the errors that children make when learning L1 and those made by L2 learners, but his work, along with Chomsky's assertion that children must have syntactic knowledge in advance of experience, set off an explosion of theoretical and experimental research. Many errors were indeed found to be developmental, which is to say that they reflect general language acquisition processes such as (over)generalization and creative but non-target-like usage (also found in L1 learners, c.f. Brown 1973), and seemed to indicate that there is a 'natural order' of language acquisition (c.f. Dulay and Burt 1974a). Vygotsky's 'Zone of Proximal Development' (Vygotsky 1978) construct was applied to SLA, to suggest that language items can only be acquired when the learner is developmentally ready to acquire them. (For a rejection of this application of the ZPD construct, see Lantolf and Dunn 1998). Considerable research has been done in order to determine what this order might be (see Wode et al. 1978; Meisel et al. 1981; Lightbown and Spada 1999; Goldschneider and DeKeyser 2001), but there are as yet no conclusive findings that have significantly influenced teaching methodology. Nevertheless, this has not stopped pronouncements of an authoritarian nature. For example, Scarcella and Oxford (1992: 18) claim:

[Learners]... acquire these grammatical structures in similar sequences – such that some features of the language (like the modal auxiliary *can*) are acquired before other features of the language (like the modal auxiliaries *could*, *would*, and *should*).

This could, of course, mean no more than 'textbooks generally present *can* before *could*, *would*, and *should*,' especially given the indisputable fact that the primary (and sometimes only) input to which the vast majority of learners of English as an L2 are exposed comes from textbooks. This has led some to posit the occurrence of 'transfer-of-training' errors, in other words, non-target linguistic forms which are pedagogically induced (Selinker 1972: 218) either by teaching materials or by teachers themselves. Thus Han and Selinker (1999: 267)

assert:

Inaccurate input, in the sense of input that conveys false notions of the target language, we think are [sic] common in textbooks and should be explored in an empirical pedagogy. Such input [in their study] not only appeared to stimulate the importation of an L1 pragmatic word order, but also disguised it under transfer-of-training.

### 2.2.3 L1–TL interface

A far better known (and researched) source of non-target forms, however, is L1 ‘interference’ or transfer. Han and Selinker (1999: 249) in fact assert: ‘Among various possible SLA factors that have stabilizing effects, language transfer has been singled out as the principal one.’

There have been many attempts to predict ‘typical’ pitfalls for particular groups of learners (e.g. French 1949; Paulovsky 1949; Burt and Kiparsky 1972; Crewe 1977), but these generally do not take into account individual variation, based as they are on some version of the Contrastive Analysis Hypothesis (CAH) elaborated by Lado (French 1949; Paulovsky 1949; Burt and Kiparsky 1972; Crewe 1977; ), which emphasizes the role of the first language in the acquisition of a second. CAH posits that similarities in structure and function between L1 and L2 would facilitate acquisition of L2, while differences would hinder acquisition. Errors in this view are seen largely as the result of negative transfer from L1 (e.g. ‘Don’t say: We will hope *it*. (Das wollen wir hoffen.) Instead, say: Let’s hope so.’ Paulovsky 1949: 41, original emphasis). R. Ellis (2008a: 349–403) provides a comprehensive account of the study of language transfer, and concludes that while transfer is no longer seen as even a major factor in the acquisition process, few would argue that it does not play any part at all (see James 1980, 1994 for overviews.)

While CAH as a predictive theory of SLA has been largely abandoned, there has recently been a resurgence in interest in Contrastive Analysis, with studies of syntactic transfer (e.g. Chan 2004) and crosslinguistic influence (e.g. Odlin 2005) providing insight

into the ‘cultural faultlines’ (Kramsch 1993) that separate cognitive frameworks and constrain the ultimate TL attainment of second language learners under most conditions. However, it should be noted that studies which attempt to describe or quantify the L1 origins of L2 errors generally agree that transfer cannot account for the majority of errors (e.g. Richards 1970; George 1972; Picard 2002; Harrat 2011).

#### **2.2.4 Sociocultural perspectives**

It is of course not true that all SLA scholars uncritically accept the interlanguage (IL) construct, nor even the abstracted construct of ‘native speakers’ and ‘non-native speakers’ of English. Sridhar, for instance (1994: 802), bemoans the ‘duplicative competence model in L2 pedagogy and SLA research’, which, he charges, characterizes ‘the overwhelming majority of L2 acquirers and users (all but a mere 5%) as speakers of *interlanguages* (Selinker, 1992), that is, as failed monolingual rather than successful bilingual.’ Rampton (1987) and Seidlhofer (2011) argue that learner language may differ from TL forms deliberately, for sociocultural reasons such as expressing solidarity (e.g. by code-switching), to signal to listeners that the speaker *is* a learner, or as a form of resistance to perceived cultural and linguistic encroachment. Larsen-Freeman (2000: 170) thus argues: ‘At the least, an emic perspective may sort out the motivation behind the use of such ‘deviant’ forms... non-native speakers have multiple social identities, being a learner is just one of them.’ The essential point here is that the ‘target’ language may well not be that of the NS at all.

While it may be challenging to envision an approach to learner language analysis that does not use NS norms as units of analysis, it is virtually impossible to conceive of instructional methodologies without such norms. And yet, such conceptions are not uncommon in the literature. For instance, Larsen-Freeman asserts that complete convergence with the target language (TL) is not an expected outcome in language pedagogy since, as we

have seen, L2 learners may not wish to emulate native-speaker norms, which in any case are never static systems themselves, there being ‘no fixed, homogeneous target end state to language evolution or development’ (Larsen-Freeman 2006: 592). In IL analysis, TL norms are viewed with considerable wariness:

In-depth contextual interpretation [of learner production] is therefore necessary in order reliably to establish regular form–function correspondences. Once an interpretation has been established, the surest way of missing learner-language regularities is to imagine a ‘corresponding’ utterance in another language – the target language or the source language – then attribute its organization back to the learner’s utterance (cf. the ‘closeness fallacy’; Klein and Perdue, 1989). One cannot rely on TL sentence-internal functions such as ‘subject’, ‘object’, as this would amount to analysing the learner’s language as if it were (imperfect) target language. (Klein and Perdue 1997: 311)

Klein and Perdue recognize the attraction of the ‘target deviation perspective’, and acknowledge its pedagogical roots, but view it as unhelpful ‘when we want to know something about how the human language capacity functions and which principles determine the acquisitional process’ (Klein and Perdue 1997: 307. See also Bley-Vroman 1983).

One reason, therefore, for the declining emphasis on the importance of CF is that the identification of NS norms as well as the relevance of these to IL research are in question.

Thus Han affirms:

Current SLA researchers are, in my view, confronted with two daunting tasks: the first is to systematically construct a developmental perspective on the native speaker or, more specifically, on [what one individual shares with another], and the second is to develop a parallel understanding of successful L2 users. Both lines of research are pivotal, not just to solving the native speaker conundrum, but more importantly, to establishing a scientific basis for SLA research practice as well as for L2 teaching, learning and testing. (Han 2004: 185)

This research, in particular Chapters 3 and 5, represents an attempt to address these tasks from the perspective of both NS and NNS perception of error.

## **2.3 Fluency**

In Module II of this research, Brumfit’s (1979: 115) definition of fluency as the learner’s

‘truly internalized grammar’, as represented in the learner’s ‘natural language use, whether or not it results in native-speaker-like language comprehension or production’ (Brumfit 1984: 56) was employed as a starting-point for the investigation of a methodology for fluency development, and it was shown that the methodology did indeed permit such ‘natural language use’. In addition, distinctions were made between more or less fluent speakers, distinctions which pointed to a conceptualization of fluency that went beyond Brumfit’s. For instance, Levelt points to the necessity, in fluent speech, for several systems to operate in parallel:

Most of the components underlying the production of speech, I will argue, function in a highly automatic, reflex-like way. This automaticity makes it possible for them to work in parallel, which is a main condition for the generation of uninterrupted fluent speech. (Levelt 1989: 2)

This automaticity, in second-language speaking, must also be an indicator of the learner’s ‘truly internalized grammar’, in other words, of IL systematicity. Gatbonton and Segalowitz similarly note:

In a more psychological sense, automaticity refers to the operation of those mechanisms underlying performance that function quickly, without interference from other ongoing cognitive processes, and that draw relatively little or no attentional resources away from other concurrent processing activities. (Gatbonton et al. 1988: 474).

They also suggest a distinction between ‘skills concerned with the selection of utterances (knowing what to say, to whom, and when) and skills concerned with the actual production of these utterances (producing them rapidly and smoothly, without hesitations and pauses)’ (p. 473) and note that while both are integral to a definition of fluency, the two can develop independently. Thus their descriptions of fluency range from a narrow ‘being able to execute a basic repertoire of commonly needed phrases with little effort’ (p. 476), to a more general ‘rapid, effortless speech production, or automatization’ (p. 478). They make the connection between this repertoire and automatization quite explicit:



Utterances that are automatized in the senses just described resemble in many ways what others have called formulaic speech, or speech forms produced as unanalyzed wholes, prepatterned expressions, or routinized utterances. [...] Many authors have recognized that such forms of automatic speech have a natural place in both L1 and L2 development (Gatbonton et al. 1988: 474).

Empirical research has attempted to establish operational definitions of fluency as used in the field and to investigate the relationship between the quantifiable temporal factors in speech production and the subjective assessment of fluency by human judges, especially in the context of proficiency testing. For example, Lennon (1990: 404-405), building on work by Levelt (1983; 1989), identified twelve features, broadly grouped into two categories which he calls ‘temporal components’ and ‘vocal dysfluency markers’:

Temporal components:

- a. Words per minute (both raw count and ‘pruned’ of filler and repetitions)
- b. Total unfilled pause time as % of total delivery
- c. Mean word count of ‘runs’ between pauses
- d. % of T-units (independent clauses, accompanied by any associated dependent clauses) followed by pause (filled and unfilled)
- e. % of total pause time at all T-unit boundaries (filled and unfilled)
- f. mean pause time at T-unit boundaries (filled and unfilled)

Vocal dysfluency markers:

- g. repetitions per T-unit
- h. self-corrections per T-unit
- i. filled pauses per T-unit
- j. % of repeated and self-corrected words
- k. Total filled pause time as % of total delivery

Lennon found significant correlations between three of these features (words per minute, % of T-units followed by pause, and filled pauses per T-unit) and the scores assigned by human judges. He did not dismiss the others as potential fluency indicators, but interestingly concluded that self-correction seemed to be ‘a very poor fluency indicator’ (p. 412) because

‘the increased ability to reformulate, monitor, and self-correct production on-line’ may be a part of fluency development in advanced learners (p. 413) and in practice helps to make oral production generally more comprehensible, and therefore perceptibly more fluent to judges. It is unclear whether this finding can be stretched so far as to mean that increasing self-correction can be interpreted as a sign of increasing fluency, particularly in the case of the stimulated recall, elicited imitation, or correction task employed in Chapter 4. At the very least, however, self-correction can be seen as evidence of increasing accuracy or awareness of error.

Subsequent research in this area has refined Lennon’s taxonomy, largely from the perspective of identifying temporal features which correlate most highly with the ratings for fluency given by human judges (e.g. Chambers 1997; Kormos and Dénes 2004; see R. Ellis and Barkhuizen 2005: 139-164 for an overview). Elements of such taxonomies clearly apply only to ‘natural language use’, or at least only to spontaneous speech, which the Running List Test (RLT) employed in Chapter 4 is not. For instance, we would not expect filled pauses to play much part in a test in which what is to be communicated is predetermined (see, for example, Clark and Fox Tree 2002) and the content planned (Skehan 2009) to a greater or lesser extent. For the same reason, it makes little sense to count the mean length of runs and T-units, or to use either as reference points in the measurement of other (dys)fluency markers. These measures would apply, of course, to the Small Talk conversations themselves, but not in the context of the RLT. On the other hand, unfilled pauses during a timed test can be assumed to represent a lack of automaticity, and therefore of fluency. The following, then, are the criteria that will be used to measure the fluency of reformulations in Chapter 4:

- a. Speech rate (speed of delivery)
- b. Total unfilled pause time as % of total delivery
- c. Repetitions or false starts (distinct from self-corrections)

## 2.4 The connection between complexity, accuracy, and fluency

Recent research into complexity, accuracy, and fluency (CAF) has revealed interesting interrelations between perceptions of these constructs, and has suggested a pivotal role for formulaicity in the development of each (Kormos and Dénes 2004; N. Ellis et al. 2008; R. Ellis 2009; Housen and Kuiken 2009; Norris and Ortega 2009; Robinson et al. 2009). Kormos and Dénes, for instance, note that increased L2 proficiency depends on the acquisition of ‘automatic sequences’:

Our research suggests that accuracy also plays an important role in fluency judgements and sometimes overrides the effect of temporal factors on listeners... The correlations between the temporal and linguistic variables also reveal that accuracy is positively related to temporal variables that are influential in fluency judgements. In other words, it seems that those students who were fluent in terms of speed and pace also produced accurate output. In psycholinguistic terms this means that one is only able to speak fluently if speech production mechanisms are largely automatic and if automatic sequences are memorised, retrieved and used accurately (see Schmidt, 1992 for a review). Low-proficiency students generally cannot rely on a sufficient number of automatic sequences and apply conscious rule-based mechanisms, and if they strive to be highly accurate, their speech becomes very slow. Thus in certain cases especially among less competent speakers, speed and accuracy might be in inverse relationship with each other. (Kormos and Dénes 2004: 158–60)

Interestingly, this view reverses the commonly held assumption that formulaic utterances are a ‘crutch’ for L2 learners (e.g. Larsen-Freeman 2009: 579), and instead makes conscious, rule-based production the hallmark of the beginner:

This information-processing model assumes that, in the initial stages of learning, controlled processes are adopted and used to perform accurately and are in effect the ‘stepping stones’ for the development of subsequent automatic processes. (McLaughlin 1990: 620)

And thus:

Instead of relying, then, on ‘unconscious’ knowledge of an elaborate, formal grammar of rules, one can think of linguistic behavior as the product of a rather heterogeneous compilation of associated memories. (McLaughlin 1990: 624)

This view is reflected in the ‘idiom principle’ postulated by Sinclair:

A language user has available to him or her a large number of semi-preconstructed phrases that constitute single choices, even though they might appear to be analyzable

into segments. To some extent this may reflect the recurrence of similar situations in human affairs; it may illustrate a natural tendency to economy of effort; or it may be motivated in part by the exigencies of real-time conversation. (Sinclair 1991: 110)

In the same vein, Widdowson (1989), for example, examines the ‘packaging view’ that learners approach the L2 not with generative systems but with ‘chunks, prefabricated routines, or unopened packages’ (p.135). He reasons that

communicative competence is not a matter of knowing rules for the composition of sentences and being able to employ such rules to assemble expressions from scratch as and when occasion requires. It is much more a matter of knowing a stock of partially pre-assembled patterns, formulaic frameworks, and a kit of rules, so to speak, and being able to apply the rules to make whatever adjustments are necessary according to contextual demands. Communicative competence in this view is essentially a matter of adaptation, and rules are not generative but regulative and subservient. This is why the Chomsky concept [of competence] cannot be incorporated into a schema for communicative competence. (Widdowson 1989: 135)

Or, as Bateson (1979: 17) succinctly put it, ‘context shaping is just another term for grammar.’

This research, while not denying the possibility of innate and universal constraints on syntactic development, assumes that the primary language acquisition mechanism in SLA is related to pattern detection and recall, and that the abstraction of regularities as ‘rules’ is a metalinguistic, rather than linguistic, skill. Thus it concurs with Wray’s view:

The skill of the syllabus writer and teacher lies in adequately juxtaposing the learners’ ability to accumulate linguistic repertoire through the observation of language in use, with their predilection to apply conscious analysis. It is observation which will best support the developing sense of what ‘sounds right’ in a given context. But it is analysis that will make up the shortfall between what the classroom context can provide and the creative linguistic knowledge which the learner needs to develop. (Wray 2000: 484)

Since this research concerns CF, it must also acknowledge the unfortunate fact that L2 learners are judged by standards which may be more rigorous than those customarily applied to native speakers (NS). Lennon, citing a British TV sports commentator’s metaphor-mixing gaffe, ‘I think the Italians are going to have their hands cut out tonight,’ cautions that ‘native speaker fluency may be purchased only at the price of errors that would be unacceptable in

the nonnative speaker' (Lennon 1990: 396). He cites Baars' (1980) hypothesis that such 'blends' in fluent speech production originate in parallel and competing speech plans, a hypothesis which could as reasonably be applied to L2 speech production as to L1. The difference is, presumably, that the NS slip would be recognized as such if pointed out, whereas the same cannot be said with certainty of the NNS.

Lennon's use of 'unacceptable' may be overstating the case: most language teachers are aware of the distinction between slips and more systematic errors and can provide CF accordingly. However, this may be an argument for delayed CF over the immediate kind generally offered through recasts, repetition, and so forth in the course of classroom activities: pointing out a slip during an interaction may be quite disruptive, especially in the case where the speaker is unaware of having made one. Doing so forces the speaker to focus explicitly on language which, as far as she is concerned, was perfectly acceptable, but which is now being questioned. In contrast, delayed CF, situated outside of the immediate communicative context, allows for evaluation of and reflection on language use and allows the speaker to compare the produced form with her own intuitions as well as the provided reformulation (c.f. Sheen 2004; R. Ellis and Sheen 2006), and causes no loss of face (Kobayashi 1995). Presumably, 'slips' will be immediately recognizable as such by the speaker and far more easily 'correctable' than systematic errors. This hypothesis will be tested in Chapters 4 and 5.

## **2.5 Methodological issues**

### **2.5.1 Grammaticality judgements**

While each study in the research that follows includes a review of the related methodological literature, the following sections will discuss specific methodological issues related to the Timed Grammaticality Judgement tasks in Chapters 3 and 5.

The use of grammaticality judgement tests (GJTs) as evidence for implicit linguistic knowledge in native speakers remains controversial owing to a number of legitimate concerns about confounds such as test modality, sentence parsability and processability, semantic interpretation, memory limitations, and so forth (Chaudron 1983; Birdsong 1989; Schütze 1996; Murphy 1997; Rimmer 2006; Dąbrowska 2010). In addition, by some definitions they are not judgements of *grammaticality* at all:

The notion ‘acceptable’ is not to be confused with ‘grammatical.’ Acceptability is a concept that belongs to the study of performance, whereas grammaticalness belongs to the study of competence... Grammaticalness is only one of many factors that interact to determine acceptability. (Chomsky 1965: 11)

Acceptability has been defined as ‘the feelings speakers have about the well-formedness of sentences in their language’ (Newmeyer 1983: 51). I will not attempt to debate this point, which has been amply discussed elsewhere (Chaudron 1983; Birdsong 1989; Schütze 1996; Mandell 1999; Riemer 2009). In this research, it is accepted that these judgements are a performance task; that is, they do not offer a direct window onto L2 competence. Instead, as Carroll suggests,

[i]t seems far more reasonable to assume that these judgements involve an interaction between metalinguistic knowledge (encoded in the conceptual system), the internalized grammar, and the learner’s perceptual systems. (Carroll 2001: 186)

The use of grammaticality judgement tests (GJTs) as a source of data on IL development in second-language learners is even more controversial since several of the confounds listed above become even more serious when no claim can be made as to homogeneity of competence, as would certainly be the case with individual ILs (Schachter 1976; Gass 1983; R. Ellis 1991; Han 2000; El-dali 2010). Nevertheless, GJTs remain a common tool in SLA research because the data are so readily available, and because of an intuitive sense (or perhaps blind faith) that metalinguistic performance, whatever its origins, means *something*:

It is quite easy to imagine systematic grammar-based communication systems which are very poorly designed for the task of making grammaticality judgements on

arbitrary strings of words. In the case of human language (viewed as an idea-expressing and communicating system), it is hard even to see what particular (evolutionary?) functionality this ability might have. Thus its existence might be considered a fortunate accident. (Bley-Vroman and Masterson 1989: 209)

Perhaps the key to understanding the true utility of GJTs, as well as their controversial status in the field, lies behind the choice of the phrase ‘arbitrary strings of words’ in the above quotation: it is indeed difficult to see what evolutionary advantage might ensue from the metalinguistic ability to judge the grammaticality of such strings as:

*?The plane that the pilot that the police questioned flew crashed.* (Rimmer 2006: 253)

*\*What does Mary want to know whether John has already sold?* (Bley-Vroman et al. 1988: 8)

These kinds of ‘arbitrary strings’ are designed to test the boundaries of syntactic competence and are truly ‘underdetermined by the input’ – and deliberately concocted for that reason. Nobody, or precious few at least, actually says things like *The plane that the pilot that the police questioned flew crashed*, least of all second-language learners, and this fact is frequently used as evidence for innate syntactic competence when those same learners are able to correctly identify grammaticality without prior exposure. However, if one looks at less arbitrary strings, for instance the kinds of strings that language learners – whether L1 or L2 – actually *do* produce, one can very quickly see the evolutionary advantage of an ability to recognize grammaticality:

*I’m going to choose it off.* (a candy cane from a Christmas tree) (Bowerman 1982: 112)

*I’m patting her wet.* (Bowerman 1982: 113)

These examples, from NS English children (aged 3:11 and 4:00 respectively) show an already developed ability to identify constructions (Goldberg 1995; 2006; 2011) and to creatively build novel ones, but novel ones which are endearingly unlike adult patterns. Without the ability to judge her own strings against the data of the linguistic environment, the child would have no cause to abandon these items (and other, non-target-like constructions), and would

not develop language resembling that of her speech community. In fact, there could be no speech communities at all if humans did not share this ability to evaluate their (novice) language output against that of the (expert) speakers around them. And, with the exception of phonology, this is presumably not an ability with a critical or sensitive period, since the need to blend in, to belong, or conversely to mark one's distinctiveness and otherness, lies at the very heart of human social organization.

It is in part the search for this type of metalinguistic awareness that motivates this research. The underlying question is whether it applies to L2 learning, and if so to what extent. The motivation to identify with the target speech community is far more complex in L2 than in L1 learning (Norton Pierce 1995; Pavlenko and Blackledge 2004), and the differences presumably go some way towards explaining differential end-state achievement. But equally important must be the availability of linguistic input with which to fine-tune one's production. Since this input may indeed be restricted in the case of L2 acquisition, and since the linguistic environment of the TL and the local speech community are often not the same (e.g. EFL situations), it would be surprising if a mechanism such as that described above did not produce some very *un*-target-like forms.

However they are processed and whatever they are named (this research follows R. Ellis 1991 and many others in referring to them as 'grammaticality judgements'), there seems little debate that these judgements do reflect our intuitions about well-formedness, and as such can contribute to our understanding of learner knowledge, as Gass suggests:

[I]ntuitional data, as a reflection of metalinguistic awareness, are important in second language research both in and of themselves for what they reveal about language learning, and also because they provide us with a crucial aspect of a learner's knowledge, an aspect without which we cannot hope to gain a complete picture of the second language acquisition process. (Gass 1983: 287-288)

In the pilot study in Chapter 3, a set of language items, taken from Small Talk sessions and



recorded by a NS, is used to explore the responses of NS and NNS to error and well-formedness. In the study in Chapter 5, learners were asked to judge the well-formedness of their own utterances, recorded by themselves, an approach not reported in any published study to date.

### 2.5.2 Reaction time

According to Juffs (2001: 207-208), reaction time (RT) measures have only recently begun to feature in SLA research, and studies have employed them to investigate a variety of behaviours such as *wh*-extraction (Juffs and Harrington 1995), developmental changes in the nature and timing of sentence interpretation (Devescovi et al. 1999), and transfer of the 'Functional Categories Parameter' (Bley-Vroman and Masterson 1989). In theory, one advantage of RT measures is that they permit a degree of differentiation between more and less automatic judgements, or between judgements based on implicit as opposed to explicit linguistic knowledge (Han and R. Ellis 1998; R. Ellis 2005b; Cook 2009). Bley-Vroman and Masterson propose two ways in which this might work:

One plausible theory is that when the sentences are grammatical, the language processing system immediately and automatically produces a unified high-level representation of the examples; and identity can be determined on the basis of comparing unitary representations at this level. When the examples are not grammatical, however, no high-level representations can be computed, and matching must be done by other less efficient strategies, for example, by a word-by-word comparison... A second possible explanation is that when an example is ungrammatical, the mind constructs two representations: one of the example as is, another of a corrected version. In this second explanation it is mental correction rather than ungrammaticality per se which slows down matching time for ungrammatical examples. (Bley-Vroman and Masterson 1989: 214-215)

This research follows these authors in their conclusion that what is crucial is not the explanation of the phenomenon but its reality. It is hypothesized that learners will judge items which sound correct to them (whether they are well-formed or not) faster than items about which they are unsure. Figure 2 shows an example of one participant's judgements from the

study in Chapter 3, which have been standardized and arranged from fastest at the top to slowest at the bottom. The horizontal axis shows the mean (0) and increments of one-half standard deviation (*SD*) above and below the mean. Thus, the participant judged item 10 (at the top) at a speed 1 *SD* faster than the mean, while item 3 at the bottom was 3 *SD* slower than the mean. In this case, one item (8) was incorrectly judged at a speed close to the mean,

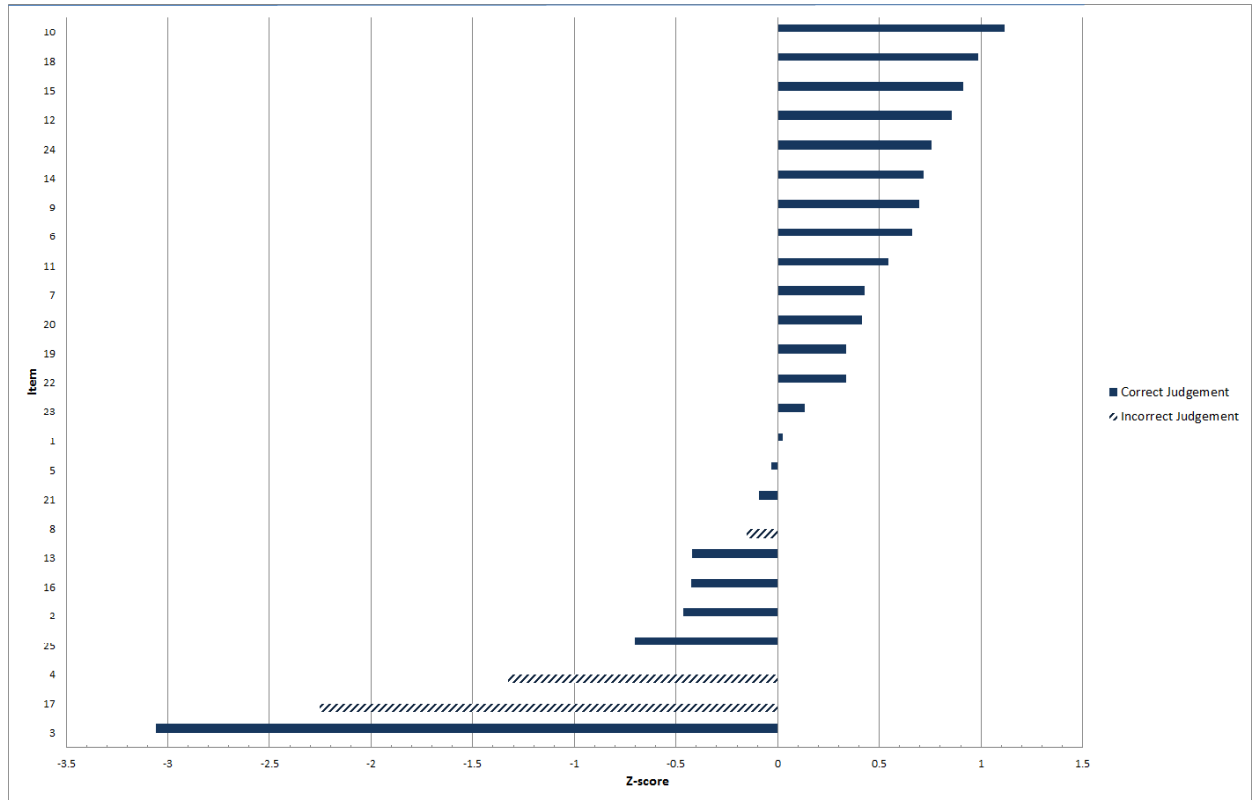


Figure 2: Example of standardized scores from timed grammaticality judgement test

and two incorrect judgements (4 and 17) were made at speeds greater than 1 and 2 *SD* below the mean, respectively. According to the hypothesis above, it is assumed that the participant is unsure of the well-formedness of items 4 and 17, but is more confident (but incorrect) about item 8. This approach to the measurement of RT is also used in the measurement of fluency in the elicited imitation test (Chapter 4) as it allows participants' performance on each item to be measured against a baseline of their own mean fluency.

### 2.5.3 The ecological validity of teacher research

While validity is, or should be, an overriding concern in any research, a central concern throughout this research has been its ecological validity. In essence, any research into pedagogical issues by teachers is by definition *action research* (Freeman 1998; Burns 2010), and the present study is no exception. Therefore, a primary goal of this study is to test the strengths and weaknesses of the methodologies described herein, and to make adjustments where necessary which are grounded in observation and data collection rather than intuition alone. Put another way, while the intuitions of experienced teachers are by no means to be dismissed as being too subjective or lacking validity, there has to be a reasonable level of corroboration from other data sources in order to support or test those intuitions.

As the ELT/TESOL field has matured, it has grown wary of theory-driven instructional methodologies, having experienced a half century of competing ‘best’ instructional methods (Kelly 1969; Howatt 1984; Richards and Rodgers 1986). Many in the field have adopted what Kumaravadivelu calls a ‘postmethod pedagogy’ (Kumaravadivelu 2003; 2006), characterized by a resistance to top-down (theorizer-to-practitioner) methodological innovation, by teacher autonomy, and by what he calls ‘principled pragmatism’, meaning an instructional approach based on teachers’ own experience as learners and teachers, and through professional development and consultation and collaboration with peers.

This is not to say that the products of ‘principled pragmatism’ cannot themselves be the object of investigative research. They not only can, but should: however much teachers may intuitively ‘know’ that a given pedagogical approach is effective, empirical investigation is always necessary and should always be welcome. Without critical scrutiny, our pedagogy can only develop in limited and perhaps arbitrary ways. This is the motivation for calls by

many in the field (e.g. Burns 2010; Edge 2001; 2011; Edge and Richards 1993; Freeman 1998; ) for greater involvement by teachers in the research of theoretical and especially instructional issues.

Because the challenge described in Section 1.1 to track individual language development systematically in teaching contexts primarily concerns pedagogy rather than theory development, the overarching concern of this research has been its ecological validity and sustainability (van Lier 2008: 602; Tudor 2003; Kramsch and Steffensen 2008). The ecological perspective views the classroom as having a pedagogical and social relevance beyond that of the research agenda, and attempts to focus attention ‘on the subjective reality which various aspects of the teaching–learning process assume for participants, and on the dynamic interaction between methodology and context’ (Tudor 2003: 1). In terms of empirical research, this position would argue for perspectives and methodologies that not only acknowledge this subjective reality and the complexity of the teaching and learning context, but which also contribute to and ideally arise from the ongoing pedagogical activities. I would go one step further and argue that because CF itself is utterly context-dependent, ecologically valid CF research can only be achieved through classroom-based research. At the same time, empirical confirmation of the sort described above underpins the contribution to and development of second language acquisition theory.

The context-dependence of ecological research has been recognized as a limitation by some:

An ecological research approach offers more internal validity (appropriately called ecological validity) but less reliability and inordinately less generalizability or external validity. (Kramsch and Steffensen 2008: 26)

However, replication of studies in diverse contexts is a potential solution to the lack of generalizability engendered by ecological research. To this end, the studies in Chapters 4 and

5 were conducted in two very different ESL programs in which Small Talk and the CF methodology were part of the regular curriculum.

---

## CHAPTER 3: PILOT STUDY OF A TIMED GRAMMATICALITY JUDGMENT TEST

---

### 3.1 Introduction

In order to establish appropriate design and procedures for the study in Chapter 5, a pilot study of the timed grammaticality judgement test (TGJT) was conducted. This also helped to establish the level of agreement between proficient English speakers in judging individual utterances grammatical or ungrammatical, and to investigate the range of reaction times within and between individuals. A further purpose was to expose any response bias that might exist (Birdsong 1989) that is, a tendency for participants to favour one type of judgement (e.g. grammatical) over the other – see Section 3.5.2. The pilot study took place in two stages: the first involved the design and implementation of the online test platform and software, the selection of items for the TGJT, pilot testing with volunteer participants, and exploratory analysis of the results; in the second stage, a subset of items was chosen based on the level of consensus between the proficient (teacher) volunteers in stage one. Adjustments were made to the test platform in order to eliminate technical problems and to clarify instructions.

Volunteers were solicited through the TESL-L listserv, and the online test was deployed for two months, from March to May 2011.

### 3.2 Research questions

The pilot study specifically addressed the following questions:

- 1) Is there a mean difference in either accuracy or reaction time between native speakers (NS) and non-native speakers (NNS) in the identification of errors presented in aural

format?

It is predicted that NS will be both faster and more accurate in their judgements than NNS. However, since most of the participants are teachers, and since the range of test items is restricted, proficient NNS should be indistinguishable from NS in the accuracy, if not the speed, of their responses. It should be noted that the definition of 'native speaker' in Applied Linguistics is controversial, particularly in the case of bilinguals and English varieties around the world (Rampton 1990; Leung et al. 1991; Kachru 1992; Davies 2003; Han 2004; Mahboob 2005). In addition, research has demonstrated that far from having homogenous linguistic competence, native speakers differ greatly in their grammatical knowledge (Dabrowska 2010) and that highly educated non-native speakers can outperform less educated native speakers of English in comprehending grammatically challenging English sentences (Chipere 1998).

- 2) Do proficient speakers show consensus in their judgements of correct and incorrect samples of learner English?

The extensive literature on grammaticality judgements indicates that a number of factors, from presentation order and modality to response bias, can threaten the reliability of metalinguistic performance (see Chaudron 1983; Birdsong 1989; R. Ellis 1991; Schütze 1996; Mandell 1999 for reviews). The hypothesis is that in judging learner language, there will be a high degree of consensus between NS and proficient NNS.

- 3) Is there a measurable relationship between the accuracy and speed of judgements?

The hypothesis is that longer reaction times will be associated with more accurate judgements for both proficient speakers and learners, as confident judgements should be reflected in longer reaction times.

### **3.3 Participants**

In stage one, the participants were six teachers and three ESL students at Gonzaga University's English Language Center (ELC). The teachers are all native speakers of English, with ELT experience ranging from five to twenty years. The students have upper-intermediate level proficiency, all being in the final level of the ELC, with an estimated IELTS total score of 6.0-6.5. (One student in fact took the Academic IELTS during this time period and scored 6.5.)

In stage two, a call for volunteers went out to approximately 4,000 current subscribers to the TESL-L listserv, an electronic discussion forum for professionals in the field of English language education (<http://www.hunter.cuny.edu/~tesl-l/about.html>), and 52 people volunteered to take the TGJT online. Volunteers were asked to self-identify as native or non-native speakers, and while there is no way to be certain whether volunteers were truthful, responses indicate that 38 NS and 14 NNS took the test.

### **3.4 Design and procedures**

#### **3.4.1 Timed judgement tests**

Timed grammaticality judgement tests are commonly used in psycholinguistic research as a means to elicit judgements based on implicit rather than explicit knowledge (Birdsong 1989; R. Ellis 1991; Cook 1994; Han 1996; Reinders 2005; Tremblay 2005). The general approach is to restrict the amount of time that a participant has to make her judgement, either by hiding the item after a given time or by categorizing it as 'no response' if the predetermined time has elapsed. The time allocated varies from 3.5s (Han 2000: 183) to 20s (Tremblay 2005: 143). The problem with this approach is that it assumes that all participants react at the same speed, and it must be remembered that 'react' here means:

- a) reading (or more rarely, listening to) the test item



- b) interpreting it, which entails mapping surface form onto likely linguistic function
- c) verifying one's intuitions that the form and function can be paired
- d) checking again (where possible) that one hasn't mis-read or heard
- e) introspecting to find possible alternative explanations and/or applying prescriptive rules
- f) performing whatever action is required to register a judgement

Although each of these steps can happen extremely quickly, individual variation in steps a) and f) alone make it impossible to know at which point a participant has started to use explicit knowledge (steps d and e), if that is what one is investigating.

An alternative approach is to allow unlimited time but to measure reaction time (RT) as an additional source of information about the process (Bley-Vroman and Masterson 1989). This also allows the researcher to establish a baseline RT for each participant against which each judgement can be compared. The mean and standard deviation (*SD*) of the RT for each participant, along with the mean and *SD* of groups of participants (e.g. NS and NNS, Arabic L1 and Japanese L1) constitute very important psycholinguistic data that have largely been eschewed by researchers in the field, possibly because of the technical challenges and expense involved. As we will see below, commonly available technologies make this far less of a consideration today, and there now seems no reason not to make the measurement of RT a standard practice in TGJT research.

### **3.4.2 Selection of test items**

In stage one, 30 items were chosen from 1,600 utterances collected during Small Talk conversations held between January and June, 2010. The speakers were all students from a public women's college in the United Arab Emirates. The sentences were selected in the

following way: teachers chose ten sentences that they had previously identified as ‘All Do’ on worksheets, in other words, that they had required all students in their class to correct, irrespective of who had said them (see Section 4.1 for details). The criterion for selection of the ten sentences was whether the teacher considered each item to be a ‘typical’ error for his or her students. From the resulting list of 50 sentences, three teachers by consensus eliminated all that were either covert errors (superficially well formed) or that were difficult to interpret without contextual knowledge. This resulted in a list of 30, shown in the left-hand column in Table 1. For half of these items, the reformulations provided by the teachers who had noted the error during the original conversation in which it occurred (the Small Talk/CF context)

Table 1: *Items for pilot TGJT*

Original sentence from worksheet	Reformulation
1. That’s something, يعني, I hate it.	That’s something, you know, I hate.
2. That’s make you happy?	—
3. There is some shops which sell this.	There are some shops which sell this.
4. You don’t mind to marry a smoker?	—
5. That’s will affect their grades.	—
6. Anything that you use it in daily life.	—
7. Can you tell us what does the gift mean to you?	—
8. I will buy something I know they like it.	I will buy something I know they like.
9. I know her from school.	I have known her since school.
10. Nobody knows who is Batman.	Nobody knows who Batman is.
11. In the past, the womans wear the traditional clothes.	—
12. Yeah, actually I’m agree with you.	—
13. Do you think it’s help reduce the traffic?	Do you think it helps reduce the traffic?
14. Because it’s reduce the traffic problem.	—
15. Each person in the family have one car.	Each person in the family has one car.
16. The government should encourage locals using public transportation.	The government should encourage locals to use public transportation.
17. I think it will have a big change in my life.	I think it will have a big effect on my life.
18. They do stuff that it’s not allowed here.	—
19. I think love is much important than money.	I think love is much more important than money.
20. The fees it’s very expensive.	—
21. There are too much building.	There are too many buildings.

Table 1 (*cont.*)

22. What you mean, ‘crime’?	What do you mean, ‘crime’?
23. They also affects in our children.	—
24. It’s make them happy.	—
25. We had some agree and some disagree.	We had some agreements and some disagreements.
26. It’s show that I am relaxed.	It shows that I am relaxed.
27. What it’s mean?	—
28. Why you don’t believe this?	—
29. Do you think he will say for you the truth?	Do you think he will tell you the truth?
30. Why the people are doing this?	—

were substituted for the error. The reformulations are shown in the right-hand column in Table 1.

### 3.4.3 Linguistic profile of the pilot test items

The original items were analysed according to the phonological, morphological, syntactic, lexical, and discourse features that make them non-target-like. The purpose of this step was not only to determine what kinds of errors the teachers had selected as being most typical, but also to ascertain whether these kinds of errors are uniquely or predominantly characteristic of L1 Arabic speakers. If this were the case, it would limit the generalizability of the findings but might lend additional support to the substantial body of research on language transfer in interlanguage development. The analysis is shown, with relevant errors underlined, in Table 2.

Table 2: *Analysis of 30 typical errors from intermediate L1 Arabic speakers*

error domain/error type	examples
<b>phonological</b>	[none]
<b>morphological</b>	
V inflection	That’s <u>s</u> make you happy? That’s <u>s</u> will affect their grades. Yeah, actually I’m agree with you.

Table 2 (cont.)

	<p>Do you think <u>it's</u> help reduce the traffic?          Because <u>it's</u> reduce the traffic problem.  <u>It's</u> make them happy.  <u>It's</u> show that I am relaxed.          What <u>it's</u> mean?</p>
<b>syntactic</b>	
NP determination	<p>There are <u>too much</u> building.          In the past, <u>the womans</u> wear <u>the traditional clothes</u>.          Why <u>the people</u> are doing this?</p>
Direct question in place of Noun Clause	<p>Can you tell us what <u>does</u> the gift <u>mean</u> to you?          Nobody knows who <u>is Batman</u>.</p>
Topic/Subject duplication	<p>The fees <u>it's</u> very expensive.</p>
S-V agreement	<p>There <u>is</u> some shops which sell this.          Each person in the family <u>have</u> one car.</p> <p>The fees <u>it's</u> very expensive.          There <u>are</u> too much building.          They also <u>affects</u> in our children.</p>
V complementation	<p>You don't mind <u>to marry</u> a smoker?          The government should encourage locals <u>using</u> public transportation.</p>
Transitivity	<p>They also affects <u>in</u> our children.</p>
Missing AuxV in question	<p>What <u>___</u> you mean, 'crime'?</p>
S-V inversion in question	<p>Why <u>you don't</u> believe this?          Why <u>the people are</u> doing this?</p>
Adjective phrase	<p>I think love is <u>much important</u> than money.</p>
Adjective clause [Subj Pn retention] [Obj Pn retention]	<p>They do stuff that <u>it's</u> not allowed here.          That's something, <u>يعني</u>, I hate <u>it</u>.          Anything that you use <u>it</u> in daily life.          I will buy something I know they like <u>it</u>.</p>
<b>lexical</b>	
word choice/collocation/idiom	<p>I know her <u>from</u> school.          I think it will have a big <u>change in</u> my life.          Do you think he will <u>say for</u> you the truth?</p>
word form	<p>In the past, the <u>womans</u> wear the traditional clothes.          We had some <u>agree</u> and some <u>disagree</u>.</p>

Table 2 (*cont.*)**discourse**

Tense sequencing

I know her from school.In the past, the womans wear the traditional clothes.

code-switching

That's something, يعني, I hate it.

The error domains and types are part of the error taxonomy discussed in Chapter 6. For the present, it is important to note that since the data are authentic learner production, there are seven items, 23% of the total, in which there is more than one error type. In one case, sentence 11, *In the past, the womans wear the traditional clothes*, there are three distinct error types, one of which (NP determination) occurs twice. It is common practice in tests of grammaticality to create test items which not only target one structure only but which also contain no other errors (Chaudron 1983), and with good reason. If an item with multiple errors is judged to be incorrect, the researcher has no way to determine which of the errors triggered the judgement. However, this study is concerned less with the investigation of specific structural features in interlanguage grammars than with the hearer or learner's ability to detect ungrammaticality at all. Therefore, of these seven items with multiple errors, the four that were not reformulated were left with all errors intact.

For the purpose of investigating whether the thirty items represent uniquely Arabic L1 errors, the database compiled from Small Talk errors (see Chapter 6) was queried. In all cases except two, equivalent examples from speakers of other languages and at other proficiency levels could be identified. Equivalent errors from speakers of other L1s are shown with examples in Table 3.

Table 3: *Examples of equivalent errors to the test items found from speakers of L1s other than Arabic in the Small Talk database*

<b>error type/ sub-categorization</b>	<b>examples of equivalent errors</b>	<b>L1s</b>
<b>morphological</b>		
V inflection	It is help me. I say ‘yes,’ it’s mean ‘yes.’	Korean, Mandarin, Japanese, Spanish
<b>syntactic</b>		
NP determination	Sometimes the womans are very crazy.	French, Korean, Mandarin, Japanese, Spanish
Direct question in place of Noun Clause	Do you know what’s the story about Christmas?	French, Korean, Mandarin, Portuguese, Japanese, Spanish
Topic/Subject duplication	Your parents they no successful in school.	French, Spanish
SV agreement	There is seven people.	Korean, Mandarin, Japanese, Spanish, Vietnamese
V complementation	They don’t mind their kids to be heterosexual.	Korean, Japanese
Transitivity	How the past life affect on now.	all
Missing AuxV in question	What your mom said?	all
S-V inversion in question	What you did in restaurant?	all
Adjective phrase	When Mami went to the Bali for trip, the merchant gave a price much expensive than the regular price.	Korean, Japanese, Thai
Adjective clause	[subject pronoun retention – none]  This is a thing that we hit it, and shoot opponent’s goal.	[none]  Korean, Japanese, Spanish
<b>lexical</b>		
word choice/collocation/ idiom	I am angry when the people say me something wrong.	Korean, Japanese, Spanish
word form	We have to avoid being addictive or dependence.	all

Table 3 (*cont.*)**discourse**

Tense sequencing

We started English since junior high school.  
In the past, we can use computer to send money to other people.

Korean, Mandarin,  
Japanese, Spanish

code-switching

[none]

[none]

The fact that no code-switching examples could be found in the database is not surprising given the multilingual composition of the students in the English Language Center, from whom most of the Small Talk data has been gathered. This is not to say that students never code-switch for purposes such as filled pauses or other communicative strategies, rather that teachers in this program do not seem to put such items on CF worksheets. The other error type that could not be located in the database from non-Arabic L1 learners was adjective clause subject retention (*They do stuff that it's not allowed here*). From a language transfer perspective (e.g. Kharma 1987; Swan and Smith 2001; Harrat 2011; see also Lewkowicz 1971 for a discussion of topic-comment sentences embedded as adjective clauses in Arabic) this is not a surprising finding, as this error is widely acknowledged to be attributable to Arabic syntax. However, it should be noted that the absence in the database of a particular error type in the production of speakers of any particular L1 could also be attributable to sampling methodology (discussed in Hunter 2012) and cannot at this point be construed as evidence in support of or against any theoretical position.

The remaining error types, as Table 3 shows, also occurred in the production of other L1 learners. This implies that these errors are commonplace enough to be appropriate material for a TGJT measuring error sensitivity.

#### **3.4.4 Recording of the test items**

The choice of presentation modality – written or aural – might reasonably be said to put certain learners at a disadvantage, especially those with limited exposure to oral input. However, to be consistent with the CF methodology, the TGJT presented items to participants in aural, not written format. Another motivation for this choice is a study by R. Ellis et al. (2009: 112), in which the authors speculate: ‘The online processing required for the oral modality would arguably encourage learners to draw upon their implicit L2 knowledge, while avoiding the possible stress of a computerized timed GJT.’ Unfortunately, they do not elaborate on the latter point, but as outlined below, the participants in this study did not report stress of any kind. It is worth noting in passing that very few researchers use an aural format on GJTs in SLA research: in Chaudron’s (1983) survey, only five out of 23 studies conducted between 1960 and 1980 used an aural format, and more recent surveys (e.g. R. Ellis 1991; Reinders 2005) indicate that this trend has not changed.

The 30 test items, and an additional two example sentences, were recorded by a native speaker of British RP English with a slight American accent. Care was taken to enunciate clearly without putting undue emphasis on erroneous items. The Audacity audio software version 1.3.2 was used for the recordings, which were made in 32-bit format with a sample rate of 44.1KHz, and encoded into 30 separate mp3 files in 16-bit format with a bitrate of 128. This represented a compromise between audio clarity and file size, which could affect data transfer and retrieval speed online, but no participant reported lack of audio clarity as being a confounding factor. The audio files were uploaded to a web server.

#### **3.4.5 Design of the testing platform**

The platform for the TGJT was deployed online to enable teacher and learners in diverse locations to participate. This entailed creating a server-hosted MySQL database to store the



data from the test; a web-based interface to present the items to the users and log their judgements and reaction times; and server-side .php scripting to communicate between the two. The interface was designed using the Adobe Flash CS2 authoring software and exported to the web as Shockwave Flash (.swf) file.

The MySQL database table that stored the TGJT data contained an autonumber ID field and fields for Username, URL (web address) of the test item sound file (which also served as an identifier for the test item itself), the respondent's judgement (0 for Incorrect and 1 for Correct), the response time for that item (in milliseconds), and an automatically-generated timestamp. The timestamp was included so that should any participant attempt the TGJT twice, only the first attempt would be counted to counteract any practice effect.

There was a concern that measuring reaction times online using Flash might introduce distortions in the timing which would invalidate the results. However, Reimers and Stewart (2007: 17) found that Flash thus deployed did 'not appear to introduce significant random error to RT [reaction time] measurements', but was on average 30-40ms slower than baseline conditions. This was felt to be an acceptable distortion since reaction times in SLA research are typically measured in 100ms increments (e.g. Bley-Vroman and Masterson 1989: 222).

The Flash web interface<sup>1</sup> consists of four frames: the first asks users to log in using a self-chosen username and prearranged password, and asks them to specify whether they consider themselves a native speaker of English. Also included on the first frame is a button which produces a simple sine wave tone when pushed, together with instructions to users to test their sound before continuing. Once they have done this, the second frame appears which gives instructions for the TGJT (Figure 3).

---

<sup>1</sup> The interface can be seen at: <http://www.celticnotes.com/SMALLTALK/errorstestset.html>; any name can be used as the Username, and the password is 'elc'.



*Figure 3: Instructions for the TGJT*

After the user clicks the Start button, the third frame is presented, which itself has a Start button. This is to ensure the user is not surprised by the presentation of the first example sentence. Once this is clicked, the first mp3 file (example sentence 1) is streamed from the server and played, and as soon as the end of the file is reached, the 'Incorrect' and 'Correct' buttons appear, and the `getTimer` function in Flash is called. This scripting function gets the current time in hours, minutes, seconds, and milliseconds from the local computer on which the .swf file is running. As soon as the user clicks either button, the `getTimer` function is called again, and the time difference between the two function calls is calculated. This is sent, via .php script, to the MySQL database along with the other information listed above. The 'Incorrect' and 'Correct' buttons are hidden, and replaced with a button that says 'Next', which repeats the process with the next test item. The user thus can neither skip an item nor backtrack to repeat a previous item: any attempt to do so resets the test completely to the first frame. It is important to note that in contrast to many TGJTs, this test did not limit the amount of time participants could think about an item; it merely asked them to respond 'as quickly as

possible'. This was thought to be the best way to avoid the stresses referred to by R. Ellis et al. (2009: 112) above. It should also be noted that users could not make a judgement before the entire item had been played.

At the bottom of this frame is a text box which indicates how far through the test the user is (e.g. '12 of 30'). Once the final sentence has been reached, the user is presented with the numbers of the items she judged incorrectly, followed by a detailed explanation of the errors in the sentences (Figure 4). Since only the first attempt of any user would be included in the analysis, it was felt that providing the target items would allow participants to check their memory of the sentences against a written version; for teachers, this would permit clarification of any controversial items, while for learners, this might provide useful input on common errors.

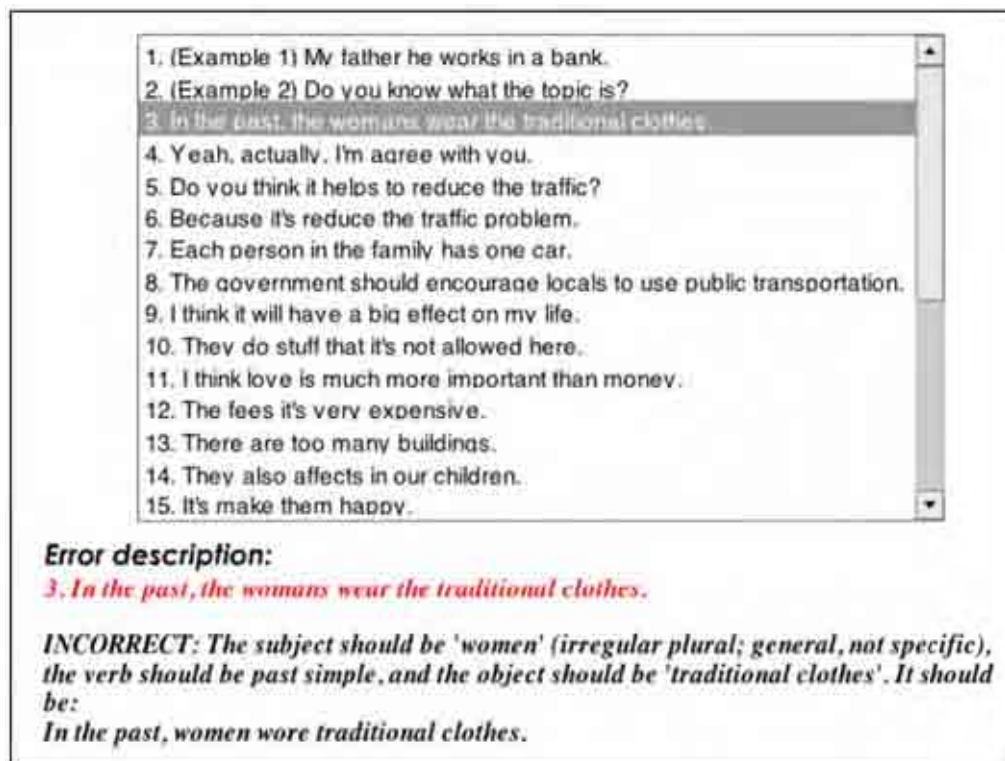


Figure 4: Feedback on the TGJT items

### 3.4.6 Elimination of unreliable items

The nine participants in stage one were asked to take the TGJT as described above. The responses from the six NS teachers were used to calculate an inter-rater reliability statistic for the test set, and it was found to have a Cronbach's alpha of .962. Nevertheless, the average accuracy score was calculated for each sentence, and any that did not reach 85% agreement were eliminated. In total, five sentences, 1, 7, 9, 10, and 22, were removed from the test items, resulting in a revised alpha of .986. Although the sentences were eliminated on a strictly statistical basis, where participants' judgements differed from the anticipated response, they were asked to explain if possible what their thinking had been. In two cases, respondents claimed to hear the items as two sentences, thus rendering them ungrammatical and grammatical, respectively:

- 1. That's something, || you know, I hate.
- 7. Can you tell us || what does the gift mean to you?

In one case, a word was misheard:

- 10. Nobody knows who Batman (misheard as '*bad man*') is.

And for the remaining two items, respondents claimed that they 'sounded odd':

- 9. I have known her since school. (*since school* thought to be non-standard)
- 22. What do you mean, 'crime'? (*What do you mean BY 'crime'?* thought to be a better reformulation)

Since not all of the participants felt the same way about these items, it was considered acceptable to simply eliminate them from the test set rather than re-record them. However, in the case of 9 and 22, a cautionary note was introduced concerning the idiosyncrasy of error *gravity* judgements: whilst the participants admitted that they did not consider the items 'technically' ungrammatical (despite having judged them so), their unfamiliarity with the locutions (both of which appear several times in both the Corpus of Contemporary American

English and the British National Corpus) led them to reject them. The tendency for NS participants to reject the grammatical yet unidiomatic is evidence for the claim that GJTs are in fact not judgements of grammaticality at all, but judgements of *acceptability* (Can 2007: 294 n5; Chomsky 1965: 11; Bard et al. 1996: 33-34). During the course of the pilot study, one participant wrote:

Have you considered specifying the variety of English by which you are judging *correct* or *incorrect* in your error recognition test? I completed the test using Edited American English as the standard; however, my answers would have been quite different if I would have been using another English variety or judging only on competence of communication. (S.Y., personal correspondence)

The fact that ‘S.Y.’ *did* use ‘Edited American English’ as the standard by which to judge the items confirms that more than grammaticality is at stake; but by the same token, her choice of this variety presumably was informed by her awareness that it *is* a standard, possibly the same one she teaches and expects her students to conform to. The final proposition, ‘judging only on competence of communication’ is an interesting one in the context of language teaching: one would be hard pushed to claim that *any* of the items in Table 1 are ‘incorrect’ by the standards of ‘competence of communication’, so long as the latter is taken to mean only that they communicate *something*. This is thin ice, pedagogically speaking, since one would have to agree that ‘communication’ is at least partially in the ear of the listener, which by definition makes it not a standard; furthermore, anything that met this standard would no longer be in need of instruction.

The judgements from three ESL students were not used in the reliability calculation (since it was thought that the variation in their judgements could reasonably be attributed to lack of competence rather than disagreement over the status of the items); nevertheless, group means were compared using Student’s t-test to verify if the set of test items would produce significantly different responses from the two groups. As anticipated, the response time (in

milliseconds) of the teacher group ( $M = 1043$ ,  $SD = 1005$ ) was significantly faster than the student group ( $M = 1965$ ,  $SD = 2051$ ),  $t(1) = -3.06$ ,  $p = .003$ . In addition, the accuracy (max. = 1) for the teacher group ( $M = .98$ ,  $SD = .14$ ) was significantly higher than the student group ( $M = .64$ ,  $SD = .49$ ),  $t(1) = 4.9$ ,  $p < .001$ . Thus, the six teachers had a mean RT and  $SD$  of RT of just over one second, and achieved 98% accuracy, while the three students had a mean RT and  $SD$  of RT of approximately two seconds and 64% accuracy. A note is required here concerning the  $SD$  figures and their purpose: it is not unusual for  $SD$  figures to exceed the mean (this is, after all, what a Normal Distribution z-table does, setting the mean at 0 and  $SD$  at 1), but it can indicate extreme variation or the presence of outliers in the data. In many studies, these outliers would be eliminated from calculations or replaced with the mean, on the assumption that they are anomalous (for instance, a participant gets momentarily distracted, causing a response time three times her average). However, in a test where participants are asked to judge the well-formedness of an item, extreme figures might not be outliers at all, especially if several participants take longer to judge the same item. In addition, this study proposes that the  $SD$  of reaction time is a measure of an individual's metalinguistic proficiency (see below). Therefore, once acceptable reliability was established, outliers were retained.

### 3.4.7 Summary of stage one

In addition to eliminating five sentences that posed a small threat to the reliability of the test item set, this stage of analysis indicated that the TGJT would be sufficiently robust to discriminate between proficient and less proficient participants. On the basis of the above analysis, it can be anticipated that proficient speakers will not only be faster and more accurate in their judgements, but that the variance in their reaction times (as indicated by the Standard Deviation figures) will also be less extreme. This point is taken up in the discussion

of the second stage, below. Finally, the first stage of the trial revealed no problems with the testing platform itself.

### **3.5 Stage two data collection and findings**

As mentioned above, volunteers for stage two were solicited through the TESL-L listserv, and the online test was deployed for two months, from March to May 2011. Fifty-two participants completed the TGJT, resulting in 1300 judgements. In addition to t-tests to reveal group differences, descriptive statistics for each respondent as well as for each item were calculated in order to facilitate a fine-grained analysis of how test items differentially affect judgement speed and accuracy for individuals.

#### **3.5.1 Anomalous responses by NS group**

Given that a perfect inter-rater reliability score was not attained on the test set ( $\alpha = .986$ ), it was not surprising to discover that of the 950 judgements by the NS group, 32 (3%) were contrary to the expected response (Table 4). It is conceivable that in a few of these cases the user had misheard or accidentally clicked the wrong button, but an analysis of the responses showed that in six cases, more than one NS judge had chosen an unanticipated response. It is reasonable to ask what factors could have resulted in misjudgements by the NS group, who should in theory have no difficulty in distinguishing grammatical sentences from ungrammatical.

One strong possibility, given the modality of the test, is that participants might have misheard sentence 4 as *Because it's reduceD the traffic problem*. A similar problem might have occurred with items 2 and 8. Recall that the items were read by a NS with careful enunciation, so if the original NNS speaker recordings had been used, the effect would doubtless have been exacerbated. In item 16, sticklers might have objected to *which* as a

subject relativizer, a prescriptive rule in American English<sup>2</sup>. Low parsability (Schütze 1996: 163; Juffs and Harrington 1995) could also have accounted for some of the erroneous

Table 4: *Divergent responses by NS group*

Item number	Sentence	Anticipated response	Frequency of divergent response (percentage of all responses for this item)
16	There are some shops which sell this.	Grammatical	6 (16%)
03	Do you think it helps to reduce the traffic?	Grammatical	4 (11%)
19	I will buy something I know they like.	Grammatical	4 (11%)
23	Why you don't believe this?	Ungrammatical	4 (11%)
04	Because it's reduce the traffic problem.	Ungrammatical	4 (11%)
17	You don't mind to marry a smoker?	Ungrammatical	2 (5%)
01	In the past, the womans wear the traditional clothes.	Ungrammatical	1 (3%)
02	Yeah, actually I'm agree with you.	Ungrammatical	1 (3%)
08	They do stuff that it's not allowed here.	Ungrammatical	1 (3%)
12	They also affects in our children.	Ungrammatical	1 (3%)
14	We had some agreements and some disagreements.	Grammatical	1 (3%)
20	It shows that I am relaxed.	Grammatical	1 (3%)
22	Anything that you use it in daily life.	Ungrammatical	1 (3%)
25	Why the people are doing this?	Ungrammatical	1 (3%)
<b>Total</b>			<b>32 (3%)</b>

judgements. For instance, *In the past, the womans wear the traditional clothes* could be considered a 'garden path' sentence (Pinker 1994; Ferreira et al. 2001), especially if one misheard *the womans wear* as *the womenswear*. It is difficult to say which if any of these explanations account for the divergent responses, and harder still to determine if the same confounds might have affected the NNS group. However, since even the highest count of divergent responses, occurring on item 16, represented only 16% of all NS responses for this

<sup>2</sup> See, for example, Tuten and Swanson (2003): 'Attorneys are taught to use *which* for nonrestrictive clauses and *that* for restrictive clauses so as not to cause a misreading in legal documents.'



item, it was deemed acceptable to include these anomalous items in subsequent analyses, despite the common practice in the psycholinguistic literature of replacing reaction times for inaccurate responses with mean RT (see, for example, Bley-Vroman and Masterson 1989; Murphy 1997; see Juffs and Harrington 1995: 499 for an alternative approach, that of eliminating reaction times on inaccurate judgements from analyses). In any case, because of the increased number of NS judges, Cronbach's alpha for the test set actually increased to .996 even with the anomalous judgements included.

### **3.5.2 Response bias**

An effect documented in the literature on grammaticality judgement tests (GJTs) by NNS participants is an overall response bias towards rejecting grammatical items as ungrammatical (R. Ellis 1991; Davies and Kaplan 1998). Birdsong (1989) suggests that such a bias could threaten the validity of GJTs when used with NNS. This is possibly true in the case of research intended to test a specific linguistic hypothesis, such as the operation of the 'subjacency principle' in SLA (Bley-Vroman et al. 1988). In such cases the researcher cannot be certain whether the response bias or the variable under investigation is responsible for rejection of a grammatical item. In contrast, in this study what is being measured is the generalized ability to distinguish grammatical from ungrammatical items and response time in doing so, and since there were in fact more ungrammatical items than grammatical (14 to 11) in the test set, such a bias should have worked in favour of the NNS.

It was found that NS participants were as likely to judge an incorrect item as correct as to do the opposite, suggesting that miscomprehension of the items (Table 4, above) was the cause. In contrast, the NNS participants were three times more likely to judge an incorrect item as correct than to do the opposite (Table 5), implying a tendency to view erroneous forms as correct (a response bias, in other words).

Table 5: *Summary of incorrect judgements by NS and NNS groups*

NS		
Anticipated response	Judgement	n
Ungrammatical	Grammatical	16
Grammatical	Ungrammatical	16
NNS		
Anticipated response	Judgement	n
Ungrammatical	Grammatical	58
Grammatical	Ungrammatical	20

To determine whether this response bias is significant, a chi-square test for goodness-of-fit was applied, using the anticipated count of correct and incorrect items in the test set as the *expected* frequencies and the actual judgements made as the *observed* frequencies. It was found that the NNS participants were more likely to judge items as grammatical ( $X^2 (1, N = 14) = 16.74, p < .001$ ). With the 38 NS participants, no response bias was found ( $X^2 (1, N = 38) = 0.19, p = .66$ ). A likely explanation for this finding, which contradicts the studies mentioned above, is that to the less proficient NNS in the study, many of the sentences ‘sound right’, as will be discussed in Section 3.6.

### 3.5.3 The relationship between item length and judgement data

The test items are of differing complexity (in terms of clause embedding, complementation, and so on), and it was hypothesized that more complex structures would correspond to longer reaction times as well as less accurate judgements, at least for the NNS. However, it is possible that these effects could be obscured by short-term memory limitations during parsing and judging. To investigate this possibility, correlations between the length of the audio recording for each item and the accuracy or speed of judgements were calculated (Table 6). A weak but significant positive correlation was found between item length and reaction time for the NS, but not for the NNS. Otherwise, longer recordings did not affect either accuracy or

reaction times any more or less than shorter ones.

Table 6: *Correlations of length of audio recording ( $M = 2.7s$ ,  $SD = 0.78s$ ) with Accuracy and RT of responses, by NS status*

		Accuracy	Reaction Time
NS	<b>Length of audio recording</b>	$r = -0.14$ $p = .670$ $N = 950$	$r = .118$ $p < .000$ $N = 950$
NNS	<b>Length of audio recording</b>	$r = 0.08$ $p = .159$ $N = 350$	$r = .001$ $p = .980$ $N = 350$

### 3.5.4 The relationship between judgement accuracy and reaction time

While there was no relationship between the length of the audio recordings and accuracy or RT, it was found that the longer a NNS judge takes to make a judgement, the more likely it is to be incorrect, leading to a rejection of the hypothesis in research question 3 (Section 3.2). Using a mean score for accuracy, for RT, and for the standard deviation of RT for each NNS participant, a strong, negative association was found between accuracy and reaction time for this group ( $r = -.597$ ,  $p = .031$ ,  $df = 11$ ). In addition, accuracy correlated highly with the standard deviation of reaction times for the same group  $r = -.675$ ,  $p = .011$ ,  $df = 11$ . These strong, negative correlations suggest that in general those NNS whose judgements were faster and more consistently so were also more accurate, and vice versa. This finding was confirmed by performing a t-test comparison of means for incorrect and correct judgements, which showed NNS participants as a group were slower when they judged inaccurately, and faster when they were accurate, by 1,570 ms ( $t = 3.51$ ,  $p = .001$ ; see Table 7).

Table 7: *Mean reaction times for incorrect and correct judgements, NNS participants*

	Judgement	n	$M$ (ms)	$SD$ (ms)
RT	Incorrect	78	4161	3634
	Correct	247	2591	2760

No significant correlations were found for the NS group between accuracy, RT, and SD of RT. Interestingly, however, a t-test showed that the NS group had almost exactly the same mean difference in the speed of their correct and incorrect judgements, 1,488 ms,  $t = 3.47$ ,  $p = .001$  (Table 8) as the NNS group. Thus, while no significant relationship exists between

Table 8: *Mean times for incorrect and correct judgements, NS participants*

	<b>Judgement</b>	<b>n</b>	<b><i>M</i> (ms)</b>	<b><i>SD</i> (ms)</b>
RT	Incorrect	32	2550	2282
	Correct	918	1062	1004

accuracy and RT (which is not surprising as the mean accuracy level for the NS participants is so high that insufficient variance exists for such correlations to be significant), it is still the case that when a participant is correct, her judgements are made more quickly than when she is not, regardless of whether she is a NS or a NNS. Slower speeds on incorrect judgements indicate a lack of confidence, and it is possible that the inclusion of a ‘Not Sure’ option would have helped to distinguish uncertainty from inaccuracy. However, as explored in the following section, *what* a participant is confident about depends to a significant degree on whether she is a NS or a NNS.

### 3.5.5 Judgements as ungrammatical and grammatical

To investigate this, the mean difference in RT when participants judged items as correct in contrast to incorrect was calculated (Table 9). This time it was the NS group that showed a small but significant difference in the speed of their judgements, but surprisingly it was

Table 9: *Mean reaction times for judgements as ungrammatical and grammatical, NS participants*

	<b>Judgement</b>	<b>n</b>	<b><i>M</i> (ms)</b>	<b><i>SD</i> (ms)</b>
RT	Ungrammatical	532	984	904
	Grammatical	418	1275	1296

the *grammatical* items which took slightly longer to judge, by 290 ms ( $t = -2.79, p = .005$ ). This finding contradicts those in Murphy (1997: 51), where grammatical sentences were judged more rapidly by all groups in both aural and written modalities, except for the ‘ESL’ (NNS) group in the aural modality. The NNS group in the pilot study showed no significant difference in this means comparison, suggesting that as a group, the NNS approach the judgement task differently from the NS group.

A reasonable hypothesis is that NS participants simply listened for anything ‘odd-sounding’, a holistic, intuitive process which begins as soon as the item begins to play. With an incorrect item, they may already have formed a judgement that it is incorrect by the time the audio has finished playing (consider: *That’s will affect their grades.*), permitting a quick response time when the button choices appear on the screen. In the case of correct sentences, when the audio has finished playing and nothing obviously wrong has been encountered, it is possible that at least some of the participants will mentally replay the sentence to be certain. Another possibility is that even native speakers bring explicit, metalinguistic knowledge into play during the judgement process, at least in the case of undecided items. For instance, item 16 from Table 4: *There are some shops which sell this*, resulted in the highest average RT of all items (Table 10). If the speculation concerning prescriptivism (see section 3.5.1) were

Table 10: *The five items with highest mean reaction times for NS group*

Item number	Sentence	Anticipated response	Mean RT (ms)	SD of RT (ms)
16	There are some shops which sell this.	Grammatical	1993	2227
03	Do you think it helps to reduce the traffic?	Grammatical	1952	2327
04	Because it’s reduce the traffic problem.	Ungrammatical	1725	2350
05	Each person in the family has one car.	Grammatical	1397	909
19	I will buy something I know they like.	Grammatical	1374	937

accurate, one would indeed expect a longer reaction time as metalinguistic knowledge is called into play, as was found in this case. Given that four of the five items in Table 10 were also among the items for which NS participants gave unanticipated responses (Table 4), it is safe to assume that many participants found these items to be borderline cases. In other words, they may have sounded odd, but not clearly ungrammatical. This goes some way towards explaining why the teachers in the replica-pedagogic task in Module II did not reach a greater degree of consensus in their notation of errors: learner language, especially at lower and intermediate levels, is characterized by borderline utterances such as those in Table 10. Thus one element of the teachers' decision whether to note down an error is the appraisal of how acceptable it really is. This, as we have just seen, takes time – although the high SD figures in Table 10 would seem to indicate that some participants reached their decision much faster than others.

For the NNS participants, the negligible and non-significant difference between RT on judgements as grammatical ( $M = 2901$ ,  $SD = 3284$ ) and ungrammatical ( $M = 2722$ ,  $SD = 2673$ ),  $t = -0.555$ ,  $p = .793$ ,  $df = 348$ , coupled with the fact that they are faster when they are correct in their judgements, suggest that they use a different strategy: rather than screening for 'odd-sounding' language, they might be matching the input against an internal template of 'correctness' to make a judgement based on whether the item 'sounds right'. Additionally, they might employ explicit metalinguistic knowledge in the form of pedagogical rules to assist in this process (analogous to the use by NS of prescriptivist rules). Of course, what 'sounds right' will depend on proficiency, L1 influence, exposure to input, degree of fossilization and so forth, so we could anticipate that at any given point in the acquisition process, a learner might make judgements which are:

- a. fast and correct – indicating a language item which is acquired, familiar, ‘right-sounding’
- b. fast and incorrect – indicating an acquired, familiar, ‘right-sounding’ item which is in fact a fossilized, erroneous form
- c. slow and correct – indicating a item which is not yet proceduralized and automatic, and which therefore needs to be considered more carefully, possibly with the assistance of explicit metalinguistic knowledge, L1 translation, analogy with acquired forms, and so on
- d. slow and incorrect – indicating an item which is either completely unfamiliar or not familiar enough to benefit from the assistance of explicit metalinguistic knowledge, L1 translation, analogy with acquired forms, and so on

What constitutes ‘fast’ and ‘slow’ will naturally be highly idiosyncratic, irrespective of native-speaker status, but it is reasonable to predict that as L2 proficiency increases, so does overall judgement speed. In addition, with items which have not been fully acquired, whether correctly or not, learners will show fluctuation in judgement speeds, gradually becoming faster and, it is to be hoped, more accurate. This point is taken up in the discussion below.

### 3.5.6 Group mean differences

As anticipated, there was a mean difference in both accuracy and speed between NS and NNS participants. NS were more accurate ( $M = 0.99$ ,  $SD = 0.11$ ) than NNS ( $M = 0.79$ ,  $SD = 0.41$ ),  $t = 7.67$ ,  $p < .001$ ,  $df = 257$  and made faster judgements: NS ( $M = 973$ ,  $SD = 715$ ), NNS ( $M = 2887$ ,  $SD = 3104$ ),  $t = -9.6$ ,  $p < .001$ ,  $df = 255$ .

Group means, of course, conceal individual variation which contributes to a richer and more complex picture of language proficiency. Given that the participants were contacted through an online discussion forum for ELT professionals, it is highly likely that there is a range of proficiencies among the self-declared NNS. For this reason, the mean accuracy and reaction time, as well as the standard deviation of reaction times, were plotted on a 3-D

scatterplot matrix (see Figure 5; circles represent NS and triangles, NNS).

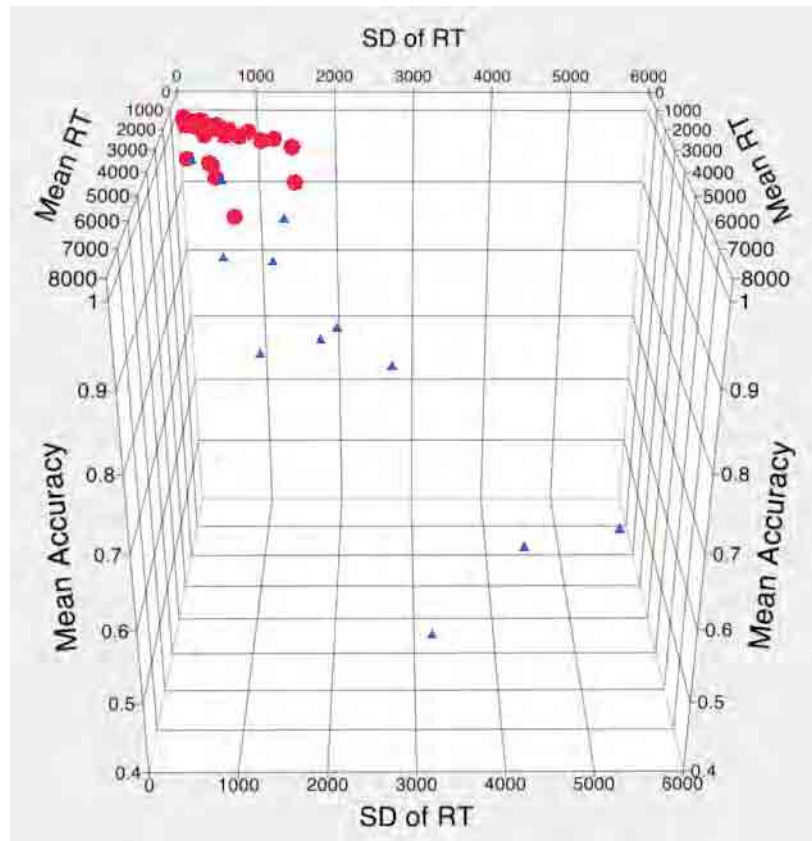


Figure 5: 3-D scatter plot of mean score, mean RT, and SD of RT

There is a far more pronounced clustering of NS participants than NNS, who in fact seem to be potentially separable into subgroups. However, it should also be noted that at least two NNS participants show scores that are indistinguishable from those of the main cluster of NS. A further complicating factor is that several of the NS participants show average scores and reaction times that are similar to those of some of the NNS. To investigate this further, a composite standardized score was calculated for each participant by subtracting the z-scores for mean RT and SD of RT from the z-score for accuracy. This *ad hoc* calculation is admittedly somewhat arbitrary, combining three variables which are assumed to represent discrete psycholinguistic features. However, it does permit a group-wide comparison against



the mean for this construct, as shown in Figure 6 (where 0 is the mean and each unit on the horizontal axis is 1 *SD*).

As can be seen, participants 9, 16, and 27 ('uelf', 'Halia', and 'Gloria') performed on a par with NS counterparts, while NS participant 42, 'cm' did no better than participant NNS 41, 'mexico'. This kind of ranking indicates that the TGJT is not, or is not solely, a measurement of underlying linguistic competence. Even if it were, the range of linguistic forms in the test items – coming as they do from the production of high-intermediate learners of English – is sufficiently restricted that there is no reason to suppose that a highly proficient NNS could not judge as accurately and fast as an average NS, and even more so, if she were familiar with not only the forms in question but also with the kind of errors that intermediate learners make, errors which she herself perhaps once made.

Although NS as a group are faster and more accurate than NNS in detecting well-formedness on a range of structures, the ability to do so is a trainable skill; assuming that NS competence is relatively homogeneous, the variation in combined accuracy, speed of judgements, and consistency of judgement speed would seem to suggest that metalinguistic awareness itself could become procedural or automatized. In other words one can get better at spotting and evaluating errors with practice. Thus some NS would show greater facility in this task than others, and some NNS would show similar facility to that of some NS. Given that this kind of metalinguistic awareness is precisely what language teachers would call on while

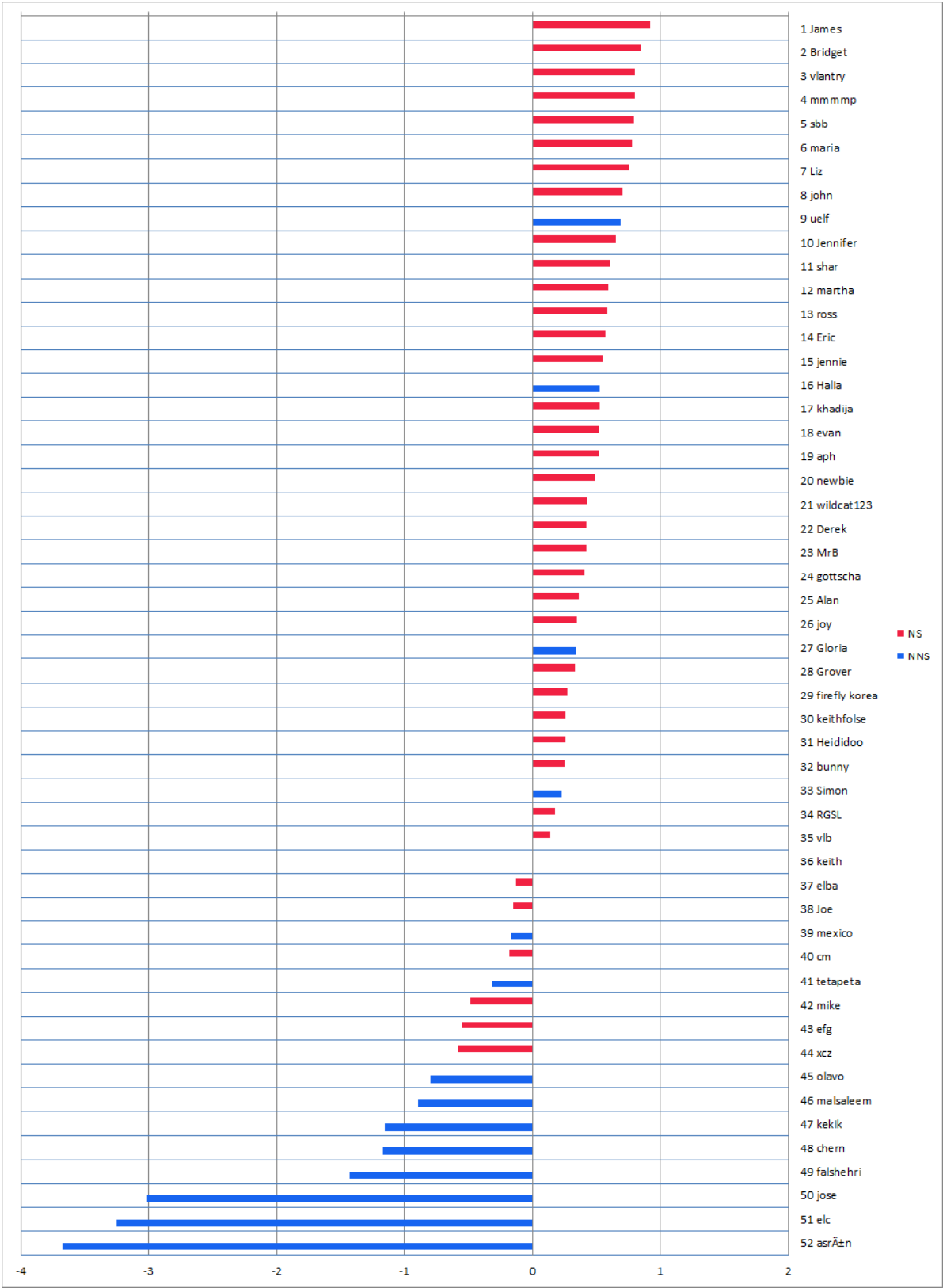


Figure 6: All TGJT participants, ranked by combined accuracy, RT, and SD of RT

monitoring students and providing CF, it would be a reasonable supposition that ‘uelf’,

Halia', and 'Gloria' are experienced NNS English teachers. However, this is pure conjecture, and to substantiate such a claim would require the collection of demographic data beyond simple (self-reported) native-speaker status.

### 3.6 Differential performance by two NNS participants

What is somewhat more certain is that different participants (whether NS or NNS) are using different kinds of knowledge (implicit vs. explicit, or intuitive vs. metalinguistic) to approach the test or even different items within the test, as suggested in section 3.5.5. In this section, two NNS participants' judgement data will be closely examined for internal consistency and to determine, if possible the nature of 'fast' and 'slow' judgement for individual judges.

Figure 7 shows the judgement data for NNS 'Gloria', whose judgement speed, accuracy, and consistency put her in approximately the median position in Figure 6. The data have been sorted from fastest judgement speed (top) to slowest, with incorrect judgements shaded darker. 'Gloria' evidently has high proficiency in English (or at least in the English represented by the range of items in the test set), as evidenced by her ranking in Figure 6. She made three erroneous judgements, and all three were made with slower than average speed: one is slightly below her mean judgement speed, and the other two are well below average, in fact more than one *SD* below. In this case, determining what constitutes 'slow' and 'fast' judgements, as described in section 3.5.5 above, is relatively straightforward: if we take the range from 1 *SD* above to 1 *SD* below as 'normal', then the three judgements below -1 *SD* (items 4, 17, and 3), would be considered 'slow'. By the same token, the judgement at the top (item 10), centred on +1 *SD* would be considered 'fast'. A possible interpretation of this, in the absence of corroborating proficiency data, would be that the linguistic items judged correctly and 'fast' are acquired and automatic for 'Gloria'. This is not to say that we can

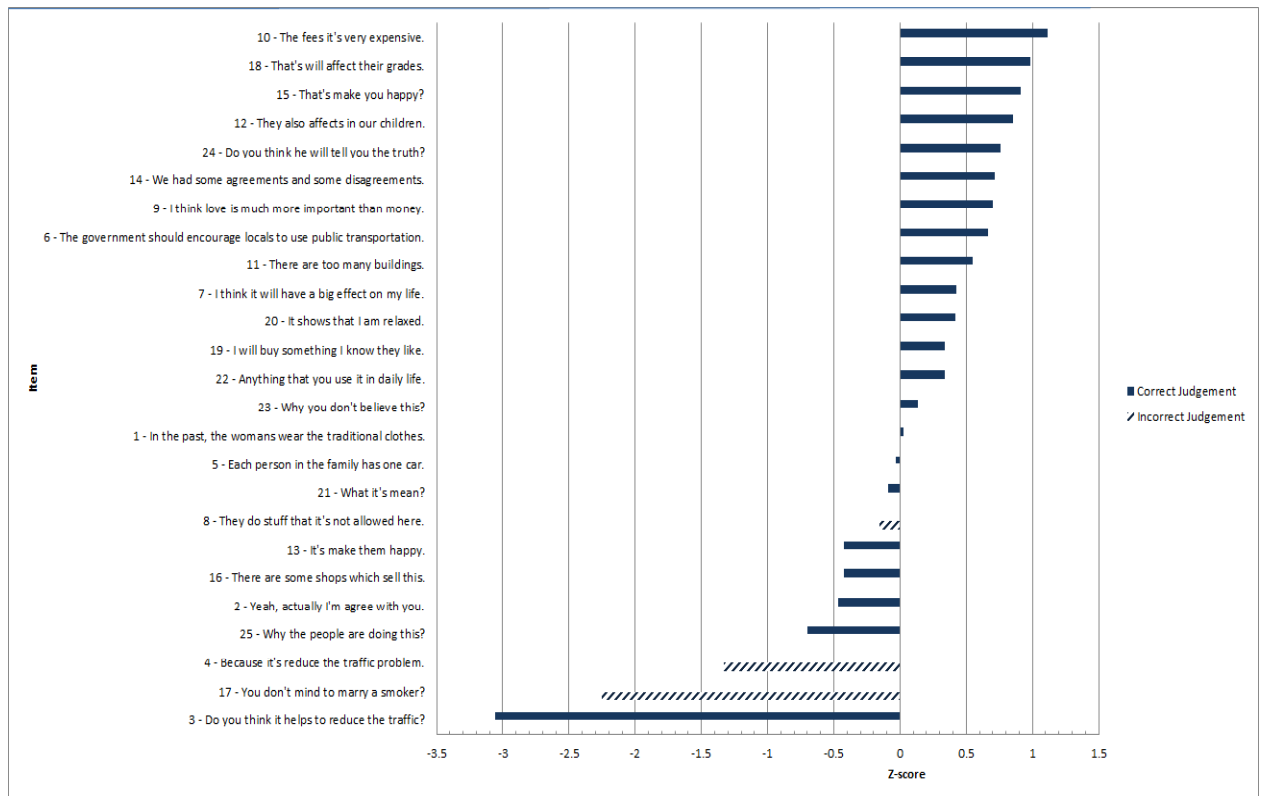


Figure 7: Judgement data from 'Gloria' (NNS)

predict her *production* of these forms (or their correct equivalents) to be equally automatic, but that she can effortlessly discern the correct from the incorrect.

On the other hand, 'Gloria's' incorrect judgement of item 8, *They do stuff that it's not allowed here*, is anomalous, in that she correctly judges the analogous subject relative clause item 16, *There are some shops which sell this*. A similar contradictory judgement occurs with items 4 and 13, *Because it's reduce the traffic problem* and *It's make them happy*. Assuming that these anomalous judgements do not result from the kind of confounds suggested in section 3.5.1, it would be possible to infer that these structures are not yet fully acquired nor proceduralized. A similar inference could be made about the 'slow' judgements on items 3 and 17, *Do you think it helps to reduce the traffic?* and *You don't mind to marry a smoker?*, both of which involve verb complementation. We can rule out the possibility that 'Gloria' objects to the question formation in item 17 since she rejected item 23, *Why you don't believe*

*this?* with faster than average speed. In pedagogical terms, then, it would seem reasonable to assume, pending disconfirming evidence from other sources, that ‘Gloria’ would benefit from instruction or at least CF targeting subject relative clauses and gerund/infinitive verb complements.

Turning to a somewhat less proficient NNS participant, ‘Falshehri’, we see a different picture (Figure 8). Two items, 23, *Why you don’t believe this?* and 25, *Why the people are*

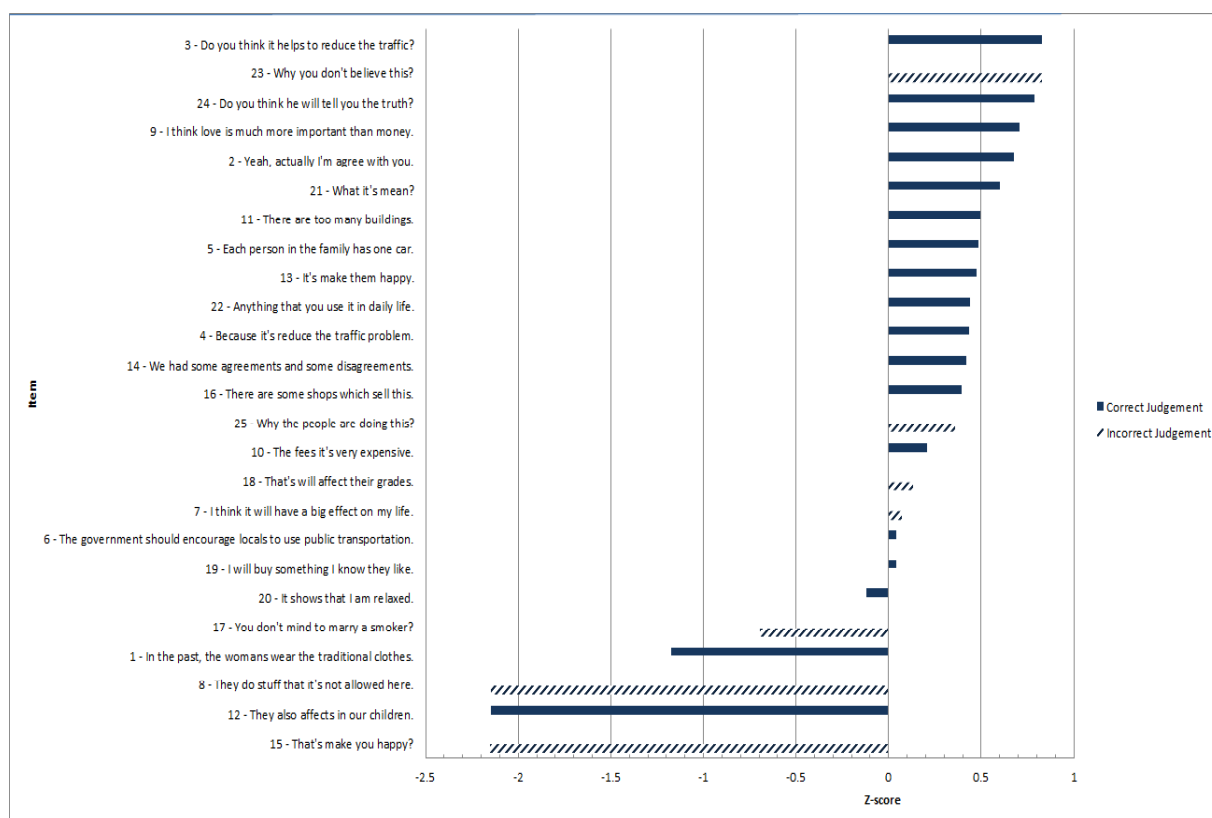


Figure 8: Judgement data from ‘Falshehri’ (NNS)

*doing this?* are both incorrectly judged and rapidly so, suggesting that this question formation sounds intuitively correct to her. If so, it would be a strong candidate for a fossilized interlanguage (IL) form, which might be highly resistant to remediation (Han and Selinker 1996; 1999; Lakshmanan and Selinker 2001). Another systematic error can be seen in item 15 *That’s make you happy?* and 18 *That’s will affect their grades*. In this case, however, a

correct judgement on a similar item, 13 *It's make them happy* might suggest that this form has a transitional status in 'Falshehri's' IL and could be more responsive to pedagogical remediation, as might items 8 and 16 discussed in 'Gloria's' case above.

### 3.7 Implicational hierarchies

One further point is worthy of comment here. In at least one of the cases mentioned, there seems to be an implicational hierarchy among all NNS responses, such that if a particular item is judged correctly, others are very likely to be correctly judged as well, as is

Table 11: *Implication hierarchy of NNS judgements of items 18, 15, and 13*

Item 18 <i>That's will affect their grades.</i>	Item 15 <i>That's make you happy?</i>	Item 13 <i>It's make them happy.</i>
+	+	+
+	+	+
+	+	+
+	+	+
+	+	+
+	+	+
+	+	-
-	+	+
-	+	+
-	+	-
-	-	+
-	-	+
-	-	+
-	-	-

the case in all but two judgements in Table 11. The probability of this occurring by chance is so small as to make it worthy of further study. Given that the test items for the TGJT were chosen by different teachers from the production of different learners, it is surprising and encouraging to find such an effect in the responses of randomly selected NNS judges.

Whether the explanation stems from a universal acquisition order (Bailey et al. 1974; Dulay and Burt 1974a), a frequency effect in input processing (N. Ellis 2002), a meaning-based

approach (Bley-Vroman 2002), or some as yet undiscovered cause, the very existence of judgement patterns such as these implies some degree of systematicity in IL development and merits further investigation.

### **3.8 Discussion**

What the foregoing has attempted to do is to perform a sort of qualitative ‘regression analysis’ based on a very limited sample of language items, considered to represent typical language problems of certain upper-intermediate learners. Needless to say, in its present state it is a blunt instrument, as it uses only this limited test set and the judgement speed and accuracy of participants to purportedly measure linguistic knowledge. In addition, whatever it purports to and actually does measure, it can provide only a cross-sectional measurement. However, it has demonstrated that participants do respond differently to different items, and that there may be some systematicity to their responses, which is what one would hope for in a tool intended to assist in pedagogical decision-making, although the set of test items is at this stage far too restricted and generalized to be of use to individual learners.

Given the number of assumptions made above, it is essential to establish what kind of evidence would support or disconfirm such pedagogical hypotheses. Foremost among these would be evidence from each learner’s own production – after all, the language items in the test set and the IL forms they represent certainly do not represent any one participant’s IL, even though there may be overlaps. If it could be established that, say, ‘Gloria’ herself were making errors of this type, the exercise would naturally be more informative. In addition, it is important to note that the participants in this pilot study were not exposed to a written version of the items until after the TGJT was completed, nor could they listen to any item more than once. Permitting the latter would greatly complicate the investigation of whether target items were being judged using implicit or explicit knowledge, although it might help to rule out

erroneous judgements caused by mishearing. This potential confound could also be eliminated by using recordings made by the participant herself, since familiarity with her own speech patterns might reduce ambiguity at that level. This possibility will be explored in Chapter 4. (There currently exists no research on differential comprehension of one's own *speech* vs. that of others; however, studies by Wilcox (1978) and Smith and Bisazza (1982) suggest that in general, one's own *accent* is preferred over that of others.)

Using written items in the TGJT would, of course, introduce not simply a different presentation modality but also different judgement criteria – such as spelling and punctuation – and a different receptive skill, with concomitant processing considerations (Chaudron 1983; Birdsong 1989; Schütze 1996; Murphy 1997). However, using written items in an elicited imitation/correction stage would permit far greater confidence in the analysis of IL forms. This point will also be taken up in greater detail in Chapter 4.

### 3.9 Summary of findings

In addition to testing the functionality and viability of a web-based TGJT platform, the pilot study sought answers to three research questions. In answer to the first, *Is there a mean difference in either accuracy or speed between NS and NNS participants?*, it was found that overall, large and significant differences exist between the two groups. However, it was also found that in the case of the linguistic competence represented by the limited set of test items, certain NNS participants performed on a par with the NS group (Figure 6). Further, a small but significant difference was found in reaction times to grammatical versus ungrammatical items in the NS group but not the NNS group, suggesting a differential approach to the detection of ungrammaticality.

The second research question, *Do proficient speakers show consensus in their judgements of correct and incorrect samples of learner English?*, addressed the issue of TGJT



reliability within the NS group. It was found that 3% of the NS judgements were contrary to the anticipated responses, but overall, these had little effect on the reliability of the test set (which actually increased to  $\alpha = .996$  owing to the increased number of NS judges), indicating that NS judges were very capable of identifying grammatical versus ungrammatical items in this type of TGJT. This finding supports the claim made in Module II that teachers are consistent in their identification of errors.

Finally, in addressing the third research question, *Is there a measurable relationship between the accuracy and speed of judgements?*, no significant association was found for NS between the average accuracy of their judgements and the time they took to make them, whereas it was found that for the NNS group, accuracy and reaction time were strongly and negatively correlated: those who took longer to judge were more likely to be incorrect. Also strongly and negatively correlated for the NNS group were accuracy and the standard deviation of RT, so that judges whose judgements were consistently closer to the mean were also more accurate overall. This makes sense, since the greater one's linguistic proficiency, the more able one is to make fast and accurate judgements on the grammaticality of a range of items, and the less time one will require to judge any particular item. The potential for this type of test to be harnessed in a diagnostic capacity at the group and individual learner level forms the focus of the following two chapters.

---

## CHAPTER 4: ELICITED IMITATION AND CORRECTION TESTS

---

### 4.1 Introduction

There is evidence in the CF literature that successful uptake, which according to Loewen ‘involves students in either repairing their erroneous utterances or demonstrating an understanding of a linguistic item’ (Loewen 2004: 156), is related to gains in overall proficiency (R. Ellis et al. 2001; Williams 2001; Loewen 2002; Sheen 2004). However, uptake occurring as a response to a recast in the course of classroom dialogue can be interpreted as a discourse phenomenon, and thus its status as evidence of acquisition is ambivalent at best (R. Ellis and Sheen 2006; Long 2007). The CF methodology under investigation here addresses these issues by employing delayed CF and by measuring acquisition in terms of the accuracy and fluency of error reformulation.

This chapter therefore documents the investigation into the effectiveness of delayed CF by measuring the learners’ ability to correct (reformulate) the language errors produced by themselves and peers during conversational interaction (Small Talk). To do this, an elicited imitation and correction task is employed. This task, called a Running List Test (RLT), is used to establish baseline measures of accuracy and fluency for individual participants, against which to compare subsequent performance.

### 4.2 Research questions

This stage of the investigation addressed the following research questions:

- 1) To what extent are learners able to reformulate their errors accurately and fluently in a delayed test?

The strongest evidence of acquisition is, of course, consistent target-like production under conditions of fluency (i.e. when the learner's attention is on meaning not form), but measuring this is not a practical option. Instead, the ability to reformulate an error under time pressure is adopted as an indication of automaticity, and offered as a robust alternative to 'uptake' as it is customarily used.

2) Are learners consistent in their ability to reformulate errors?

Consistency over time would seem to be a reasonable indicator that acquisition is taking place, although we must acknowledge that correct performance on a test does not guarantee correct performance in fluent production. However, if there is a significant level of backsliding, it is difficult to claim that the CF is effective.

3) Do learners find it more difficult to reformulate their own errors or those of peers?

It is common pedagogical practice, especially in monolingual teaching contexts, to assign CF tasks such as correction of 'key' errors from a fluency activity (Hedge 2000: 292), but little research has been done on the differential ability of students to correct their own and others' spoken errors. According to some (e.g. Long and Porter 1985), other-correction in group work is rare, occurring in response to only 1.5% of peer errors; on the other hand, these other-corrections are usually accurate (99.7% accurate, according to the same authors, p. 216). However, when a more proficient language learner (or teacher) can identify the error during interaction, learners are able to repair approximately 75% of peer errors (Raof and Razali 2010). This investigation looks at whether a similar level of repair will occur in a delayed test, and whether the source of error has any effect on the fluency and accuracy of reformulations.

### 4.3 Description of 'Running Lists'

The description of the Small Talk methodology in Chapter 2 of Module II (see also the

published version, Hunter 2012) will be briefly summarized here in order to familiarize the reader with the way in which corrective feedback (CF) is provided to students.

During the course of a Small Talk session, the teacher makes notes of ungrammatical or non-target-like language used by the students in the conversations. These are entered into a database which produces a worksheet for the students, and the teacher makes an audio recording of reformulations of the items. Both the worksheet and the audio file are made available to the students through online course management software. Teachers customarily select specific items on the worksheets for all students to correct, regardless of the actual speaker ('All Do' items). Students thus have to attend to corrections on five to ten items per worksheet, and generally receive one worksheet every week. A variety of activities are used to focus attention on the CF provided by worksheets, from quick warm-up activities to focused grammar instruction, usually at the initiative of the teacher but occasionally at the request of one or more students who do not understand or wish to know more about the formal or lexical content of the reformulations. In this way, this stage of the CF process is more akin to 'guided noticing' than Focus on Form.

The students periodically practise and are tested on their own worksheet items in class, to promote automatic use of the targeted forms. Each student thus keeps a 'Running List' of the errors (with no other markings), and practises these in class with other students, for example by giving a copy to a partner and saying the corrected versions to see if the partner can hear the differences. Thus the Running List activities combine features of discrimination, elicited imitation, and correction tests: in the first stage, the student listens to the audio reformulations recorded by the teacher in order to identify differences between what she hears and the errors on the worksheet. In the second stage, the student sees the original worksheet item and tries to reformulate it correctly. The exact version from the teacher's

recording is not the only acceptable reformulation, of course, and any version that conveys the intended meaning in standard English is acceptable. This stage thus constitutes a type of delayed stimulated recall and correction task, both of which are widely used in SLA research to measure IL development or linguistic competence (Hu 2002; Révész 2002; R. Ellis et al. 2006, 2009).

The database is used to keep track of students' Running Lists. In the case where the teacher has assigned 'All Do' items, these can be manually or randomly selected and added to either the individual or whole class Running List. Thus each student's Running List is to a large extent individualized. Just as there is no way to know at this stage whether a student's incorrect utterance is a systematic error or a 'slip', there is no way of knowing whether an 'All Do' item is at the student's current stage of IL development, an established part of her competence, or beyond it; however, the Running List Test (RLT) can reveal linguistic forms that are more or less easily elicited, which gives some indication of the degree of acquisition.

#### **4.4 The role of memory in elicited imitation**

In both pedagogical practice and research, there is a legitimate concern that performance on elicited imitation tasks depends to a greater extent on working and phonological memory than on linguistic competence, or at least that the former are a potential confound in the exploration of the latter. In other words, if learners 'simply memorize' the reformulated items, how can we be certain that subsequent successful recall of those reformulations reflects their competence, and not their ability to reproduce rote-memorized 'chunks' of language without necessarily understanding them or being able to manipulate the individual linguistic components that constitute them? Tomita et al. (2009: 346) caution researchers to 'ensure that the performance of [elicited imitation tasks] is not greatly influenced by participants' rote repetition abilities' or by 'participants' capacity to store or hold information (i.e., short-term

memory abilities).’ In addition, they suggest that if the task is intended to measure implicit knowledge, participants must be concentrating on meaning, not form. R. Ellis’ (2005b) elegant solution to this was to embed the target forms in truth statements to which participants were asked to respond before imitating the form.

Turning first to the issue of working and phonological memory, there are two safeguards suggested in the literature to control for these: one is to introduce some delay between the presentation and the elicitation, as Ellis did. The second is to increase item length to fifteen syllables or longer (Mackey 2005). In the case of the RLT, even if participants were to listen to the teacher’s reformulations immediately before taking the test, there would be approximately 30 items to store in working memory. This points to the essentially reconstructive nature of the task (Erlam 2006): participants are presented with the original errors as prompts, but still have to know what the correct form should be, which is unlikely to be something they can hold in working memory for the duration of the test. Items appearing earlier on the test might be subject to such effects, but there is no guarantee that earlier items are ‘easier’ than later ones. A second point about memory confounds is that phonological memory (PM) may in fact be an important contributing factor to acquisition. O’Brien et al. (2007) for instance found that in learners of intermediate competence, PM predicted gains in oral fluency over the course of a semester of study, regardless of whether the study was in a domestic or study-abroad context. They claim that ‘the relationship between phonological memory and L2 oral fluency observed... suggests that it might also be related to L2 grammar skill’ (O’Brien et al. 2007: 576), but they do not say whether PM is a constant within each learner or whether it is something which can be developed and expanded over time. A review of related studies by Hummel and French (2010: 380) concludes: ‘although basic phonological memory capacity may indeed be a fixed trait, the relative processing efficiency

underlying this capacity appears sensitive to the effects of training.’ It may be, then, that students with better phonological memory will do better on tasks like the RLT, but it is equally possible that tasks like the RLT and related activities contribute to the development of PM, which in turn may contribute to development of proficiency. Clearly, further research is needed to establish the relative contributions of each.

The role of long-term memory (LTM) in SLA is no less controversial. Certainly, rote memorization has little place in the ELT methodologies of the last half-century (see, for example, van Lier 2008: 603) and the notion that the significant portions of the syntactic system of a language can be acquired in this way has been dismissed by most SLA theorists in the West, even while the essential role of memory in lexical acquisition and retention is acknowledged by all. This view is grounded in the ontological separation of syntactic and lexical systems: rules are rules, and words are words, and the two systems are stored and processed separately (Pinker 1999). Sinclair (1991, cited in Hunston and Francis 2000), on the other hand, claims that syntax and lexis are interdependent and must be described together. Presumably, then, an account of language production would also have to treat the two as interdependent. This is a point echoed by Skehan:

What is now possible... is to view fluency as partly dependent on the role of memory in performance, and the way in which the unit of performance may vary, sometimes functioning at the morpheme level, but often drawing on larger units beyond the level of the word, and including very large numbers of ready-made phrases. The consequence of this is to recast the performance problem not simply as a computation problem (that underlying rules are applied in interaction with simple lexical elements) but also as a retrieval problem, where chunks of language are orchestrated, in real time, to achieve fluent performance. (Skehan 1998: 285)

It is very probable, then, that fluent oral production does indeed involve and necessitate the automatic retrieval and contextualized adjustment (i.e. making the syntactic and morphological changes required to achieve grammaticality in the context) of prefabricated chunks of language in real time (N. Ellis 1996; Hunston and Francis 2000; Hunston et al.

1997; Widdowson 1989: 135; Willis 1990; Wray 2000; 2002). The competing explanation, that oral production is a creative process of construction wherein it is the syntactic and morphological rules that are automatically retrieved and deployed along with (presumably small) lexical items in order to achieve meaning (Krashen 1981), should not be discounted, however. It is highly likely that in practice we use both strategies during communication depending on a wide variety of factors such as the complexity of what we are trying to express, the context of the speech event, our state of mind, and, of course, our proficiency in the language we are speaking, as Kormos and Dénes point out:

Low-proficiency students generally cannot rely on a sufficient number of automatic sequences and apply conscious rule-based mechanisms, and if they strive to be highly accurate, their speech becomes very slow. Thus in certain cases especially among less competent speakers, speed and accuracy might be in inverse relationship with each other. (Kormos and Dénes 2004: 160)

This is probably also true *within* language learners, as well. For some utterances, we will have *les mots justes*, the grammatical and acceptable (i.e. idiomatic, or at least not noticeably unidiomatic) way to express the intended meaning. For others, we will have no idea how to achieve this and will have to rely on a variety of other strategies, from translation to the bottom-up assembly of words and rules, or as close an approximation as we can make, to get our meaning across – a process which will almost certainly require greater conscious attention and therefore have a noticeable effect on processing speed. It should be possible, then, to use fluency, to the extent that it constitutes a measure of lexico-syntactic retrieval and processing speed, as an indicator of which items a learner can produce automatically. Items which require conscious attention to form, which the learner has to ‘think about’, should be produced with noticeably less fluency. Thus as with the recognition of items on the TGJT (section 3.5.5), there should be a range of items in the production of any learner, from those which can be produced fast and accurately to those which are produced slowly, with greater



hesitation, and perhaps inaccurately.

#### **4.5 The Running List Test (RLT)**

The test is in the form of a timed version of the practice. The class as a group decides on a target number of items per minute (usually ten to twelve) and each student is given a clean copy of her Running List from the database (see Appendix 4 and Appendix 5). She is then recorded as she reformulates the erroneous sentences within a given time limit, usually two to three minutes depending on the level of the students. The recordings are done in a language lab, using Dill software (Taylor et al. 2010) on Mac computers, which allows for synchronized, timed recording of up to 20 individual students. The teacher listens to these recordings and gives a grade based on ‘fluency’ and ‘accuracy’: a student who can correctly say one or two sentences per minute may be highly accurate but not very fluent (see Section 4.7.1). Conversely, a student who can get through 20 items a minute but is only correct on some of them is very fluent but not very accurate.

The RLT thus constitutes an essential component of the CF methodology. Its use in this study therefore has a twofold purpose, first to ensure a greater degree of ecological validity than would be possible were other, non-instructional measures used; and second, to validate the decision to operationalize fluency and accuracy in this way.

#### **4.6 Participants**

For this stage of the research, data came from two intact groups of ESL students in different English for Academic Purposes (EAP) programs in the United States. The first group was an advanced class in the English Language Center of Gonzaga University (N = 13; henceforth ‘ELC’). This is not the same group of students described in Module II, however, since those students had left the program by the time the data collection process described below was initiated. The second group was an intermediate class of students from a Japanese women’s

university on a 14-week Intensive English Program in the United States, where the university has a branch campus (N = 13; henceforth 'IEP'). The average TOEIC score for the group was 407 (approximately IELTS 4.0), with individual scores ranging from 370 to 435. Both groups used Small Talk as the methodology to develop oral accuracy and fluency as part of their ongoing coursework, and were aware of and consented to the use of their data in this stage of the research, which took place during the spring semester, lasting from January to May 2011. The purpose of including a second group was to establish whether the effects observed for one group would parallel those for the other, and if so to counteract the threats to external validity and generalizability of the ecological research approach adopted (Kramsch and Steffensen 2008). Although intact classes were used for the research, the two groups differed in several important ways in addition to having different classroom teachers, as summarized in Table 12. In essence, the learning situation of the IEP group can

Table 12: *Summary of differences between ELC and IEP groups*

	ELC	IEP
English proficiency level	advanced	intermediate
English study program:		
hours per week	18	18
classes	Listening & Speaking, Writing, Reading, Grammar	Conversation, Writing, Reading, American Studies, Grammar
diversity:		
L1	Arabic, Spanish, Korean, Mandarin	Japanese
age	18–38	19–20
sex	women & men	women
socio-economic	mixed (business and academic professionals and religious from working-class and privileged backgrounds)	middle class; second tier Japanese university
purpose of study	undergraduate or graduate study in US	compulsory part of English BA degree in Japan
exposure to target language outside of class	unlimited (housed with American room mates or other L1 speakers; unrestricted access to local community)	limited (housed with classmates; one-weekend home stay with American family; very restricted access to local community)

be thought of as more akin to EFL, in terms of being monolingual, monocultural, and having limited exposure to the target language.

#### **4.7 Methodology**

During the course of the semester, both groups were given two Running Lists Tests. This permitted both cross-sectional and longitudinal analyses of the effects of the CF provided. Both groups were given a practice RLT early in the semester in order to familiarize them with the procedure and were given a grade on their performance on the RLTs, which counted towards their final grade for the course, in both cases grades on the RLTs representing approximately 10% of the final grade. The decision to include various aspects of Small Talk, such as participation, leader skills, and RLT performance in the course grade was a curricular decision that predated and was completely independent of this research; however, it was fortuitous as it ensured that most of the students would participate.

Because the content of the RLT is taken from students' language production, each student's test was different from every other student's, and the first RLT was also different from the second for each student. Therefore, this is not a pre- and post-test design, nor does it seek to ascertain the effectiveness of the CF on the development of a particular linguistic form (e.g. past simple) with a particular group of students; rather, the objective is to discern what changes take place in the accuracy and automaticity of a range of forms for a range of students.

##### **4.7.1 Pedagogical use of the RLT**

As mentioned above, teachers score the RLT based on 'accuracy' (the ability to reformulate the worksheet item in correct English) and 'fluency' (the ability to reformulate the target number of worksheet items in the given time). For each student, the teacher enters the number of sentences attempted and the number correct in a spreadsheet, and a grade is calculated by

averaging these two scores (Table 13).

Table 13: *RLT grade calculation spreadsheet*

Running List Grades: 108B Spring 2011			Time limit: 3 minutes Target # of sentences per minute: 10 Target number for this test: 30				
Name	Count of reformulations attempted	Count of correct	Time	Fluency (attempted / target)	Accuracy (% correct)	Grade	Letter
S1	18	17	3	60%	94%	77%	C+
S2	24	22	3	80%	92%	86%	B
S3	40	34	3	100%	85%	92.5%	A-
S4	29	24	3	97%	83%	90%	A-

A student who can reformulate more than the target number of sentences in the time allowed (e.g. ‘S3’ in Table 13), cannot, however, achieve a fluency score higher than 100%, in order to prevent weighting speed over accuracy. The effect of reporting both grades to students is to draw their attention to the need to practice both, resulting in a marked improvement in grades between the practice and the first RLT, two weeks later (Table 14). In most cases, the

Table 14: *ELC student grades on practice and RLT 1 (in percentages)*

Practice				RLT 1		
Student	Fluency	Accuracy	Avg.	Fluency	Accuracy	Avg.
S1	70	59	65	97	83	90
S2	37	40	39	100	85	93
S3	100	74	87	70	95	83
S4	44	50	47	73	82	78
S5	58	65	62	77	87	82
S6	78	73	76	70	100	85
S7	55	67	61	90	93	92
S8	58	76	67	100	98	99
S9	79	95	87	100	91	96
S10	58	71	65	100	87	94
S11	60	86	73	90	93	92
S13	60	75	68	100	91	96
<b>Average</b>	<b>62</b>	<b>69</b>	<b>65</b>	<b>90</b>	<b>91</b>	<b>91</b>

students’ grades for both accuracy and fluency have increased; however, in the case of ‘S3’,

who was fast but inaccurate on the practice test, the effort to improve accuracy on RLT1 resulted in a marked decrease in output, and therefore a lowered fluency score. From a teaching perspective this is in fact a desirable outcome, especially for students whose characteristic output is fluent yet marked by random or systematic errors.

These grades do not, of course, confirm acquisition of the target forms tested, but they do demonstrate the training effect of the practice test. If anything, the increase in average grade for this group from 65% to 91% might be grounds for questioning test validity. This would be true if the primary purpose of the RLT were to assess overall language proficiency, which it is not and which accounts for the relatively small contribution of the RLTs to the student's overall course grade. Instead, the purpose is to raise learner awareness, first of the need for increased accuracy (or fluency, or both), and second of the learner's ability to attend to these consciously. An additional benefit of this kind of instructional procedure is that it creates a database of language samples and accompanying audio recordings, for individual learners as well as L1 and proficiency groups, which can be investigated for much more specific evidence of language acquisition, as described below and in Chapter 6.

#### **4.7.2 Operationalization of accuracy and fluency**

While the procedure described above for scoring the RLT was adequate for instructional purposes, the operationalization of *accuracy* as 'grammaticality of reformulations' and especially *fluency* as 'number of reformulations in a given time' merits explanation and justification. For instance, should the ability to reconstruct or reformulate an utterance quickly be taken as evidence of *fluency*? Furthermore, if a reformulation is a grammatical sentence in English but does not convey the meaning of the original worksheet error – the 'interlanguage intention', to use Lakshmanan and Selinker's term (2001: 394) – or does, but in a way that is awkward or infelicitous, should it be considered *accurate*? Should only target-like

pronunciation be considered correct? In the context of real-time CF in classrooms there are few clear-cut answers to such questions, which goes some way towards explaining the reported lack of consistency in CF practices (Fanselow 1977; Truscott 1999); but delayed CF allows for greater circumspection on the part of the teacher, as well as reflection on the strengths and weaknesses of individual students, just as it permits greater reflection and noticing on the part of students (Lynch 2001; Lynch and McLean 2003; Kindt 2004; Stillwell et al. 2009).

In order to test the validity of the RLT as it is used to provide measurements of fluency, audio recordings from the RLT (ELC group) were analysed according to standard measures of fluency, where appropriate to what is essentially a non-communicative situation (see Chapter 2). These were:

- a. Speed of delivery, measured by
  - i. words per minute (WPM)
  - ii. speech rate: number of syllables  $\div$  total time
  - iii. articulation rate: number of syllables  $\div$  phonation time
- b. Phonation ratio: phonation time  $\div$  total time
- c. Repetitions and false starts (distinct from self-corrections)

The last of these, repetitions and false starts, was a manual tally of any case in which a word or phrase was repeated with no correction. For instance, a student reformulating the sentence *All the religion in the world is same thing* produced:

*All the religions...All the religions in the world are same.*

This is counted as a false start/repetition for this sentence. As this measure is a ratio of sentences with false starts or repetitions to those without, each sentence was counted only once if it contained either. The student then said,

*No...are the same.*

This is noted as a self-correction, but did not exempt the item from being counted as a false start since the error corrected was not part of the repetition. In contrast, cases of self-correction alone were not counted as repetition.

An important consideration with some of these measures (e.g. counting syllables and timing pauses) is that they are much too time-consuming to be practicable in an ongoing teaching context. Teachers in general have to be content with a holistic, intuitive ‘feel’ for student fluency, rather than systematic and perhaps more objective quantitative measurement of temporal features such as articulation rates (Fulcher 2003). If, however, ways can be found to expedite such measurement without adding to teacher workloads, and if such measurements in fact contribute useful information, then their inclusion in day-to-day practices is to be welcomed. The following sections will therefore discuss the automatic calculation of WPM as a measure of fluency before the description of the study resumes in Section 4.7.4.

### **4.7.3 Automatic calculations of words per minute**

The database of worksheet items automatically calculates a word count for each RLT item and stores the length of the mp3 recording (calculated from file size) for each item. It is thus possible to calculate WPM for each item, as well as an average WPM over the whole test for each student, assuming that the student’s reformulation on the RLT has approximately the same word count as the worksheet item. On the face of it, this assumption is questionable since it is to be hoped that the students’ eventual reformulations will *not* be the same as their worksheet errors. Two factors make the assumption of word count equivalence quite reasonable, however. First, experience has shown that in the vast majority of cases, students’ reformulations are, or attempt to be, verbatim reproductions of the teachers’ reformulations. In general, students accept the reformulations as being appropriate forms for the intended

meanings, and so have no reason to deviate from them on the RLT. Second, as was shown in Module II, teacher reformulations tend to be minimally distorting, so that as many of the student's original words are preserved as possible in achieving target-like accuracy and preserving the intended meaning.

There are, naturally, cases in which a student's utterance is not only erroneous but also quite unidiomatic, especially in the case of lower-proficiency students. Three examples of this type of error are given in Table 15, along with word counts from the worksheet,

Table 15: *Comparison of items and word counts between worksheet item, teacher's reformulation, and students RLT recording*

Worksheet item	Teacher's reformulation	Student's RLT recording
If your son gonna die – I'm sorry about that (9)	If your son were going to die – God forbid. (9)	(same – 9)
Think in logical way! (4)	Think logically! (2)	(same – 2)
This is not good deed as they mean here. (9)	That is not the kind of good deed they mean here. (11)	(same – 11)

teacher's reformulation, and student's reformulation. As can be seen, the syntactic and morphological features of the reformulations can be quite different from the 'interlanguage intention' of the original, introducing complexity without changing the semantic content of the message or, relevant to the present discussion, the average word count.

In order to test the assumption of word count equivalence, 50 worksheet items were chosen at random from the database, along with audio recordings of both teacher reformulation and student RLT. Word counts for the worksheet items were calculated automatically in the database, and those for the teacher reformulations and student RLT were counted manually. As the data were not normally distributed, they were analysed using non-parametric tests (Kruskal-Wallis Chi-Square), which revealed no significant differences in the medians of the three data sets ( $\chi^2(2, 50) = .174, p = ns$ ). In addition, Spearman correlations



were calculated to establish the strength of the association between the three data sets (Table 16). As anticipated, the three are very highly correlated, with an  $r = .933$

Table 16: *Correlation matrix for word counts on worksheet, teachers' reformulations, and students' RLT recordings*

	1	2	3
1. Worksheet items	--	.928	<b>.933</b>
2. Teachers' reformulations		--	.970
3. Students' RLT recordings			--

Note: Correlations are significant at  $p < .001$  ( $N = 50$ )

correlation between the automatic word counts from the database and the students' reformulations. Therefore, it was considered acceptable to use the item length (in words) from the database in WPM calculations (the length of the recorded RLT item divided by the word count for that item), which greatly simplified the process of assessing student fluency on the RLT. One important consequence of this measure is that since the WPM is an *a priori* calculation, if a speaker repeats all or any part of an item it will greatly reduce the reported WPM value. This is in fact advantageous, since it is assumed that repetitions and false starts are indicators of lack of automaticity, and therefore these items will be easily distinguished from those more fluently delivered.

#### 4.7.4 Correlation of fluency measures

Since several measures of fluency are under investigation, it is instructive to see how they relate to one another and to the putative measure of fluency provided by the RLT (see section 4.7.1). For this purpose, the first RLT audio recording from the thirteen students in the ELC group was processed in PRAAT (Boersma and Weenink 2011), a software program for the phonetic analysis of speech, to find average measurements for phonation time, speech rate, and articulation rate; the total word count for all attempted items was used to calculate WPM, as described above; and finally, the repetitions and false starts were manually counted. The

RLT recordings for the IEP group could not be used as the level of background noise on their recordings made accurate detection of syllables impossible (see Appendix 1). Table 17 shows the inter-correlations of each of these variables and the ‘fluency’ score from the RLT (section 4.7.1). The results should not be taken as

Table 17: *Correlation matrix for measures of fluency*

	1	2	3	4	5	6
1. Phonation ratio (phonation time ÷ total time)	--	.936 <i>.000</i>	-.005 <i>.988</i>	.158 <i>.642</i>	.682 <i>.021</i>	.591 <i>.055</i>
2. Speech rate (number of syllables ÷ total time)		--	.289 <i>.390</i>	.166 <i>.626</i>	<b>.758</b> <b>.007</b>	<b>.675</b> <b>.023</b>
3. Articulation rate (number of syllables ÷ phonation time)			--	-.035 <i>.920</i>	.349 <i>.292</i>	.354 <i>.286</i>
4. Repetition/False start ratio (sentences with false starts ÷ total number of sentences)				--	-.397 <i>.227</i>	-.540 <i>.086</i>
5. WPM					--	<b>.975</b> <b>.000</b>
6. RLT <i>fluency</i> score						--

Note: significance shown in italics

definitive, since only thirteen samples were used (each 3 minutes long, containing approximately 30 items); however, several interesting findings emerge. The first is that the RLT ‘fluency’ score correlates very strongly with WPM ( $r = .975$ ) and strongly with speech rate ( $r = .675$ ). RLT ‘fluency’ score correlations with two other measures, phonation time and the ratio of repetitions and false starts to items attempted, approach significance. These findings are consistent with those of Kormos and Dénes (2004), Lennon (1990), Riggensbach (1991) and others. In other words, these findings substantiate the content validity of the RLT ‘fluency’ score as a measurement of oral fluency. At the same time, however, they suggest that articulation rate does not contribute much to this model of fluency. The lack of significant correlation between articulation rate and any other measure could be attributable to random error in the automatic detection of syllables in Praat (see Appendix 1). However, this

same error would affect any measure that includes syllable count, and yet speech rate, which also relies on syllable count, correlates very strongly and significantly with WPM ( $r = .758$ ) as one would expect: the greater the number of words uttered in a given time, the greater the number of syllables also uttered. (This also explains the very strong ( $r = .936$ ) correlation between phonation time and speech rate.) An alternative explanation is that articulation rate, which is the number of syllables uttered during non-silent stretches, is simply a measure of speed of articulation, and perception of ‘fluency’ is only tangentially connected to the speed at which one talks. Language learners sometimes misunderstand this point, assuming that they need to ‘speak faster’ in order to sound more fluent. What is suggested by this exploratory investigation is that while fluency is characterized in part by how much language is produced in a turn (as evidenced by correlations of phonation ratio, WPM, and speech rate), it is the absence of dysfluencies (pause time and repetitions or false starts) that is a better indicator of fluent speech, defined as ‘the extent to which language produced in performing a task manifests pausing, hesitation, or reformulation’ (R. Ellis 2003: 342). Simply producing more syllables in a given time does not equate with fluency.

Instead, fluency may have more to do with the *appropriate* pausing that results from not having to pause to apply grammatical rules. As Chambers puts it:

Developing automatised mechanisms contributes to diminishing the processing load; as long as conscious efforts are required to produce accurate morphology, less space is available for other planning tasks and this is reflected in ‘choppy’ utterances (Chambers 1997: 537–8).

This also points to the connection between fluency and the automatic retrieval of lexical items and formulaic ‘chunks’, which is certainly a substantial part of what is being measured in the RLT. Ejzenberg (2000), for instance, compared the use of formulaic language by speakers rated fluent and non-fluent by human judges, and found that the more fluent used prefabricated chunks more efficiently, whereas non-fluent speakers frequently used formulae

inappropriately.

While there is no guarantee that fluent production of items on the RLT will translate to fluency in general oral production or in the production of analogous forms, for example that the reformulation *All the religions in the world are the same* will become the template for subsequent production of sentences with the structure *All the [NP] in the [NP] are the same*, there is evidence that approaching this type of error as a formulaic ‘chunk’ with high re-use potential (N. Ellis 1996; Gatbonton and Segalowitz 2005; Wray 2000; 2002) may be more beneficial than expecting the learner to analyse the constituent syntactic and morphological parts of the phrase or clause and to re-assemble these on the fly during conversation.

#### **4.7.5 Fluency baselines for individual participants**

The measurements of fluency calculated for each speaker provide an individual baseline fluency rating against which any RLT item can be compared. The three measurements found to be most predictive of fluency (phonation ratio, WPM, and speech rate) can be combined and standardized, by creating z-scores of each measure for each item and combining these, which gives an indication of the relative distance, either positive or negative, from the mean. Thus, for example, the RLT data for one participant can be examined for more or less fluent production, as well as inaccurate performance (Figure 9). Whether or not they are accurate, items which are reconstructed with greater fluency (e.g. item 49403 at the top of the figure) can thus be assumed to be more automatized. Conversely, hesitant or otherwise dysfluent items (e.g. item 49612 at the bottom) can be assumed to be indicative of more ‘conscious effort’, in Chamber’s words, mirroring the differential competence described in the context of TGJTs (section 3.6, above).

In Figure 9, the scale of the horizontal axis shows the distance above or below the mean (shown as 0), in units of  $\frac{1}{2} SD$ . Once again, the range of  $\pm 1 SD$  from the mean will be

assumed to represent ‘normal’ performance, since the  $\pm 1$  SD range in a standard normal

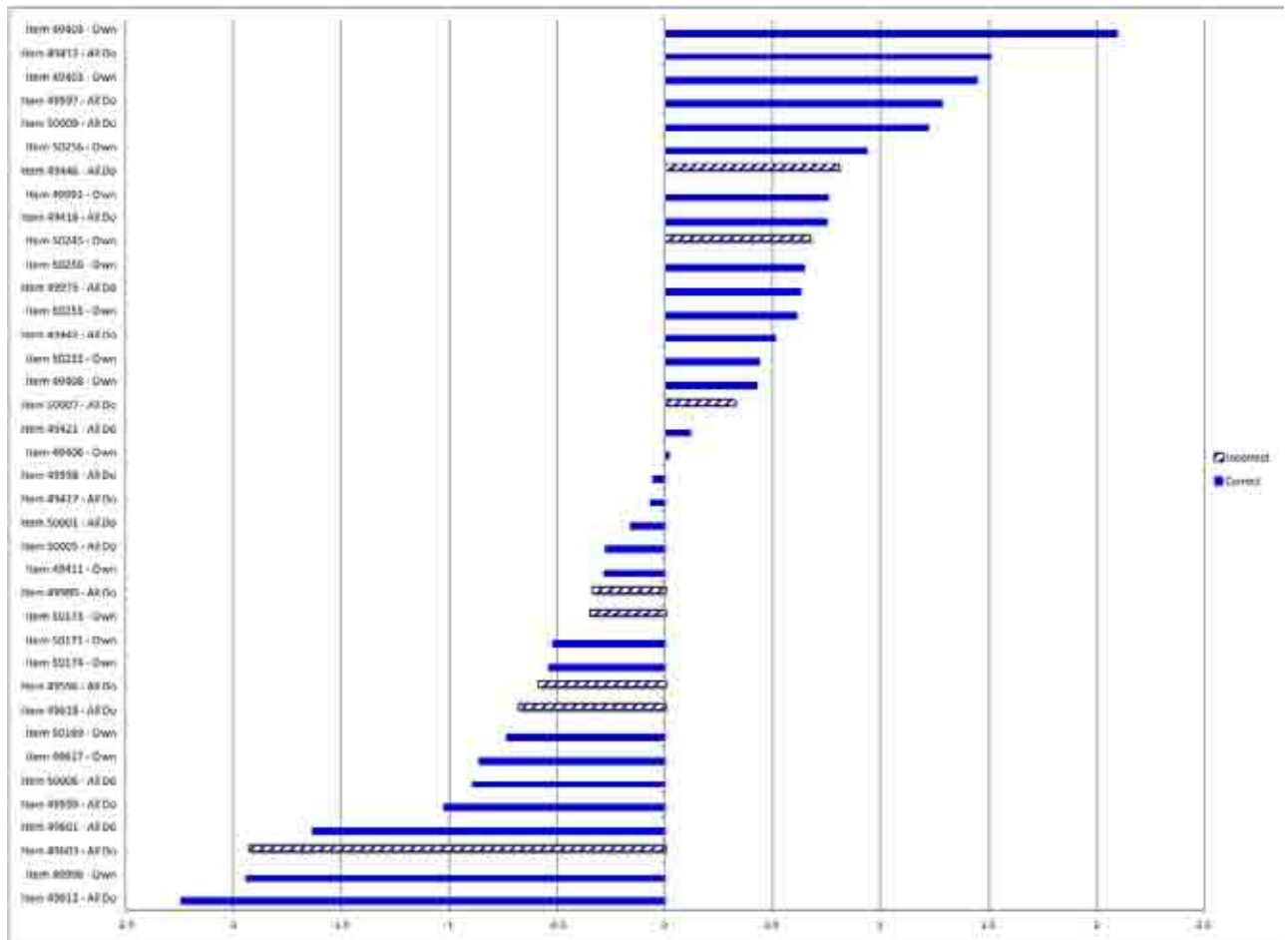


Figure 9: Standardized z-scores for RLT 1 items for one participant

distribution encompasses approximately 70% of the data, which is to say that the linguistic forms are neither fully automatized nor necessitate fully conscious control. If this is a reasonable assumption, item 49403 should be fully automatized. In fact, the occurrence of this item is serendipitous, since the speaker accidentally recorded it twice: for the RLT, the students receive a clean copy of both their own Running List and the ‘All Do’ list, and if an item is on both lists, they are asked to ignore it the second time. This student recorded the item, which was her own error and which happened to be on the ‘All Do’ list (it is still her error, so appears as ‘own’ in Figure 9 both times), and then recorded it again later in the test, about a minute and a half later. What is interesting is that the topmost of the items in Figure 9

is the second recording, which gives some indication of the moderate effect that practice has on the fluency of her delivery. It is also confirmation that this methodology for measuring fluency is robust, since it places the two recordings close to one another. One final observation about this item is that the worksheet error from the Small Talk conversation five weeks prior to the RLT was *Do you know what will happen at future?*, which was reformulated to *Do you know what will happen in the future?* Since the speaker was clearly able to produce the correct form, and fluently, on the RLT, it is likely that the item was a ‘slip’ in the first place. It would be premature to draw any conclusions about the effectiveness of the CF methodology based on this evidence, but it would be logical to assume that no further remedial effort is necessary for this item.

In contrast, the five items at the bottom of Figure 9 which are 1 *SD* or more below the mean can probably be assumed not to be automatic for the speaker, and should remain active on her Running List, as should any items within the  $\pm 1$  *SD* range which are incorrect. Since the items are tracked in the database, it is a relatively simple step for teachers to mark items for inclusion or exclusion as they listen to them, thus creating specific focus on forms tailored to the needs of the individual learner.

#### **4.7.6 Accuracy – grammaticality and acceptability**

Measures of accuracy are generally considered less complex than those of fluency, but as discussed above (see especially section 3.4.6), they are not without controversy and are probably as susceptible to contextual considerations as fluency measurements. In the context of delayed CF, much of this variation is fortunately controlled for, since the context in which the ‘interlanguage intention’ was produced is known, as is the target reformulation. There remains the question of whether the RLT item represents ‘performance which is native-like through its rule-governed nature’ (Skehan 1985) or, to paraphrase Lennon, (1991: 182) a

linguistic form which, in the same context, would likely be produced by the speaker's native speaker counterparts. Perhaps a useful addendum to these definitions would be: *and which would, in the same context, be comprehensible to the speaker's NS counterparts with minimal negotiation*. This revised definition allows for minor differences from NS performance, of the same order as the sort of differences a NS from one part of England might encounter in conversation with a NS from another. Thus, as a reformulation for the first item in Table 15, *If your son is going to die – God forbid...* would be acceptable because a NS might well prefer the immediacy of a present unreal (first) conditional over the past unreal; but *If your son going to die – God forbid...* would not be. Similarly, non-target-like pronunciation that causes the listener to think of one or more alternatives to the intended form, or forces the listener to engage considerable resources in decoding the intended form, would not be acceptable. Clearly, the subjective judgement of the listener plays a large part in such decisions, but as the high reliability score of the pilot TGJT (Chapter 3) demonstrates, there is considerable agreement as to what constitutes an error, even when contextual information is not known. To verify this in the context of RLT production, the two teachers involved judged each other's students RLT performances as well as their own. Reliability of the accuracy scores was found to be very high ( $\alpha = .96$ ).

Finally, as we saw in Section 3.4.3, the fact that the RLT consists of authentic learner production means that items are likely to contain more than one non-target-like form. This is both a weakness and a strength, as it adds considerable complexity to the diagnostic process, yet yields finer-grained information about learners' ILs than would a test designed to measure competence in a single linguistic feature. From an ecological perspective, in any case, overall accuracy – not simply accuracy on forms pre-selected for research purposes – is of overriding concern; for this reason, items are judged inaccurate if they contain *any* inaccuracies, even if

the original errors have been successfully repaired.

#### **4.7.7 Diagnostic use of accuracy and fluency measures**

Given the range of fluency values for the speaker in Figure 9 and the relatively even distribution of incorrect items within the ‘normal’ range, what we would hope for in future RLTs, and of course in her fluent production both in and out of the classroom, is that the incorrect items would be corrected and gradually become more fluent; and that none of the correct items would conversely become incorrect or significantly less fluent. Were SLA a linear process, such a hope might be justified. However, it is well established in the field that language acquisition is far from linear, progress instead being characterized by a U-shaped curve and frequent ‘backsliding’ (Bowerman 1982; Kellerman 1985; Larsen-Freeman 2006; R. Ellis 2008a). There are a number of plausible reasons for an IL form to change over time, and such changes could make production of the form more automatic or less, and the form itself more target-like or less. Nonetheless, if we are correct in assuming that the highly automatized (i.e. fluently delivered) items are fully acquired, it would be surprising to see them become less so over a period of a few months; this would also hold for both target-like and non-target-like items, the latter being, presumably, coterminous with ‘fossilized’ forms. Skehan provides a succinct explanation of how this might function:

An instance-based approach also provides an interesting theoretical interpretation of the phenomenon of fossilization, in that one can now regard such an outcome as the premature product of a rule-based system which is then made available as an exemplar in future language use. In other words, there is no requirement that what are created as exemplars are correct. In beneficial circumstances rule-created exemplars may be supplanted by other exemplars which are created when the underlying rule-based system has evolved more. But if the underlying system does not so evolve, and if communicative effectiveness is achieved, the erroneous exemplar may survive and stabilize, and become a syntactic fossil. In this case, paradoxically, it is the usefulness in communication of a premature lexicalization that is the source of the enduring problem. (Skehan 1998: 61)

This account, if correct, is precisely why CF is so important in communicative



methodologies: without it, ‘premature lexicalization’ of any item that is communicatively adequate, target-like or not, might take place simply because it results in greater communicative efficiency. This is not to deny a role for novel, rule-based production, but rather to recast it in a supporting rather than leading role. As Lamb (1998: 169, cited in Wray 2002: 10) puts it:

Linguists seem to underestimate the great capacity of the human mind to remember things while overestimating the extent to which humans process information by complex processes of calculation rather than by simply using prefabricated units from memory.

## 4.8 Findings

As described above (Section 4.7), two groups of students (ELC and IEP) were given two RLTs during the 14-week semester, one six weeks into the semester and the other approximately eight weeks later. During these eight weeks, of course, the students’ Running Lists continued to grow as more items were added (Table 18). This means that the two tests

Table 18: *Descriptive statistics for RLT test data*

Group	N	Item Count RLT 1	Item Count RLT 2	Total	Items common to both RLTs
ELC	13	380	419	799	65 (8%)
IEP	13	311	381	692	163 (23%)

did not contain the same items each time, although there was some overlap – 65 items (8% of the combined total on RLT 1 & 2) for the multilingual ELC group and 163 items (23%) for the monolingual IEP group. Overall, both groups accurately reformulated 70–80% of their original errors, in both cases becoming more accurate on RLT 2.

### 4.8.1 Overall differences between RLT 1 & 2

#### ELC group

To determine performance differences between the two RLTs, each student’s fluency and accuracy scores on each test were compared using a t-test comparison of means. (In this case,

the combined fluency score could not be used as it is mean-dependent for each student; therefore, the three separate measures of *fluency* were compared – see Table 19). Significant differences were found in the means for all measures except *accuracy*, although this difference approached significance. The small increase in fluency on the second RLT

Table 19: *T-test comparison of RLT 1 and RLT 2, ELC group*

	<b>RLT 1 Mean (SD)</b>	<b>RLT 2 Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b><i>t</i></b>	<b>sig.</b>
Accuracy	0.76**	0.81	0.05	766	1.716	.09
WPM	118.5 (43)	126 (43)	7.5	793	2.474	.01
Phonation ratio	0.70 (0.15)	0.76 (0.15)	0.06	793	5.559	< .0001
Speech rate	2.35 (0.69)	2.65 (0.70)	0.3	793	6.029	< .0001

*\*unequal variances assumed; \*\* accuracy score max = 1*

could be the result of increased familiarity with the test format; alternatively, the new items on the second test might have been easier on average for the students. A similar analysis was therefore performed on only those 65 items on RLT 1 which were repeated on RLT 2 (Table 20), in effect making RLT 2 a post-test. Once again, no significant difference was found for

Table 20: *Paired-sample t-test of mean differences between RLT 1 and RLT 2, ELC group*

	<b>RLT 1 Mean (SD)</b>	<b>RLT 2 Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b><i>t</i></b>	<b>sig.</b>
Accuracy	0.77	0.75	0.02	64	0.275	.78
WPM	118.7 (48)	143 (45)	24.3	64	3.644	.001
Phonation ratio	0.72 (0.15)	0.83 (0.14)	0.11	64	3.755	< .0001
Speech rate	2.24 (0.77)	2.82 (0.65)	0.58	64	6.19	< .0001

*accuracy*, but an inspection of the distribution of inaccuracies revealed that only five incorrect items from the first test remained inaccurate on the second, while seven items which had been correct on RLT 1 became incorrect on RLT 2. At the same time, there was a small but significant increase on all measures of fluency, indicating that the effect of the CF was to increase the automaticity with which students could reformulate the target forms. On the face

of it, this is a disappointing result, but as will be explained in Section 4.8.2, the loss in accuracy was often accompanied by an increase in complexity, here defined as ‘[t]he extent to which the language produced in performing a task is elaborate and varied’ (Ellis 2003: 340). This definition thus follows current practices in regarding complexity as more than the degree of subordination (e.g. mean number of clauses per T-unit), but as including ‘the size, elaborateness, richness, and diversity of the learner’s linguistic L2 system’ (Housen and Kuiken, 2009: 464) or more simply, ‘structural variety and sophistication’ (Norris and Ortega 2009: 567).

### IEP group

As discussed above (Section 4.6), the IEP group was included in the research in order to determine whether the effects of the CF methodology would differ with different learners. The IEP group had fewer Small Talk sessions, and thus had fewer RLT items than the ELC group, but the provision of CF was done in exactly the same way. Just as with the ELC group, the IEP students were successful in producing target forms on the first RLT at approximately the same rate of (76% for the ELC group, 69% for the IEP group). A comparison of means between RLT 1 and 2 (Table 21)

Table 21: *T-test comparison of RLT 1 and RLT 2, IEP group*

	<b>RLT 1 Mean (SD)</b>	<b>RLT 2 Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b><i>t</i></b>	<b>sig.</b>
Accuracy	0.69	0.79	0.1	690	3.066	.002
WPM	112.8 (46)	128 (57)	15.2	690	3.864	< .0001

revealed significant increases in both accuracy and fluency (as measured by WPM): IEP students were 10% more accurate and 15 WPM faster on their reformulations on the second RLT. These results are somewhat more impressive than those for the ELC group (who

showed only a 5% gain in accuracy and 7.5 WPM gain on the second RLT – Table 19), but in both cases, the majority of the original errors were successfully corrected and could be more fluently delivered on the second RLT. As with the ELC group, an analysis of only those items common to both RLTs was performed in order to create a post-test effect, and it was found that the group increased in accuracy by 9% and fluency by 13 WPM (Table 22). For this group, there were many more items common to both RLTs, which the classroom teacher attributed to the smaller number of Small Talk sessions and to the monolingual composition of the group. The gains on this subset of items parallel those of the total sample so closely that

Table 22: Paired-sample *t*-test of mean differences between RLT 1 and RLT 2, IEP group

	<b>RLT 1 Mean (SD)</b>	<b>RLT 2 Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b><i>t</i></b>	<b>sig.</b>
Accuracy	0.70	0.79	0.09	162	4.763	.04
WPM	115.1 (46)	128 (57)	13	162	2.042	< .0001

it is reasonable to assume that the effects of the CF can be generalized to any set of items.

#### **4.8.2 Individual differences between RLT 1 & 2**

##### ELC group

The small decrease in accuracy (Table 20) for the ELC group deserves closer inspection since the overall goal of CF is to increase the accuracy of production – so increased fluency on *inaccurate* forms is clearly not a desirable outcome. Experience would suggest that different learners respond differently to CF, as do different types of error within the IL systems of different learners. To investigate these differential effects, each student's performance on items repeated on RLT 1 and 2 was analysed as a subset, allowing for comparison of production fluency against the mean. An example is shown in Figure 10. In every case the items were produced more fluently on the second RLT. In one case (Item 49986), the student

was correct on RLT 1 but then introduced a new error on the second RLT:

original Small Talk error: *\*It's childhood innocent.* ('it' = the gullibility of the young)

reformulation on RLT 1: *It's childhood innocence.*

reformulation on RLT 2: *\*It's a childhood innocence.*

On the surface, the introduction of an NP determination error here is suggestive of confusion between count and non-count nouns; however, since this is an Arabic L1 learner, it is reasonable to hypothesize that he is overgeneralizing to all singular nouns the use of the indefinite article for generic referents, as Arabic would use no article in this context (Harrat

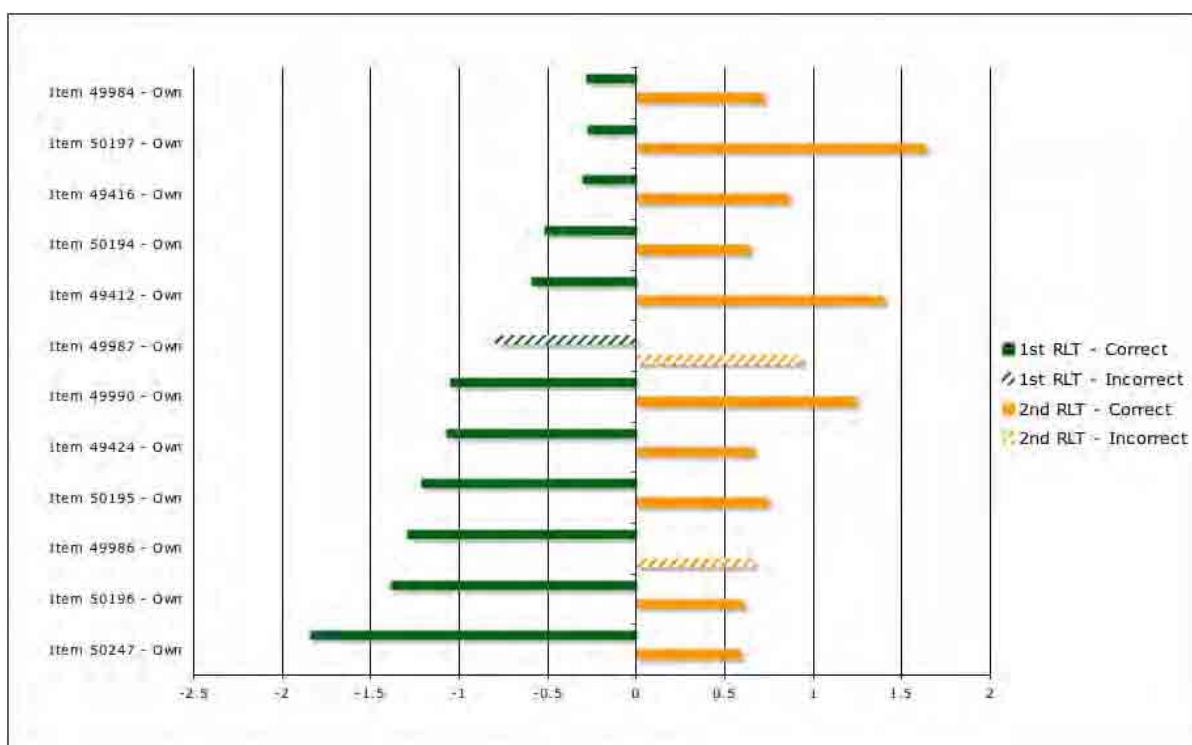


Figure 10: Comparison of RLT 1 & 2 for 'Falshehri'

2011). Support for this hypothesis comes from the same student's use of the indefinite article in another sentence (not in the dataset in Figure 10) *\*If you are a poor, you cannot show yourself*, which follows the same pattern in Arabic as the generic NP above. These examples illustrate the complexity involved in the identification of error type and cause – which leads Truscott (1999) and others to question the ability of language teachers to provide effective CF – and underscore the need for CF which is tailored to the individual learner.

The second inaccuracy in this dataset, Item 49987, is perhaps more troubling since not only is the RLT version inaccurate, but it is more fluently produced. Once again, the speaker introduced a new error:

original Small Talk error: \* *It's changing when you get older.* ('it' = childhood innocence)

reformulation on RLT 1: \* *It changes /tʃeɪndʒs/ when you get older.*

reformulation on RLT 2: \* *It changed when you get older.*

In this case it was the pronunciation of the 3<sup>rd</sup>-person inflection following a fricative final verb stem consonant that was (perhaps unfairly) identified as an error on RLT 1 by the teacher, but the student may have assumed that there was still a problem with the tense sequencing and opted for a past form. There are no examples of a similar use of the past simple in this student's production (in fact, he correctly produced a parallel sentence: *When you are famous, your name is known by everyone.*), and therefore this could be considered a 'slip' rather than an error. However, drawing it to his attention is still worthwhile, as the attentional focus would help to reveal any systematic misunderstanding that might have resulted in this new inaccuracy.

In many cases (see Appendix 3 for all RLT data), the pattern of change between RLT 1 and 2 was not as clear as that of 'Falshehri'. More typical is the sort of variation shown in Figure 11, where some items become more fluent while others become less so. Once again, it is not the case that the original errors are simply being repeated; rather, the original errors are successfully corrected and new errors are introduced. An example from the data set in Figure 11 (Item 50249) illustrates this:

original Small Talk error: \* *You think the man who does a plastic surgery is gay?*

reformulation on RLT 1: *Do you think that a man who has plastic surgery is gay?*

reformulation on RLT 2: \* *Do you think that a man who has a plastic surgery is a gay?*

'Ralhassan', is also an Arabic L1 speaker, and her original sentence shows the influence of

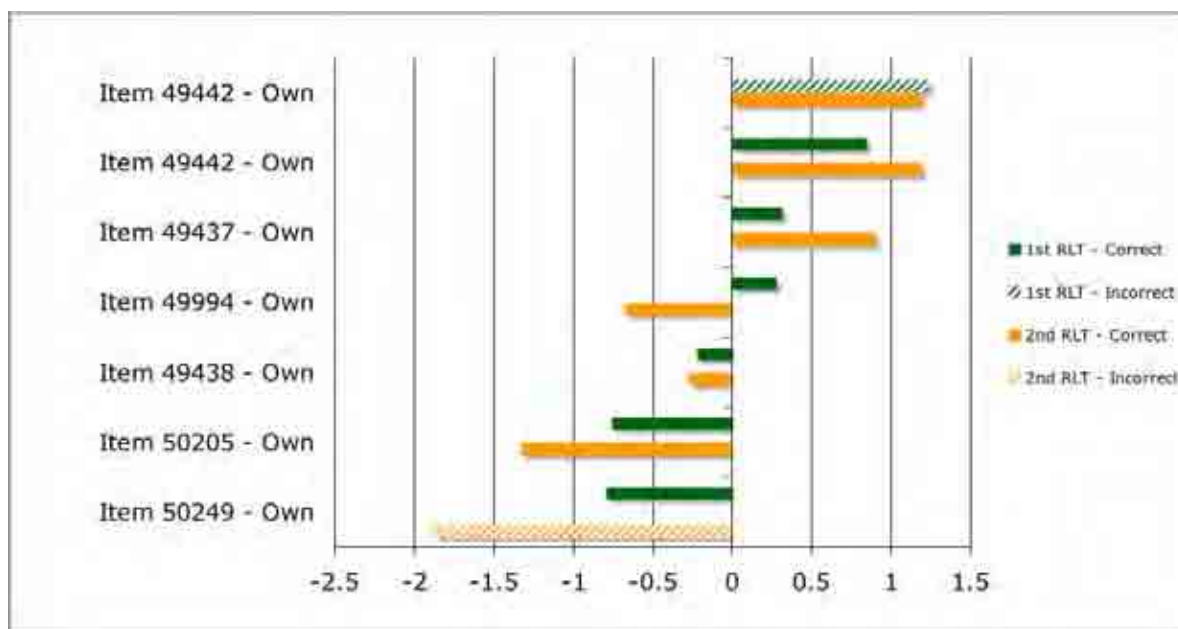


Figure 11: Comparison of RLT 1 & 2 for 'Ralhassan'

Arabic with the generic definite article along with the collocation of *does* with *surgery*. Once again we see the (re)introduction of indefinite articles as though, under the pressure of time, the speaker is applying a blanket article rule to all NPs. This is very useful information for a teacher, as it informs decisions as to which forms to focus on and, possible, how to approach the task of helping the speaker to further her understanding of those forms and the functions they serve.

### IEP group

Naturally, there are considerable individual performance differences within this subset, as there were for the ELC group (see Appendix 6). For instance, in some cases, the reformulations were averagely faster and more accurate on RLT 2 (Figure 12), showing unambiguous evidence for the effectiveness of the CF provided. As with the ELC students, where an inaccuracy was present on both RLTs, it was usually a different error, as happened with item 3-19:

original Small Talk error: \* *I heard last talking.*

reformulation on RLT 1: \* *I heard from last talking.*

reformulation on RLT 2: \* *I heard from last group I talked to.*

This introduction of new errors can again be seen not as a failure of the CF, but as evidence for developing language. In this example, the original error is a literal translation from Japanese: 先き聞いた話 (*saki kiita hanashi* – ‘before heard talking’). The teacher’s reformulation of this sentence was *I heard that from the last group I talked to*, and in her reformulation on RLT 1, the student can barely recall any of this. By RLT 2, she is much

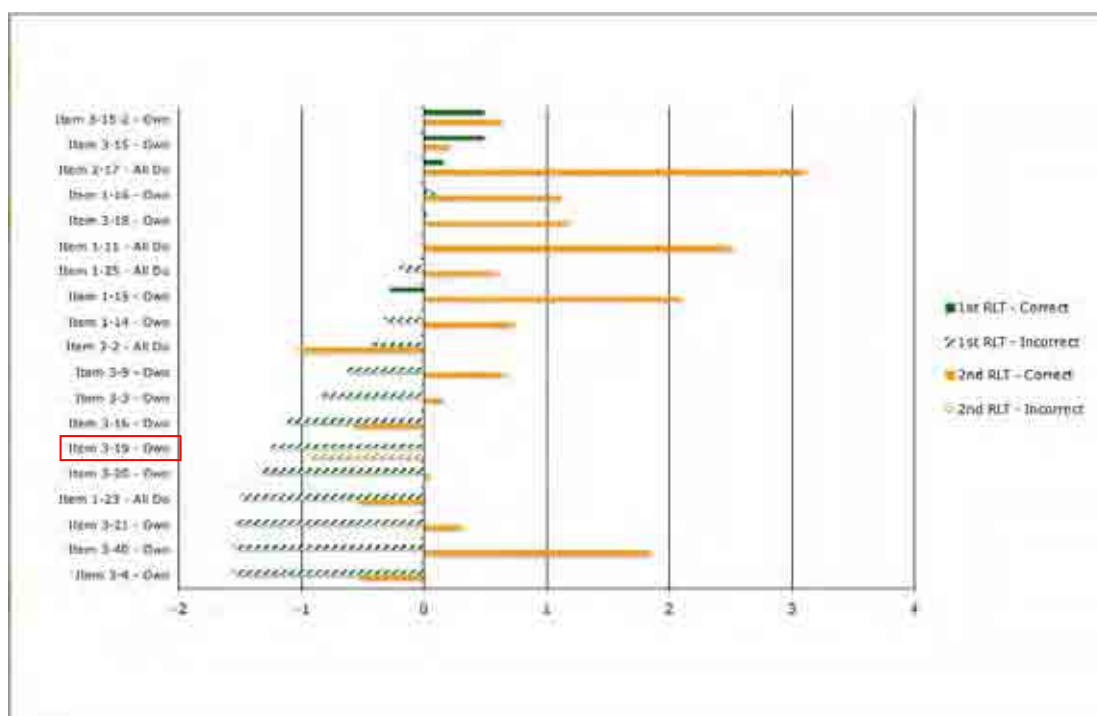


Figure 12: Comparison of RLT 1 & 2 for 'Riko'

closer, but still misses the object and a definite article. In terms of raw accuracy counts, this is still a failure; but in terms of language development, the CF has pushed the speaker towards greater complexity – and indeed, accuracy, since the final version conveys so much more meaning than the original utterance. It is not surprising that a Japanese speaker should have omitted an article, and an analysis of the teacher’s reformulation shows that the syllable /ðə/



is spoken in 127 ms, compared with 407 ms for /læst/, making it considerably less salient.

When this was pointed out to the teacher, she confirmed that students often have difficulty hearing such features, and said that she consequently includes the additional step of having them write the reformulation on the worksheet as they listen to the teacher's recording and then check the reformulation with her before the final RLT. For learners with limited exposure to oral input, this may be an essential adaptation.

One final observation can be made here regarding performance 'slips' and systematic errors: one would anticipate that a 'slip' would be easily correctable, so that the reformulation would be both fast and accurate. The data from 'Riko' in Figure 12 provide an example of this in item 3-15, which the student recorded once on RLT 1 and (accidentally) twice on RLT 2. The original error was \* *You bought many toy?*, and in all three reformulations, she clearly says 'toys'. The fluency of all three reformulations suggest that the original worksheet item was simply a 'slip' – although in marking it as 'All Do', the teacher clearly thought that it would be beneficial for all students to correct.

### 4.8.3 Relationship between accuracy and fluency

#### ELC group

In the preceding description, we have seen a number of examples of variation in performance from original Small Talk utterance to reformulation on RLT 1 and 2. This variation could be attributable to variable competence (Tarone 1983; R. Ellis 1985, 1989; see also Gregg 1990 for a critique of the variabilist position): for instance, 'Ralhassan' might have the parallel forms 'gay' (*adj*, far more common in US English) and 'a gay' (N, still common in British English). In this case, we could expect her to use the two forms fairly interchangeably in free variation. Owen (2007: 333) characterizes *free variation*, at least in NS production, as 'arguably a euphemism for 'here the linguist gives up the search for explanation' ', while

Gregg (1990: 379), arguing from the orthodox Chomksyan position that linguistic *competence* alone is the concern of the linguist, rejects the notion that such variation constitutes an *explanandum* of SLA theory at all: ‘Variation is... a fascinating and puzzling phenomenon... but one that it is not the duty of an acquisition theorist to explain.’

An alternative explanation for the apparent variation in ‘Ralhassan’s’ production is that she has no established ‘form’ at all in English, i.e. that she has a semantic item in mind, with all of its associations and L1 forms, but only a vague notion of how to express this in English, and so in the context of fluent conversation (or a timed production test), it comes out variably as ‘gay’ and ‘a gay’. The point here is that where such uncertainty exists, production should be measurably less fluent. The assembly of constituent elements required to express the desired meaning takes some conscious effort; words have to be selected and grammatical rules, whether the product of explicit training or of private analysis (conscious or unconscious), have to be applied *because* there is no ready-made ‘chunk’ available. There is no guarantee, of course, that a ready-made ‘chunk’ would be target-like, but it stands to reason that non-target-like forms, at least with this sample of students, who have not had prolonged exposure to and experience in the L2, should be marked by lack of fluency to a greater extent than target-like forms. In other words, accurate forms, in general, should be more fluently produced.

To test this hypothesis, an ANOVA was performed on the data from RLT 1, using *accuracy* as the between-groups effect (effectively treating correct and incorrect responses as different groups), and a significant effect was found for WPM,  $F(1, 379) = 5.74, p = .017$ , and Speech Rate,  $F(1, 379) = 4.05, p = .045$ , but not for Phonation Ratio. In other words, when participants got the reformulations wrong, they were averagely less fluent than when they got them right. There was no difference in the amount of pausing, however – hence the lack of

significant difference for Phonation Ratio. The same analysis on the data from RLT 2 yielded slightly different results. This time, a significant effect was found for WPM,  $F(1, 418) = 7.52$ ,  $p = .006$ , and Phonation Ratio,  $F(1, 418) = 9.17$ ,  $p = .003$ , but not for Speech Rate. That the incorrect reformulations on the two tests should have affected different aspects of the *fluency* construct is surprising, particularly since a closer investigation of the data also revealed that on both tests, there was a significant difference in the overall duration of the reformulations on incorrect items. This means that participants generally took longer to reformulate when they were incorrect, but on RLT 1 this was due to slower delivery, while on RLT 2 it was due to greater pausing. A possible explanation is that on RLT 2 the participants had reached their plateau in terms of how fast they could utter the words, perhaps as a result of more intensive practice beforehand, and consequently ‘slowing down’ to reformulate items about which they were not confident was manifested in greater pausing rather than slower speech. Whatever the explanation, on both tests the WPM difference is significant, and so a comparison can be made with the performance of the IEP group (see below). The WPM difference between correct and incorrect items was not attributable to a difference in word counts, however. It is not the case that the items that participants got wrong also had more words to reformulate, as no significant difference was found on the word counts of these items and ones which participants reformulated correctly.

The overall decrease in fluency observed on items which participants reformulated incorrectly suggests that fossilization (as represented by incorrect items fluently delivered) is not a significant issue with this population. This is a good thing, since there is evidence that learners who show a high degree of confidence on both accurate and inaccurate production make slower progress than less confident learners (Sorace 1985). It is possible, but unlikely, that the participants in the ELC group do have fossilized forms that simply did not appear on

their worksheets and therefore on the RLTs. However, as was discussed in Module II, persistent errors are highly likely to be picked up by teachers in the Small Talk CF process, so one would expect precisely such items to appear on the Running Lists of any group. There are examples in the data of items fluently delivered which are consistently incorrect (as opposed to introducing new inaccuracies), but they are rare. The following is item 50088 from ‘Falkharashi’ (see Appendix 3), which was delivered within the speaker’s ‘normal’ fluency range on both RLTs:

original Small Talk error:	* ... <i>by giving you advices</i> . (how parents help their children)
reformulation on RLT 1:	* ... <i>by giving you an advice</i> .
reformulation on RLT 2:	* ... <i>by giving you an advice</i> .

Since ‘Falkharashi’ received feedback on which items remained incorrect on RLT 1, it can be posited that the feedback he received was not specific enough to draw his attention to the source of the problem, i.e. that a non-count noun not only cannot be pluralized, but also cannot be determined by certain determiners, the indefinite article being one. An alternative explanation is conceptual transfer from L1 (Odlin 2005), in which the count status of the noun *advice* has ‘binding power’ for this learner (obviously ‘advice’ can be plural, since my parents give me lots of it!) even in the face of direct and repeated evidence to the contrary. This is one instance in which a ‘think aloud’ protocol (Birdsong 1989) would be of great benefit to the investigation, and future research should incorporate this element, at least for this type of error. In practical terms, such items should be singled out and discussed explicitly with the student.

### IEP group

The discovery of a relationship between inaccuracy and fluency in the RLT data of the ELC

group prompted a similar investigation into the IEP group's data. On RLT 1, a significant difference was found for both WPM,  $F(1, 309) = 16.1, p < .0001$ , and Duration  $F(1, 309) = 19.78, p < .0001$ . On RLT 2, as with the ELC group, the difference was less marked but still significant for WPM,  $F(1, 377) = 6.16, p = .014$ , and Duration,  $F(1, 377) = 6.92, p = .009$ . Unfortunately, it was not possible to compare differences in pausing and speech rate, as this data could not be calculated for the IEP owing to background noise on the recordings (see Appendix 1). Nevertheless, these results confirm the finding that fluency and accuracy are interdependent, which in turn suggests that the language production of these students is not marked by considerable levels of fossilization.

#### **4.8.4 Comparison of performance on own and 'All Do' errors**

##### ELC group

A random element of the RLT is the inclusion of 'All Do' items (see Appendix 5), errors from individual students which are selected by the teacher for all students to correct owing to a) the usefulness of the form/function, b) the typicality of the error, or c) to compensate for avoidance strategies by some students. One example above was the sentence *If I had hit that barrier, I would have died*, which was selected as an 'All Do' sentence for reasons a) and c). For the ELC group, approximately 20% of the worksheet items were selected as 'All Do'. To determine whether including these items made an impact on overall fluency or accuracy, mean differences were calculated between the two types of item (Table 23). (For the purposes of this comparison these categories are exclusive, so 'All Do' means only those items not counted as 'Own' for each student.) No significant differences were found for any measure except WPM, which was found to be slightly lower for the 'All Do' items. The most likely source of this difference is the speaker's lack of familiarity with the intended meaning of at least some 'All Do' items: although every student would have seen every item on the

Table 23: *T-test of performance measures for own and others' errors, both RLTs, ELC group*

	<b>Own errors Mean (SD)</b>	<b>'All Do' errors Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b>t</b>	<b>sig.</b>
Accuracy	0.76	0.79	0.03	713.8*	1.045	.29
WPM	127 (46)	119 (41)	7.6	793	-2.479	.02
Phonation ratio	0.74 (0.15)	0.73 (0.14)	0.01	793	0.595	.55
Speech rate	2.55 (0.69)	2.47 (0.73)	0.07	793	1.384	.17

\* *unequal variances assumed*

Worksheet, they might not have been in the speaker's group when the error was originally made, and would thus first have to try to imagine what meaning was intended before reformulating the item. The conclusion, however, is that on average students are no more or less accurate and fluent when reformulating their own errors than those of peers of approximately equal proficiency.

#### IEP group

For the IEP group, approximately 25% of the worksheet items were selected as 'All Do'.

Once again, a comparison of means revealed no significant differences in either the accuracy or fluency of participants' corrections of their own and peers' errors (Table 24). This confirms the finding

Table 24: *T-test of performance measures for own and others' errors, both RLTs, IEP group*

	<b>Own errors Mean (SD)</b>	<b>'All Do' errors Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b>t</b>	<b>sig.</b>
Accuracy	0.77	0.73	0.04	682.39*	1.261	.21
WPM	121 (52)	121 (54)	7.6	691	0.004	.997

\* *unequal variances assumed*

that students are able to reformulate the errors of peers, and given the accuracy and fluency with which they can do so confirms the intuition that the inclusion of peer errors for the reason suggested above is beneficial.

#### 4.8.5 Item learning and rule learning

A final point must be made about the assumption of systemic acquisition underlying the research questions in this investigation. Those successful reformulations which can be identified as acquired linguistic knowledge could be merely examples of item learning, in contrast to rule learning (Hulstijn and De Graaf 1994; R. Ellis 1999), and therefore of limited future use. In other words, the student who learns the item *If I had hit that barrier, I would have died* may not be able to extract systemic rules from the item which she can generalize to other past unreal hypothesizing. An apparent example of the lack of rule-learning came from the data from ‘Malhuzaim’ (Appendix 3):

original Small Talk error:   *\*I agree with school uniform for three reason.*

reformulation on RLT 2:   *\*I agree with school uniforms for three reason.*

On the surface, it seems evident that this learner cannot automatically apply target-like pluralization rules, so if there were a tendency for memorized chunks to be analyzed for generalizable rules, it is surely not on display here. But perhaps it is: the functions of the pluralization the two NPs in this item are quite different, even though the form is similar. In Arabic (‘Malhuzaim’s’ L1), the first NP would not be plural at all, instead being marked with the generic definite article, while the second would be. For whatever reason, this student’s focus appears to be on the generic plural, and as a result, he fails to pluralize *reason*. Thus it is impossible to say, from this example, that system learning has taken place, just as it is impossible to say that it hasn’t. More evidence would need to be amassed, both from production in Small Talk and performance on RLTs, as well as from other sources (as will be investigated in the following chapter) before a more confident diagnosis could be made. In any case, in the context of the RLT, this would be considered unsuccessful as item learning, since the entire item has to be accurate, so it would remain on his Running List as remedial CF.

#### 4.9 Discussion

This chapter has provided one perspective on the effectiveness of delayed CF by means of a test which combines features of an elicited imitation test (Erlam 2006) and a correction task. Specifically, it addressed three aspects of the CF process: the accuracy and fluency with which learners can correct their own spoken errors, the differential ability to do so with the errors of peers, and their consistency over time. Criteria for the subjective judgement of accuracy were established and were found to be highly reliable ( $\alpha = .96$ ). Two automatic measures of temporal aspects of speech production (*speech rate* and *phonation time*) were found to correlate highly with automatic calculations of words per minute (WPM), and these three measures combined to provide a standard measure of fluency for the tests. The goal of using automatic measures of fluency was to investigate the possible use of such measures as part of the normal pedagogical process; that is, the information about individual IL development that they provide could be very helpful in guiding pedagogical choices, but only if they do not substantially increase typical teacher workloads. It was found that given high-quality audio recordings with minimal background noise, the automatic measures could be used to process large quantities of data (799 files for the ELC group, approximately 1 hour of audio) in a few minutes with tolerable inaccuracy. Unfortunately, it was impossible to use two of these measures with the IEP group owing to the high level of background sound on the recordings. Instead, only WPM was used as a measure of fluency for that group. The very strong correlation between the three measures makes this an acceptable compromise, but this is acknowledged as a major weakness of the data collection process.

In addressing the first research question, *To what extent are learners able to reformulate errors correctly and fluently in a delayed test?*, it was found that learners in both groups were able to correct the spoken errors at an average accuracy level of 75% (ELC) and



69% (IEP) on the first test. On the second test, which contained largely different items for both groups, the average accuracy levels increased to 81% and 79% respectively. The average level of fluency also increased for both groups on the second test. The most conservative interpretation of these facts is that both groups simply tried harder on the second test, despite the fact that scores from both tests were included in students' overall grades for the class. In any case, the ability to correct approximately 70% of their errors is evidence that the CF is having a beneficial effect on the learners' acquisition. By way of comparison, typical 'repair' rates resulting from recasts and elicitation in immediate CF are in the region of 20% and 45% respectively (Sheen 2004: 268).

For both groups, the average fluency with which they were able to reformulate the test items also rose, more markedly for the IEP group than the ELC. More importantly, the data provided by the tests made it possible to generate individual fluency profiles, against which progress on specific language items could be measured. A recommendation from this research is that these individual profiles, which can largely be calculated and stored automatically in an electronic database, be used to monitor the development of IL forms within the individual learner. For instance, items which are either incorrectly reformulated, or which display greater than average dysfluency (indicated by being more than 1 *SD* from the mean for that individual) should be retained for future testing of that individual.

An example of this tracking potential was possible in this study as a number of items were reformulated on both tests by students in both groups, addressing the second research question: *Are learners consistent in their ability to reformulate errors?* For the ELC group, it was found that there were substantial gains in fluency on the 65 items that were repeated from the first test to the second, but the analysis for accuracy was inconclusive, possibly owing to the small number of repeated items. Some learners made gains in accuracy, while others

seemed to be equally inaccurate on these items on both tests. However, closer analysis of specific items showed that in many cases, incorrect items on the first test were inaccurate for different reasons on the second test. The remaining inaccuracies, such as the ‘advice’ example (Section 4.8.3) are themselves a potential source of information about developing ILs, as they provide specific information for form-focused instructional activities. For the IEP group, the results are more conclusive: in terms of both accuracy and fluency, there was a significant overall increase on the second test. The conclusion therefore is that there is consistency in the participants’ ability to reformulate errors, but there is a level of individual variation which adds considerable complexity to the question of *which* errors. Further research is needed to illuminate which learners, and which forms, are most responsive to this type of CF.

The final research question in this section asked: *Do learners find it more difficult to reformulate their own errors or those of peers?* In neither group was there a significant overall difference in fluency or accuracy contingent on the source of the original error, although the ELC group showed a slight decrease in WPM when reformulating peers’ errors. Three reasons were given as to why teachers might choose to assign peer errors to student: usefulness of the form, typicality of the error, and compensation for avoidance. The results of this analysis suggest no reason to discontinue this practice. Where the linguistic forms exemplified in peer errors are already acquired by a learner, she will have no trouble reformulating them, and where they are not, the error and teacher reformulation represent both input and negative evidence.

The RLT data provide two types of information. First, by comparing individual performance on items which were repeated on both tests, it is possible to track developing competence on those language forms. Theoretically, the items offer a window onto the learners’ IL, which is assumed to be continually evolving. Furthermore, the state of any

individual learner's IL is not known *a priori* so only by collecting sufficient clues as to which aspects might have stabilized can we discern where pedagogical intervention would be most beneficial and which items no longer need CF. In addition, by observing which items are quickly and accurately reformulated, it is possible to say with some confidence which represent systematic error and which may be performance 'slips'. Second, since the analysis is done at the individual level, fluency data from each RLT contribute to an overall average for the individual; thus linguistic factors which do not pertain to second-language performance, such as individual speech rate and other idiosyncrasies which might otherwise confound the investigation of competence, can be controlled for. Put differently, an individual's aggregate RLT data provide a baseline measure of performance fluency against which subsequent performance on specific language items can be measured. This is important because without such a baseline, it is impossible to measure the progress of acquisition of such items. This investigation, therefore, has contributed one possible answer to the challenge posed by Corder, R. Ellis, and others (see Chapter 1) of determining what the individual learner knows.

---

## CHAPTER 5: TIMED GRAMMATICALITY JUDGEMENT TESTS

---

### 5.1 Introduction

In the discussion of the pilot timed grammaticality judgement test (TGJT) in section 3.8, it was suggested that if learners were asked to perform a TGJT with items drawn from their own production, the data might help to refine pedagogical decisions about which areas might most benefit from instructional focus on form, targeted practice, and so on. It was hypothesized that participants take longer to judge items which are less well established in their ILs, and conversely that well-established items, whether target-like or not, are more quickly judged. A parallel hypothesis was presented in Chapter 4 with regard to fluency of production on the RLTs: items which are under automatic control, whether target-like or not, are produced more fluently than those which require conscious attention to syntax, lexical retrieval, and so on. In this chapter, the relationship between accuracy, complexity, and fluency is investigated from the perspective of learners' recognition of grammaticality in their own production.

Few scholars acknowledge that fluency is a concept that applies to receptive as well as productive skills. Gatbonton and Segalowitz are an exception:

In one sense, [fluency] refers to the speed and ease of handling utterances; the greater the automaticity, the faster the *recognition* and production of grammatically correct and communicatively appropriate utterances. (Gatbonton et al. 1988: 474; emphasis added)

This is an intriguing notion, and one which goes to the heart of the question of how (second) language is produced. The assumption underlying the investigations in Chapters 3 and 4 is that fluency is a measure of automaticity, of proceduralized knowledge (McLaughlin 1987; Towell et al. 1996), or of 'acquisition' (Krashen 1981; 1982; 1985; Schwartz 1986; Zobl

1995). Evidence was presented in the previous chapter that fluency can be measured on a gradient, so that some IL forms are produced more fluently and others less so, even under the identical conditions of a primarily form-focused test for which participants have prepared. It has been suggested that less fluent production requires the use of explicit knowledge ('learned' knowledge, in Krashen's epistemology), in other words the conscious application of rules, but that this process can become automatized and more implicit with time and practice. It has also been suggested that formulaic language plays an integral role in this process, in that more proficient learners make more use of formulae than do novices. This does not only apply to idiomatic formulae such as *to tell you the truth* or *the more... the more*, but to any prefabricated 'chunk' of language that the learner uses to reduce processing load, such as *I mean....*

The relationship between explicit and implicit knowledge is complex, and has been muddled by the proposal of separate and dissociated epistemologies in language learning ('acquisition' and 'learning'). There are valid reasons to believe that much in language comprehension and production involves 'mental processes that are far beyond the level of actual or even potential consciousness' (Chomsky 1965: 8), i.e. that cannot be verbalized or even explicitly thought about. But this does not mean that there is no interaction between conscious and unconscious systems at all, as N. Ellis points out:

Krashen (1985) was correct to the extent that, as he termed it, acquisition and learning are different things; in psychological vernacular, explicit and implicit knowledge are distinct and dissociated; they involve different types of representation and are substantiated in separate parts of the brain (N. Ellis, 1994c, 1996; Schacter, 1987; Squire & Kandel, 1999). Paradis (1994) was correct in stating that explicit knowledge does not become implicit knowledge, nor can it be converted to it. Nevertheless, there is interaction. However unlike they are, these two types of knowledge interact. The interface question...has motivated a wide range of empirical research over the last 30 years, and the weight of the subsequent findings demonstrates that language acquisition can be speeded by explicit instruction. (N. Ellis 2005: 307)

On this point, Krashen has submitted the following explanation: 'Occasionally, we learn

certain rules before we acquire them, and this gives us the illusion that the learning actually caused the acquisition.’ (Krashen 1979: 158). Elsewhere, Krashen elaborates on his *post hoc ergo propter hoc* objection, suggesting that the incidental input made available as a result of discussions and negotiations during explicit rule-learning is itself what causes the acquisition (Krashen 1992). This hypothesis is almost certainly unfalsifiable, however, since there is no way to determine which input is used by which processes, and so others have taken a less rigid position on the ‘interface question’ (see R. Ellis et al. 2009: 20–23 for an overview). For instance, N. Ellis explains that even in an ideal pedagogical context, the input is too restricted to be sufficient:

Adult acquisition of an L2 is a different matter in that what can be acquired implicitly from communicative contexts is typically quite limited in comparison to native speaker norms, and adult attainment of L2 accuracy usually requires additional resources of consciousness and explicit learning. (N. Ellis 2007: 18)

Thus the prevailing view in the literature (Krashen excepted) is that linguistic forms, explicitly focused on, thought about, practised, and gradually automatized, can become implicit knowledge (Gatbonton and Segalowitz 2005; DeKeyser 2007). We may not be able to examine the resulting implicit knowledge, but this does not mean it did not derive from what was once in our awareness (McLaughlin 1990; Schmidt 1990; Hulstijn and Schmidt 1994).

In fact, we try to examine our implicit knowledge all the time: such thinking – metalinguistic introspection and intuitions about grammaticality – has become a standard source of ‘empirical’ data for linguistics, despite the warnings of Chomsky himself that judgements based on introspection are themselves performance data. When a linguist tests a rule such as Subjacency, she does so by probing the limits of her internal representations of syntax, her ‘I-language’ in Chomsky’s (1986: 15-24) terms; in general, she does this by trying out sentences to see (or rather, to *feel*) which conform to her internal representations and

which do not. Those that do not, such as *\*Who did you believe the rumour that Mary saw?* (or *\*That's make you happy?*) are then said to delineate the boundaries of grammaticality:

We often intuitively judge the grammaticality of a sentence or the legality of a move or the propriety of an act without conscious access to the formal syntax of the domain. But let us turn the tables somewhat. It is an interesting possibility that each of those intuitions is one of a set of informal rules of limited scope and perhaps imperfect validity. The intuitions seem quite conscious. We know something that seems right or wrong, even when we don't think of or know the proper rule from a formal system. (Dulany et al. (1984), cited in McLaughlin 1990: 622)

Labov (1975; 1996) has warned against using the products of an internal system as evidence for a general theory of that system – against '[producing] theory and data at the same time'.

But if what one is interested in *is* the system of representations – the I-language – of the learner, then her intuitions are an important piece of the puzzle. From a pedagogical perspective, the lexico-grammar of the learner's IL is what is of interest, and the learner's intuitions about grammaticality are therefore a potential source of data about her IL.

R. Ellis (1991) and Han (1996; 2000) have pointed out the lack of reliability in learners' grammaticality judgements, and the lack of reliability in the judgements of naïve judges is well documented (Birdsong 1989; Schütze 1996). In the case of second language learners, it is hardly surprising that there are elements (especially the more arcane elements) of the target language (TL) about which they have no intuitions, and this is why the use of grammaticality judgements in second-language research often makes little sense.

Furthermore, the use of grammaticality judgement tests (GJTs) to elicit data for investigations into acquisition mechanisms such as Universal Grammar is doubly risky, since one can never be sure what effects the L1 might have (although researchers are generally very careful to eliminate this confound), what naturalistic input one's participants have been exposed to prior to the investigation, and what they might have been explicitly taught. The blithe assertion that the target structure is 'underdetermined in the input' (e.g. Bley-Vroman 1990; White 1990;

Hawkins 2001; Cook 2003) is far more common than it should be.

In contrast, when the input is known and the goal is to establish the nature and form of the learner's grammatical knowledge, the use of GJTs seems justified. This is not to say that the method is without complications, as discussed below, but these are no more insurmountable than many others in psycholinguistic research.

## 5.2 Research questions

In this chapter, the investigation focuses on learners' awareness of the well-formedness of their own oral production (in contrast to Chapter 3, in which a generic test set of items were presented to the participants), and was motivated by three questions:

- 1) Can learners identify the grammatical status of their own utterances?

This is a deceptively simple question, since the ontological status of a systematic 'error' is problematic. If a particular structure of interlanguage is a true representation of the competence or mental grammar of a learner, then it cannot logically be erroneous within that system; (Corder 1967; Bley-Vroman 1983). According to Corder, if learners judge their own production as inaccurate, this production must be a non-systematic 'mistake' in performance; any other production, whether conforming to TL norms or not (i.e. an 'error'), must be 'grammatical' in their IL. Another view, proposed here, is that learners do systematically produce non-target-like forms but may be able to recognize that they are not well formed.

- 2) What is the relationship between TL accuracy in production and TL accuracy in judgement of grammaticality?

For the reason outlined above, it is hypothesized that learners should be able to recognize any 'slips' and certain types of systematic errors.

- 3) What is the relationship between fluency in production and reaction time in



judgement?

If both fluent production and intuitions of grammaticality reflect automatic, implicit knowledge, there should be a strong association between measures of the two. Thus reaction time (RT) on the TGJT (see Chapter 3) should correlate closely with measures of fluency on the RLT.

### **5.3 Participants**

For this stage of the investigation, the participants formed subsets of the ELC and IEP groups (section 4.6). The TGJT data collection was not a part of the regular instructional program (unlike the RLT), and so participation was on a strictly voluntary basis. Both groups were told about the research and asked to participate, and 11 of the 13 students (85%) in the ELC group did. With the IEP group, 9 of the 13 (70%) participated. Given that more than two-thirds in each group participated, it was thought that sufficient data was collected for the analysis to be statistically robust.

### **5.4 Methodology**

#### **5.4.1 Administration of the TGJT**

The data elicited in RLT 1 for each participant formed the set of test items for the same participant in the TGJT. In other words, students listened to their own recordings and were asked to judge the grammaticality of each item as quickly as they could, in exactly the same procedure as that described in Chapter 3. Each group of participants took the TGJT approximately two weeks after doing the RLT, so that enough time would have elapsed for them not to remember exactly what they had recorded as reformulations, but not so much time that their ILs could have changed or stabilized (Reinders 2005). During the TGJT, participants wore headphones, to maximize sound quality and isolation, and had no other materials in front of them. The tests took place in a language lab on iMac computers. Before

taking the TGJT, participants were asked to practise using the test set of items described in Chapter 3 in order to familiarize themselves with the procedure. Items were presented in a randomized order. No attempt was made to balance the number of incorrect and correct items, or to screen the items for specific structures or lexis.

As described in Section 3.5.4, the TGJT was administered using an online Flash interface. Each item was played to the participant, and immediately after the sound file finished playing, two buttons appeared on the screen, allowing for a choice of Correct or Incorrect, and a timer function was called in the background. As soon as either button was clicked, the timer function was called again and the difference between the two times was sent to the MySQL database along with the participant's name, her judgement, the time stamp, and the ID of the sound file.

#### **5.4.2 Scoring the TGJT items**

In the RLT stage (Section 4.7.1) all items had already been assigned a status as grammatical or ungrammatical according to whether they conformed to TL norms. The TGJT judgement for an item thus scored 1 if it agreed with the RLT score for that item and 0 if it did not.

#### **5.4.3 Data analysis**

Data analysis took part in two stages. First, the response patterns for each group were analysed in order to determine response bias and differential response times on correct and incorrect items. This analysis was necessary to eliminate or control for factors which might confound the identification of psycholinguistic variables (Birdsong 1989; Schütze 1996). Second, the RT and response data were compared with the RLT fluency and accuracy data on the same items from the same participants, in order to determine the relationship, if any, between the two types of performance.

#### 5.4.4 Response bias

The conceptual issues behind response bias in the context of acceptability judgements on one's own second-language production are inchoate, since this is not an area which has received much attention from the SLA community. A survey of the literature on the use of grammaticality judgements in SLA reveals very few examples of learners judging their own errors (see Chaudron 1983: 358-361 for an overview), and none in which learners' own oral production has been used. One significant reason for this is suggested by White (1989), who points out that judgement data can be tailored to suit the researcher's purposes, which might include the investigation of structures that occur rarely, if ever, in natural learner output. This is undoubtedly true if the purpose of the investigation is to test the learner's intuitions about, say, putative rules of Universal Grammar, since the goal is to measure whether the learner's sense of possible grammaticality tallies with the native speaker's. Gass, on the other hand, suggests a more pragmatic goal:

[M]etalinguistic awareness has an important function for second language learners, allowing them to make comparisons between NL and TL, self-correct, and perhaps even monitor their output. Investigating a learner's ability to judge grammaticality is therefore essential to an understanding of a learner's development. (Gass 1983: 277)

Gass acknowledges the difficulties involved in eliciting grammaticality judgements from learners on their own production, suggesting that learners will naturally judge their own production as accurate, since presumably it *is* accurate according to their ILs: in other words, they would not deliberately produce inaccuracies. This explains why (second-language) writers find it so hard to spot the inaccuracies in their own written work (Cohen 1983). As Gass puts it:

When asking for judgments from adult native speakers (even linguistically unsophisticated ones), one can assume that most of the time there is at least an approximate equivalence in a speaker's ability to produce utterances, to comprehend utterances, to parse utterances, and to judge utterances. For L2 learners this is not necessarily the case since there is often a large discrepancy in one's abilities in these areas. (Gass 1983: 275)

Assumptions such as this seem to have prematurely closed off a potentially fruitful avenue of enquiry. Without serious investigation, it is impossible to know whether learners can recognize ungrammaticality in their own production.

In the few studies of this kind which have been attempted, results have been mixed. White (1977) found that her participants were able to identify approximately 60% of their (TL) ungrammatical production, with advanced learners showing no superior ability over intermediate learners. Cohen and Robbins (1976) took a qualitative approach, investigating learners' reactions to and explanations of their own written errors. Consequently they did not look at the extent to which learners were successful in correcting their errors. Gass' own study showed that intermediate learners correctly (i.e. conforming to NS rules) judged 71% of their errors, and advanced learners 68% (Gass 1983: 281), with intermediate learners slightly better at identifying the grammatical items than advanced learners, and also better at identifying the grammatical items than the ungrammatical. This is what Birdsong (1989: 101-107) means by a response bias: the tendency of NNS judges to accept ungrammatical items. In an investigation of IL competence, for instance that of Bley-Vroman et al., (1988), this kind of bias is a serious confound. But if what is being investigated is metalinguistic awareness, the bias itself is the object of study and what it reveals is potentially highly informative.

To give an example, suppose an intermediate learner is presented with 50 of her own utterances (or written sentences), produced when she was a beginner, half of these items judged by NS judges to be grammatical in standard English and the other half not. If the (now intermediate) learner can correctly identify which items are grammatical and which are not, she not only is not showing a response bias, but more importantly is demonstrating progress: her proficiency and/or metalinguistic awareness has now developed to the point where she

can accurately judge the grammaticality of her earlier production in spite of the errors it manifested. Now the same learner, presented with a selection of 50 items from her *current* output (again half (TL) grammatical and half not), might well be expected to judge more of them grammatical. In fact, Gass (1983: 280) found that intermediate learners slightly overestimated their accuracy (judging an average of 55% of their items as grammatical) and advanced learners slightly underestimating it (46%). Gass' test consisted of only twelve items per participant, and she did not perform any statistical analysis on the data, so it is difficult to know whether these results could have been achieved by chance. Nevertheless, if the goal of such research is to understand the learner's development, then Gass is correct in insisting that such measures of metalinguistic awareness of this kind are an essential source of data.

#### **5.4.5 Comparison of production and recognition performance**

The score and reaction time (RT) for each TGJT item was compared against the accuracy and WPM measures on the RLT 1 for each group (see Appendix 6 for raw data). To ensure valid cross-group comparisons, only the WPM measure was used as a 'fluency' score for the ELC group as this was the only measure available for the IEP group (Section 4.7.4). However, for the ELC group, bivariate correlations were calculated between all fluency measures and RT. For each participant, a z-score was calculated for RT and WPM for each item to permit comparison. To facilitate the interpretation of data, the standardized scores for RT were inverted (subtracted from zero), since greater RT values should be associated with less automaticity, and vice versa, whereas with WPM the converse is true.

### **5.5 Findings**

#### **5.5.1 Relationship between measures of accuracy**

##### ELC group

The correlation between accuracy on the RLT and TGJT was found to be very high, at  $r =$

.767,  $p < .0001$  (Table 25). Since accuracy of a TGJT judgement is ultimately determined by accuracy of that item on the RLT (not whether the participant *thinks* it is grammatical) it is somewhat surprising that the correlation is not stronger. Some of the variation could be

Table 25: *Correlation matrix for measures of fluency and accuracy on RLT and TGJT, ELC group*

	1	2	3	4	5	6
1. TGJT Reaction Time (RT)	--	-.027 <i>.629</i>	-.123 <i>.030</i>	.096 <i>.090</i>	-.155 <i>.006</i>	-.092 <i>.103</i>
2. TGJT Accuracy		--	.171 <i>.003</i>	.021 <i>.721</i>	.139 <i>.014</i>	<b>.767</b> <b>.000</b>
3. RLT WPM			--	.315 <i>.000</i>	.406 <i>.000</i>	.133 <i>.018</i>
4. RLT Phonation ratio				--	.509 <i>.000</i>	-.035 <i>.538</i>
5. RLT Speech rate					--	.114 <i>.045</i>
6. RLT Accuracy						--

Note: significance shown in italics

attributable to the moderate response bias shown by the group (see below). In fact, three participants did not judge any of their items as incorrect (Appendix 6, ‘Response’ column). Assuming that they were not simply being uncooperative, the conclusion must be that every one of their items *sounded* correct to them. This would decrease the strength of the relationship between the accuracy measurements on the two tests, as would cases in which participants judged an item as ungrammatical when it was in fact grammatical. The remaining variation between RLT and TGJT accuracy scores was possibly caused by participants’ identification of non-systematic ‘slips’. This point will be revisited in Section 5.5.2.

IEP group

For the IEP group, the association between accuracy measures on the RLT and TGJT was weaker than for the ELC group, but still strong at  $r = .593$ ,  $p < .0001$  (Table 26). Again the

Table 26: *Correlation matrix for measures of fluency and accuracy on RLT and TGJT, IEP group*

	1	2	3	4
1. TGJT Reaction Time (RT)	--	-.125 .085	-.072 .318	-.222 .002
2. TGJT Accuracy		--	.116 .108	<b>.593</b> <b>.000</b>
3. RLT WPM			--	.170 .018
4. RLT Accuracy				--

Note: significance shown in italics

interpretation is that at least some students recognized ‘mistakes’ in their own production, as discussed above. Further support for the interrelation of the constructs *accuracy* and *fluency* is hinted at by the weak but significant correlation between RT and RLT accuracy ( $r = -.222$ ): the more accurate their reformulations (by TL norms), the faster their judgements. To confirm this hypothesis, the bivariate correlation between RT and judgement (whether the participant *thought* an item was grammatical) was calculated. The result ( $r = -.313$ ,  $p < .0001$ ) is suggestive of the interplay between these variables and another, or possibly others, such as general reaction speed or confidence. Further research will be necessary to determine the precise nature of this interaction.

### 5.5.2 Relationship between measures of fluency

The more surprising discovery in the correlation data (Table 25), however, was the very weak relationship between RT and measures of fluency. More fluent production, measured in WPM

or speech rate, did correlate with faster RT ( $r = -.123$  and  $-.155$ , respectively) but so weakly as to suggest that fluency and RT are only tangentially related. In fact, the *accuracy* scores on the TGJT correlated just as strongly with the same measures of fluency ( $r = .171$  and  $.139$ , respectively), meaning that the more fluent their RLT production, the more likely participants were to judge it accurately – or more likely, the *less* fluent they sounded to themselves, the greater the chances they would judge the item ungrammatical and be correct in that judgement. But such tentativeness was not generally reflected in the RT data (hence the low correlation coefficient), meaning that fluency in production is not paralleled by fluency in recognition.

To exemplify, Figure 13 shows a comparison of the RLT *fluency* data (in green) and TGJT reaction time data (in blue) for ‘Falshehri’, sorted in descending order of fluency on the RLT. Once again, the incorrect reformulations and judgments are shaded lighter.

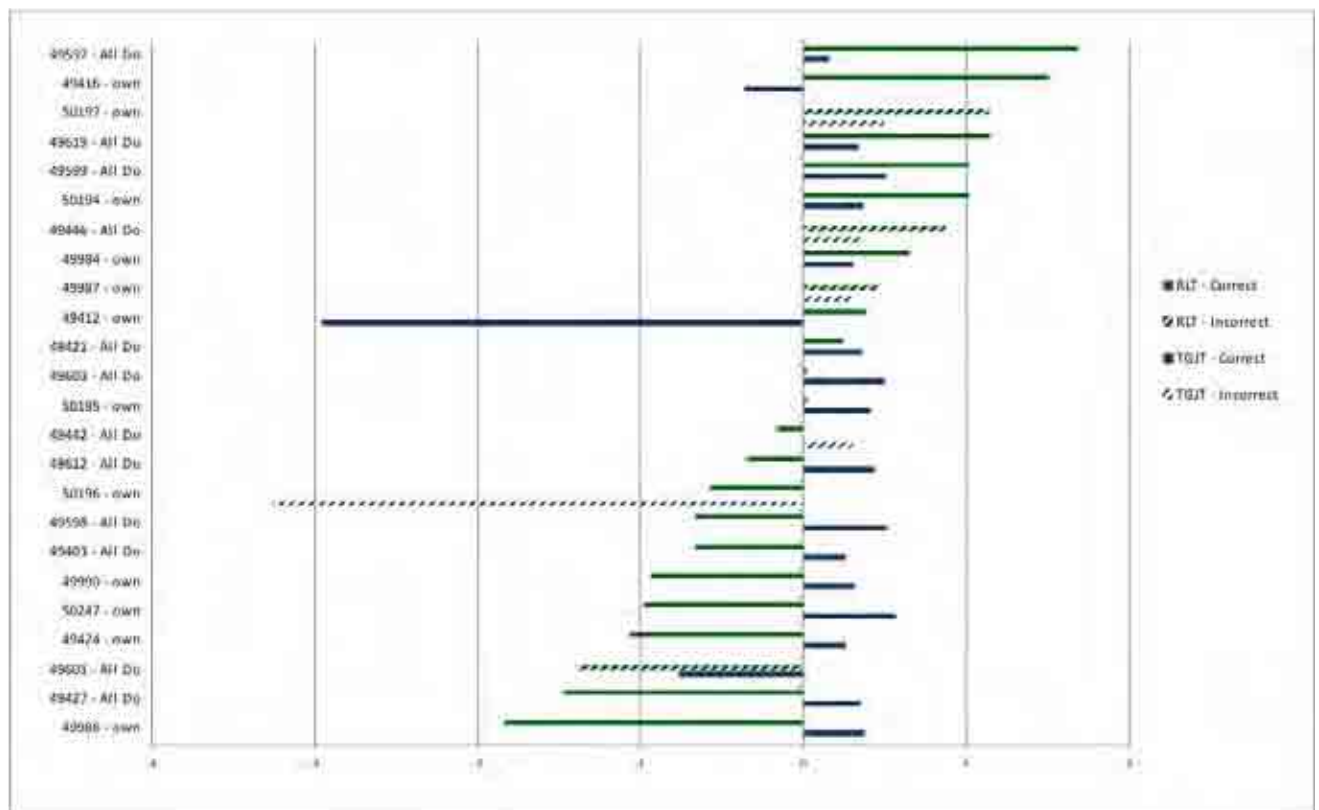


Figure 13: Comparison of RLT and TGJT data for ‘Falshehri’



be seen, most of his reaction times on the TGJT items are within 1 *SD* of the mean, regardless of RLT fluency. It is, in short, difficult to discern any association between the two measures, which would imply that the two types of fluency are motivated by different cognitive systems. This is a possibility which deserves closer scrutiny than space will allow here, and future research will address the implications of this finding.

There are two items, 49412 and 50196, on which ‘Falshehri’s’ reaction time was approximately 3 *SD* slower than his average of 821ms. In much psycholinguistic research, these extreme ‘outliers’ would probably be discarded or replaced with the mean or next highest scores (e.g. N. Ellis et al. 2008: 383; O'Brien et al. 2007: 568). However, this procedure was not followed here, as the raw RTs for these items were 3272 and 3515ms respectively, which is hardly long enough to prompt a suspicion of distraction or confusion. Instead, these items should be scrutinized with particular attention, as they may provide insight into the learner’s IL. Taking the more extreme of the two first, item 50196 (3515ms), it turns out to be an item which includes an indirect question (noun clause with *wh-*), a structure which is challenging for most learners and particularly so for Arabic L1 learners:

original Small Talk error:      \* *You never know who is your real friend.*

reformulation on RLT 1:      *You never know who your real friends are.*

In spite of the fact that the RLT 1 reformulation was correct, the relatively slow RT seems to indicate that ‘Falshehri’ had to think about this item, and still judged it as ungrammatical. This does not come as a surprise: many students find indirect questions in English to be counter-intuitive and highly confusing; ‘Falshehri’ may have been able to produce the item on the RLT, somewhat slowly but not overly so, but it still apparently *sounds* wrong to him, and he hesitates when judging.

The second item with an unusually long RT was item 49412 (3273ms). This was the first item on ‘Falshehri’s’ TGJT, so it could be argued that his RT was slower for this reason;

but given that he had just done a practice test, and given the lack of discernable pattern among his peers for the first items to be averagely slower than any others, it was concluded that the item itself had caused the slow RT:

original Small Talk error:      \* *But it might be something in future.* (meaning: anything can happen)

reformulation on RLT 1:      *But something might happen in the future.*

The speculation here is that the introduction of the verb *happen* caused him to pause and think. This verb seems to cause particular difficulty for many learners, who often passivize it. (See for example Oshita (2000). Although Oshita looks mainly at Asian languages in his study, the Small Talk database contained 8 uses of *be + happen* out of 25 total instances of *happen* (32%) for Arabic speakers at this proficiency level; 3 out of 14 (21%) for Mandarin speakers; 4 out of 15 (26%) for Korean speakers; and 9 out of 48 (18%) for Japanese speakers. The fact that this seems to be a more extensive problem for Arabic speakers at this level than speakers of other languages does not necessarily undermine Oshita's arguments, but it does speak to the need for extensive and publicly available production error data against which such arguments can be tested.) It is quite conceivable that for this reason, the 'chunk' *something might happen* does not intuitively sound right to 'Falshehri', and takes him longer to judge for that reason. Once again, the need for triangulation by means of a think-aloud protocol or interview becomes apparent, although others have warned about the inconsistency of this type of learner data as well (Cohen and Robbins 1976; R. Ellis 1991; Han 2000). In terms of pedagogical procedure, the atypical RT on this item would mark it for additional attention, as in fact occurred, so that it would remain on 'Falshehri's' Running List. This happened with both items discussed above, and as can be seen in Figure 10 (p. 95), both items were correctly and more fluently delivered in RLT 2.

### 5.5.3 Response bias

#### ELC group

‘Falshehri’ is unusual in the number of items he judged ungrammatical; as mentioned above, three ELC participants judged all of their items grammatical. To investigate the degree of response bias, a Chi-square test was performed using the (TL) grammatical status of the items as the expected frequencies, and it was found that the participants were indeed more likely overall to judge items grammatical ( $X^2 (1, N = 338) = 30.38, p < .0001$ ) than ungrammatical.

#### IEP group

Similarly, it was found that the less proficient IEP participants were more likely overall to judge items as grammatical ( $X^2 (1, N = 192) = 13.79, p < .0001$ ) than ungrammatical.

However, this response bias was not as pronounced as that found for the more proficient ELC group, which runs counter to the intuitive wisdom that with greater proficiency comes a greater ability to detect ill-formedness. This point will be taken up in the discussion below.

Given that a response bias was also found in the pilot TGJT, it is reasonable to conclude that certain errors do ‘sound right’ to learners, which in itself is a strong argument for CF: in the absence of repeated and systematic CF, these errors are likely to continue to ‘sound right’ since the learner’s own production of them is likely to be more frequent – or at least more salient – than the target forms in the input. (See also Butler Platt and MacWhinney 1982: 412 for a description of a similar mechanism in child L1 acquisition.) This might seem counter-intuitive, but the fact is that a learner’s total *grammatical* linguistic repertoire in the target language is a minute subset of that language; her *ungrammatical* repertoire is not a subset of the target language, but is similarly minuscule by comparison. Therefore, *any* aspect of the learner’s repertoire is statistically far more likely to occur in her own production – not

to mention thoughts or ‘inner speech’ (De Guerrero 1994) – than in the input.

As concerns judgements on one’s own production, where this has been considered at all in the literature, it has generally been assumed that learners will judge their own production as grammatical, as discussed above; but this is far too simplistic a picture, since it ignores the issue of ‘slips’, and it assumes that accurate production and accurate recognition are motivated by the same cognitive mechanisms, which is not at all clear from the current data or from Gass’ (1983: 280) – although Gass’ data are apparently incomplete (observed and expected values do not match) so no response bias calculations can be performed. Gass’ conclusion is that learner intuitions of grammaticality become more analytical as proficiency increases:

As Bialystok notes, ‘sentences sound right for reasons that may be completely obscure and in these cases justifications for the decisions can rarely be found’ (1981:37). The results presented here corroborate this finding. Sentences ‘felt’ wrong to the students without their having an accurate idea of why they were wrong. It is suggested here that part of what is involved in becoming more proficient in a second language is the progression from more gestalt-like to analytical analyses. We might further speculate that indeed the analyzed aspect is a necessary precondition for fluency in an L2, more so than for an L1. (Gass 1983: 285)

This speculation, however, runs counter to the evidence presented in Section 3.5.5, where it was seen that NS and NNS participants seemed to approach (TL) grammatical and (TL) ungrammatical items differently, with even proficient NNS taking longer to judge ungrammatical sentences and the NS taking longer to judge grammatical ones. If one adopts Wray’s proposal (2000; 2002; see also Sinclair 1991), that ‘gestalt-like’ analyses derive from implicit knowledge of frequency of occurrence in input, which NS can rely on and NNS cannot, this knowledge can be seen as the backdrop against which the unusual (i.e. ungrammatical) will stand out; in contrast, ‘analytical analyses’ derive from explicit knowledge of ‘rules’, which NNS may rely on to a far greater extent than NS but which take longer to process. With the addendum that the learner’s own output, whether target-language

grammatical or not, becomes grist for the gestalt mill, Wray's position is one whose explanatory power *vis à vis* error production in SLA is greater than any other so far elaborated. If anything, then, what is involved in becoming more proficient in a second language may be the opposite of what Gass suggests: the progression from analytical analyses to more gestalt-like analyses – or at least a balance between the two.

#### **5.5.4 Differential reaction time when judging items as grammatical and ungrammatical** ELC group

If the above hypothesis is correct, there should be a measurable difference in RT when NNS participants judge an item to be ungrammatical (whether it actually is or not according to standard English): an item which is 'grammatical' in the IL will sound right, and one which 'sounds wrong' will have to be checked against conscious rules. A t-test of the ELC group's responses showed a significant mean difference in RT when they judged an item as grammatical ( $M = 1600$ ,  $SD = 3982$ ) in contrast to ungrammatical ( $M = 3495$ ,  $SD = 3437$ ),  $t(1) = 2.75$ ,  $p = .009$ . Thus the group on average made their judgements almost two seconds faster when they thought items were grammatical than when they did not. A caveat must accompany this finding, however: although the finding is statistically significant and takes into account unequal variances, the extreme imbalance between the number of judgements as grammatical (315: 92%) and as ungrammatical (28: 8%) is a potential threat to the robustness of the analysis and urges caution in the interpretation of the finding.

#### IEP group

A t-test of the IEP group's responses showed a significant mean difference in RT when they judged an item as grammatical ( $M = 1834$ ,  $SD = 1533$ ) in contrast to ungrammatical ( $M = 3407$ ,  $SD = 2628$ ),  $t(1) = 3.18$ ,  $p = .003$ . Here again, the group on average made faster

judgements when they thought an item was grammatical, by approximately 1.5 seconds. With this group, the imbalance between the number of judgements as grammatical (162: 84%) and as ungrammatical (30: 16%) was not as extreme as with the ELC group, but nevertheless argues for caution in interpretation.

Tentatively, the findings from the two groups suggest that when an item ‘sounds right’ (a gestalt analysis), learners are faster in their judgements than when it ‘sounds wrong’, in which case an analytical analysis must be performed, resulting in a slower judgement. Presumably, when a sufficient stock of language has been amassed, the speed at which ungrammaticality can be discerned decreases, as was found for the NS group in the pilot study (Section 3.5.5).

#### **5.5.5 Errors and mistakes**

Pedagogically, teachers require information which will enable them to home in on items that likely require focused attention in various forms, and conversely to ignore items which seem to need no further remediation. This refers as much to systematic errors no longer made as to ‘mistakes’. According to Corder (1967: 167):

Mistakes are of no significance to the process of language learning. However the problem of determining what is a learner’s mistake and what a learner’s error is one of some difficulty and involves a much more sophisticated study and analysis of errors than is usually accorded them.

The TGJT investigation represents one such attempt at more sophisticated study and analysis.

The first question to ask is whether any ‘mistakes’ in Small Talk conversations would have been cleared up in RLT 1. In other words, if the original Small Talk ‘error’ were in fact a slip or ‘mistake’, then the student presumably would have been able to reformulate it correctly. This is not to say that students could not produce (new) mistakes on the RLT itself: even though the focus is on accuracy, the time pressure might cause inadvertent additional slips, which presumably they would identify as such on the TGJT. In section 4.9 it was

suggested that items correctly and quickly reformulated could be considered ‘mistakes’ and not systematic errors. To investigate whether participants responded differently to these items on the TGJT, a comparison of means was performed on a subset of items, defined variously as ‘mistakes’ and ‘errors’ according to these criteria:

mistakes: ‘Own’ items which were correctly reformulated at above average speed ( $>1\ SD$  above the mean for that student).

errors: ‘Own’ items which were incorrectly reformulated (at any speed)

The assumptions here are that any item which a student can correct quickly must not be a systematic error, and conversely that if she cannot correct it at any speed, it must be a systematic error.

#### ELC group

For the multilingual ELC group, these criteria yielded 23 ‘mistakes’ and 28 ‘errors’. The hypothesis was that the ‘errors’ would show a longer RT than the ‘mistakes’, which should be quickly recognizable as such. The analysis revealed a significant difference for accuracy of judgement of ‘mistakes’ ( $M = .96, SD = .21$ ) and ‘errors’ ( $M = .25, SD = .44$ ),  $t(1) = 48.9, p < .0001$ . Curiously, however, there was no significant mean difference in RT on these items (‘mistakes’  $M = 1883, SD = 2134$  and ‘errors’  $M = 1856, SD = 2534$ ,  $t(1) = 0.038, p = .97$ ). Thus as anticipated, students’ judgement on ‘mistakes’ were highly accurate, and their judgements on ‘errors’ far less so; however, their RT in these judgements showed no mean difference.

#### IEP group

The data from the monolingual IEP group were examined for examples of ‘mistakes’ and ‘errors’ from RLT 1, using the same criteria elaborated above. This yielded 20 ‘mistakes’ and

14 ‘errors’ A comparison of means for this set of data once again revealed a significant difference for accuracy (‘mistakes’  $M = 0.85$ ,  $SD = 0.37$  and ‘errors’  $M = 0.43$ ,  $SD = 0.51$ ,  $t(1) = 2.67$ ,  $p = .015$ ) but not for fluency (‘mistakes’  $M = 1733$ ,  $SD = 1307$  and ‘errors’  $M = 2185$ ,  $SD = 1077$ ,  $t(1) = -1.06$ ,  $p = .296$ ).

It is of course possible that students introduced new ‘mistakes’ during their RLT performance, which would complicate the analysis above. To determine whether this might be the case, one student was selected at random, and only those items which were incorrectly reformulated on the first RLT were examined (i.e. those on which she might have introduced new ‘mistakes’ – shown in Figure 14). By analysing her TGJT performance on those items, it

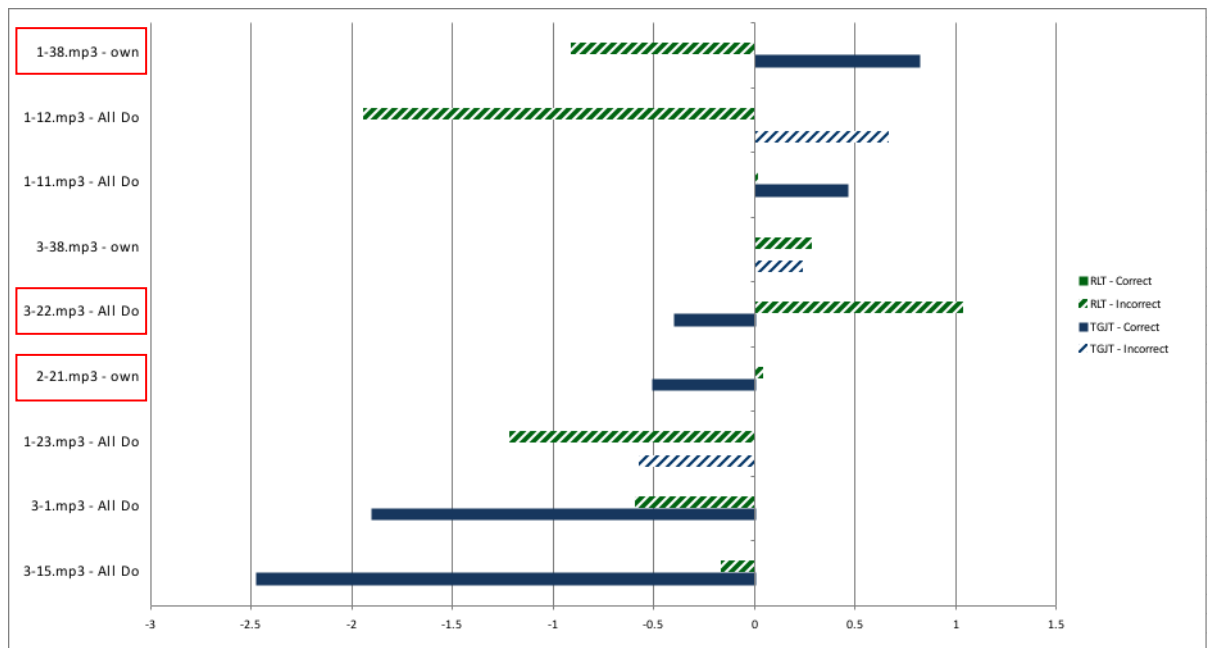


Figure 14: RLT-TGJT comparison for ‘Ayako’, ungrammatical RLT items only

should be possible to determine whether she introduced new ‘mistakes’ during the reformulation stage. Of the six items that she correctly identified as ungrammatical, four fall within the  $\pm 1$   $SD$  range, implying that she was able to recognize the ungrammaticality within her normal reaction time; and of these, three (marked in Figure 14) also appeared on her second RLT, allowing us to see whether she eventually corrected her ‘mistakes’ or not



(Figure 15). Since ‘Ayako’ recognized that these items were ungrammatical, the

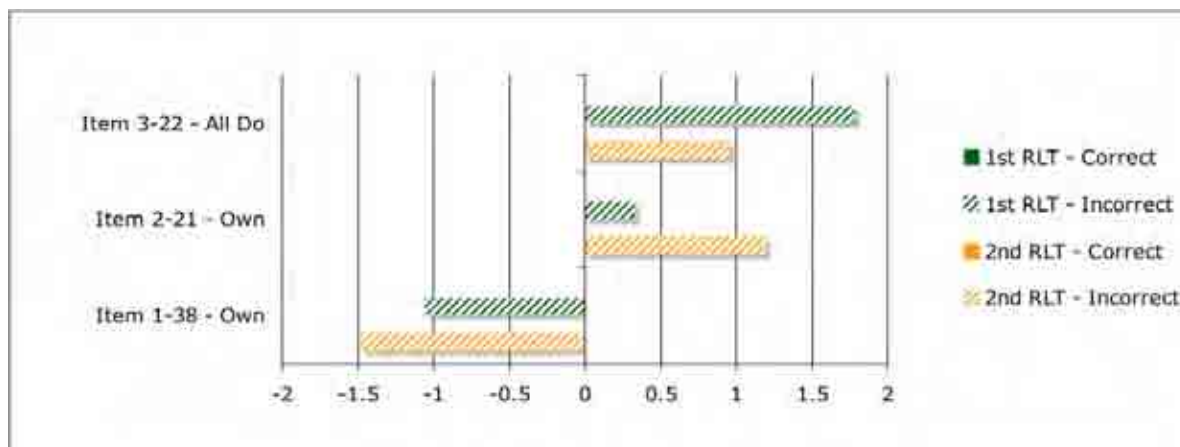


Figure 15: Three ungrammatical items from ‘Ayako’, RLT 1 and 2

mistake/error hypothesis would predict that she would correctly reformulate any ‘mistakes’ on her second RLT attempt. Since she did not, they will be examined in turn. (Note that the mean RTs in Figure 14 and Figure 15 are calculated from different sets of items, therefore the z-scores for items also differ. Regardless, relative differences between performances on the same items are preserved.)

#### Item 1-38

original Small Talk error: \* *I wish I can communicate another country people.*

**Teacher reformulation:** **I hope I will be able to communicate with people from other countries.**

reformulation on RLT 1: \* *I hope I would be able to communicate...people from another countries.*

reformulation on RLT 2: \* *I hope I will comm...I hope...I wish I comm...I can communicate another country people.*

The teacher reformulation indicates four errors: 1) the context requires an expression of future rather than present desires; 2) ‘communicate’ is intransitive and is therefore followed by a PP not a DO; 3) complex modifiers cannot be used attributively in English as in Japanese, so ‘from other countries’ must be used post-nominally (Payne 2010: 4) the speaker is not referring to a single country but countries in general. The multiplicity of error types makes

this a challenging item, and the evidence from her RLT 2 attempt indicates that ‘Ayako’ is confused to the point of giving up. Clearly, this is not a simple case of recognizing a ‘mistake’. However, the lexical choice of *hope* over *wish* does seem to have been consistently corrected, and might therefore have been a ‘mistake’ originally.

#### Item 2-21

original Small Talk error: \* *I want to go trip a lot of country.*

**Teacher reformulation:**     **I want to take trips to a lot of countries.**

reformulation on RLT 1:     \* *I want to take trip a lot of countries.*

reformulation on RLT 2:     \* *I want...take trips a lot of countries.*

The teacher reformulation indicates four errors: 1) the collocation ‘take a trip/trips’; 2) the likely need for plural ‘trips’ owing to the context of ‘a lot of countries’; 3) the need for an adverbial prepositional phrase for adjunct, rather than what appears to be a DO; 4) the need for a plural noun with ‘a lot of’. The RLT 2 attempt indicates partial success again: ‘Ayako’ remembers the collocation and the plural ‘trips’ but omits the ‘to’ in ‘want to’. (The ‘to’ for the adverbial was missed on both attempts). Therefore one might tentatively conclude that on the TGJT she recognized a ‘mistake’ with ‘trip’. Nevertheless, the introduction of a new ‘mistake’ with ‘I want take..’, and the persistent error with the adverbial render the mistake/error distinction rather irrelevant: there is still work to be done here regardless of the IL status of any one of these errors.

#### Item 3-22

original Small Talk error: \* *I think American food is very strong taste.*

**Teacher reformulation:**     **I think American food has a very strong taste.**

reformulation on RLT 1:     \* *I think American foods are very strong taste.*

reformulation on RLT 2:     \* *I think American food has very strong taste.*

This time the teacher reformulation identifies two errors: 1) the use of ‘have’ instead of ‘be’ for NP predicates when describing a quality (as opposed to equating two concepts); 2) the

need for an indefinite article in NPs with count noun heads. The second RLT attempt indicates partial success (the correct verb) but there is no way to tell which of the three errors on the first RLT attempt ‘Ayako’ in fact recognized, so we cannot conclude that on the TGJT she recognized a ‘mistake’, only that she quickly recognized that the item as a whole was ungrammatical.

As these examples have made clear, the utterances learners produce in genuine communicative interaction are likely to be an amalgam of acquired forms, partially acquired forms, and guesswork. It is reasonable that adult L2 learners will rely on the highly automatized forms (even in translation) of their L1 when TL forms are not available, and this is very evident in the IEP data. Distinctions such as Edge’s (1989: 9–11) between ‘slips’, which learners can self-correct, ‘errors’ (which they cannot), and ‘attempts’, which are ‘a guess or when neither the intended meaning nor the structure is clear to the teacher’ may be more useful than Corder’s simple dichotomy, since the evidence here is that learners can correctly reformulate items and yet still judge them as ungrammatical, and vice versa. One is reminded here of Willis’ claim that ‘it is the learners’ attempts to mean that pave the way for learning’ (2003: 110-111).

### **5.5.6 Comparison of judgements of own and ‘All Do’ errors**

#### ELC group

For the ELC group, 52% of the TGJT items were ‘All Do’, in other words were errors originally made by other speakers. To determine if TGJT performance was affected by the source of the item, a t-test was performed on the data, which revealed no significant differences in student judgements of their own and peers’ errors (Table 27). This confirms the finding in Section 4.8.4 that students are no slower or less accurate with their own errors than with those of peers.

Table 27: *T-test of TGJT measures for own and ‘All Do’ errors, ELC group*

	<b>Own errors Mean (SD)</b>	<b>‘All Do’ errors Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b><i>t</i></b>	<b>sig.</b>
Accuracy	0.80	0.77	0.03	341	.737	.461
RT	1965 (5330)	1558 (1978)	406	341	.948	.344

IEP group

As with the ELC group, a comparison of means for judgments of participants’ own items and those of peers revealed no significant differences in either accuracy or RT (Table 28).

Table 28: *T-test of TGJT measures for own and ‘All Do’ errors, IEP group*

	<b>Own errors Mean (SD)</b>	<b>‘All Do’ errors Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b><i>t</i></b>	<b>sig.</b>
Accuracy	0.82	0.73	0.09	190	.888	.130
RT	1931 (1350)	2172 (2078)	241	190	1.57	.376

This confirmatory finding suggests that the practice of assigning ‘All Do’ items for pedagogical reasons is not in any way detrimental for the group as a whole.

**5.5.7 Comparison of TGJT performance by both groups**

Since the set of TGJT items for each group was different, it is not possible to make a cross-group comparison in order to rank participants, as was done in the pilot study. Given the difference in proficiency between the groups, the complexity of the language – and errors – each group produces is also different, as an inspection of the test items in Appendix 6, column D reveals. A CF methodology that did not account for these differences would be seriously flawed. Put another way, IL development is a moving target, and CF has to accommodate not only differences between groups but also within members of the groups. It follows therefore that the CF methodology should have approximately similar effects irrespective of learner

proficiency.

In Section 4.8.1, evidence was presented that the CF methodology does not place differential demands on different proficiencies, instead providing individual benchmarks for IL development. In this chapter, the investigation has focused on RT as a corollary of accuracy in grammaticality judgement, so it would be informative to discover whether the TGJT data show mean differences between the groups. The comparison of means in Table 29 shows that while there is no significant difference in accuracy (again, it must be stressed

Table 29: *T-test of TGJT measures for ELC and IEP groups*

	<b>ELC Mean (SD)</b>	<b>IEP Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b><i>t</i></b>	<b>sig.</b>
Accuracy	0.78	0.77	0.02	497	.42	.675
RT	1377 (1737)	2079 (1832)	702	497	4.3	.000

that the linguistic proficiency of the two groups is not being compared here, only their ability to recognize errors in their production), the ELC group was averagely seven-tenths of a second faster in their judgements than the IEP group. This finding is possibly explained by the proficiency difference between the two groups: the ELC group are more proficient and have greater exposure to the TL (see Table 12), and despite the greater complexity of the language, their average reaction time suggests that they are able to rely more on their implicit knowledge than the IEP group. We would expect this to be the case, but there are other possible explanations for the average RT difference. A likely candidate is general listening proficiency: the ELC group consists largely of Arabic and Spanish L1 learners whose educational experiences and learning styles generally prioritize auditory learning (e.g. Reid 1987: 96, Table 3). However, in this TGJT the participants are listening to themselves, and both the form and content of the utterances are known beforehand, which should offset at least some of the effects of greater listening proficiency. Further research comparing similar populations at different proficiency levels is needed in order to determine whether RT in this

context is affected by proficiency – but if it were, this type of test, which is inexpensive and easy to administer, might be an interesting source of data for placement and proficiency assessment.

#### **5.5.8 TGJT influence on post-test**

A number of items from both groups ( $N = 97$ ) appeared on RLT 1, then on the TGJT and finally on RLT 2. In addition, since some participants did not take the TGJT (see Section 5.3), it was possible to discern what effect, if any, the TGJT itself might have had on the ultimate accuracy of the items which appeared on RLT 2. Because it was possible, even likely, that the more invested students would have volunteered to take the TGJT, the mean accuracy and fluency scores on the RLT 1 of those who (subsequently) took the TGJT were compared with those of the students who did not (item  $N = 72$ ). Significant differences were found between these two groups on both measures, with the former averagely 17% more accurate and 22 WPM faster than the latter ( $p < .005$ ). This initial imbalance between the groups makes the findings of the post-RLT 2 means comparison questionable, but they are reported here as they were rather surprising.

The group of participants who took the TGJT showed no increase in overall accuracy between RLT 1 and 2 (Table 30). In other words, taking the TGJT was neither beneficial nor detrimental to their accuracy, which remained at around 80%. At the same time, their fluency as measured by WPM increased somewhat, perhaps indicating greater confidence. But as the non-TGJT group increased in their WPM by a greater amount, and furthermore increased 21% in accuracy, one has to wonder whether the TGJT had a deleterious effect. However, it should be noted that both groups attained an RLT 2 accuracy score of around 80%, which is highly reminiscent of the accuracy scores achieved overall by both the ELC and IEP groups on RLT 2 reported in Section 4.8.1. Thus the conclusion is that any increases in accuracy and

Table 30: *T-test comparison of means on RLT 1 and RLT 2 according to TGJT participation*

<i>Participated in TGJT</i>						
	<b>RLT 1 Mean (SD)</b>	<b>RLT 2 Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b><i>t</i></b>	<b>sig.</b>
Accuracy	0.79	0.79	0.00	95	.000	1.000
WPM	125 (45)	136 (60)	11	95	-.299	.003
<i>Did not participate in TGJT</i>						
	<b>RLT 1 Mean (SD)</b>	<b>RLT 2 Mean (SD)</b>	<b>Difference</b>	<b>df</b>	<b><i>t</i></b>	<b>sig.</b>
Accuracy	0.60	0.81	0.21	70	-2.821	.006
WPM	103 (43)	119 (48)	16	70	-3.87	.000

fluency are attributable to practice rather than the presence or absence of TGJT. Those who took the TGJT were also those who already had reached, on average, the highest accuracy score possible through practice.

From the ecological viewpoint the TGJT, which is not currently part of the CF methodology, thus had no detrimental effect on students' ultimate performance. But neither did it serve to raise awareness of error, as might have been predicted. It is possible that more extensive feedback from the TGJT might do so: at present, the only feedback provided is the number of items students got correct – unlike the pilot TGJT, which gave very specific feedback on participants' judgements (see section 3.4.5). Informal comments from participants indicated that they found the TGJT 'interesting' and 'fun' (and many commented on how strange their voices sound to them!), and it may be that with much more specific feedback – say, a list of the items with the TL grammaticality alongside their judgements for comparison – might better focus students' attention on problem areas. Future research should address this question in order to determine whether TGJTs can play a facilitating role in acquisition.

## 5.6 Discussion

The evidence presented in this chapter has been used to address three questions. In answer to the first, *Can learners identify the grammatical status of their own utterances?* it was found that both groups were biased towards judging their production as grammatical, indicating that many of their errors sounded correct to them. Only 19% of the ungrammatical items were judged ungrammatical, initially prompting the hypothesis that only non-systematic ‘mistakes’ in their own performance can be judged ungrammatical by learners. It was further hypothesized that these ‘mistakes’ would be characterized by faster RTs in judgement. Several possible candidates were identified on this basis, but the complexity caused by multiple errors within a single item render this element of the research inconclusive. A provisional conclusion, however, is that learners can in fact recognize some ungrammaticality that is neither a ‘slip’ nor a systematic ‘error’, but may be instead what Edge (1989: 9–11) calls an ‘attempt’.

In addressing the second question, *What is the relationship between TL accuracy in production and TL accuracy in judgement of grammaticality?* as noted in section 4.9, learners in both groups were able to reformulate their spoken errors at an average accuracy level of 75% (ELC) and 69% (IEP), resulting in accuracy correlations with the TGJT of  $r = .767$  and  $r = .593$  respectively. This indicates that the relationship between TL accuracy on the two measures is fairly strong. But it is more than likely that this question is too broad, and should be more narrowly focused: the evidence indicates that overall, the participants correctly judged the grammaticality of 78% (ELC) and 77% (IEP) of the items. Both groups judged approximately 5% of the grammatical items as ungrammatical, and approximately 18% of the ungrammatical items as grammatical, indicating that learners at these two proficiency levels are likely to be inaccurate in their judgements approximately 20% of the time. It is precisely



this inaccuracy which is of pedagogical interest, since these are the forms which most likely represent errors or ‘attempts’ in need of further remediation.

What is somewhat more reliable is the finding that there is no simple way to distinguish between ‘errors’ and ‘mistakes’ on the basis of the metalinguistic judgements of learners, which contradicts the finding in section 4.9. It is possible that the original data (the Small Talk items) are in fact mostly errors and not mistakes, since there is a level of judgement and screening by the teacher at the moment of data collection (see Module II). However, the response bias analysis in Section 5.5.3 revealed that the less proficient IEP group was better able to recognize their own inaccuracies, which might lead to the conclusion that they make more ‘slips’, a phenomenon documented by Poulisse:

The large number of L1-based slips in beginner’s L2 speech can be explained [by the fact that] L1 procedures are largely automatized, while L2 procedures are not yet. Thus it is relatively hard for beginning L2 speakers to activate the L2 procedures (it takes much attention), while the L1 procedures must be suppressed. As a result, sometimes the L1 procedures will accidentally take the place of the required L2 procedures. (Poulisse 2000: 145).

Poulisse found that proficiency-related differences emerge at the lexical level (mostly substitutions) followed by phonological and morphological level (mainly verb forms) (Poulisse 1999). Hence the IEP group simply produced more ‘L1-based slips’ (such as ‘Ayako’s’ *to go trip*) and were more able to recognize them – but not necessarily correct them; the ELC group produced fewer and recognized fewer, and thus the remaining ungrammaticality can be assumed to be more systematic. Therefore, the hypothesis that all learners should be able to recognize only correctable ungrammaticality (Corder’s ‘mistakes’) must be adjusted to incorporate L2 proficiency.

The final research question in this chapter, *What is the relationship between fluency in production and reaction time in judgement?* is difficult to answer with any degree of confidence. There certainly does not appear to be a strong, linear relationship between fluent

production and the speed of judgements of grammaticality. This supports findings such as those of Coppieters (1987) and Birdsong, who concludes:

Inasmuch as metalinguistic performance reflects idiosyncratic skill parameters, which vary across task and across individuals, it cannot, in any rough-and-ready manner, reflect grammatical or linguistic competence presumably possessed by all speakers of a language. (1989: 61)

Simply put, there are too many individual and contextual factors for metalinguistic performance to be seen as a window onto linguistic competence, or indeed, linguistic performance. Instead, the speculative conclusion is that speed of reaction time in judgement of an item depends more on whether a participant judges it to be grammatical or not than on factors of fluency in production. If the item ‘sounds right’, the RT will be much faster, by an average of 1.5 seconds. (Compare this with the finding for NS in the pilot study that items which ‘sound wrong’ are judged faster.) The analysis behind this finding is based on comparatively little data, since the majority of TGJT items happened to be (TL) grammatical, and in addition the response bias found in both groups resulted in a great imbalance between judgements as grammatical and as ungrammatical. Additional research, in which a better balance of (TL) grammatical and ungrammatical items could be planned – which would potentially permit an increased number of judgements as ungrammatical – would enable a more valid investigation of the origin of L2 metalinguistic intuitions. In addition, it would be informative to design a comparison experiment in which reaction time is measured from item onset, rather than from item end, to participant judgement; this might allow more precise measurement of fluency in recognition, especially at higher proficiency levels.

In terms of CF, the implications of this investigation are modest but contribute to our overall understanding of the various choices before teachers. One fairly unequivocal finding is that learners’ metalinguistic judgements of their own output are no more accurate or faster than their judgement of peer output, which supports the conclusion from the RLT

investigation that the practice of using peer error as a source of CF input is not harmful. It is hard at this point to say whether or how the TGJT is beneficial, but there is no reason to reject the assumption that it can raise awareness by addressing avoidance and drawing attention to useful language forms, especially if more specific feedback on erroneous judgements could be provided.

This exploratory TGJT investigation is one of a very few studies of learner intuitions which uses the learner's own production as data, and the only such study to date that uses audio recordings of the learners themselves. Whether this line of investigation ultimately turns out to be rewarding will depend on refinements in the research methodology such as those suggested; however, the creation of a database of learner errors together with accompanying audio recordings of learner reformulations will almost certainly prove valuable for a number of SLA investigations. The pedagogical utility of such a database is that it will permit on-going refinements of the CF methodology by allowing comparison of students' current and former production, comparison by students of other students' reformulation attempts, and so forth. This will be taken up in the following chapter.

---

## CHAPTER 6: INVESTIGATION OF THE SMALL TALK DATABASE

---

### 6.1 Introduction

In Chapters 4 and 5, evidence was provided that delayed corrective feedback (CF), even in the context of fluent oral production, is effective in pushing learners toward greater accuracy and complexity. It was found that in both reformulation and recognition, learners achieved an average accuracy level of 70-80% and analysis of the remaining inaccuracy pointed to increasing complexity (i.e. newly-introduced errors) as a major contributing factor. Simply put, learners do not achieve 100% accuracy through CF because they are in the process of learning. What the CF methodology under investigation can do, however, is track this learning at the level of both the cohort and the individual. The mechanism for this tracking is the database referred to in Figure 1 and throughout. This chapter describes this database and discusses the applicability and contribution of concepts and techniques from corpus linguistics to the systematic CF approach described herein.

### 6.2 Research questions

In contrast to the preceding chapters, which adopted an experimental approach to the investigation of CF, this chapter represents an exploratory line of investigation into the utility of the error database in terms of pedagogy and research. It therefore addresses the following two broad research questions:

1. Can an error database provide a representative monitor of linguistic development in individuals and groups?
2. What are the pedagogical applications of such a database?

### 6.3 Rationale for a learner error database

For reasons which will be outlined below, an error database may be unsuitable for many types of research, and may ultimately prove unsuitable even for most kinds of IL analysis.

Nevertheless, the Small Talk database has certainly provided authentic data on spoken learner error for research purposes. For instance, Harrat (2011) used a stratified sample of errors from Arabic L1 speakers at different proficiencies to investigate the relative frequency of developmental and transfer errors; Wagner et al. (2009: 480) used a random sample of 4500 items from the Small Talk database to test their automatic error detection system; Green (2006) used a subset of the database to analyse the spoken errors of Japanese learners and proposed an instructional syllabus based on the frequency, difficulty, and utility of the intended forms. But perhaps the application with the potential for most widespread use is pedagogical: teachers, and especially trainee teachers, need ways to check their intuitions about (inter)language use, intuitions which can greatly contribute to the design and direction of the instructional syllabus, and data-driven research is beginning to provide empirical tools for this purpose. Furthermore, learners can benefit from a systematic sampling of their developing forms as well as intentional focus on the ways in which their (and their peers') language differs from the TL.

These objectives are generally shared by learner corpus researchers, and for this reason, this investigation has adopted many of the perspectives and techniques common in such corpus research. However, the Small Talk database differs from a corpus in several important ways. Gilquin and De Cock, for instance, assert:

When using corpus data, one question worth asking is how prototypical one's corpus is. Prototypical corpora are characterised by the fact that they have been produced in a natural communicative setting, which sets them apart from more experimental data like acceptability judgements, word association tests or measurements of reaction times (Gilquin & Gries 2009:6). In this respect, so-called 'corpora of speech errors'... do not qualify as corpora in the corpus linguistic sense. (Gilquin and De Cock 2011:

163)

Thus, since all of the language contained in the Small Talk database has been deliberately selected for its ungrammaticality or awkwardness, it lacks some criterial features of a corpus on which most corpus linguists seem to agree. For instance, Cowan et al., building on work by Granger (1998) and others, suggest the following essential characteristics of a learner corpus ‘for determining persistent grammatical errors of L2 learners’:

It should (a) encompass different levels of proficiency, (b) consist of extensive samples of learner language that facilitate analysis of grammatical errors caused by phenomena beyond the boundaries of the sentence, (c) be labeled so that researchers and materials developers can determine whether the total number of errors of a given type is produced by a small number of learners or by many different learners, and (d) be fairly large. (Cowan et al. 2003: 452)

However, even a database of errors would have to conform to such criteria as far as possible in be maximally representative and useful. For this reason, the considerations detailed by Cowan et al. will be examined in relation to the design and composition of the Small Talk database.

The first of these points is addressed in Section 6.5 below, although it needs to be said that spoken data are difficult to gather from low proficiency students, particularly fluent spoken data. It is generally felt that CF of the type described in this research is inappropriate for very low-level learners, who essentially cannot produce any language fluently. Nevertheless, the little data that have been collected are still quite revealing and it is to be hoped that more will be added. Izumi et al. make an additional point concerning the need for longitudinal data:

[T]he real significance of EA [Error Analysis] cannot be identified without using diachronic data in order to describe learners’ developmental stages. The types and frequencies of errors change with each acquisition phase. Without longitudinal data of learner language, it is difficult to obtain a reliable result by EA. (Izumi et al. 2005: 73)

There are two main reasons why context is considered important in learner corpora, in addition to that given by Cowan et al. (point b, above). First, researchers will quite reasonably

be interested in what learners do *successfully* (Kindt and Wright 2001; Granger et al. 2002). This refers not only to output which is target-like, but also to successful interaction and communication, whether or not it meets target norms. Second, much of the work with learner corpora has related to the observed frequency of grammatical and lexical forms, comparing these to corpora of NS output and identifying ‘overuse’ and ‘underuse’, as well as ‘misuse’ (e.g. Granger and Tyson 1996; Hinkel 2002; Barlow 2005; for critiques of this approach, see Galloway 2005; Tan 2005). Without a representative sample of learner production, it would be difficult to describe ‘use’ at all, and a representative sample would have to include the entirety of what can be observed and recorded at any point, not merely a restricted subset of errors.

The pedagogical application of corpus data (Johns 1991) has been criticized for providing a bottom-up, overly inductive, and decontextualized view of language (Widdowson 2002; Flowerdew 2009; Scheffler 2011). There is insufficient space here to review this debate, which has dogged corpus linguistics for many years; however, the question of context is highly relevant to the data collection and pedagogical use of the Small Talk database, and needs to be addressed. Here, de Beaugrande’s (2001: 114) comments are pertinent:

I would submit that a real text cannot be decontextualized, that is, removed from any context; we can only shift it into a different context, which is an ordinary transaction not just in language classrooms, but in most reports or discussions of what somebody said. With real text, you cannot help getting implicated in interpreting it.

To illustrate, a random sample of utterances from the Small Talk database is presented in Table 31. As CF, these items were transcribed by classroom teachers during Small Talk sessions, an act of interpretation in itself but one which has been shown to remain faithful to the spirit, and most often the letter, of the speakers’ actual words. (This point was covered more fully in Module II of this research.) The ‘Context/Vocabulary’ and ‘Topic’ columns contribute to the reformulation process – and frequently (lines 1, 2, 6, 8–10) indicate the

intended meaning. Next, they were presented to the speakers and their peers, which constitutes a second level of interpretation as the participants reconstruct the intended meaning and compare this with the teacher's reformulation. They might also form part of a 'focused worksheet' (for example, expressions

Table 31: *Random sample of ten utterances from Small Talk database*

	<b>L1</b>	<b>Utterance</b>	<b>Context/Vocabulary</b>	<b>Topic</b>	<b>Level</b>
1.	Spanish	If I get married with you, it's only with you.	can't get re-married	Divorce	105
2.	Arabic	But it take many time.	a long time	Capital Punishment	104
3.	Arabic	I took a lot of exercise.	Referring to activities he took part in.	Vacation	102
4.	Korean	Did you think about the crime, last night?	The leaders today are Hyunchae and Maryam.	Crime	107
5.	Arabic	There is a teacher I want to meet him.	30 minutes away from his home, by car	Food Check-in	104
6.	Arabic	After my father shout at me, he asked me to be in his shoes.	put myself in his shoes	Mistakes	106
7.	Japanese	I think it's not Japanese story.	after surgery, her personality changed (was it a heart transplant?)	Reincarnation	108
8.	Thai	I want to be a lawyer.	in the past	Turning point	105
9.	Korean	In my country we just study grade.	To get good grades	Education Systems in the World	105
10.	Arabic	When I was child, I lost in the mall.	get + V	Getting in Trouble	104

with *time*, simple past VPs, adjective clauses, and so on) for focus-on-form instruction.

Finally, they might be used for research purposes (see below). In these last two applications, the original speaker might not be available to clarify meaning, and the words that preceded and followed the utterance are similarly absent. However, for instructional purposes this is not a drawback since the intended meaning can nearly always be deduced from the utterance itself, the contextual gloss provided, and finally, by the teacher reformulation. Even without the latter, it is quite easy to understand the intended meaning of the utterances in Table 31, and experience with Small Talk conversations demonstrates that the full context of the



conversation does not necessarily contribute to our understanding. For instance, in Module II, a portion of the transcript for a conversation between three students was presented, of which a section is reproduced here:

- 1) S9 What does mean?
- 2) S6 Like changes something.
- 3) S12 Change to um I, I don't know how to say. Like a little bit, change bigger than [inaudible] (looks at T1)
- 4) S9 Custom?
- 5) S6 You mean transport?
- 6) S9 Do you change your car?
- 7) S12 You drive fast driver, fast drive um
- 8) S6 Explain [(inaudible)]

The underlined portions are the utterances that were entered into the database, along with topic and contextual information. The point here is that in none of the three cases would the full context contribute to an understanding of the intended meaning and its divergence from a TL version. Granted, the utterances on their own say nothing about the pragmatics of the interaction, and valuable information about the phonological or discourse features has been lost; however, from S9's point of view, the relevant information is that *\*What does mean?* is non-target-like and that a proficient speaker would probably say something like *What do you mean?* or *What does that mean?* And if the purpose were to investigate learners' (mis)uses of *mean*, for example, the surrounding context would not contribute more to one's understanding than would the teacher's reformulation: *What does that mean?*

The labelling or annotation of the database is the main focus of this chapter, but several preliminary observations should be made. Annotation generally refers to two types of labelling: the first is 'metadata' (Krishnamurthy and Kosem 2007: 368–9), which is essentially descriptive information about the participants and circumstances that gave rise to the speech event. The second type of labelling is tagging (Atwell 1987; Leech 1993; McEnery and Wilson 1996; Granger 2003), which refers to the annotation of linguistic features in the data, most commonly parts of speech (POS), types, tokens, and lemmas, in a form which is

machine-readable but unobtrusive. Tagging is critical, as it refines the searchability of the texts beyond simple string matching (the procedure used in the *Find* function of a word processor), allowing the user to identify, for example, all instances of a passive VP.

The conventional wisdom in corpus linguistics concerning corpus size is that bigger is better (Sinclair 1991). Statistically, larger samples are more likely to be representative of actual language use than smaller ones, but in addition, larger samples will likely contain a greater diversity of sources, registers, and styles. Granger notes that the question of size is relative, however:

A corpus of 200,000 words is big in the SLA field where researchers usually rely on much smaller samples but minute in the corpus linguistics field at large where recourse to mega-corpora of several hundred million words has become the norm rather than the exception. (Granger 2003: 465)

In fact, this situation is rapidly changing: only two of the ten corpora surveyed by Pravec (2002) are under 500,000 words, and most are much bigger. Even though the Small Talk database consists entirely of samples of learner production characterized by their non-target-like nature, it stands to reason that here, too, representativeness (of non-target-like production) is to a great extent a product of sample size: the larger the database, the more likely that the error patterns detected therein will accurately reflect the developing interlanguages of the individuals and groups sampled.

The research and pedagogical value of learner corpora is becoming less controversial and has been extensively argued for in recent literature (e.g. Kindt and Wright 2001; Granger et al. 2002; Meunier 2002; Gilquin et al. 2007). The online *Learner corpus bibliography* (Paquot 2009) currently lists over 440 books, journal articles, and dissertations related to learner corpus research (LCR) published between 1993 and 2011: like its parent field, corpus linguistics, LCR is a research field which has now established its legitimacy. In particular, Granger notes that computer-aided error analysis is ‘a key aspect of the process which takes

us towards understanding interlanguage development and one which must be considered essential within a pedagogical framework' (Granger et al. 2002: 13–14). It should be noted, however, that the majority of learner corpora consist of written rather than oral production, chiefly owing to the relative ease of compiling the former (Römer 2008) and the considerable resources required to compile the latter (Pérez-Paredes 2003; Barlow 2005). Furthermore, there are very few spoken learner corpora compiled from interactive conversation. Most commonly the approach is to transcribe interviews (e.g. the Louvain International Database of Spoken English Interlanguage (*LINDSEI*), De Cock et al. 2010) or from oral assessment tests (e.g. the Japanese Learner English Corpus, Izumi and Ishihara 2004). Perhaps the most developed oral interaction corpus to date is the Multimedia Adult ESL Learner Corpus (Reder et al. 2003), but this has not yet been systematically tagged, let alone error-tagged, and of the 3600 hours of video and audio data collected, only 150 hours have so far been transcribed.

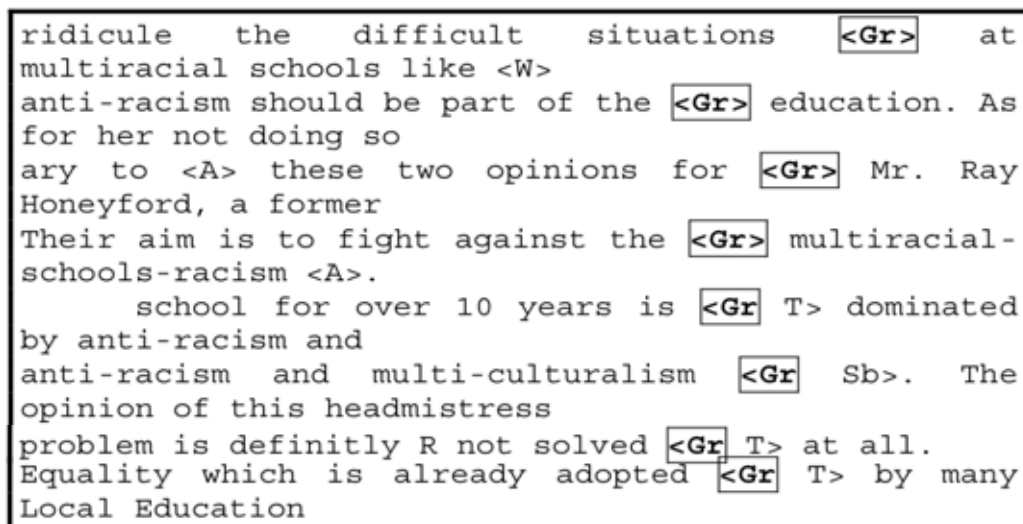
The reality of an essentially pedagogically-driven data collection project is such that the resources that would enable the amassing of entire conversations, transcribed and annotated, are far beyond the reach of most researchers, let alone teachers. Thus contextualization of the data in the error database described here is restricted to the utterances themselves, whatever contextual information can be included in the course of the Small Talk conversations, and the reformulations of the teacher. Subsequent interpretation of the error database data relies on the intuitions of the participants (in this case learners, teachers, and analysts), rather than being 'objectively' established by full discourse context. As Leech and others have pointed out, the objectivity of corpus analysis is always relative:

Recent corpus users have accepted that corpora, in supplying first-hand textual data, cannot be meaningfully analysed without the intuition and interpretative skills of the analyst, using knowledge of the language (qua native speaker or proficient non-native speaker) and knowledge about the language (qua linguist). In other words, corpus use is seen as a question of corpus plus intuition, rather than of corpus or intuition. (Leech 1991: 74, in Partington 1998: 2)

In the case of the Small Talk error database, the absence of full contextualization is offset by its utility and applicability in teaching and CF, as will be explored in Section 6.7 below.

#### 6.4 Error taxonomies and analytical frameworks

The error analyses of learner corpora are often underdeveloped and lack the specificity which would make them valuable to learners, teachers, or SLA linguists. For instance, Mukherjee and Rohrbach (2006: 226) offer the following example (Figure 16) of a concordance of grammar errors which ‘should be taken by the student as a starting-point for a revision of some general rules in English grammar from which she deviated, e.g. in the field of article usage (cf. lines 2 and 4) and with regard to the distinction between past tense and present perfect (cf. lines 5, 7 and 8)’.



ridicule the difficult situations <Gr> at  
 multiracial schools like <W>  
 anti-racism should be part of the <Gr> education. As  
 for her not doing so  
 ary to <A> these two opinions for <Gr> Mr. Ray  
 Honeyford, a former  
 Their aim is to fight against the <Gr> multiracial-  
 schools-racism <A>.  
 school for over 10 years is <Gr> T> dominated  
 by anti-racism and  
 anti-racism and multi-culturalism <Gr> Sb>. The  
 opinion of this headmistress  
 problem is definitely R not solved <Gr> T> at all.  
 Equality which is already adopted <Gr> T> by many  
 Local Education

Figure 16: An example of syntactic error tagging from Mukherjee and Rohrbach, 2006

Although the authors do not elaborate, it is clear from the excerpt above that there is some level of sub-categorization in this tagging: three examples are tagged with ‘<Gr T>’ to refer to misuse of tense. The authors are to be commended on their efforts – the tagging of errors in learner corpora is still done largely manually – but it must be recognized that the result is a

very blunt instructional tool: first, opinion in the SLA literature has converged recently on the opinion that explicit feedback is of greater use than implicit (N. Ellis 2005; R. Ellis et al. 2006; Li 2010; Sheen 2010), and pointing out a ‘grammar’ error, or even a ‘tense’ error, may not be explicit enough. Second, any attempt to quantify error levels, say in order to target specific areas in need of remediation, would require a greater level of precision than can be provided by a schema such as that of Mukherjee and Rohrbach. This is somewhat ironic, since one of the authors’ purposes is to narrow the ‘widening gap and a growing lag between on-going and intensive corpus-linguistic research on the one hand and classroom teaching on the other’ (Mukherjee and Rohrbach 2006: 205–6). They rightly claim that

the focus on their own students’ output will involve many more teachers in corpus-based activities and that, secondly, the exploration of learner data by the learners themselves will motivate many more learners to reflect on their language use and thus raise their foreign language awareness. (Mukherjee and Rohrbach 2006: 228)

Perhaps these authors are implicitly making the point that complex tagging systems are too unwieldy for use by the average learner (or teacher) or perhaps they were simply unaware of the work that has already been done in developing error taxonomies or classification systems. Granger (2003), for example, offers an error tagset of 11 major categories and 54 subcategories. Díaz-Negrillo and Fernández-Domínguez (2006) review no fewer than twelve corpora with error tagging systems either in place or under development, ranging from tagsets of around 30 items to upwards of 100. They acknowledge that tagsets are frequently developed for specific purposes and emerge from specific theoretical positions, but lament:

[E]rror taxonomies tend to account for diverse dimensions of error classification, encoding conventions and annotation models, which only shows that there is no standard of error annotation. Tools do not seem to be shared by the research community and, when commercialized (e.g. the Louvain error tagging system) they do not appear to be used as widely as it would be expected. (Díaz-Negrillo and Fernández-Domínguez 2006: 97-98)

While there may be no standard, several researchers have proposed design guidelines for annotation schemes. Granger proposes four general characteristics of a maximally effective

system:

- 1) *informative* but *manageable*: it should be detailed enough to provide useful information on learner errors, but not so detailed that it becomes unmanageable for the annotator;
- 2) *reusable*: the categories should be general enough to be used for a variety of languages;
- 3) *flexible*: it should allow for addition or deletion of tags at the annotation stage and for quick and versatile retrieval at the postannotation stage; and
- 4) *consistent*: to ensure maximum consistency between the annotators, detailed descriptions of the error categories and error tagging principles should be included in an error tagging manual. (2003: 467, emphasis in original)

These points are arguably common sense, but it must be said that the design of corpora and annotation schemes involves a degree of technical expertise for which many teachers and linguists are not trained. For example, a menu-driven tagging system such as the one employed by the Louvain team (Dagneaux et al. 1998) allows for increasingly fine-grained analyses in the tagging process without overwhelming the user; but if the user decides that a new tag is necessary she may well be unable to add it on the spot. Similarly, if a tag is to be deleted, or more likely, reassigned or subsumed under a different tag, there should be a mechanism for locating and reassigning all instances of that tag applied up to that point. These are programming considerations that must be decided in the early stages of design.

Tono (2003: 804) follows the approach proposed by James (1998) to include two levels of analysis, '(a) linguistic category classification (e.g. [grammar] – [verb] – [morpheme] – [tense]) and (b) target modification taxonomy (e.g. [omission/addition/misinformation/etc.])'. It is the second of these that is perhaps the main cause of the lack of standardization in error tagging schemes: the very real problem, discussed in Module I and throughout this research, of the uncertainty of error type (Milton and Chowdhur 1994; Tono 2003). This certainly affects the reliability of analyses, which in turn influences the design of the tagging schema. This problem is particularly acute in the analysis

of writing, where no appeal can generally be made to the author for explanation. But even when the error has been noted and analysed by the same person, there are times when it is difficult to be sure what the correct interpretation is. (Recall the ambiguous *So, every season changes color in the mountain* discussed in Module II.) Even when an error can be unambiguously interpreted, there may be a number of possible reformulations, none of which is necessarily ‘more accurate’ than the others. In these cases, the best way to proceed is to confer with the speaker and with other annotators (Fitzpatrick and Seegmiller 2004: 227), and the best outcome is to include multiple analyses and tags (Tono 2003). Leech’s (1997: 7) maxim that an annotation scheme be ‘based as far as possible on consensual or theory-neutral analyses of the data’ thus remains an ideal to strive for. One cannot help but agree with Milton and Chowdhur’s conclusion that ‘[t]agging a learner corpus allows us, at least and most, to systematize our intuitions’ (1994: 129).

## **6.5 Description of the error database**

The Small Talk database currently contains approximately 38,000 utterances, representing over 300,000 words gathered during classroom Small Talk sessions between 1994 and 2011. The distributions of utterances and words by L1 and English proficiency are given in Table 32 and Table 33. As can be seen, these distributions are quite uneven, with some L1 backgrounds and proficiency levels much better represented than others. These distributions can be attributed to economic factors such as the recruitment of students to the English Language Center, educational factors such as the English language education in students’ home countries, and random factors such as the willingness of teachers to contribute data to the Small Talk database: as with any instructional technology there is a degree of training required, and some teachers are simply not willing to invest the time and energy to learn how

Table 32: *Breakdown of utterances in error database by L1*

<b>L1</b>	<b>% of utterances</b>	<b>Word count</b>
Japanese	29.65%	88,847
Arabic	25.40%	74,250
Korean	17.63%	51,639
Mandarin (Chinese)	11.30%	32,238
Spanish	9.33%	27,408
Thai	1.42%	4,464
Portuguese	0.76%	2,423
Russian	0.91%	2,372
Ukrainian	0.73%	2,151
Vietnamese	0.59%	1,694
French	0.41%	1,147
Farsi	0.32%	1,097
Amharic	0.44%	1,049

Table 33: *Breakdown of utterances in error database by proficiency*

<b>ELC Level</b>	<b>Approximate IELTS score</b>	<b>% of utterances</b>	<b>Word count</b>
099	0 – 1.5	0.06%	79
100	1.5 – 2.5	0.17%	303
101	2.5 – 3.5	0.50%	973
102	3.5 – 4.5	2.49%	5,180
103	4.0 – 4.5	8.77%	22,610
104	4.5 – 5.0	16.16%	43,473
105	5.0 – 5.5	11.71%	32,662
106	5.5 – 6.0	12.59%	36,638
107	6.0 – 6.5	29.06%	92,804
108	6.5 – 7.5	19.43%	62,461

to use it. In addition, the technology itself and theoretical underpinnings of the CF methodology are experimental, and some teachers may not be convinced that the enterprise is worthwhile.

The Small Talk database was compiled using Microsoft Access to facilitate future migration to a SQL server with web-based front-end interfaces. The use of a relational database to store text and hyperlinks to related files (e.g. audio recordings) is increasingly preferred over the storage of multiple annotated text files for reasons of retrieval speed and scalability (Davies 2009: 164 & ff), and in the case of the Small Talk data, the relative ease with which data could be entered, annotated, analysed, and retrieved in a database application



made Access an attractive choice for the error database. The primary limitation of Access is the maximum database size of 2GB, but web migration is relatively straightforward once the database architecture is in place. These limitations aside, the database is representative of the error production of lower intermediate to advanced ESL learners from five language backgrounds (Japanese, Arabic, Korean, Mandarin, and Spanish) engaged in fluent oral communication.

## **6.6 Annotation procedures**

As mentioned in above, the Small Talk error database has been compiled as a relational database, and therefore there is no annotation or tagging of the text itself. Instead, utterances and annotations are stored in separate tables, permitting multiple analyses and annotations. This is particularly useful for teacher-training purposes, since a cohort of trainees can be assigned a subset of the data and can perform independent analyses without affecting the ‘official’ analysis or each other’s. The ‘tagsets’ are also stored in tables, which means that additions, deletions, and updates (i.e. re-categorization) of all related data can be quickly and easily made.

Worksheets are entered using the Worksheet Entry form (Figure 17), on which the minimum required information is the ‘Expression’ and the ‘Speaker’. This form and the Analysis form (Figure 18) both feature a link to the web-based Corpus of Contemporary American English (*COCA* – Davies 2009) so that those entering the data or performing the analyses can check their intuitions about grammaticality or acceptability when necessary. After the worksheet has been entered, teachers customarily record their reformulations, on which they base their error analysis. The analysis is done by selecting all or part of the utterance (the ‘extent’ of the error, or ‘the rank of the linguistic unit, from minimally the morpheme to maximally the sentence, which would have to be deleted, replaced, reordered,

Worksheet Entry Form

Worksheet Entry Form: **Spring I, 2011** **107 A**

Expression:

Pronunciation:

Context/Vocabulary:

**All Do** ☒


Check this box if you want to mark this sentence for all students to correct.

**Check a corpus**

Worksheet #:  Speaker:

Date:  include other section ☐

Topic:

 Delete this sentence.

**Spellcheck** **Analyze this Worksheet** **Preview Worksheet** **Preview Student Error Count**


Record: 14 of 44  Filtered Search

Figure 17: Small Talk database Worksheet Entry form

**Sub Analysis Form** Current Files: [Teachers&ClassesID] = 1903 AND [Worksheet#] = 1

Expression: Think in logical way | Grammar/Vocabulary: | Register: Spring 1 2011 | Worksheet #: 1 | Date: 1/19/2011 | Teacher: Hunter

Register: aalmuqati | Level: 107 | Term: | Teacher: Hunter

Error count: 5 of 43

Word ☐ Phrase ☐ Clause ☐ Sentence ☐ Text ☐ Paragraph ☐

Word form ☒ Choice ☐ collocation ☐ stress ☐ register ☐

noun P ☒ Verb P ☐ Adj P ☐ Adv P ☒ Prep P ☐ Particle ☐

SVC ☐ & ☐ TO ☐ OTH ☐ link ☐ complex ☐ Neg ☐ SV split ☐ Q ☐ Voice ☐ Exist ☐ Modal ☐

Word comp ☐ Sent bound ☐ Cls ☐ Claj ☐ Clav ☐ Clamp ☐ Cl-coord ☐ Cl-adv ☐ Apposition ☐

Topic Seg ☐ Num Seg ☐ Form Seg ☐ W Seg ☐ Det Seg ☐ Predicator ☐

Segment 1  Segment 2

☐ Count back to

Instructions: Highlight an error from the "Expressions" box above, select a category from "Syntax", and then choose a sub-analysis on the right. Save the analysis, or use "Reset" to start again.

Error: | Syntax: |

Note: |

Error	Syntax	Sub-analysis	Notes	Analysis
logical way	NP	Count	a logical way	Hunter
in logical way	wf	PP > Av	logically	Hunter

Record: 11 4 5 of 43 |  Search

Figure 18: Error analysis and tagging form for Small Talk database

or supplied in order to repair production’ – Lennon 1991: 191), and then choosing the first level of analysis, or error ‘domain’. In Figure 18, the utterance *Think in logical way!* has multiple (two) reformulations, and so the analysis proceeds on the assumption of either an incomplete NP or as a word form (wf) error in which a prepositional phrase has been used in place of a (much more common) adverb. In both cases, the target form is also supplied, along with the name of the analyst, so that the rationale behind the analysis can be questioned later. After the error domain (e.g. NP) has been selected, a second drop-down menu appears (not visible in Figure 18) with the syntactic, morphological, or semantic categories pertaining to the target modification relevant to that domain (in the case of NP, categories such as *specified, non-specified, count, non-count, determiner agreement*, and so on). The middle section of this form shows an array of feature tick boxes which are automatically ticked when the analysis is made below. This permits quick visual confirmation of the analysis, but also allows for fast searches of similar features as a way to increase the consistency and reliability of analyses.

### **6.7 An example of an error analysis: conditional clauses**

While considerations of space will not permit explanation of each domain and its sub-categories, one example should serve to illustrate some of the theoretical and procedural issues. The choice of conditional clauses is motivated by the following factors: first, they are relatively common in both NS and NNS production, and therefore constitute an area of pedagogical interest; second, they are present at all proficiency levels in the database, so can illuminate developing complexity and accuracy; third, as they can contain all clausal elements present in independent clauses, their analysis illustrates the procedural choices involved in systematic categorization; and finally, conditional constructions can span multi-clausal units and therefore involve particular pronoun and verb form sequencing issues. The description

and commentary below is primarily informed by Quirk et al., (1985), but two pedagogical grammar texts, the *Collins COBUILD English Grammar* (Sinclair 2005) and Carter and McCarthy (2006) were also consulted.

### 6.7.1 Comparison of *if* conditionals in native speaker data and Small Talk data

In native speaker (NS) English, timed conditional clauses are identifiable by the presence of subordinators such as *(even) if, unless, as/so long as, assuming (that), given (that), in case, in the event (that), just so (that), on condition (that), provided/ing (that), supposing (that), whether... or*, and the *-ever* conditional-concessive subordinators (e.g. *whoever, whatever*), and expressions such as *no matter wh-*. In formal English, a hypothetical present or future conditional can be signalled by an inversion of the subject and operator (*Had I known...*), and subjunctive *were* and tentative *should* (*Were you to..., Should he call,*). In informal English, the coordinating conjunction *and* can also signal a condition, as in *Do that again and I'll tell!* In the spoken section of the Corpus of Contemporary American English, currently containing 90 million words, *if*-clauses are by far the most frequent, with approximate distributions as shown in Table 34. This distribution is only approximate, since searches of this kind include non-conditionals (e.g. *I don't know if I'm going to hire him*) and accurately distinguishing between past and perfect uses of *had*, for instance, is challenging. However, the relative frequencies of the conditional types (zero, first, second, and third conditionals, respectively) have been calculated here for purposes of comparison with learner production, not in order to get an accurate picture of NS use, which is beyond the scope of this research. The traditional nomenclature is used here for convenience only (c.f. Sinclair 2005: 350), and it is acknowledged that these types exclude other legitimate conditional structures.

Table 34: *Distribution of if-clauses in the Corpus of Contemporary American English*

	Type	Example	N	Number per 1000 words	Percentage of all if- clauses
0	if + verb base form/3 <sup>rd</sup> person, present VP in main clause	<i>But sure, if you want to call it a cult, that's fine.</i>	150,950	1.68	58%
1 <sup>st</sup>	if + verb base form/3 <sup>rd</sup> person, future VP in main clause	<i>If we let them go, they'll just keep trying to get us.</i>	25,796	0.29	10%
2 <sup>nd</sup>	if + verb past	<i>At least if you knew, you might be able to do something about it, right?</i>	63,887	0.71	25%
3 <sup>rd</sup>	if + verb past participle	<i>If you had been there you would not have noticed.</i>	18,785	0.21	7%

The Small Talk database was searched to identify *if*-clauses, and these were examined to eliminate reported questions containing *if*. This yielded 2,131 utterances of the conditional type. However, at lower proficiency levels, utterances which have a clear conditional meaning often do not contain a subordinating conjunction of any kind:

*For example, I with Fahad talk anything, you cannot interrupt.*

*Someone have power, graduate is very easy.*

The conditional intent of these utterances is often only identifiable in the communicative context, and teachers generally signal this by writing ‘conditional’, ‘hypothetical’, ‘if’, and so on in the Context field (see Figure 17). With these included, the count of conditional sentences rose to 2,192, with distribution across the levels as shown in Figure 19. All of these utterances, and any others that included subordinating conjunctions that indicated a conditional clause, were identified by means of the ‘Semantics’ fields (Figure 19) so as to permit analysis of both accurate and inaccurate uses.

At first glance, the comparison between the Small Talk data and the *COCA* data seems to suggest heavy overuse of *if* conditionals: the Small Talk data show a total across all levels

of 7.41 per 1000 words, while *COCA* shows 2.88 per 1000 words. However, it should be noted that the *COCA* data do not contain transcripts of informal conversations except where

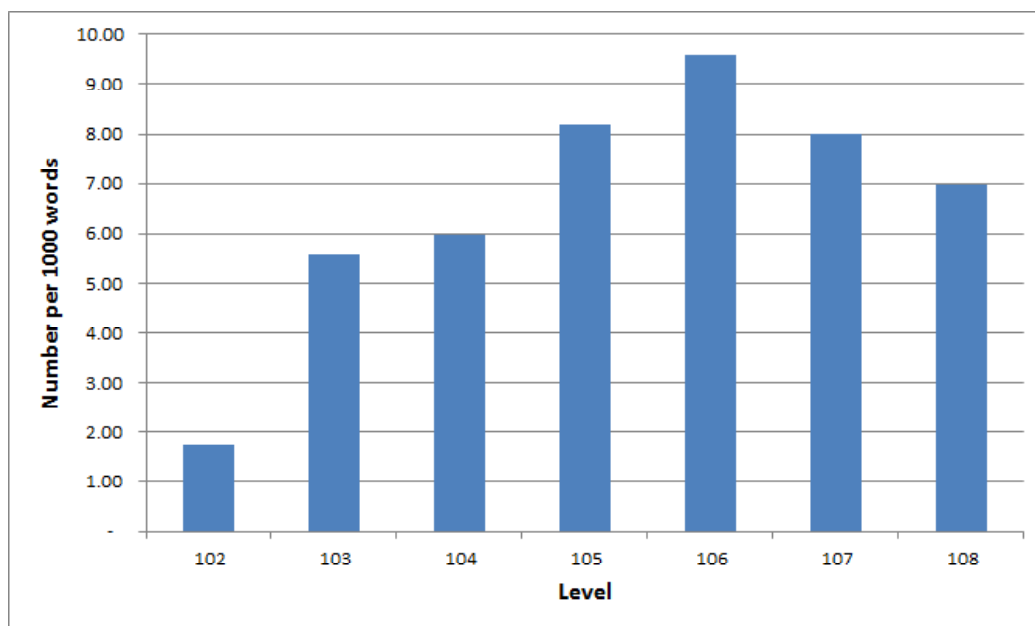


Figure 19: Distribution of conditional sentences in Small Talk database, by level

these occur in the context of radio and TV broadcasts. Furthermore, conditional structures are fairly complex and learners often produce non-target-like forms, which would mean that they would appear more frequently in the error database. The following sections will therefore explore the kinds of use and misuse of the structures evident in the data, with a view to establishing pedagogical recommendations.

### 6.7.2 Target and non-target uses of conditionals

In the analysis of conditional-like structures, as with any structure, it is important to establish criteria by which to judge correctness. For instance the sentence *If both you and your husband have job, you will share [the housework]?* could be considered inaccurate for several reasons: the lack of determiner (or plural) for *job*, the intransitive use of *share*, and possibly the lack of subject-verb inversion in the question. These errors need to be brought to the attention of the speaker, but it is not the conditional structure itself which is erroneous. As a ‘first’

conditional, it contains the requisite elements: the subordinating conjunction, a present tense with future meaning, and a future tense in the independent clause. Therefore, in this analysis, the structure is considered to be target-like, whatever other elements may be erroneous. Furthermore, if there is any indication that the VP in the subordinate clause is ‘present-like’, the conditional structure is considered successful, as in *If I smoking, you will get my smoking*. In contrast, if the VPs in the two clauses are mismatched, for example in *If you were woman, do you want to makeup?* the structure is considered unsuccessful.

A more controversial decision in the analysis concerns the type of conditional structure that learners should be using, and whether they should be using a conditional structure at all. For instance, in the context of a conversation on the topic, ‘If I had a million dollars’, when a learner says *If I get one million dollar, I can do many things in my country*, should this be judged inaccurate because it does not acknowledge the hypothetical nature of the condition? Generally, teachers seem to think that it should, depending on the level of the student and other factors such as whether a proficient speaker might conceivably not use a second conditional in these circumstances. In the following examples, the teacher reformulations or comments indicated that they felt a second conditional was called for:

*If somebody kill your family, what will you do?*

*If I have a baby, I will do abortion.*

*If you have any chance to go to hospital to take care of old people, will you go?*

*Do you want to be willing to disclose to other people if you get [AIDS]?*

In other cases, a conditional structure does not seem appropriate at all, *when* being more likely:

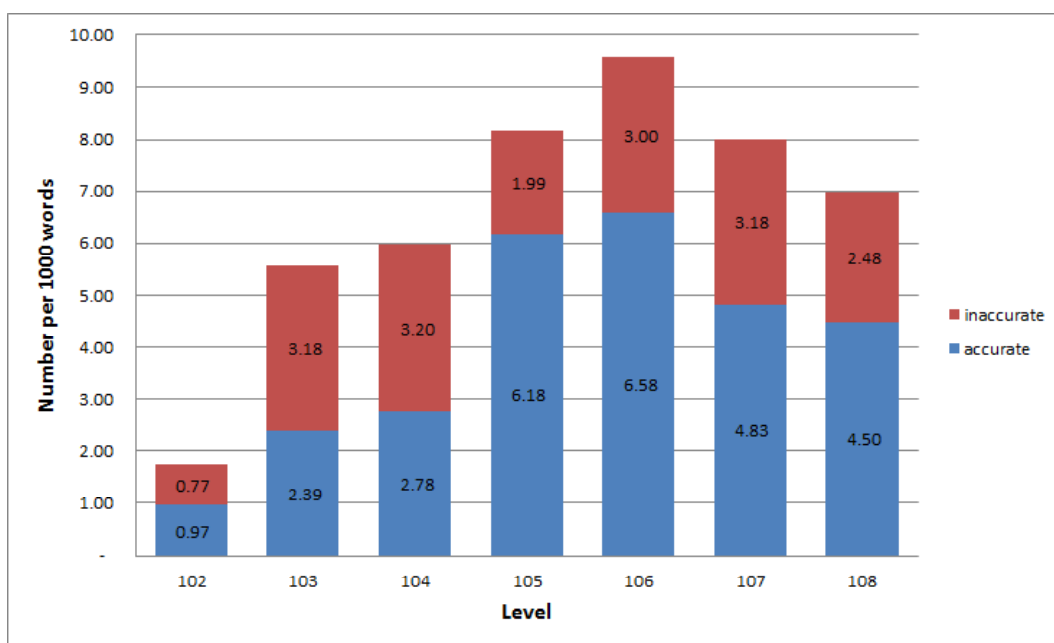
*If I get my children, I want to be like my mom.*

*If Spring will come, it's my most exciting things [time].*

*Now I usually listen to the pop or rock, but if I'm 27 or 28, I will listen to another kind of music.*

These decisions are arguably too subjective to yield an accurate picture of the use of these structures. Nevertheless, the business of CF *is* subjective and intuitive, and while some critics of the practice see this as a reason to abandon it completely (Krashen 1994; Truscott 1999), teachers and more importantly students do not generally agree (see, for example, Katayama, 2007; Schulz 2001). Furthermore, experience with the Small Talk CF methodology seems to indicate a tacit understanding on the part of the students that if a teacher reformulates an utterance in a certain way, and the basic meaning is preserved, then that is a reasonable target.

Once the inaccurate uses have been identified, the distributions of accurate and inaccurate uses of conditional structures can be mapped across proficiency levels, as shown in Figure 20.



*Figure 20: Distribution of accurate and inaccurate conditional structures, by level*

This is still too coarse-grained a picture to be of much value, but it does imply stages of greater and lesser control, for example between levels 104 and 105. It also shows that even in a database of errors, there is much that is successful. But to be truly informative, much more detail about the nature of learner attempts, as well as successes and failures, is required. For



instance, do learners have an awareness of functions of the various conditional forms? Are they more successful in their attempts at certain types of conditional sentences? What are the most commonplace errors they make in their attempts?

To address the first question, all attempts (correct or otherwise) that could be readily identified as one of the four conditional types were analysed by level (Figure 21; the lowest levels, 099 and 100, did not have sufficient data to merit inclusion). As can be seen, on average the learners have a tendency to overuse the zero and first conditional and underuse the second and third by comparison to the NS speaker usage (from *COCA*). However, it must

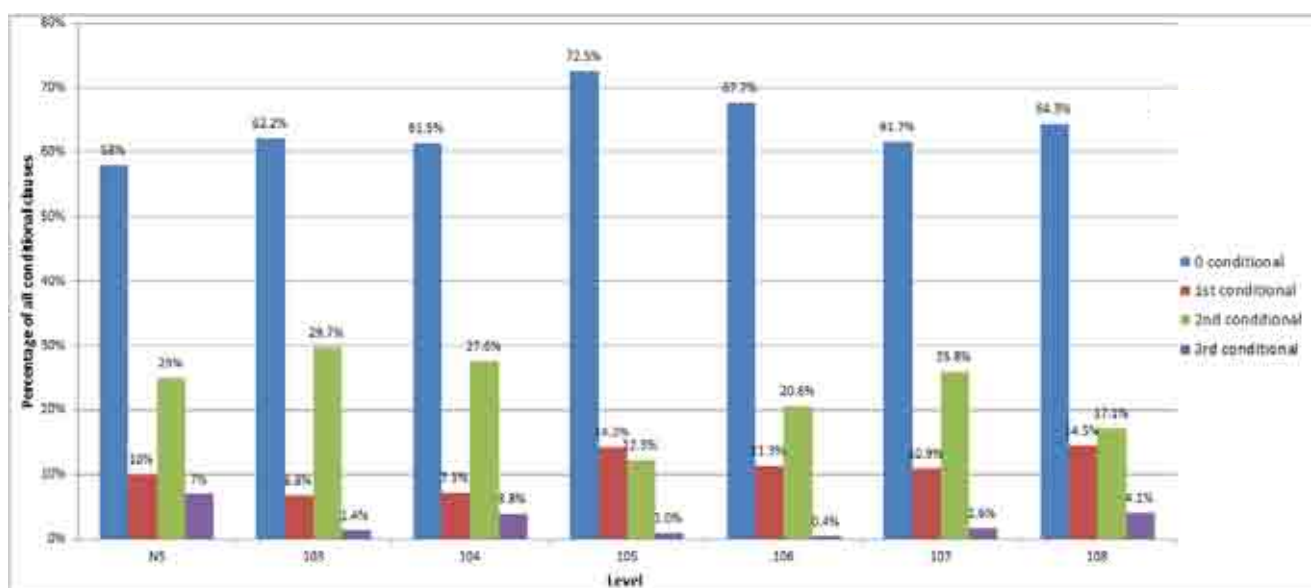


Figure 21: Distribution of all conditional clauses (accurate and inaccurate) by level, compared with NS production

be remembered that the *inaccurate* attempts at, say, a second conditional reflect teachers' interpretations of the intended meaning, which is to say the teachers' intuitions about what the target form would likely be. For instance, the sentence *If somebody kill your family, what will you do?* demonstrates accurate use of the first conditional structure, but was counted as an error because it would likely be spoken as a second conditional by a proficient speaker in the context in which it was spoken. This is a clear example of the comparative fallacy (Bley-

Vroman 1983), that is, the fallacy of analysing interlanguage forms according to target norms; but it also highlights both the pedagogical irrelevance of the comparative fallacy and its theoretical weakness: if we assume that the intention is to discuss a hypothetical situation, then the fact that the learner can produce the (arguably) communicatively adequate first conditional is irrelevant; she still needs to know that she has failed to signal the hypothetical nature of the question.

Furthermore, in terms of the systematicity of her IL (which is the focus of Bley-Vroman's concerns) we cannot be sure from this example whether she is aware that hypothetical conditionals are expressed using a form which is distinct from future conditionals (and simply didn't produce it) or whether she is assuming they are identical; and without such knowledge, we are no closer to an accurate understanding of her systematic IL. Thus the weakness of the comparative fallacy construct lies in attempting to find systematicity without knowledge of intention, knowledge which can easily be furnished by processes such as the CF methodology described here. For instance, by examining her production of conditional structures (all produced within a two-month period while she was in levels 107–8), we can confirm that the speaker intended to speak hypothetically but may have been unaware of the different form:

*If I divorce, and my children want to meet my husband, I let my children meet him.*

*If somebody kill your family, what will you do?*

*If execution doesn't exist, I think the criminal increase.*

*I'm suffering if my mother or my father be in hospital a long time.*

*If one of my family is euthanasia, I agree.*

*If your friend sing in karaoke bar, do you want to listen?*

With the possible exception of the first and last sentences, all of these would be expressed as hypothetical present/future conditions by a proficient speaker of English. Since the speaker

here is a Japanese learner at high-intermediate proficiency, it is extremely unlikely that she has never encountered a hypothetical conditional in English; in contrast, it seems likely that she does not know that the form refers to the hypothetical present/future. She might, for instance, believe that the sentence *If somebody killed your family, what would you do?* refers to a *past* hypothetical situation; many learners of English do assume this (see, for example, Norris 2003), which would explain not only their relatively infrequent use of the second conditional – recall that NS use of past hypothetical *if*-clauses accounts for only 7% of the data in Table 34 – but also the complete absence of attempts to produce third conditionals (see below).

Turning to the structures that learners at each level successfully produce (Figure 22), it can be seen that their ability to do so is overwhelmingly attributable to their control of the

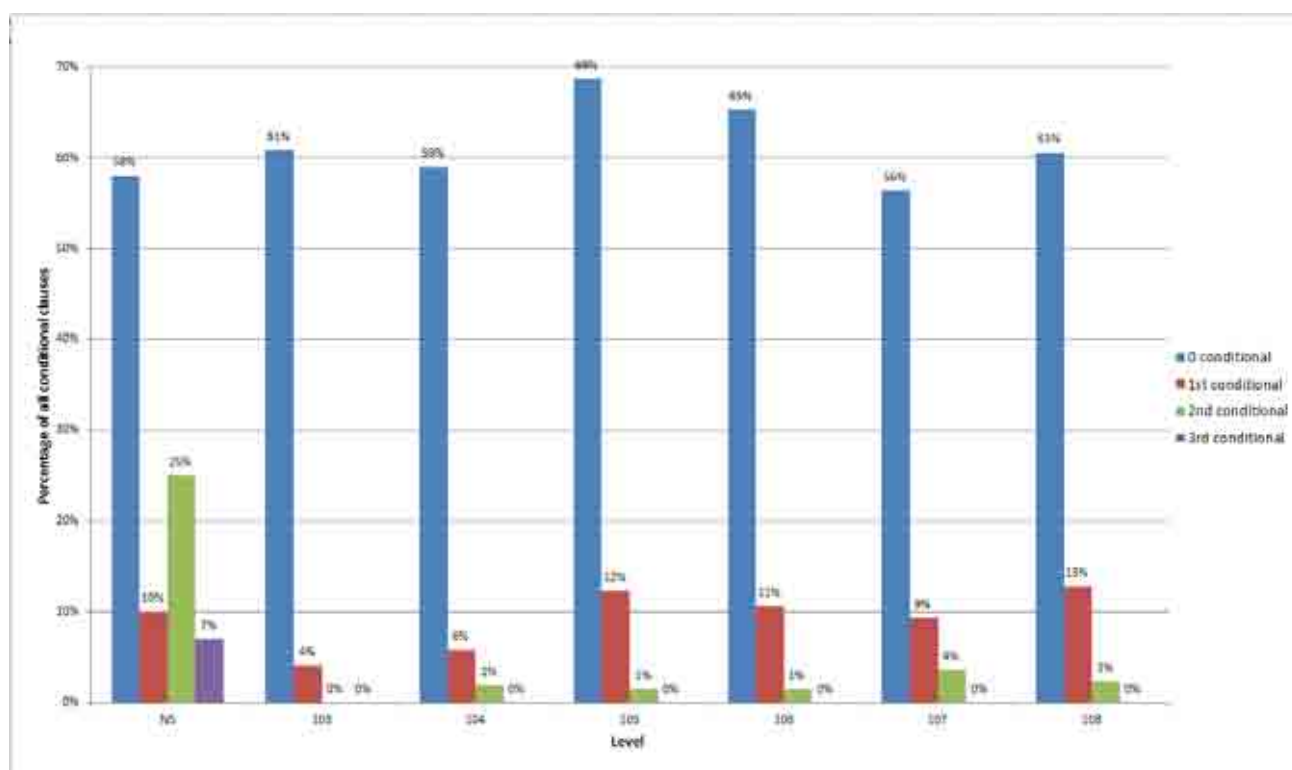


Figure 22: Accurate use of conditional clauses by level, compared with NS production

zero conditional (e.g. *If your girlfriend says about her problem, do you feel you have to solve*

*that problem?*). In contrast, first and second conditionals account for a much smaller proportion of their correct use, and no example of the third conditional is present in the data. In this sense, then, the learners are under-using second and third conditionals by comparison with NS data, while their use of zero and first conditionals is approximately target-like in frequency. This kind of evidence is precisely the argument for CF and form-focused teaching in general: those who argue that input alone will result in target-like production fail to see that learners need to be convinced, through CF and other methods, that the language they are using, while accurate in some contexts, is not necessarily target-like in others. This can clearly be seen in the analysis of inaccurate production (Figure 23), in which errors in or avoidance of the second conditional becomes very apparent. Here an interesting feature of IL development emerges: at the intermediate level (105), there is a marked decrease in attested errors of this type, which could be a statistical anomaly, caused by the data collection methodology (but see Figure 20, in which overall conditional use for this level is above average), but may plausibly

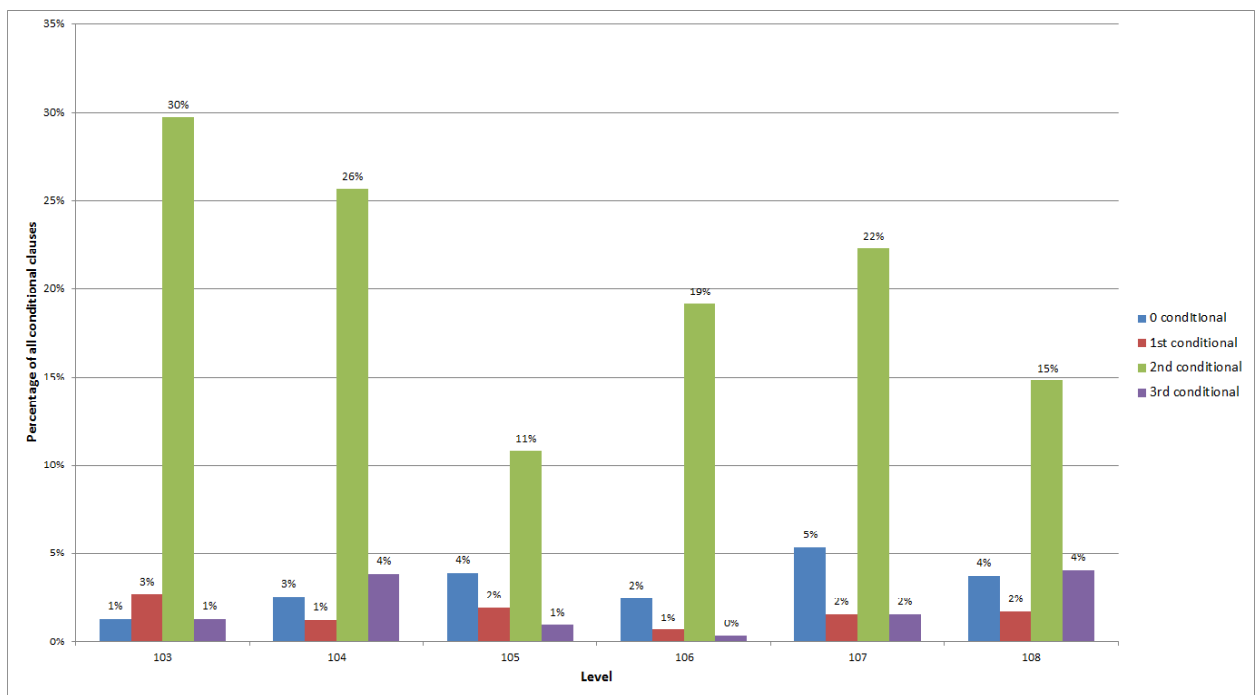


Figure 23: Conditional clause errors, by level

be associated with a greater reluctance by learners at this level to take risks, which might itself be a contributing factor in the ‘intermediate plateau’ (Cullen 2008). This is entirely speculative, however, and more research would be needed to investigate the possibility.

A further finding that is worthy of comment is the similarity in the distribution of error types across L1 groups. For this analysis, six types of error were isolated: first, second, and third conditional forms for both the subordinate (sub) and main clauses (Figure 24), grouped by L1 background regardless of proficiency level. Errors in zero conditional, few as they are, were ignored. The striking similarity in distribution across L1 groups, especially Spanish, which has the greatest similarity to English as concerns these forms, speaks to the particular challenge that learners of all backgrounds face in control of VP tense in the subordinate clause of first conditionals (namely, not using a future VP), and particularly both the subordinate and main clause VPs in second conditionals, which constitute the overwhelming majority of errors. Whether this can be explained by the markedness of the structure, the counterintuitive nature of the form–function relationship, or perhaps a particular difficulty in resetting a parameter of Universal Grammar is beyond the scope of this research; nevertheless, the particular challenge offered by this structure argues for a more systematic focus in the syllabus on this particular form. Thus, while the CF provided may have helped to fine-tune the individual learner’s IL, the aggregate data from the error database can additionally contribute to pedagogical decisions for all learners, a point which will be taken up in section 6.8.

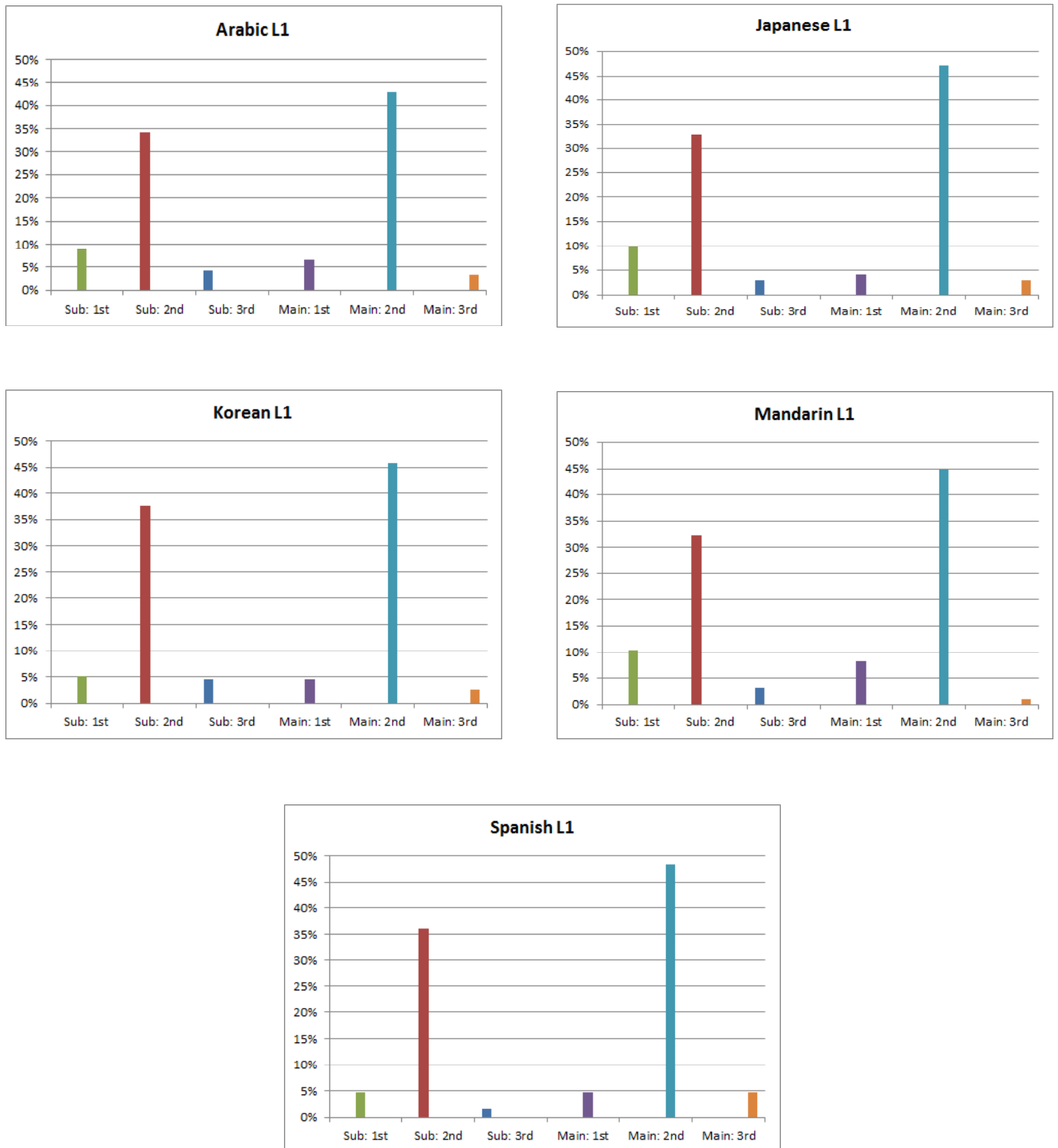


Figure 24: Conditional clause errors, by L1 background

### 6.7.3 Conjunction errors in conditional clauses

A significant proportion of learner errors with conditionals have to do with the selection (or omission) of conjunctions and the coordination of adverb clauses. The analysis yielded six error types, which highlights the inadequacy of error analyses that use such terms as ‘omission’, ‘substitution’, ‘syntax’, ‘morphology’, and so forth (e.g. Burt and Kiparsky 1972; Guntermann 1978; James 1998; Truscott 2001). As

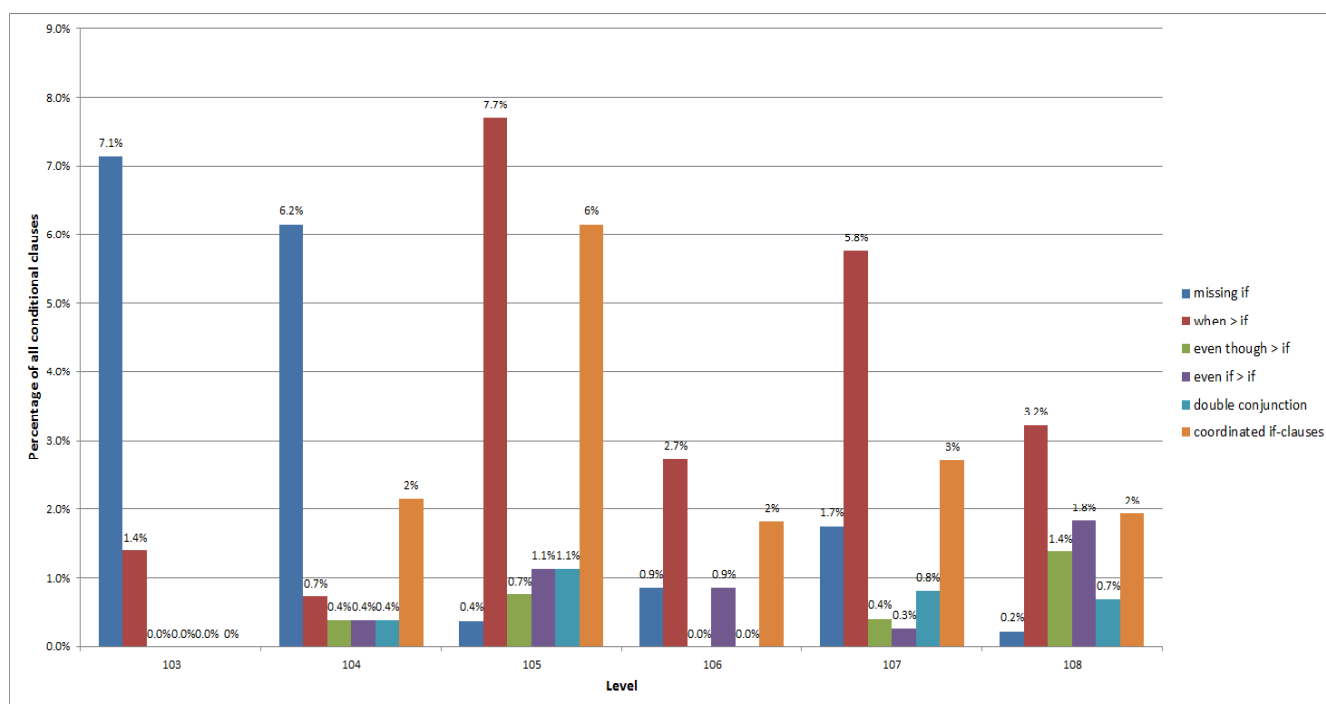


Figure 25: Subordinating conjunction errors in conditional clauses, by level

can be seen (Figure 25), the three ‘substitutions’ (when/even though/even if → if) are very differently distributed across the levels, and an error classification system that looked only at ‘substitutions’ or ‘lexical errors’ would completely miss the mark: these *particular* substitutions are worthy of separate identification and therefore constitute distinct error categories in the present analysis. This section will therefore briefly exemplify and comment on these error types in turn.

Missing conjunction

*I go back million years, I want to see the life.*

*She collect some fruit, she have many babies.*

This type of error is most prevalent at lower levels, where learners simply juxtapose independent clauses and rely on context to communicate the relationship (temporal, conditional, etc.) between them. CF at this level focuses more on establishing this relationship by means of the appropriate conjunction, so although the reformulations do specify the concomitant VP forms, it is unlikely that learners would be able to reproduce them consistently, even on a Running List test.

if in place of when

*If I get my children, I want to be like my mom.*

*If a person become adult, the parents don't care [care for them], just give suggestion.*

Particularly evident in the production of intermediate learners, this error is quite characteristic of those with L1 Japanese or Korean, presumably because in both languages the same VP form communicates both a temporal and a conditional relationship. As noted above, the choice of *when* or *if* is not always clear in English.

if in place of even if

*If the information hasn't proven, people believe it.*

*If they don't live each other, they can pregnant!*

(even) if in place of even though

*He asked me ten dollars, even if the meter said five.*

*She called him every day again, even if they broke up.*

The difference between concession and condition is sometimes very nuanced (particularly with present tense VPs: see Quirk et al. 1985: 1098–9). In the first set of examples,



concession would be more appropriate than strict condition; but in the case of past events, *even if* and *even though* communicate information about the telicity or recurrence of the action: the second set of examples above, as stated, both imply an unintended recurrence. The sophistication of the distinction between the conjunctions means that these errors are rare, and occur mostly at the advanced levels.

#### Double conjunction

*Even if we thought that we are not receiving some benefits, but we are part of the United States now, so we have to pay, as an American people.*

*If I don't have this [Valentine's] day, and then I really respect them, but in my daily life, I can't say thank you.*

In these examples, both subordinate and main clauses are marked with conjunctions, which is particularly common among speakers of Asian L1s (the combination *even though... but* is also common), in which the second conjunction is obligatory. Students have also reported that the subordinating conjunction does not feel adequate, especially in longer clauses.

#### Coordinated clause

*If you walk the streets, you see the snake across the street, you can't hit the snake.*

*If I get 29 or 30, I didn't have girlfriend, can you get married with me?*

A more sophisticated version of the 'missing conjunction' error, these involve a second subordinate clause which is assumed to be 'covered' by the subordinating conjunction. Most prevalent at the intermediate level, this type of error becomes less frequent as learners become comfortable handling multiple clauses.

Although none of these errors is especially frequent in the data compared to VP errors in these clauses, this is another area of possible avoidance like the underuse of second and third conditional forms (Figure 22). There are, of course, competing explanations: students might simply make few errors of this type, teachers might not be present to hear many of the errors they do make, or teachers might simplify the worksheet items to focus on erroneous

portions only. The first alternative is unlikely, given the types of errors which are plentiful in the data. Secondly, while teachers are certainly not present at every conversation, the random sampling methodology (see Module II) should mean that the aggregate data are generally representative of the production of these learners. As the project continues and more data are added, it should be possible to draw firmer conclusions on this point. Finally, an analysis of mean utterance length in the database shows a consistent increase with increasing proficiency ( $r = .977, p < .0001$ ), as shown in Table 35.

Table 35: *Mean utterance length in database, by level*

<b>Level</b>	<b>Mean utterance length (words)</b>
099	3.43
100	4.73
101	5.15
102	5.61
103	6.79
104	7.15
105	7.34
106	7.66
107	8.41
108	8.46

The extremely high correlation between level and word count seems to imply that learners produce consistently longer utterances as their fluency and control increases (as one would expect), and that teachers' collection of data faithfully reflects this. The alternative explanation is that teachers consistently simplify and shorten utterances as they note them down, in close correlation to the level of the students. This seems very unlikely and is not supported by the findings in Module II. We can therefore conclude that avoidance is a real possibility, and that methods that raise learner awareness of both problematic and avoided forms are necessary, as the following section will explore.

## 6.8 Pedagogical implications

The analysis of conditional clauses described above represents only a fraction of the possible pedagogical foci in the language acquisition process. In many of the examples presented, there are other types of systematic errors that have not been discussed here, but which could be brought to the attention of the learner – and which in fact were, as reformulations in the CF process. The assumption made throughout this research is that such reformulations can be thought of as formulaic ‘exemplars’ (Skehan 1998), items which, as target-like versions of the learner’s own interlanguage intentions, have greater consciousness-raising and retention potential than the arbitrary syllabus and input of the course book. In addition to these exemplars, however, an annotated error database can potentially contribute to the learner’s understanding of the systems underlying the structures, providing a form of ‘data-driven learning’ (Johns 1991).

To be useful in this process, the data must be annotated in a variety of ways, as described above, and must be easily accessible to teachers. In the Small Talk database, data queries have been facilitated by means of the ‘Focused Worksheet Maker’ form (Figure 26), which allows the user to select combinations of grammatical ‘tags’, isolating specific semantic functions, L1 backgrounds, proficiency levels, and so forth. In the example below, items with subordinate clause errors of the second conditional type have been selected. The results show the utterance, context, topic, speaker’s L1 and level, each of which can be filtered out if desired. Thus it is possible to produce a list of items very specifically tailored to individual learners or groups of learners, for instance Arabic speakers at a low-intermediate level.

The results can be printed out as they are or copied to a word-processing program for further manipulation. In the case of smaller, discrete forms such as indefinite article use with

count nouns, the results can be used as they are for noticing/correction activities. For errors with a larger extent, such as conditional clauses, teachers can select a number of prototypical

**Focused Worksheet Maker**

SYNTAX: CLAV AND AND IN LEVEL OR

CLAV Substituted: present/future (2nd) condition

SEMANTICS OR

LOOK FOR A WORD OR PHRASE

SMALL TALK TOPIC

SEARCH CRITERIA: CLAV: present/future (2nd) condition unreal

RESULTS (144) PRINT THIS WORKSHEET COPY RESULTS TO CLIPBOARD

(Double-click on an Expression to see details)

Expression	Context/vocabulary	Topic	First Language	Law
If she's alive now, she's supposed to be 20 years old.	She died in the accident.	Medical Mistakes	Arabic	108
If smoke in front of my father he will kill me.		New Year	Arabic	104
If somebody have a car accident, you allow him to do a plastic surgery?		Plastic Surgery	Arabic	107
If somebody kill your family, what will you do?		Crime	Japanese	107
If someone do something bad, I would want them do that for me.		Apologizing	Arabic	106
If someone need (one of your organs), do you want to give him?	(hypothetical)	organ donation	Arabic	107
If the baby is born and they knew the baby have a mother but they took me		abortion	Korean	107
If the good deed is very common, it will limit the crime and make the work		good deeds	Arabic	105
If the government allow euthanasia, maybe when he turn old, he will this c		Euthanasia	Mandarin (Chinese)	107
If the government make rules like the drinking age...	limiting internet use by children	connecting: telephor	Arabic	106
If the law don't protect your rights--	human rights; civil rights; the ri	Getting Married	Spanish	104
If there is more money, what you want to do with more money.		\$1,000,000	Arabic	104
If there is no border or passport, the world will be as one country.		4-3-2	Arabic	106
If these kind of illness is in your body, what would you like to do?	Topic: cancer, terminal illness	terminal illness	Spanish	107
If they didn't get money, they will be punished.	gangs that make children beg	homelessness	Mandarin (Chinese)	107
If they do racing together and they have accident they shoot each other.		My hometown	Arabic	104
If they don't want, what would you do?		forceful parents	Korean	106
If they don't allow smoking, you will go outside and think "Why am I gambli	(unreal?)	gambling	Arabic	108
If they get divorced, all the business will destroy.	Sometimes, love comes after m	in Korea, how can yo	Arabic	108
If they love each other, they got married.	two orphans, who are brother a	abortion	Mandarin (Chinese)	107
If they will say that the doctor don't want more--	process, procedure	Medical Mistakes	Portuguese	108
If this movie is continue, if Gail's come, Don't know	Entertainment: Paul	Person: Paul	Mandarin (Chinese)	107

Records: 144 1 of 204 Filtered Search

Figure 26: The 'Focused Worksheet Maker' form in the Small Talk database

errors of a certain type (say, second conditionals where the hypothetical premise is very obvious, as in *If execution doesn't exist, I think the criminal increase.*) and correct half of them. Additionally, a teacher may choose to correct any other errors that do not relate to the target structure. The results can be used in class as a noticing/correction task, in games and warm-ups, and so forth. Alternatively, the selected set of items can be used in a grammaticality judgement task, presented in written or oral format, as described in Chapter 3. These kinds of activities can be very helpful in raising learner awareness of challenging

forms, are easily produced, and have the advantage of being both authentic and graded to the learner's proficiency (see Appendix 7 for an example activity). The availability of this kind of activity addresses the second research question, *What are the pedagogical applications of such a database?*

A second, hitherto less explored use of the error database involves tracking the production of specific language features by individual learners and groups of learners. For instance, of the 713 individual learners whose data are represented in the database, 528 (74%) have utterances which were tagged for conditional clauses of at least one type. Many of these students remained in the ESL program for several months and up to a year and a half in some cases. By selecting only students who remained in the program for at least two semesters (seven months), it is possible to ascertain when the forms begin to appear and whether any change occurred in their production during this time. Eight students fit these criteria, and the data from one are presented in Table 36, with likely second conditional sentences shown in bold.

Table 36: *Conditional clause sentences spoken by 'falsoaimy', by level*

Level	Utterance
<b>103</b>	<b>Maybe this mean is, if I forgot my key?</b>
104	If homeworks it's free time,--
104	If you have some homeworks, it's not free time.
<b>105</b>	<b>If I learn, I will learn only Spanish or Italian.</b>
105	Saudi Arabia, if you mean to use body language is very rude.
<b>106</b>	<b>If I wear and hang out I'm going to be weird.</b>
107	For example, when a wife die, her husband will get her pension, even if she is die.
107	If he work for the government, he has to retire when he get 60 or 65.
<b>107</b>	<b>If he's in the sea and he needs help, I'm gonna (risk my life for him).</b>
108	For me, if she wanna study and work, it's fine.
108	If my father is not an educated person, he might do something to his daughter.
<b>108</b>	<b>If you didn't pass an exam, he would go to the professor and explain to him.</b>

It should be noted that, unlike in the case of aggregate data as argued above, the data

collection methodology is very unlikely to be representative of the production of any individual, and that all of these items were collected because they contain errors of some kind. Logically, it is possible that this learner did, in fact, produce correct second conditional sentences which were not collected. Nonetheless, two observations can be made about this sample: first, the speaker made attempts to produce the form throughout his study in the ESL program. This is worth mentioning because the published teaching materials to which he would have been exposed during this time do not generally discuss the second conditional until about the intermediate, 104–5 level. For example, the *Touchstone* series (McCarthy et al. 2005) introduces the structure in Unit 8 of Book Three, intended for intermediate learners. The evidence from the Small Talk database, however, shows that learners want to speak about hypothetical situations long before this level (in fact the database shows Small Talk topics, all chosen by students themselves, that demand second conditional forms, such as *If you could be an animal, what animal would you be?*, starting at the 102 level), and while it could be argued that they simply do not have the control to produce such language prior to the intermediate level, it could also be argued that they can and should approach the problem formulaically, building up a stock of second conditional ‘exemplars’ through the process of attempting to communicate their ideas and receiving target-like reformulations.

Related to this point is the fact that most published teaching materials, as well as grammatical syllabi developed by language programs, generally present structures only once. This is to be expected, since there is much ground to cover and some form of organization is needed to allow ease of reference. But the likely result of this is that teachers simply work through the textbook or syllabus one structure at a time, regardless of when the students happen to be attempting those structures. As the data in Table 36 show, a recursive approach to the syllabus would match the learner’s internal syllabus, but such a syllabus would be so

specific to individual learners, or possibly cohorts of learners, that it would be impossible without the systematicity afforded by a tracking system such as the Small Talk database. With such a system, however, there is a much greater chance of approaching Brumfit's vision of 'allowing people to operate as effectively as they could, and attempting to mould what they produced in the desired direction, rather than explicitly teaching and expecting convergent imitation' (Brumfit 1984: 50). In this sense, the syllabus becomes reactive, guided by the actual production of learners and by the reformulations of teachers. There is no reason to believe that the same issues of avoidance and coverage of formal language features would not apply, but there is no question that the relevance of the syllabus to the learner would greatly increase.

Even were one not to abandon the prescriptive or *a priori* syllabus (Johnson and Johnson 2011) completely, an error database could still be a valuable guide in tailoring the syllabus to one's own students. If the composition of a class is mostly Arabic speakers, for example, the kinds of structures, lexical items, discourse and pragmatic skills, and so forth that require repeated or concentrated instruction will differ somewhat from those required by a class of Korean speakers, yet an *a priori* syllabus rarely specifies the weight that various elements should be given. The evidence from this analysis would suggest that in the case of conditional structures, little to no instructional effort is needed with the zero conditional at any level; the first conditional could be introduced and periodically reviewed at the lower levels; the second and third conditionals should be introduced as soon as students attempt to make hypothetical statements, even if only as 'All Do' items in the CF process. In this way, students are more likely to notice the structure (or misuse of it) in the input around them, including in their own and peers' production (Thornbury 1997).

## 6.9 Discussion

This chapter has described a database-driven approach to CF and syllabus design, by arguing for the utility of an error database such as the Small Talk database, which can be employed to systematize the collection and annotation of learner production. Through the examination of conditional structures, the analysis has shed light on both the distribution and frequency of learner attempts to use these forms, as well as the nature of the errors they make in doing so. Space has not permitted a more thorough inventory of linguistic forms, and so the focus on conditional structure errors should not give the impression that these errors are more frequent than other types. In fact, adverb clause and tense sequence errors, of which conditional clauses are a small subsection, combined account for less than 8% of the errors encountered. However, the analysis has shown the process by which the errors are annotated, and through this, how the ‘tagset’ is being developed. Currently, this taxonomy (see Appendix 8) contains 32 ‘error domains’ (such as ‘Clav’) and over 180 error types (such as ‘when > if’). It is not known at this point whether the taxonomy will grow as more data are analysed and existing error types fail to account for the data, or whether error types will be conflated as the analysis proceeds, and it is quite likely, given the exotic nature of learner production, that both will occur.

As has been noted, the assignment of error types can be quite subjective, and there is plenty of room for disagreement between annotators, not simply over the most appropriate classification of error type in any particular instance, but also over the plausible reformulations on which such analysis is performed. To a certain extent, the taxonomy thus represents a multiplicity of working hypotheses, some more controversial than others and some better supported by the evidence of the data than others, in the sense that either the data fit the interpretation better or there is simply more data available. In this sense, this research



and the pedagogical approach fall into the empirical methodological approach of corpus linguistics, and one categorized by

the common assumptions that linguistic theorizing should be driven first and foremost by (representative samples of) authentic language data, and that a solid linguistic hypothesis and theoretical claims should be based on a thorough description of these data with regard to the phenomenon under investigation (Römer and Wulff 2010: 100).

The database has been designed to allow for multiple analyses, and annotators periodically meet to review the more controversial areas and to attempt to reach consensus. When this occurs, the reassignment of items to another, or several other error types is a simple matter, even with hundreds or thousands of items.

In areas where there is greater consensus and confidence in the hypotheses, the error database provides a valuable way for teachers, and especially trainee teachers, to check their intuitions about (inter)language use, intuitions which greatly influence instructional approaches. For instance, the data on individual and aggregate use of conditionals presented in this chapter indicate that while zero and first conditionals are early acquired, second and third conditionals (and combinations of all of these) are attempted at low proficiency levels and probably require repeated instructional focus. Furthermore, L1 background does not seem to affect error rates on these forms (in contrast to, say, the acquisition of NP determination, which manifests clear L1 background effects in the data). This addresses the first research question, *Can an error database provide a representative monitor of linguistic development in individuals and groups?* While space has not permitted a full description of the functionality of the Small Talk database, it should be clear that tools such as the ‘Focused Worksheet Maker’ provide an easily accessible reference of interlanguage forms which can be consulted to inform teachers in their efforts to find effective instructional approaches. The database thus represents a small step towards the systematic CF technology foreseen by

Hendrickson over 30 years ago:

The computer printout of these error clusters essentially would ‘map’ developmental phases through which the students had passed as their speaking and writing proficiency increased over time. The information provided by such cognitive mappings could serve as a basis for improving foreign language curricula... The massive amount of data systematically collected, recorded, and analyzed may reveal some useful discoveries about language learning universals. (1979: 364)

This ‘mapping’ and the pedagogical training and materials that have derived from it are one answer to the second research question, *What are the pedagogical applications of such a database?*

Currently the Small Talk database is housed on a shared network drive<sup>3</sup>, which permits access to faculty and teacher trainees in the English Language Center. The next stage in development will migrate the database to an online platform, permitting non-ELC teachers to enter and analyse their own students’ error production, whether collected in Small Talk or other communicative activities. This will increase the size and scope of the database beyond the limited set of languages currently represented. Whether or not others wish to contribute, the existing data alone is a source of authentic information on learner error for teachers and researchers. This is a much-needed resource, since even with the growing availability of learner corpora, SLA scholars still rely heavily on data collected by researchers as much as 40 years ago, when learner corpora were in no way comparable to those available today. R. Ellis et al. (2009) are typical in their reliance on Burt and Kiparsky (1972) and Dulay and Burt (1974b). The point is not that this data might be inaccurate, simply that it is hard to believe that it can adequately inform SLA research which makes claims such as the following:

First and foremost, an attempt was made to select target language structures that were *known to be universally problematic to learners* (i.e. to result in errors). To this end, the SLA literature on error analysis was consulted (e.g. Burt & Kiparsky, 1972). (R. Ellis et al. 2009: 42, emphasis added)

---

<sup>3</sup> The database application (but not the utterances or student data) is available for download at [http://www.gonzaga.edu/Academics/International-Students/TESOL\\_programs\\_research.asp](http://www.gonzaga.edu/Academics/International-Students/TESOL_programs_research.asp)

Even more common is reference to the ‘morpheme studies’ of the 70s, which to this day have not been replicated with learner corpus data (two exceptions being McEnery et al. 2006: 247-263 and Izumi and Ishihara 2004, both of which look at Japanese learner data only):

[r]egular past tense *-ed* is typically introduced in elementary and lower intermediate textbooks, but it is not among the morphemes acquired early (Dulay & Burt, 1974; Makino, 1980). (R. Ellis et al. 2006: 351)

As the studies cited in section 6.3 show, (Green 2006; Wagner et al. 2009; Harrat 2011), there is a real need for attested learner error data, and it is hoped that the Small Talk database will fill this need.

---

## CHAPTER 7: CONCLUSION – IMPLICATIONS AND RECOMMENDATIONS

---

### 7.1 Overview of the research

Chapter 1 began by questioning the contribution of SLA and Applied Linguistics to language teaching and by highlighting one of the more intractable challenges to language acquisition research and pedagogy, that of discovering what the individual learner knows. The research has documented three approaches to the investigation of this question, with the primary pragmatic goal of validating a hitherto little-explored approach to the provision of corrective feedback (CF) in second-language oral production, but with the equally important goal of testing theoretical assumptions and hypotheses concerning CF and the role of negative evidence. It has shown, I hope convincingly, that delayed CF of this kind is effective in developing complexity, accuracy, and fluency in adult second-language learners, and that the methodologies and technologies described are well within reach of most practitioners.

### 7.2 Summary of principal empirical findings

The pilot study in Chapter 3 established that teachers and proficient non-native speakers were able to identify well-formedness in learner language with very high reliability. This finding addresses a weakness of the research in Module II, which sought to identify the consensus with which different teachers identify CF for the same learners. The fact that they did not always do so could have been attributable to issues of perception, in other words a lack of recognition of well-formedness. However, the inter-rater reliability finding in Section 3.5.1 of .996 would indicate that perception of error was likely not at issue. Instead, time pressure and

pedagogical decisions were more likely to have caused the different teacher responses on the task.

Perhaps the most encouraging finding from the research is that delayed CF is effective, since participants are able to reformulate their errors with approximately 80% accuracy, which is much more promising than the typical rates of uptake and even repair in immediate CF (Section 4.9). This finding, like any in research on uptake and repair, is open to challenges, primary among which would be the charge that correct reformulation and correct fluent use are not at all the same thing. The investigation in Chapter 4 accounted for this by proposing measures of fluency as one indicator of acquisition and automatization, and similarly the timed grammaticality judgement test (TGJT) in Chapter 5 sought to establish degree of acquisition by using reaction time as a measure of automatization. Interestingly, while there appears to be a strong correlation between participants' accuracy on these two measures, there seems to be little association between the fluency with which participants produced utterances and their reaction time in recognizing the well-formedness of their utterances. This was an unexpected finding which deserves closer scrutiny in future research, but the tentative conclusion at this stage is that the psycholinguistic motivators for production and recognition are only tangentially-related performance variables.

Another encouraging finding, at least for the participants in these investigations, is that fossilization, or 'premature lexicalization' (Skehan 1998: 61), is not a major concern. As judged by the number of fluent reformulations and recognitions of erroneous forms, the learner data did not seem to indicate a significant presence of persistent or irremediable errors (Section 4.8.3). Having said that, analysis of the data seems to indicate that certain non-target-like forms sound right to learners, resulting in a 'response bias' in the TGJTs (Section 5.5.3). This argues strongly for the systematic focus on form and consciousness-raising provided by

delayed CF in order to prevent such fossilization. One element of the systematicity in the provision of CF afforded by the approach described is the practice of assigning learners peer errors to correct based on, for example, the perceived usefulness of the form or typicality of the error, and to compensate for avoidance. The analyses in Section 4.9 and Section 3.5.5 indicate that this practice does not affect performance on the tasks in any significant way, and there is therefore no reason to discontinue it. In contrast, in its present form the TGJT did not seem to contribute at all to the accuracy of subsequent reformulations (Section 5.5.8). While this task is not currently part of the CF methodology, the recommendation from this research is that more specific feedback should be incorporated into the future design of the TGJT in order to promote more accurate use.

The investigation of the error database in Chapter 6 resulted in several interesting findings. First, learners attempt conditional clauses at approximately the same rate as revealed by an investigation of *if*-clauses in a native speaker corpus. Second, while these attempts are often inaccurate, there appears to be a marked decrease in error rates at the intermediate level which is not explained by the sampling methodology (Section 6.7.2). This could be explained by a reluctance to take risks at this level. Further analysis of error rates on different types of structure will show whether this is a statistical anomaly in the data or a general trend, perhaps evidence of the existence of an ‘intermediate plateau’. Third, the production of errors in the 1<sup>st</sup>, 2<sup>nd</sup>, and 3<sup>rd</sup> conditional structures is remarkably consistent across five L1 backgrounds, suggesting that the problems posed by these structures have less to do with cross-linguistic influence and more to do with linguistic or cognitive factors such as form-function pairings, processing of increasingly complex verb phrases, and so forth. Finally, and relatedly, it was shown that mean utterance length seems to be very closely correlated with proficiency, which may have implications for learnability and acquisition orders (Section 6.7.3) and for the

systematicity in IL development hinted at by the implicational hierarchy described in Section 3.7.

### **7.3 Theoretical implications**

This investigation has primarily concerned itself with the pedagogical issues surrounding delayed CF, and it has therefore adopted a relatively neutral stance with regard to the more polarized theoretical positions in Second Language Acquisition. That said, the approach to CF described is based on the assumption that much of language learning involves the acquisition of formulae, patterns, routines, and constructions, and that fluent production necessitates the retrieval and contextual shaping of such prefabricated items, more than the ability to create novel utterances from small lexical units and internalized rules. Second language learners generally do not have the luxury of exposure to the quantity of language, nor the time to formulate and test hypotheses about its regularities and peculiarities, that native speakers enjoy, which accounts for the overuse of rule-based production, and concomitant lack of idiomaticity, in much learner language (Wray 2002; see also the differential responses by NS and NNS in Section 3.5.5). Nevertheless, one of the aims of language instruction should be to provide ways to make salient to learners the idiomatic, target-like versions of their attempts, and it has been argued here that delayed CF represents one possible approach.

The evidence for such attempts in this research has been the presence of non-target-like production which the learner recognizes as such, but which is not subsequently correctable. This indicates that Corder's distinction between 'errors' and 'mistakes' is not precise enough. Learners also produce language which is exploratory and this may be a significant catalyst in the acquisition process. Many teachers certainly think it is; hence their insistence that learners 'take risks' in their production (Hurd and Murphy 2005: 56). Learners may recognize that such production is ill-formed but not be able to correct it. Thus it is

neither a performance ‘slip’ nor competence ‘error’, but rather an ‘attempt (to mean)’.

Understanding this theoretical distinction would help teachers to be less anxious about student production of ill-formed language, instead seeing it as a fruitful avenue for exploration and instruction.

This view of language acquisition and instruction presupposes the means to identify and monitor individual language production and development, in other words the ‘highly individualized approach to cater to differences in developmental level among the students’ identified by R. Ellis (2008b: 3). As this investigation has demonstrated, this approach is not as unattainable as has been assumed to date. First, the Small Talk methodology (or any mechanism for creating genuine interaction) provides the opportunity for collection of production data. Second, the analysis of this data gives a preliminary view of the type and frequency of the ill-formed items. Third, the fluency of the elicited imitation and reformulation of the target-like versions, as measured by WPM – proposed as an acceptable measure of fluency in oral production (Section 4.9) – can be used to generate individual profiles against which to measure accuracy and automaticity of specific language items. Finally, the learner’s accuracy (and perhaps automaticity) in recognizing the well-formedness of these reformulations can be used to identify items more or less in need of further attention. This is an approach to the problem of determining what the individual learner knows that has both theoretical validity and proven ecological validity, in that it can be integrated into a variety of instructional settings.

#### **7.4 Technological developments**

This research has entailed the development of two technologies which may prove very useful to language research and teaching. The first is the web-accessible Flash-based TGJT platform (Section 3.4.5). It remains to be seen whether such a tool can contribute to language



acquisition itself but for psycholinguistic research, the low cost and wide compatibility of this technology should make such a platform very attractive for research employing reaction time and grammaticality judgement data. Indeed, there seems to be no reason for reaction times not to be standard data in investigations of implicit linguistic knowledge (Juffs 2001), especially since web accessibility can greatly expand the potential pool of participants at very low cost.

The second, and perhaps more exciting technology is the Small Talk database architecture (Chapter 6), which as we have seen can function as a simple tracking system for monitoring learner language, or with annotation as an error database for research and pedagogical applications. Only a fraction of the full functionality of this tool has been described here, but it is hoped that as more teachers and researchers become aware of its availability and utility, this technology and the approach to second language teaching and study that it represents will become more commonplace.

## 7.5 Recommendations for future research

If, as argued, delayed CF represents not simply uptake but *intake*, available either as evidence confirming or disconfirming IL hypotheses or as phraseological exemplars for the instantiation of future utterances and for the development of intuitions of well-formedness (and there is no reason to suppose that these are mutually exclusive possibilities), then we could anticipate that subsequent fluent production (and recognition) would show some indication of this. In this research, *fluency* has been measured according to temporal features in speech production, and reaction time in recognition. However, if we adopt Brumfit's (1979: 115; 1984: 56) definitions of *fluency* as the learner's 'truly internalized grammar' and 'natural language use', then convincing evidence of a causal relationship between CF and acquisition is admittedly not easy to obtain. Nevertheless, it is possible that well-crafted experimental research could investigate the (re)appearance in subsequent fluent

conversational production of exemplars provided as CF. This is surely an area worthy of further research, as is the question of which errors respond best to this kind of CF (Section 4.9) and what level of explicit metalinguistic feedback should accompany the CF.

A second line of enquiry proposed by this research is an investigation of the psycholinguistic variables, especially the relationship between linguistic and metalinguistic ‘fluency’. This investigation has proposed that metalinguistic and linguistic competence are separate but related in potentially interesting ways. The standard deviation of reaction time has been proposed as an overall measure of metalinguistic competence (Section 3.4.6), but this deserves further study, as does the relationship between overall proficiency and reaction time on a TGJT (Sections 3.5.6 and 5.5.7). This has implications not just for the provision of CF, but potentially also for proficiency testing.

Finally, there is an urgent need for continued analysis of the error database and development and refinement of the analytical tools (Section 6.6). This in turn will require inter-rater reliability studies to ensure the highest possible degree of analytical consistency. It is hoped that the online migration of the database, by making it accessible off site, will not only accelerate the collection and analysis of data, expanding the variety of L1 backgrounds represented, but also invite closer collaboration between researchers and teachers.

---

## REFERENCES

---

- Aston, G. (1997) 'Small and large corpora in language learning'. Retrieved 27 July, 2011 from: <http://www.sslmit.unibo.it/~guy/wudj1.htm>
- Aston, G. (2000) Corpora and Language Teaching. In *Rethinking language pedagogy from a corpus perspective: papers from the Third International Conference on Teaching and Language Corpora*, (Eds, Burnard, L. & McEnery, T.) P. Lang, Frankfurt am Main; New York, 7–17.
- Atwell, E. S. (1987) 'How to detect grammatical errors in a text without parsing it'. Proceedings of the third conference on European chapter of the Association for Computational Linguistics.
- Bailey, N., Madden, C. & Krashen, S. (1974) 'Is there a "natural sequence" in adult second language learning?', *Language Learning*, 24 235–243.
- Bard, E. G., Robertson, D. & Sorace, A. (1996) 'Magnitude estimation of linguistic acceptability,' *Language*, 72 32–68.
- Barlow, M. (2005) Computer-based analyses of learner language. In *Analysing Learner Language*, (Eds, Ellis, R. & Barkhuizen, G.) Oxford University Press, Oxford, 335–357.
- Basturkmen, H., Loewen, S. & Ellis, R. (2004) 'Teachers' stated beliefs about incidental focus on form and their classroom practices,' *Applied Linguistics*, 25 (2), 243–272.
- Bateson, G. (1979) *Mind and nature: a necessary unity*. Dutton, New York.
- Bialystok, E. & Sharwood Smith, M. (1985) 'Interlanguage is not a state of mind: an evaluation of the construct for second-language acquisition,' *Applied Linguistics*, 6 101–117.
- Birdsong, D. (1989) *Metalinguistic performance and interlinguistic competence*. Springer–Verlag, Berlin; New York.
- Bley-Vroman, R. (1983) 'The comparative fallacy in interlanguage studies: the case of systematicity,' *Language Learning*, 33 1–17.
- Bley-Vroman, R. (1990) 'The logical problem of foreign language learning,' *Linguistic Analysis*, 20 (1–2), 3–47.
- Bley-Vroman, R. (2002) 'Frequency in production, comprehension, and acquisition,' *Studies in Second Language Acquisition*, 24 209–213.
- Bley-Vroman, R., Felix, S. & Ioup, G. (1988) 'The accessibility of universal grammar in adult language learning,' *Second Language Research*, 4 (1), 1–32.
- Bley-Vroman, R. & Masterson, D. (1989) 'Reaction time as a supplement to grammaticality judgements in the investigation of second language learners' competence,' University of Hawai'i *Working Papers in ESL*, 8 (2), 207–237.

- Boersma, P. & Weenink, D. (2011) 'PRAAT'. Retrieved May 24, 2011 from:  
<http://www.fon.hum.uva.nl/praat/>
- Bosher, S. (1990) 'The role of error correction in the process-oriented ESL composition classroom,' *MinneTESOL Journal*, v8 p89–101 1990, 8 89–101.
- Bowerman, M. (1982) Starting to talk worse: clues to language acquisition from children's late speech errors. In *U-shaped behavioural growth*, (Ed, Strauss, S.) Academic Press, New York, 101–145.
- Brown, R. (1973) *A first language; the early stages*. Harvard University Press, Cambridge, MA.
- Brumfit, C. J. (1979) 'Notional syllabuses – a reassessment,' *System*, 7 (2), 111–116.
- Brumfit, C. J. (1984) *Communicative methodology in language teaching: the roles of accuracy and fluency*. Cambridge University Press, Cambridge.
- Burns, A. (2010) *Doing action research in English language teaching*. Routledge, New York.
- Burt, M. & Kiparsky, C. (1972) *The Gooficon: a repair manual for English*. Newbury House, Rowley, MA.
- Butler Platt, C. & MacWhinney, B. (1982) 'Error assimilation as a mechanism in language learning,' *Journal of Child Language*, 10 401–414.
- Can, A. (2007) 'When it does not fit into the schema,' *Egitim Fakültesi Dergisi XX*, 20 (2), 283–313.
- Carroll, S. (1995) The irrelevance of verbal feedback to language learning. In *The Current State of Interlanguage: studies in honor of William E. Rutherford*, (Eds, Eubank, L., Selinker, L. & Sharwood Smith, M). John Benjamins Publishing Company, Amsterdam.
- Carroll, S. E. (2001) *Input and evidence. The raw material of second language acquisition*. John Benjamins Publishing Company, Philadelphia, PA, USA.
- Carter, R. & McCarthy, M. (2006) *Cambridge grammar of English: a comprehensive guide; spoken and written English grammar and usage*. Cambridge University Press, Cambridge, New York.
- Chambers, F. (1997) 'What do we mean by fluency?' *System*, 25 (4), 535–544.
- Chan, A. Y. W. (2004) 'Syntactic transfer: evidence from the interlanguage of Hong Kong Chinese ESL learners,' *Modern Language Journal*, (88), 56–74.
- Chaudron, C. (1983) 'Research on metalinguistic judgments: a review of theory, methods, and results,' *Language Learning*, 33 (3), 343–377.
- Chenoweth, N. A., Day, R. R., Chun, A. E. & Lupescu, S. (1983) 'Attitudes and preferences of ESL students to error correction,' *Studies in Second Language Acquisition*, 6 (1), 79–87.
- Chipere, N. (1998) 'Real language users'. Retrieved 29/5/07, from:  
<http://cogprints.org/712/00/real.PDF>
- Chomsky, N. (1965) *Aspects of the theory of syntax*. The M.I.T. Press, Cambridge, MA.
- Chomsky, N. (1986) *Knowledge of language: its nature, origin, and use*. Praeger, New York.

- Chun, A. E., Day, R. R., Chenoweth, N. A. & Lupescu, S. (1982) 'Errors, interaction, and correction: a study of native-normative conversations,' *TESOL Quarterly*, 16 (4), 537–547.
- Clark, H. H. & Fox Tree, J. E. (2002) 'Using uh and um in spontaneous speaking,' *Cognition*, 73–111.
- Cohen, A. (1983) 'Reformulating second-language compositions: a potential source of input for the learner,' *ERIC document*, ED 228866 1–23.
- Cohen, A. & Robbins, M. (1976) 'Toward assessing interlanguage performance: the relationship between selected errors, learners' characteristics, and learners' explanations,' *Language Learning*, 26 (1), 45–66.
- Cook, A. L. (2009) *An investigation into the role of implicit knowledge in adult second language acquisition*. Unpublished PhD dissertation, University of Edinburgh.
- Cook, V. (1994) Timed grammaticality judgements of the Head Parameter in L2 Learning. In *The dynamics of language processes: essays in honour of Hans W. Dechert*, (Eds, Dechert, H.-W. & Bartelt, G.) Gunter Narr, Tübingen, 15–31.
- Cook, V. (2003) 'The poverty-of-the-stimulus argument and structure-dependency in L2 users of English,' *IRAL*, 41 201–221.
- Coppieters, R. (1987) 'Competence differences between native and near-native speakers,' *Language*, 63 (3), 544–573.
- Corder, S. P. (1967) 'The significance of learner's errors,' *IRAL*, 5 (4), 161–170.
- Corder, S. P. (1981) *Error analysis and interlanguage*. Oxford University Press, Oxford.
- Cowan, R., Choi, H. E. & Kim, D. H. (2003) 'Four questions for error diagnosis,' *CALICO Journal*, 20 (3).
- Crewe, W. J. (1977) *Singapore English and standard English: exercises in awareness*. Eastern Universities Press, Singapore.
- Cullen, R. (2008) 'Teaching grammar as a liberating force,' *ELT Journal*, 62 (9), 221–230.
- Dąbrowska, E. (2010) 'Naive v. expert intuitions: an empirical study of acceptability judgments,' *The Linguistic Review*, 27 (1), 1–23.
- Dagneaux, E., Denness, S. & Granger, S. (1998) 'Computer-aided error analysis,' *System*, 26 163–174.
- Davies, A. (2003) *The native speaker: myth and reality*. Clevedon: Multilingual Matters.
- Davies, M. (2009) 'The 385+ million word corpus of contemporary American English (1990–2008+) design, architecture, and linguistic insights,' *International Journal of Corpus Linguistics*, 14 (2), 159–190.
- Davies, W. & Kaplan, T. (1998) 'Native speaker vs. L2 learner grammaticality judgements,' *Applied Linguistics*, 19 (2), 183–203.
- De Beaugrande, R. (2001) 'Interpreting the discourse of H. G. Widdowson: a corpus-based critical discourse analysis,' *Applied Linguistics*, 22 (1), 104–121.
- De Cock, S., Gilquin, G. & Granger, S. (2010) 'Louvain international database of spoken English interlanguage (LINDSEI)'. Retrieved July 29th, 2011 from:

<http://www.uclouvain.be/en-cecl-lindsei.html>

- De Guerrero, M. C. M. (1994) Form and Function of Inner Speech in Adult Second Language Learning. (Eds, Lantolf, J. P. & Appel, G.) Ablex Publishing Company, Norwood, NJ, pp. 83-116.
- de jong, N. H. & Wempe, T. (2009) 'Praat script to detect syllable nuclei and measure speech rate automatically,' *Behavior Research Methods*, 41 (2), 385–90.
- DeKeyser, R. (2007) *Practice in a second language: perspectives from applied linguistics and cognitive psychology*. Cambridge University Press, Cambridge, New York.
- Devescovi, A., D'Amico, S. & Gentile, P. (1999) 'The development of sentence comprehension in Italian: a reaction time study,' *First Language*, 19 129–163.
- Díaz–Negrillo, A. & Fernández–Domínguez, J. (2006) 'Error tagging systems for learner corpora,' *RESLA*, 19 83–102.
- Dulay, H. C. & Burt, M. K. (1974a) *Natural sequences in child second language acquisition*. *Language Learning*, 24 37-53.
- Dulay, H. C. & Burt, M. K. (1974b) You can't learn without goofing. In *Error analysis: perspectives on second language acquisition*, (Ed, Richards, J.) Longman, London, 95–123.
- Dulay, H. C., Burt, M. K. & Krashen, S. D. (1982) *Language two*. Oxford University Press, New York.
- Edge, J. (1989) *Mistakes and correction*. Longman, New York.
- Edge, J. (2001) *Action research*. Teachers of English to Speakers of Other Languages, Alexandria, VA.
- Edge, J. (2011) *The reflexive teacher educator in TESOL: roots and wings*. Routledge, New York.
- Edge, J. & Richards, K. (1993) *Teachers develop, teachers research: papers on classroom research and teacher development*. Heinemann, Oxford; Portsmouth, NH.
- Ejzenberg, R. (2000) The juggling act of oral fluency: a psycho-sociolinguistic metaphor. In *Perspectives on fluency*, (Ed, Riggenbach, H.) The University of Michigan Press, Michigan, 287–314.
- El–dali, H. M. (2010) 'Second language learners' metalinguistic ability and classroom instruction: focus on grammaticality judgments,' *Journal of Language and Literature*, 2 (8), 56–75.
- Ellis, N. (1996) 'Sequencing in SLA: phonological memory, chunking, and points of order,' *Studies in Second Language Acquisition*, 18, 91–126.
- Ellis, N. (2002) 'Frequency effects in language processing: a review with implications for theories of implicit and explicit language acquisition,' *Studies in Second Language Acquisition*, 24 143–188.
- Ellis, N. (2005) 'At the interface: the dynamic interactions of explicit and implicit language knowledge,' *Studies in Second Language Acquisition*, 27, 305–352.
- Ellis, N. (2007) The weak-interface, consciousness, and form focused instruction: mind the

- doors. In *Form focused instruction and teacher education: studies in honour of Rod Ellis*, (Eds, Fotos, S. & Nassaji, H.) Oxford University Press, Oxford, 17–33.
- Ellis, N., Simpson-Vlach, R. & Maynard, C. (2008) ‘Formulaic language in native and second language speakers: psycholinguistics, corpus linguistics, and TESOL,’ *TESOL Quarterly*, 42 (3), 375–396
- Ellis, R. (1985) ‘Sources of variability in interlanguage,’ *Applied Linguistics*, 6 118–131.
- Ellis, R. (1989) Sources of intra-learner variability in language use and their relationship to second language acquisition. In *Variation in second language acquisition: psycholinguistic issues*, (Eds, Gass, S., Madden, C., Preston, D. & Selinker, L.) Multilingual Matters, Philadelphia, 22–45
- Ellis, R. (1991) ‘Grammaticality judgments and second language acquisition,’ *Studies in Second Language Acquisition*, 13 161–186.
- Ellis, R. (1994) *The study of second language acquisition*. Oxford University Press, Oxford.
- Ellis, R. (1999) ‘Item versus system learning: explaining free variation,’ *Applied Linguistics*, 20 (4), 460–480.
- Ellis, R. (2003) *Task-based language learning and teaching*. Oxford University Press, Oxford.
- Ellis, R. (2005a) *Instructed second language acquisition*. Ministry of Education, Wellington, New Zealand.
- Ellis, R. (2005b) ‘Measuring implicit and explicit knowledge of a second language: a psychometric study,’ *Studies in Second Language Acquisition*, 27 141–172.
- Ellis, R. (2008a) ‘Principles of instructed second language acquisition,’ *CAL Digest*,
- Ellis, R. (2008b) *The study of second language acquisition*. Oxford University Press, Oxford.
- Ellis, R. (2009) ‘The differential effects of three types of task planning on the fluency, complexity, and accuracy in L2 oral production,’ *Applied Linguistics*, 30 (4) (4), 474–509.
- Ellis, R. & Barkhuizen, G. P. (2005) *Analysing learner language*. Oxford University Press, Oxford.
- Ellis, R., Basturkmen, H. & Loewen, S. (2001) ‘Preemptive focus on form in the ESL classroom,’ *TESOL Quarterly*, 35 (3), 407–423.
- Ellis, R., Loewen, S., Elder, C., Erlam, R., Philp, J. & Reinders, H. W. (Eds.) (2009) *Implicit and explicit knowledge in second language learning, Testing and Teaching Multilingual Matters*, Bristol.
- Ellis, R., Loewen, S. & Erlam, R. (2006) ‘Implicit and explicit corrective feedback and the acquisition of L2 grammar,’ *Studies in Second Language Acquisition*, 28 339–368.
- Ellis, R. & Sheen, Y. H. (2006) ‘Reexamining the role of recasts in second language acquisition,’ *Studies in Second Language Acquisition*, 28 575–600.
- Erlam, R. (2006) ‘Elicited imitation as a measure of L2 implicit knowledge: an empirical validation study,’ *Applied Linguistics*, 27 (3) (3), 464–491.
- Fanselow, J. F. (1977) ‘The treatment of error in oral work,’ *Foreign Language Annals*, 10

583–593.

- Ferreira, F., Christianson, K. & Hollingworth, A. (2001) ‘Misinterpretations of garden-path sentences: implications for models of sentence processing and reanalysis,’ *Journal of Psycholinguistic Research*, 30 (1), 3–20.
- Fitzpatrick, E. & Seegmiller, S. (2004) The Montclair electronic language database project. In *Applied corpus linguistics. A multidimensional perspective*, (Eds, Connor, U. & Upton, T.A.) Rodopi, Amsterdam, 223–237.
- Flowerdew, L. (2009) ‘Applying corpus linguistics to pedagogy: a critical evaluation,’ *International Journal of Corpus Linguistics*, 14 (3), 393–417.
- Foster-Cohen, S. (2001) ‘First language acquisition. Second language acquisition: ‘What’s Hecuba to him or he to Hecuba?’,’ *Second Language Research*, 17 (4), 329–344.
- Freeman, D. (1998) *Doing teacher research: from inquiry to understanding*. Heinle & Heinle, Pacific Grove.
- French, F. G. (1949) *Common errors in English: their cause, prevention, and cure*. Oxford University Press, London.
- Fulcher, G. (2003) *Testing second language speaking*. Pearson Education Limited, Harlow, UK.
- Galloway, I. (2005) ‘Computer learner corpora and their pedagogical application,’ *TESOL Quarterly*, 39 (2), 333–339.
- Gass, S. M. (1983) ‘The development of L2 intuitions,’ *TESOL Quarterly*, 17 (2), 273–291.
- Gatbonton, E. & Segalowitz, N. (1988) ‘Creative automatization: principles for promoting fluency within a communicative framework,’ *TESOL Quarterly*, 22 (3), 473–492.
- Gatbonton, E. & Segalowitz, N. (2005) ‘Rethinking communicative language teaching: a focus on access to fluency’. *The Canadian Modern Language Review*, 61 (3), 325–353.
- George, H. V. (1972) *Common errors in language learning: insights from English; a basic guide to the causes and preventions of students’ errors in foreign language learning*. Newbury House Publishers, Rowley, MA.
- Gilquin, G. & De Cock, S. (2011) ‘Errors and disfluencies in spoken corpora: setting the scene,’ *International Journal of Corpus Linguistics*, 16 (2), 141–172.
- Gilquin, G., Granger, S. & Paquot, M. (2007) ‘Learner corpora: the missing link in EAP pedagogy,’ *Journal of English for Academic Purposes*, 6 319–335.
- Ginther, A., Dimova, S. & Yang, R. (2010) ‘Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring,’ *Language Testing*, 27 (3), 379–399.
- Goldberg, A. (2011) *Cognitive linguistics*. Routledge, Milton Park, Abingdon; New York.
- Goldberg, A. E. (1995) *Constructions: a construction grammar approach to argument structure*. University of Chicago Press, Chicago.
- Goldberg, A. E. (2006) *Constructions at work: the nature of generalization in language*. Oxford University Press, Oxford; New York.
- Goldschneider, J. M. & DeKeyser, R. M. (2001) ‘Explaining the “natural order of L2



morpheme acquisition” in English: a meta-analysis of multiple determinants,’ *Language Learning*, (51), 1–50.

Granger, S. (1998) *Learner English on computer (studies in language and linguistics)*. Addison Wesley Publishing Company,

Granger, S. (2003) ‘Error-tagged corpora and CALL: a promising synergy,’ *CALICO Journal*, 20 (3), 465–480.

Granger, S., Hung, J. & Petch-Tyson, S. (2002) *Computer learner corpora, second language acquisition, and foreign language teaching*. John Benjamins Publishing Co, Amsterdam.

Granger, S. & Tyson, S. (1996) ‘Connector usage in the English essay writing of native and non-native EFL speakers of English,’ *World Englishes*, 15 9–29.

Green, B. (2006) ‘A framework for teaching grammar to Japanese learners in an intensive English program,’ *The Language Teacher*, 30 (2), 3–11.

Gregg, K. R. (1986) ‘Review of S. D. Krashen, the Input Hypothesis: issues and implications,’ *TESOL Quarterly*, 20 116–122.

Gregg, K. R. (1990) ‘The variable competence model of second language acquisition and why it isn’t,’ *Applied Linguistics*, 11 (4), 364–383.

Guntermann, G. (1978) ‘A study of the frequency and communicative effects of errors in Spanish,’ *Modern Language Journal*, 62 249–253.

Hammerly, H. (1991) *Accuracy and fluency: toward balance in language teaching*. Multilingual Matters, Philadelphia.

Han, Y. (1996) *L2 learners’ explicit knowledge of verb complement structures and its relationships to L2 implicit knowledge*. Unpublished PhD dissertation, Temple University.

Han, Y. (2000) ‘Grammaticality judgment tests: how reliable and valid are they?,’ *Applied Language Learning*, 11 (1), 177–204.

Han, Z. (2004) ‘To be a native speaker means not to be a nonnative speaker,’ *Second Language Research*, 20 (2), 166–187.

Han, Y. & Ellis, R. (1998) ‘Implicit knowledge, explicit knowledge and general language proficiency,’ *Language Teaching Research*, 2 (1), 1–23.

Han, Z. & Selinker, L. (1996) ‘Multiple effects and error resistance: a longitudinal case study’. Retrieved June 5th, 2006 from:  
<http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.33.4047&rep=rep1&type=pdf>

Han, Z. & Selinker, L. (1999) ‘Error resistance: towards an empirical pedagogy,’ *Language Teaching Research*, 3 (3), 248–275.

Harrat, D. (2011) *A cross-linguistic analysis of the spoken errors of Arabic L1 learners of English*. Unpublished MA dissertation, Gonzaga University.

Harris, R. (1998) ‘Making grammar instruction relevant through student-run conversations’. Workshop presented at the annual TESOL Conference, Seattle, Washington, March 16.

Hauser, E. (2005) ‘Coding “corrective recasts”: the maintenance of meaning and more

- fundamental problems,' *Applied Linguistics*, 26 (3), 293–316.
- Hawkins, R. (2001) 'Universal grammar in second language acquisition,' *Second Language Research*, 17 (4), 345–367.
- Hedge, T. (2000) *Teaching and learning in the language classroom*. Oxford University Press, Oxford.
- Hendrickson, J. (1979) 'Evaluating spontaneous communication through systematic error analysis,' *Foreign Language Annals*, 12 (5), 357–64 (reprinted 1981 by SAMEO Regional Language Center, Singapore).
- Hinkel, E. (2002) *Second language writers' text*. Lawrence Erlbaum, Mahwah, NJ.
- Housen, A. & Kuiken, F. (2009) 'Complexity, accuracy, and fluency in second language acquisition,' *Applied Linguistics*, 30 (4), 461–473.
- Howatt, A. P. R. (1984) *A history of English language teaching*. Oxford University Press, Oxford; New York.
- Hu, G. (2002) 'Psychological constraints on the utility of metalinguistic knowledge in second language production,' *Studies in Second Language Acquisition*, 24 347–386.
- Hulstijn, J. & De Graaf, R. (1994) 'Under what conditions does explicit knowledge of a second language facilitate the acquisition of implicit knowledge? A research proposal,' *AILA Review*, 97–112.
- Hulstijn, J. & Schmidt, R. (1994) 'Consciousness in second language learning,' *AILA Review*, 5–112.
- Hummel, K. M. & French, L. M. (2010) 'Phonological memory and implications for the second language classroom,' *The Canadian Modern Language Review*, 66 (3), 371–391.
- Hunston, S. & Francis, G. (2000) *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. John Benjamins Publishing Co, Amsterdam; Philadelphia.
- Hunston, S., Francis, G. & Manning, E. (1997) 'Grammar and vocabulary: showing the connections,' *ELT Journal*, 51 (3), 208–216.
- Hunter, J. (2012) "'Small talk': Developing fluency, accuracy, and complexity in speaking,' *ELT Journal*, 66 (1), 30–41.
- Hurd, S. & Murphy, L. (2005) *Success with languages*. Routledge, London; New York.
- Izumi, E. & Ishihara, H. (2004) 'Investigation into language learners' acquisition order based on an error analysis of a learner corpus,' *IWLeL 2004: an Interactive Workshop on Language e-Learning*, 63–71.
- Izumi, E., Uchimoto, K. & Ishihara, H. (2005) 'Error annotation for corpus of Japanese learner English,' *Proceedings of the Sixth International Workshop on Linguistically Interpreted Corpora (LINC 2005)*, 15 October 71–80.
- James, C. (1980) *Contrastive Analysis*. Addison Wesley Publishing Company, New York.
- James, C. (1994) 'Don't shoot my dodo: on the resilience of contrastive and error analysis,' *IRAL*, 32 (3), 179–200.
- James, C. (1998) *Errors in language learning and use: exploring error analysis*. Longman,

New York.

- Johns, T. (1991) 'Should you be persuaded: two examples of data-driven learning,' *English Language Research Journal*, 4 1–16.
- Johnson, K. & Johnson, H. (Eds.) (2011) *Encyclopedic dictionary of applied linguistics* Blackwell Publishing, Oxford
- Juffs, A. (2001) 'Psycholinguistically oriented second language research,' *Annual Review of Applied Linguistics*, 21 207–220.
- Juffs, A. & Harrington, M. (1995) 'Parsing effects in second language sentence processing: subject and object asymmetries in *wh*-extraction,' *Studies in Second Language Acquisition*, 17 483–516.
- Kachru, B. B. (1992). *The Other Tongue: English across cultures*. Urbana: University of Illinois Press.
- Katayama, A. (2007) 'Japanese EFL students' preferences toward correction of classroom oral errors,' *Asian EFL Journal*, 9 (4), 289–305.
- Kellerman, E. (1985) If at first you do succeed. In *Input in second language acquisition*, (Eds, Gass, S. & Madden, C.) Newbury House Publishers, Inc., Rowley, MA. 345–353.
- Kelly, L. G. (1969) *25 centuries of language teaching; an inquiry into the science, art, and development of language teaching methodology, 500 B.C.–1969*. Newbury House Publishers, Rowley, MA.
- Kharna, N. N. (1987) 'Arab students' problems with the English relative clause,' *IRAL*, 25 (3), 257–266.
- Kindt, D. (2004) *A systemic view of emergent course design: a multimethod exploration of the complex, dynamic nature of student engagement in an emergent EFL course*. Unpublished PhD dissertation, University of Birmingham.
- Kindt, D. & Wright, M. (2001) 'Integrating language learning and teaching with the construction of computer learner corpora'. Retrieved July 15, 2011 from: <http://www3.nufs.ac.jp/~kindt/media/corpora.pdf>
- Klein, W. & Perdue, C. (1997) 'The basic variety (or: Couldn't natural languages be much simpler?),' *Second Language Research*, 13 301.
- Kobayashi, T. (1995) 'Can retrospective feedback improve ESL speech?' *Studies in Culture*, 4 (1), 1–27.
- Kormos, J. & Dénes, M. (2004) 'Exploring measures and perceptions of fluency in the speech of second language learners,' *System*, 32 145–164.
- Kramsch, C. (1993) *Context and Culture in Language Teaching (Oxford Applied Linguistics)*. Oxford University Press, New York.
- Kramsch, C. & Steffensen, S. (2008) Ecological perspectives on second language acquisition and language socialization. In *Language and Socialization*, Springer Verlag, Heidelberg, 17–28.
- Krashen, S. (1979) 'A response to McLaughlin, "The monitor model: some methodological considerations"', *Language Learning*, 29 151–167.

- Krashen, S. (1981) *Second language acquisition and second language learning*. Pergamon Press, Oxford.
- Krashen, S. (1982) *Principles and practice in second language acquisition*. Pergamon Press, Oxford.
- Krashen, S. (1985) *The Input Hypothesis: issues and implications*. Longman ELT, New York.
- Krashen, S. (1992) 'Formal grammar instruction. Another educator comments,' *TESOL Quarterly*, 26 (2), 409–411.
- Krashen, S. (1994) The input hypothesis and its rivals. In *Implicit and explicit learning of languages*, (Ed, Ellis, N.) Academic Press, San Diego, CA, 45–77.
- Krashen, S. (2002) 'The comprehension hypothesis and its rivals'. Selected papers from the eleventh international symposium on English teaching/fourth Pan-Asian conference 395–404.
- Krishnamurthy, R. & Kosem, I. (2007) 'Issues in creating a corpus for EAP pedagogy and research,' *Journal of English for Academic Purposes*, 6 356–373.
- Kumaravadivelu, B. (2003) *Beyond methods: macrostrategies for language teaching*. Yale University Press, New Haven, CT.
- Kumaravadivelu, B. (2006) 'TESOL methods: changing tracks, challenging trends,' *TESOL Quarterly*, 40 (1), 59–81.
- Labov, W. (1975) *What is a linguistic fact?* Peter de Ridder Press, Lisse.
- Labov, W. (1996) When intuitions fail. In *Papers from the Parasession on Theory and Data in Linguistics*, (Eds, McNair, L., Singer, K., Dolbrin, L. & Aucon, M.) 77–106.
- Lakshmanan, U. & Selinker, L. (2001) 'Analysing interlanguage: how do we know what learners know?,' *Second Language Research*, 17 (4), 393–420.
- Lantolf, J. P. & Dunn, W. E. (1998) 'Vygotsky's zone of proximal development and Krashen's i+1: incommensurable constructs; incommensurable theories,' *Language Learning*, 48 (3), 411–442.
- Larsen-Freeman, D. (2000) 'Second language acquisition and applied linguistics,' *Annual Review of Applied Linguistics*, 20 165–181.
- Larsen-Freeman, D. (2006) 'The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English,' *Applied Linguistics*, 27 (4), 590–619.
- Larsen-Freeman, D. (2009) 'Adjusting expectations: the study of complexity, accuracy, and fluency in second language acquisition,' *Applied Linguistics*, 30 (4), 579–589.
- Leech, G. (1993) 'Corpus annotation schemes,' *Literary and Linguistic Computing*, 8 (4), 275–281.
- Leech, G. (1997) Teaching and language corpora: a convergence. In *Teaching and language Corpora*, (Eds, A., W., Fligelstone, S., McEnery, T. & Knowles, G.) Longman, Harlow, 1–24.
- Lennon, P. (1990) 'Investigating fluency in EFL: a quantitative approach,' *Language Learning*, 40 387–417.

- Lennon, P. (1991) 'Error: some problems of definition, identification, and distinction,' *Applied Linguistics*, 12 (2), 180–196.
- Leung, C., Harris, R. & Rampton, B. (1991) The idealised Native Speaker, reified ethnicities, and classroom realities. *TESOL Quarterly*, 31, 543–560.
- Levelt, W. J. M. (1983) 'Monitoring and self-repair in speech,' *Cognition*, 14 41–104.
- Levelt, W. J. M. (1989) *Speaking: from intention to articulation*. MIT Press, Cambridge, MA.
- Lewkowicz, N. K. (1971) 'Topic-comment and relative clause in Arabic,' *Language*, 47 (4), 810–825.
- Li, S. (2010) 'The effectiveness of corrective feedback in SLA: a meta-analysis,' *Language Learning*, 60 (2), 309–365.
- Lightbown, P. M. & Spada, N. (1999) *How languages are learned*. Oxford University Press, Oxford.
- Lochtman, K. (2002) 'Oral corrective feedback in the foreign language classroom: how it affects interaction in analytic foreign language teaching,' *International Journal of Educational Research*, 37 271–283.
- Loewen, S. (2002) *The occurrence and effectiveness of incidental focus on form in meaning-focused ESL lessons*. Unpublished PhD dissertation, University of Auckland.
- Loewen, S. (2004) 'Uptake in incidental focus on form in meaning-focused ESL lessons,' *Language Learning*, 51 (4), 153–188.
- Long, M. H. (1990) 'The least a second language acquisition theory needs to explain,' *TESOL Quarterly*, 24 (4), 649–666.
- Long, M. H. (2007) Recasts in SLA: the story so far. In *Problems in SLA*, (Ed, Long, M. H.) Erlbaum, Mahwah, NJ. 75–116.
- Long, M. H. & Porter (1985) 'Group work, interlanguage talk, and Second Language Acquisition,' *TESOL Quarterly*, 19 (2), 207–228.
- Lynch, T. (2001) 'Seeing what they meant: transcribing as a route to noticing,' *ELT Journal*, 55 (2), 124–132.
- Lynch, T. & McLean, J. (2003) 'Effects of feedback on performance: a study of advanced learners on an ESP speaking course,' *Edinburgh Working Papers in Applied Linguistics*, 12 19–44.
- Lyster, R. & Ranta, L. (1997) 'Corrective feedback and learner uptake,' *Studies in Second Language Acquisition*, 19 37–66.
- Mackey, A. (2005) 'Revisiting elicited imitation'. Retrieved July 12, 2001 from: [http://mason.gmu.edu/~asherris/Portfolio/Coursework/Ling681SLAResearchMethodology/Elicited\\_Imitation.pdf](http://mason.gmu.edu/~asherris/Portfolio/Coursework/Ling681SLAResearchMethodology/Elicited_Imitation.pdf)
- Mackey, A., Al-Khalil, M., Atanssova, G., Hama, M., Logan-Terry, A. & Nakatsukasa, K. (2007) 'Teachers' intentions and learners' perceptions about corrective feedback in the L2 classroom,' *Innovation in Language Learning and Teaching*, 1 (1), 129–152.
- Mahboob, A. (2005) 'Beyond the native speaker in TESOL,' *Culture, Context, & Communication*, 60–93.

- Mandell, P. B. (1999) 'On the reliability of grammaticality judgement tests in second language acquisition research,' *Second Language Research*, 15 (1), 73–99.
- Margolis, D. P. (2010) 'Handling oral error feedback in language classrooms,' *MinneWITESOL Journal*, 27 4–17.
- Marinis, T. (2003) 'Psycholinguistic techniques in second language acquisition research,' *Second Language Research*, 19 (2), 144–161.
- McCarthy, M., McCarten, J., Sandiford, H. & Aldcorn, S. B. (2005) *Touchstone*. Cambridge University Press, Cambridge, UK; New York.
- McEnery, T. & Wilson, A. (1996) *Corpus linguistics*. Edinburgh University Press, Edinburgh, Scotland.
- McEnery, T., Xiao, R. & Tono, Y. (2006) *Corpus-based language studies: an advanced resource book*. Routledge, New York.
- McLaughlin, B. (1987) *Theories of second language learning*. Edward Arnold, London.
- McLaughlin, B. (1990) '“Conscious” versus “unconscious” learning,' *TESOL Quarterly*, 24 (4), 617–634.
- Meisel, J., Clahsen, H. & Pienemann, M. (1981) 'On determining developmental stages in natural second language acquisition,' *Studies in Second Language Acquisition*, 3 (2), 109–135.
- Meunier, S. (2002) The pedagogical value of native and learner corpora in EFL grammar teaching. In *Computer learner corpora, second language acquisition and foreign language teaching*, (Eds, Granger, S., Hung, J. & Petch-Tyson, S.) John Benjamins, Amsterdam, 119–142.
- Milton, J. & Chowdhur, N. (1994) Tagging the interlanguage of Chinese learners of English. In *Entering text*, (Eds, Flowerdew, L. & Tong, A. K. K.) Language Centre, The Hong Kong University of Science and Technology, Hong Kong, 127–143.
- Mukherjee, J. & Rohrbach, J. (2006) Rethinking applied corpus linguistics from a language-pedagogical perspective: new departures in Learner Corpus Research. In *Planning, gluing, and painting corpora: inside the applied corpus linguist's workshop*, (Eds, Kettemann, B. & Marko, G.) Peter Lang, Frankfurt, pp. 205–232.
- Murphy, V. (1997) 'The effect of modality on a grammaticality judgement task,' *Second Language Research*, 13 (1), 34–65.
- Nassaji, H. & Swain, M. (2000) 'A Vygotskian perspective on corrective feedback in L2: the effect of random versus negotiated help on the learning of English articles,' *Language Awareness*, 9 (1), 126–145.
- Nemser, W. (1971) 'Approximative systems of foreign language learners,' *IRAL*, IX (2), 113–123.
- Newmeyer, F. J. (1983) *Grammatical theory, its limits and its possibilities*. University of Chicago Press, Chicago.
- Norris, J. & Ortega, L. (2009) 'Towards an organic approach to investigating CAF in instructed SLA: the case of complexity,' *Applied Linguistics*, 30 (4) (4), 555–578.
- Norris, R.W. (2003) 'How do we overcome the difficulties of teaching conditionals?,'

*Bulletin of Fukuoka International University*, 9 39–50.

- Norton Pierce, B. (1995) 'Social identity, investment, and language learning,' *TESOL Quarterly*, 29 9–31.
- O'Brien, I., Segalowitz, N., Freed, B. & Collentine, J. (2007) 'Phonological memory predicts second language oral fluency gains in adults,' *SSLA*, 29 557–582.
- Odlin, T. (2005) 'Crosslinguistic influence and conceptual transfer: What are the concepts?,' *Annual Review of Applied Linguistics*, 25 3–25.
- Oshita, H. (2000) 'What is happened may not be what appears to be happening: a corpus study of "passive" unaccusatives in L2 English,' *Second Language Research*, 16 (4), 293–324.
- Owen, C. (2007) 'A holistic approach to words like holistic,' *English Studies*, 88 (3), 332–350.
- Panova, I. & Lyster, R. (2004) 'Patterns of corrective feedback and uptake in an adult ESL classroom,' *TESOL Quarterly*, 36 (4), 573–595.
- Paquot, M. (2009) 'Centre for English corpus linguistics: learner corpus bibliography'. Retrieved July 30, 2011 from [http://sites-test.uclouvain.be/cecl/projects/learner\\_corpus\\_bibliography.html](http://sites-test.uclouvain.be/cecl/projects/learner_corpus_bibliography.html)
- Partington, A. (1998) *Patterns and meaning*. John Benjamins, Philadelphia.
- Paulovsky, L. (1949) *Errors in English*. Verlag für Jugend und Volk, Vienna.
- Pavlenko, A. & Blackledge, A. (2004) *Negotiation of identities in multilingual contexts*. Multilingual Matters, Clevedon; Buffalo.
- Payne, T. E. (2010) *Understanding English grammar: a linguistic introduction*, Cambridge University Press, Cambridge.
- Pérez-Paredes, P. (2003) Integrating networked learner oral corpora into foreign language instruction. In *Extending the scope of corpus-based research: new applications, new challenges*, (Ed, Granger, S.) Rodopi, Amsterdam, 249–261.
- Picard, M. (2002) 'L1 interference in second language acquisition: the case of question formation in Canadian French,' *IRAL*, 40 61–68.
- Pienemann, M. (1989) 'Is language teachable? Psycholinguistic experiments and hypotheses,' *Applied Linguistics*, 10 52–79.
- Pienemann, M. (1992) 'COALA—a computational system for interlanguage analysis,' *Second Language Research*, 8 59–92.
- Pinker, S. (1994) *The language instinct: how the mind creates language*. Harper Perennial Modern Classics, New York.
- Pinker, S. (1999) *Words and rules: the ingredients of language*. Basic Books, New York.
- Poulisse, N. (1999) *Slips of the tongue: speech errors in first and second language production*. John Benjamins, Amsterdam.
- Poulisse, N. (2000) 'Slips of the tongue in first and second language production,' *Studia Linguistica*, 54 (2), 136–149.

- Pravec, N. A. (2002) 'Survey of learner corpora,' *ICAME Journal*, 26 81–114.
- Quirk, R., Greenbaum, S., Leech, G. & Svartvik, J. (1985) *A Comprehensive grammar of the English language*. Longman, London; New York.
- Rampton, B. (1987) Stylistic variability and not speaking 'normal' English: some post-Labovian approaches and their implications for the study of interlanguages. In *Second Language Acquisition in Context*, (Ed, Ellis, R.) Prentice Hall, New York, 47–58.
- Rampton, B. (1990) Displacing the 'native speaker': expertise, affiliation, and inheritance. *ELT Journal*, 44, 97–101.
- Raof, A. H. B. A. & Razali, S. M. B. C. (2010) 'Peer-to-peer corrective feedback in a group interaction'. Retrieved from: <http://eprints.utm.my/11202/>
- Reder, S., Harris, K. & Setzler, K. (2003) 'The multimedia adult ESL learner corpus,' *TESOL Quarterly*, 37 (1), 546–557.
- Reid, J.M. (1987) 'The learning style preferences of ESL students,' *TESOL QUARTERLY*, 21 (1), 87–111.
- Reimers, S. & Stewart, N. (2007) 'Adobe flash as a medium for online experimentation: a test of reaction time measurement capabilities,' *Behavior Research Methods*, 39 (3), 365–370.
- Reinders, H.W. (2005) *The effects of different task types on L2 learners' intake and acquisition of two grammatical structures*. Unpublished PhD dissertation, The University Of Auckland.
- Révész, A. (2002) 'Task-induced content-familiarity, task-driven attention to form, and learner uptake of recasts: a preliminary inquiry'. Retrieved November 15, 2008 from: <http://journals.tc-library.org/ojs/index.php/tesol/article/viewPDFInterstitial/13/14>
- Richards, J. C. (1970) 'A non-contrastive approach to error analysis,' *ELT Journal*, 25 (3), 204–219.
- Richards, J. C. & Rodgers, T. S. (1986) *Approaches and methods in language teaching: a description and analysis*. Cambridge University Press, Cambridge; New York.
- Riemer, N. (2009) 'Grammaticality as evidence and as prediction in a Galilean linguistics,' *Language Sciences* 31 (2009) 612–633, 31 (2009), 612–633.
- Riggenbach, H. (1991) 'Towards an understanding of fluency: a microanalysis of nonnative speaker,' *Discourse Processes*, 14 423–441.
- Rimmer, W. (2006) 'Grammaticality judgment tests: trial by error,' *Journal of Language and Linguistics*, 5 (2), 246–261.
- Robinson, P., Cadierno, T. & Shirai, Y. (2009) 'Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production,' *Applied Linguistics*, 30 (4), 533–554.
- Rolin-Ianziti, J. (2006) 'Teacher corrective practices in the foreign language classroom: the effect of timing'. *Social Change in the 21st Century* 1–14.
- Römer, U. (2008) Corpora and language teaching. In *Corpus Linguistics. An International Handbook (Volume 1)*, (Eds, Lüdeling, A. & Kytö, M.) 112–130.



- Römer, U. & Wulff, S. (2010) 'Applying corpus methods to writing research: explorations of MICUSP,' *Journal of Writing Research*, 2 (2), 99–127.
- Sampson, G. (1997) *Educating Eve: the "language instinct" debate*. Cassell, London; Washington, D.C.
- Scarcella, R. C. & Oxford, R. L. (1992) *The tapestry of language learning: the individual in the communicative classroom (methodology)*. Heinle & Heinle Publishers,
- Schachter, J. (1976) 'Learner intuitions of grammaticality,' *Language and Learning*, 26 (1), 67–76.
- Schachter, J. (1988) 'Second language acquisition and its relationship to universal grammar,' *Applied Linguistics*, 9 (3), 219–235.
- Scheffler, P. (2011) 'Using corpora in the language classroom,' *ELT J*, 65 (3), 248–250.
- Schmidt, R. (1990) 'The role of consciousness in second language learning,' *Applied Linguistics*, 11 (2), 129–158.
- Schulz, R. (2001) 'Cultural differences in students and teachers perceptions concerning the role of grammar instruction and corrective feedback: USA-Colombia,' *Modern Language Journal*, 85 (2), 244–258.
- Schütze, C. T. (1996) *The empirical base of linguistics: grammaticality judgments and linguistic methodology*. University of Chicago Press, Chicago, IL.
- Schwartz, B. D. (1986) 'The epistemological status of second language acquisition,' *Second language research*, 2 (2), 120–159.
- Seidlhofer, B. (Ed.) (2003) *Controversies in applied linguistics*. Oxford University Press, Oxford.
- Seidlhofer, B. (2011) *Understanding English as a Lingua Franca*. Oxford University Press, Oxford.
- Selinker, L. (1972) 'Interlanguage,' *IRAL*, X (3), 209–231.
- Sheen, Y. H. (2004) 'Corrective feedback and learner uptake in communicative classrooms across instructional settings,' *Language Teaching Research*, 8 (3), 263–300.
- Sheen, Y. H. (2010) 'Differential effects of oral and written corrective feedback in the ESL classroom,' *Studies in Second Language Acquisition*, 32 203– 234.
- Sinclair, J. (1991) *Corpus, concordance, collocation*. Oxford University Press, Oxford.
- Sinclair, J. (2005) *Collins COBUILD English grammar, second edition*. Harper Collins, Glasgow.
- Skehan, P. (1985) 'A framework for the implementation of task-based instruction,' *Applied Linguistics*, 17 (1), 23–62.
- Skehan, P. (1998) *A cognitive approach to language learning*. Oxford University Press, Oxford.
- Skehan, P. (2009) 'Modelling second language performance: integrating complexity, accuracy, fluency, and lexis,' *Applied Linguistics*, 30 (4), 510–532.
- Smith, L. E. & Bisazza, J. A. (1982) 'The comprehensibility of three varieties of English for

- college students in seven countries,' *Language Learning*, 32 129–269.
- Sorace, A. (1985) 'Metalinguistic knowledge and language use in acquisition-poor environments,' *Applied Linguistics*, 6 (3), 239–254.
- Sridhar, S. N. (1994) 'Sources of bias in SLA research: a reality check for SLA theories,' *TESOL Quarterly*, 28 (4), 800–803.
- Stillwell, C., Curabba, B., Alexander, K., Kidd, A., Kim, E., Stone, P. & Wyle, C. (2009) 'Students transcribing tasks: noticing fluency, accuracy, and complexity,' *ELT Journal*, 64 (4),
- Swan, M. & Smith, B. (2001) *Learner English: a teacher's guide to interference and other problems*. Cambridge University Press, Cambridge; New York.
- Tan, M. (2005) 'Authentic language or language errors? Lessons from a learner corpus,' *ELT Journal*, 59 (2), 126–134.
- Tarone, E. (1983) 'On the variability of interlanguage systems,' *Applied Linguistics*, 4 142–163.
- Taylor, M., Schaefer, M. & Schneirov, Z. (2010) 'Dill: the digital language learning lab'. Retrieved from: <http://web.mmlc.northwestern.edu/projects/DiLL.shtml>
- Thornbury, S. (1997) 'Reformulation and reconstruction: tasks that promote "noticing",' *ELT Journal*, 51 (4), 326–335.
- Tomita, Y., Suzuki, W. & Jessop, L. (2009) 'Elicited imitation: toward valid procedures to measure implicit second language grammatical knowledge,' *TESOL Quarterly*, 43 (2), 345–350.
- Tono, Y. (2003) 'Learner corpora: design, development and applications'. Corpus Linguistics 2003 Conference 28–31 March, 800–809.
- Towell, R., Hawkins, R. & Bazergui, N. (1996) 'The development of fluency in advanced learners of French,' *Applied Linguistics*, 17 (1), 84–119.
- Tremblay, A. (2005) 'Theoretical and methodological perspectives on the use of grammaticality judgment tasks in linguistic theory,' *Second Language Studies*, 24 (1), 129–167.
- Truscott, J. (1996) 'The case against grammar correction in L2 writing classes,' *Language Learning*, 46 (2), 327–369.
- Truscott, J. (1999) 'What's wrong with oral grammar correction,' *The Canadian Modern Language Review*, 55 (4), 437–456.
- Truscott, J. (2001) 'Selecting errors for selective error correction,' *Concentric: Studies in English Literature and Linguistics*, 27 (2), 93–108.
- Truscott, J. (2004) 'Evidence and conjecture on the effects of correction: a response to Chandler,' *Journal of Second Language Writing*, 13
- Truscott, J. (2007) 'The effect of error correction on learners' ability to write accurately,' *Journal of Second Language Writing*, 16 255–272.
- Tudor, I. (2003) 'Learning to live with complexity: towards an ecological perspective on language teaching,' *System*, 31 (1), 1–12.

- Tuten, N. L. & Swanson, G. R. (2003) 'Which or that?'. Retrieved August 26, 2011 from: [http://www.accu-assist.com/grammar-tips-archive/GrammarTip\\_which-or-that.htm](http://www.accu-assist.com/grammar-tips-archive/GrammarTip_which-or-that.htm)
- van Lier, L. (2008) Ecological-semiotic perspectives on educational linguistics. In *The handbook of educational linguistics*, (Eds, Spolsky, B. & Hult, F. M.) Blackwell, Oxford, 596–605.
- Vygotsky, L. S. (1978) *Mind in society: the development of higher psychological processes*. Harvard University Press, Cambridge, MA.
- Wagner, J., Foster, J. & Josef, V. G. (2009) 'Judging grammaticality: experiments in sentence classification,' *CALICO*, 26 (3), 474–489.
- White, L. (1977) 'Error analysis and error correction in adult learners of English as a second language,' *Working Papers in Bilingualism*, 13 42–58.
- White, L. (1989) *Universal Grammar and second language acquisition*. J. Benjamins Publishers Co, Amsterdam; Philadelphia.
- White, L. (1990) 'Another look at the logical problem of foreign language learning: a reply to Bley-Vroman,' *Linguistic Analysis*, 20 (1–2), 50–63.
- Widdowson, H. G. (1989) 'Knowledge of language and ability for use,' *Applied Linguistics*, 10 (2), 128–137.
- Widdowson, H. G. (2002) 'Corpora and language teaching tomorrow'. Keynote Lecture delivered at the Fifth Teaching and Language Corpora Conference, Bertinoro, Italy, 29 July.
- Wilcox, G. K. (1978) 'The effect of accent on listening comprehension: a Singapore study,' *English Language Teaching Journal*, 32 118–127.
- Williams, J. (2001) 'The effectiveness of spontaneous attention to form,' *System*, 29 325–340.
- Willis, D. (1990) *The Lexical Syllabus*. Collins COBUILD, London.
- Willis, D. (2003) *Rules, patterns, and words: grammar and lexis in English Language Teaching*. Cambridge University Press, Cambridge, UK.
- Wode, H., Bahns, J., Bedey, H. & Frank, W. (1978) 'Developmental sequence: an alternative approach to morpheme order,' *Language Learning*, 28 (1), 175–185.
- Wray, A. (2000) 'Formulaic sequences in second language teaching: principle and practice,' *Applied Linguistics*, 21 (4), 463–489.
- Wray, A. (2002) *Formulaic language and the lexicon*. Cambridge University Press, Cambridge.
- Zobl, H. (1995) 'Converging evidence for the 'acquisition–learning' distinction,' *Applied Linguistics*, 16 (1), 35–56.