CORA Cork Open Research Archive
Cartlann Taighde Oscailte Chorcaí

Title	The development of oral competence: a semi-longitudinal study on English-speaking adult L2 learners of Chinese in Ireland
Authors	Guo, Rongrong
Publication date	2022-03
Original Citation	Guo, R. 2022. The development of oral competence: a semi-longitudinal study on English-speaking adult L2 learners of Chinese in Ireland. PhD Thesis, University College Cork.
Type of publication	Doctoral thesis
Rights	© 2022, Rongrong Guo. - https://creativecommons.org/licenses/by-nc-nd/4.0/
Download date	2026-11-15 02:09:03
Item downloaded from	https://hdl.handle.net/10468/13573

UCC
University College Cork, Ireland
Coláiste na hOllscoile Corcaigh

Ollscoil na hÉireann, Corcaigh

National University of Ireland, Cork



The development of oral performance: a semi-longitudinal study on English-speaking adult L2 learners of Chinese in Ireland

Thesis presented by

Rongrong Guo

for the degree of

Doctor of Philosophy

University College Cork

Department of Asian studies

Head of Department: Professor Kiri Paramore

Supervisor(s): Dr. Yanyu Guo

Dr. Carlotta Sparvoli

[2022]

Table of Contents

DECLARATION ..... **IV**

ACKNOWLEDGMENTS ..... **V**

ABBREVIATIONS ..... **VI**

ABSTRACT ..... **VII**

LIST OF TABLES ..... **VIII**

LIST OF FIGURES ..... **IX**

CHAPTER 1: INTRODUCTION **1**

1.1 L2 oral proficiency assessment 1
1.2 Speaking assessment of L2 Chinese 5
1.3 The structure of the thesis 7

CHAPTER 2: MEASURING COMPLEXITY, ACCURACY, AND FLUENCY **8**

2.1 Definitions and measurement of CAF constructs 8
2.1.1 Accuracy 9
Definitions of accuracy 9
Measuring accuracy 10
2.1.2 Fluency 11
Definitions of fluency 11
Measuring utterance fluency 12
2.1.3 Complexity 15
2.1.3.1 Lexical complexity 16
Definitions of lexical complexity 16
Measuring lexical complexity 17
2.1.3.2 Syntactic complexity 20
Definitions of syntactic complexity 20
Measuring syntactic complexity 22
2.2 CAF measures on L2 oral Chinese studies 24
2.2.1 Production units in the literature 24
2.2.2 CAF measures in L2 Chinese speaking studies 25
2.2.2.1 Accuracy measures 25
2.2.2.2 Fluency measures 27
Pause marking 28
2.2.2.3 Complexity measures 31
Lexical complexity 32
Summary of CAF measures analysed in L2 Chinese speaking studies 35
2.3 Relationships between complexity, accuracy, and fluency 36
2.3.1 Two models on the relationship between CAF constructs 37
Skehan's Limited Attentional Capacity model 37
Robinson's Cognition Hypothesis 40
The similarity and differences between the two models 42
2.3.2 Experimental studies on the relationship between CAF constructs 43
Studies on the trade-off effect and connected improvement between constructs 44
Studies on the trade-off effect and connected improvement within each construct of CAF 47
Interim summary 50
2.4 Main factors affecting L2 oral performance 54
2.4.1 Study Abroad 56
2.4.2 Other factors 62
2.4.3 Previous work on L2 Chinese oral development 67

CHAPTER 3: THE STUDY **69**

3.1 Research objectives 69
3.2 Participants 71
3.3 Methodology 72
3.3.1 Instruments 73
3.3.2 Procedure and Ethics 74
3.3.3 The study abroad period 75
3.4 The CAF measures used in this study 76
Accuracy measures 76
Fluency measures 77
Complexity measures 79
Lexical complexity 79
Syntactic complexity 80
3.5 Data transcription coding and trimming 81
3.5.1 Data trimming 81
Data trimming 1: editing out interlocutor speech 81
The dataset 83
Transcribing the pruned data 84
3.5.2 Coding accuracy 84
3.5.3 Coding fluency 85
3.5.3.1 Pause marking 86
3.5.3.2 Repairs and repetitions 91
3.5.4 Coding Complexity 92
3.5.4.1 Coding lexical complexity 93
3.5.4.2 Coding syntactic complexity 96
3.5.5 Summary of CAF measures analysed in this study 99
3.6 Research questions 100
3.7 Predictions 101
3.7.1 CAF development during pre-SA and post-SA (S1-S2) 102
3.7.2 CAF development during FI at home context (S2-S3) 103
3.7.3 Correlations between CAF measures during pre- and post-SA (S1-S2) 103
3.7.4 Correlations between CAF measures during FI at home context (S2-S3) 105

CHAPTER 4: DATA ANALYSIS AND RESULTS 107

4.1 SA related results (S1-S2) 107
4.2 FI at home maintenance results (S2 - S3) 108
4.3 Correlations between CAF constructs related to SA effects (S1-S2) 110
4.4 Correlations between CAF related to FI at home maintenance (S2-S3) 111
4.5 Individual performance during the 28 month period (S1-S3) 112

CHAPTER 5: DISCUSSION 116

5.1 SA effects on oral performance (S1-S2) 116
5.1.1 Development of complexity 116
Syntactic complexity 116
Lexical complexity 117
5.1.2 Development of accuracy 118
5.1.3 Development of fluency 119
5.1.4 Summary of CAF development 121
5.2 The effects of FI at home maintenance on oral performance (S2-S3) 122
5.2.1 Development of complexity 122
Syntactic complexity 122
Lexical complexity 123
5.2.2 Development of accuracy 124
5.2.3 Development of fluency 125
5.2.4 Summary of CAF development 127
5.3 Correlations between CAF sub-constructs related to SA effects (S1-S2) 128
5.3.1 Correlations between subconstructs within CAF after SA 128
5.3.2 Correlations between complexity, accuracy, and fluency after SA 131
Complexity and fluency 132
Summary of relationships between CAF constructs and subconstructs within CAF (S1-S2) 132

5.4 Correlations between CAF related to FI at home maintenance 133
5.4.1 Correlations between subconstructs within CAF related to FI at home maintenance 134
5.4.2 Correlations between complexity, accuracy, and fluency related to FI at home maintenance 135
5.5 Evaluating CAF indicators 136

CHAPTER 6: CONCLUSION 139

6.1 Summary of findings 139
6.2 Teaching and learning implications 143
6.3 Limitations and suggestions for future work 145

REFERENCES 149

APPENDIXES 172

Appendix A: Normality of 14 CAF measures 172
Appendix B: Scores per participant 173
Appendix C: Markings per participant 177
Appendix D: HSK Scores per participant 178
Appendix E: Tasks used in S1, S2 and S3 179

Declaration

This is to certify that the work I am submitting is my own and has not been submitted for another degree, either at University College Cork or elsewhere. All external references and sources are clearly acknowledged and identified within the contents. I have read and understood the regulations of University College Cork concerning plagiarism and intellectual property.

Signed: _____ Rongrong Guo _____

Acknowledgments

Looking back on the long journey with ups and downs, I'm so grateful for being supported and encouraged by my supervisors, family, peers, and friends throughout the whole process as a doctoral student at University College Cork. First of all, I sincerely thank my previous supervisor, Associate Professor Carlotta Sparvoli, who approved me to be a Ph.D. candidate and guided me to the threshold of the research. I must also thank my present supervisor, Dr. Yanyu Guo, for her steadfast and unlimited support during the most challenging period before I completed this study. Without her ever-present help, the completion of the dissertation would not be possible. Second, I would like to give thanks to the head of Asian Studies, Professor Kiri Paramore, who has been constantly supporting me during the journey. Then, I have been so blessed by my friends who have been continually feeding me tasty food. Those are Bonnie, Ide, Lu, Min, Shuo, Tong, Xiaozhen, Yun, Xin & Yi, Kathy & Ghee, Qin & Guang. I would also like to thank those who have encouraged me when I was going through hardships, Ann, Janaki, John, Jo, Julia, Sinéad, Tina, Zejing, Eddie, and Marguerite. Moreover, so much gratitude to my family, my brother, and my sister-in-law who have been taking care of my parents which enabled me to work on my dissertation wholeheartedly. All of them have held me accountable through the ups and downs, which have made me and my life better and have brought me to the successful completion of this dissertation. Further, "Rejoice always, pray without ceasing, give thanks in all circumstances; for this is the will of God in Christ Jesus for you (1 Thessalonians 5:16-18)." This is one of many bible verses that motivated me to continue with perseverance. Last but not least, I want to express my gratitude to two external examiners, Dr. Zhiyan Guo and Dr. Shanshan Yan, for taking their time to review the thesis and offer helpful comments.

Abbreviations

ALFP: The average length of filled pause
ALSP: The average length of silent pause
AS-unit: Analysis of Speech Unit
CAF: Complexity, Accuracy and Fluency
CSL: Chinese Second Language
C-unit: Communication Unit
MLR: Mean Length of Runs
FI: Formal Instruction at home
FP100: The number of filled pauses per 100 syllables
HSK: Hanyu shuiping Kaoshi, Chinese Proficiency Test
IM: Intensive domestic immersion context
RR100: The number of repairs and repetitions per 100 syllables
SA: Study Abroad
SR: Speech Rate
SP 100: The number of silent pauses per 100 syllables
TTR: type-token ratio
T-unit: Minimal Terminable Unit

Abstract

The semi-longitudinal study explores the impact of learning environments on the oral Complexity, Accuracy, and Fluency (CAF) of adult English-speaking learners of Chinese, investigating when and how the oral performance of instructed L2 learners changes in two contexts: Formal Instruction at-home (FI) and Study Abroad (SA). Moreover, the study discusses relationships between the CAF constructs and those between the sub-constructs, to assess the oral performance of instructed L2 learners. Two widely documented theoretical hypotheses on attention allocation and tasks, the Trade-off Hypothesis (Skehan, 2009; Skehan & Foster, 2012) and the Cognition Hypothesis (Robinson, 2001; 2003; 2005; 2011), are examined with data collected from ten English-speaking undergraduates of an Irish university from three oral tests across 28 months (including 10-month of SA experience). The results show that the students benefit from SA in terms of syntactic complexity (subordination and length of the unit), lexical sophistication as well as speed fluency with a slight decrease in dysfluency at the cost of accuracy. This is attributable to the study abroad experience as well as rehearsed monologue tasks (cf. Wright, 2020) that the participants took in the study. The SA benefits oral gains in terms of speech fluidity, syntactic complexity (length and subordination), and lexical sophistication. The factor of task design must also be taken into consideration when L2 learners' oral gains are evaluated. After coming back to the FI context for six months, a significant decrease, in general, is observed regarding FI at home maintenance on the oral gains that obtained from the SA experience. However, lexical variety reveals significant improvement. The findings suggest that learners in the FI context tend to concentrate on learning vocabulary (diversity) and syntactic complexity, at the expense of fluency, accuracy (Juan-Garau & Pérez-Vidal, 2007) and lexical sophistication in this study. Generalized from the analysis during the pre and post-SA periods, trade-off effects are observed prevailingly between CAF constructs (in particular between complexity and accuracy, and between accuracy and fluency), while simultaneous improvements are present within CAF, in particular, and between speed and breakdown within fluency, and between syntactic complexity and lexical sophistication within complexity. These results (trade-off effects) confirm Skehan's predictions that, tensions are found between control (accuracy) and risk-taking (complexity), and between focusing on meaning (fluency) and form (accuracy) (Skehan, 2009; Wang & Skehan, 2014). Task characteristics are attributed to the finding because different characteristics support different performance areas (Skehan & Foster, 2012): pre-planning is argued to elicit greater complexity and fluency. For the inter-relationship between CAF measures, after learners returned to FI at home context for six months, their performance, in general, supports trade-off effects between lexical diversity and syntactic complexity as well as between lexical diversity and fluency. The results contribute to the trade-off hypothesis that, tensions can be found between subconstructs within CAF (complexity). The prioritization of attentional resources is determined by task types and learning contexts, revealing that vocabulary development is at the cost of syntactic complexity and fluency in FI context (Juan-Garau & Pérez-Vidal, 2007). Based on the findings, the study also provides pedagogical implications and recommendations for the development of L2 Chinese oral performance in a university teaching setting.

Keywords: oral performance, L2 Chinese, semi-longitudinal design, Study Abroad (SA), Formal Instruction At-home (FI), task design

List of tables

Table 1. Review of lexical complexity measures used in L2 studies 19
Table 2. Accuracy measures in L2 Chinese speaking studies 25
Table 3. Fluency measures used in L2 Chinese speaking studies 27
Table 4. Comparisons between HSCDD (2001), CSCLCIE (2010), New HSK (2012) 32
Table 5. Lexical complexity measures in L2 Chinese speaking studies 32
Table 6. Syntactic complexity measures in L2 Chinese speaking studies 34
Table 7. Investigation of factors on oral performance in L2 studies 55
Table 8. The profile of participants 72
Table 9. Three stages of data collection 73
Table 10. Data per participant and the total 83
Table 11. The transcription standard 84
Table 12. Categorisation of lexical errors 84
Table 13. CAF measures analysed in this study 99
Table 14. SA effects on oral performance (S1-S2) 107
Table 15. FI at home effects on oral performance (S2 vs. S3) 109
Table 16. Normality distribution for each CAF measure 172
Table 17. Scores for each measure per participant 173
Table 18. Markings per participant rated by the course's teachers at S1, S2 and S3 177
Table 19. HSK Scores per participant 178

List of Figures

Figure 1. Lexical complexity at different levels (Bulté & Housen, 2012:28) 16
Figure 2. Different levels of grammatical complexity (Bulté & Housen, 2012:27) 22
Figure 3. Annotation of silent pauses in Praat 88
Figure 4. Extraction of silent pauses in Praat 88
Figure 5. Silent-only file created in Praat 89
Figure 6. Labelling filled pauses 90
Figure 7. The filled pauses with start and end times in one speech sample 91
Figure 8. Using THULAC for the first step of segmenting words in a script 95

Chapter 1: Introduction

Under the scope of oral assessment, using the CAF framework to assess L2 learners' oral performance is a growing area. The main factors, in particular, learning contexts, which affect oral development assessed by the Complexity, Accuracy, and Fluency (CAF) measures as well as the relationship between CAF measures have received sustained attention in the L2 field (see Section 1.1). However, the effects of learning contexts have not been widely applied in assessing L2 speaking Chinese. Additionally, the effects of learning environments on relationships between the CAF components and those between the sub-constructs within CAF in the L2 speaking Chinese have rarely been investigated (see Section 1.2). Therefore, this semi-longitudinal study will investigate the effects of learning contexts (formal instruction at home (FI) and study abroad (SA)) on the oral CAF of adult English-speaking Chinese learners to analyse the effects of learning contexts on relationships between the CAF components and those between the sub-constructs within CAF in the oral performance of adult English-speaking Chinese learners.

1.1 L2 oral proficiency assessment

L2 learners' oral proficiency is normally assessed by means of two general approaches: holistic scales and analytical scales (Tonkyn, 2012; Jin & Mak, 2013; Metruk, 2018, 2019; Namaziandost & Ahmadi, 2019). Holistic scoring gives an overall score based on the performance as a whole (Namaziandost & Ahmadi, 2019), which can be described as an impressionistic or global scale (Pan, 2016). This means evaluating the overall performance in a qualitative manner, considering the performance as a whole (Griffith & Lim, 2012). For example, speaking scores on TOEFL are graded holistically. Such scores are given based on an overall impression of students' abilities. Normally a 4 or 5-point scale is used. Students' performance is graded using the categories "excellent", "good", "fair", and "poor" (Griffith & Lim, 2012). In terms of analytical scoring, this refers to the separate salient features of performance where each aspect is graded individually (Metruk, 2018). The analytical approach assesses some discrete features of performance and mixes those scores to produce an overall score (Taylor & Galaczi, 2011). The assessment of speaking proficiency comprises different subcategories across

research, including fluency (e.g., Council of Europe, 2001; Pan, 2016), vocabulary (e.g., Metruk, 2018; Namaziandost & Ahmadi, 2019; Pan, 2016), accuracy (e.g., Council of Europe, 2001; Metruk, 2018; Pan, 2016); pronunciation (e.g., Metruk, 2018; Namaziandost & Ahmadi, 2019)); and grammar (e.g., Namaziandost & Ahmadi, 2019).

Together, these two scoring approaches can achieve a comprehensive assessment as they supplement each other (Metruk, 2018; Namaziandost & Ahmadi, 2019). However, the two assessment scales lack specification concerning the level of students' speaking skills, and underestimate some essential and general aspects of learners' performance (i.e., complexity, accuracy, and fluency constructs) (Alghizzi, 2017). Therefore, the key question that arises is, to what extent do certain features of learners' oral performance contribute to their proficiency levels (Riggenbach, 1991; Tonkyn, 2012). To answer this question, it requires L2 proficiency to be specified in an objective, quantitative and verifiable fashion (Nihalani, 1981). Many SLA scholars claim that L2 proficiency is not a unitary construct, but rather it is a multi-componential one (Housen, 2012; Housen & Kuiken, 2009). In this sense, components of L2 oral proficiency can be reliably captured by the concepts of complexity, accuracy, and fluency (CAF) (Housen & Kuiken, 2009; Housen, 2012; Kuiken, 2019).

Using the CAF constructs to assess L2 learners' oral performance is a burgeoning field. As dependent variables, the three dimensions of CAF have often been used together to evaluate the effects of other factors on L2 performance, such as age, instruction, aptitude, learning contexts and task types (Housen et al., 2012). For instance, Mora (2006) evaluated the effect of age on the oral fluency development of young L2 English learners and Freed (1995) conducted an empirical study that compared the oral fluency development of two groups of L2 French students studying abroad and at home respectively. Moreover, increasing attention has been paid to reveal cognitive features of CAF in L2 research. CAF has been characterised as a primary phenomenon of psycholinguistic processes and mechanisms in underlying L2 acquisition and processing (Housen et al., 2012). For instance, to understand the effects of psycholinguistic factors in L2 fluency development, Towell and Dewaele (2005) conducted a four-year longitudinal study which assessed a group of twelve L2 advanced French students' linguistic knowledge and fluency. It concluded that individual differences and the age of the L2 learners' result in different language production.

As an important and distinct dimension of L2 proficiency and performance, CAF has been supported empirically and theoretically (Housen et al., 2012). Theoretically, the three constructs of CAF have been acknowledged as revealing the major stages of change in the underlying L2 system: (i) the internalisation of new L2 structures or greater complexity, which is implied by the development of more complicated and more sophisticated L2 knowledge systems; and ii) the modification of L2 knowledge, as shown by learners' reconstruction and improvement of their L2 knowledge, including the deviant or non-target-like facets of their interlanguage so that the learners produce more accurate L2 structures; and (iii) the incorporation and proceduralisation of L2 knowledge, which is displayed by greater performance control such as higher fluency. The higher fluency results from routinisation, lexicalisation, and automatisation of more complex L2 elements (Housen et al., 2012; Alghizzi, 2017). Empirically, each dimension of the CAF triad is identified as recognisably different in L2 performance (Skehan & Foster 1997, 2001; Norris & Ortega 2009). For instance, Norris and Ortega (2009) reviewed the measurement of syntactic complexity in L2 production to illustrate the need for a more organic approach to investigate CAF in instructed SLA. On the basis of such research, it can be concluded that the CAF constructs, as a universal measure, can be applied to all possible learners and contexts. Furthermore, CAF constructs have always been measured in a specific setting for specific purposes, and they have been widely applied to assess the oral skills of language learners and to depict their proficiency level and their progress in language learning. However, the constructs remain controversial (Housen & Kuiken, 2009). For instance, several challenges to CAF have been identified by Housen and Kuiken (2012): 1) the definitions of the three constructs; 2) the nature of their cognitive, linguistic, and psycholinguistic correlates and underpinnings; 3) the interconnections and interdependencies among the CAF components in L2 performance and development; 4) the operationalisation and measurement of CAF in empirical research; and 5) the identification of factors which affect the manifestation and development in L2 use and learning.

The main factors which affect language development assessed by the CAF measures have received sustained attention in the L2 field (Housen et al., 2012). Two categories of factors are often investigated: internal linguistic features and external factors (Housen et al., 2012). Comparatively, external factors, in particular, contextual factors (e.g., Formal Instruction At-Home (FI), Study Abroad (SA) have received extensive attention. The explosion of SA research over the last two decades has been stimulated by the global popularity of SA programmes (Devlin, 2019; Tullock & Ortega, 2017;

Yang, 2016). One way to explore the benefits of SA is through studies conducted on SA-FI comparisons (e.g., Collentine, 2004; Segalowitz, Freed, & Collentine, 2007). However, such studies have been relatively few in number, demonstrated by the fact that there has only been a relatively small amount of research on this topic (Yang, 2016). Considering the major difference between the two contexts (e.g., SA is viewed as meaning focused learning conditions. In contrast, FI is more form-focused (McManus et al, 2021)), it has therefore been claimed that the key is not to compare the learning contexts with one another, but rather that each context should be comprehended in its own right (Sanz, 2014).

In terms of the benefits and outcomes of SA, a meta-analyis of 467 SA studies revealed mixed findings (Tullock & Ortega, 2017; Yang, 2016). Such findings might result from the interaction between internal factors related to learners and external factors associated with contexts (Tullock & Ortega, 2017). The four major variables that have been most widely evaluated across SA studies are: the pre-departure proficiency of SA students (e.g., Llanes, 2011; Valls-Ferrer & Mora, 2014), individual differences (e.g., DeKeyser, 2014; Sanz, 2014), the age of participants (e.g., Llanes & Serrano, 2017), and the duration of the SA programmes investigated (e.g., Lara et al., 2015; Llanes & Serrano, 2011; Serrano et al., 2016). The first three factors are associated with learners, while the programme length is directly related to SA itself. When learners are on a SA sojourn, internal factors located in the learners interact with external factors located in the context, and this interaction leads to the learners' language gains. For example, it is widely agreed that a certain onset proficiency level is necessary to enable learners to take full advantage of the SA period (e.g., Valls-Ferrer & Mora, 2014). Likewise, some researchers have revealed that individual differences can play a significant role in learners' linguistic gains during SA (e.g., Sanz, 2014). Age has also been shown to be a variable that can contribute to learners' language gains (e.g., Llanes & Serrano, 2017). Finally, it is commonly believed that greater SA benefits result from a longer stay (e.g., Llanes, Baro & Serrano, 2011). Moreover, the impact of methodological factors (i.e., task type) on SA has been investigated. For instance, some studies have explored the impact of task type on oral development during the SA period (e.g., Wright, 2018, 2020). They reveal that task-load effects might override SA impact, in particular, the rehearsed/planned monologue task (e.g., Wright, 2018, 2020). However, because the aforementioned factors vary across SA studies, this has contributed to the inconsistency of SA findings.

Another challenge for CAF research relates to the relationship between CAF constructs and subconstructs (Housen et al., 2012). Existing L2 studies which assess L2 learners' performance, in general, focus on two current competing theoretical hypotheses concerning attention allocation and tasks: the Trade-off Hypothesis (Skehan, 2009; Skehan & Foster, 2012) and the Cognition Hypothesis (Robinson, 2001, 2003, 2005, 2011). Moreover, there is an emerging trend to employ Dynamic Systems Theory (DST) to explore the correlations among CAF components because CAF itself is considered to be a dynamic system (Norris & Ortega, 2009). DST focuses on change and makes change central to theory and method. In particular, it attempts to examine dynamic and variable patterns of L2 language development (Larsen-Freeman & Cameron, 2008). For instance, Spoelman and Verspoor (2010) examined the acquisition of Finnish in a longitudinal study which applied DST. The results showed that there are complex interactions in the development of accuracy and complexity measures consisting of peaks and regressions and progress and backsliding which indicate a non-linear L2 development pattern.

1.2 Speaking assessment of L2 Chinese

In the context of L2 Chinese speaking studies, both holistic and analytic rating methods have been implemented for L2 oral Chinese assessment (Liao, 2018). For instance, the HSK (Hanyu Shuiping Kaoshi, Chinese Proficiency Test) Speaking Test (HSKK) and the Youth Chinese Test (YCT) apply holistic scoring; whereas Chinese oral tests, such as the Spoken Chinese Test (SCT), use analytic scoring. The SCT score report provides an overall score and five analytic sub-scores which describe the test-taker's competency in spoken Chinese. The overall score is a weighted average of the five sub-scores: grammar, vocabulary, fluency, pronunciation, and tone (Li & Li, 2014). Using analytic rating methods, L2 Chinese studies have investigated Chinese speaking skills in different categories, which include pronunciation, vocabulary, grammar, and fluency (e.g., Du, 2013; Jin & Mak, 2013; Wang, 2002). For instance, Jin and Mak (2013) found seven distinguishing features (target-like syllables, speech rate, pause time, word tokens, word types, grammatical accuracy, and grammatical complexity) under four categories - pronunciation, fluency, vocabulary, and grammar - which have been widely employed in evaluating L2 Chinese speaking performance.

The assessment of L2 Chinese speaking has also evolved from subjective perceptions, such as the HSKK (Liu, 1997), to a more analytical and multi-componential approach (Liao, 2018). Starting

from the first decade of the twenty-first century, the CAF framework has been increasingly widely used in TCSOL (e.g., Du, 2013; Wright & Zhang, 2014; Zhai & Feng, 2014, Wright, 2020; Wu, 2014; Ye, 2015; Zhou, 2015; Ding & Xiao, 2016; Chen & Zhou, 2016; Liu, 2017). However, not all studies have applied the three dimensions of the CAF framework. Instead, some studies have used CAF-related measures and other specific analyses at the same time (e.g., Guo, 2007). Moreover, there has been no clear distinction between CAF constructs in some L2 Chinese studies. For example, accuracy was merged into the fluency domain (e.g., Zhai & Feng, 2014). Also, among the three constructs of CAF, fluency has received the most attention when measuring L2 Chinese learners' oral performance (e.g., Du, 2013; Feng, 2018; Wang, 2018; Wright & Zhang, 2014; Wright, 2020). In certain other research, some specific CAF subcomponents have been investigated, such as one facet of dysfluencies, filled pauses (e.g., Wu, 2008; Liu, 2019; Wu & Jin, 2020).

Concerning the speech development of L2 learners of Chinese investigated by CAF measures, compared to accuracy and complexity, fluency is more likely to be developed at a higher rate (Chen, 2012). Furthermore, in terms of oral accuracy and complexity, significant improvement might be achieved when learners are at an advanced level (Ye, 2015). It is anticipated that L2 learners of Chinese achieve and maintain desired fluency at the cost of accuracy and complexity in speaking (Chen, 2012; Zhai & Feng, 2014; Ye, 2015; Liao, 2018). However, the proficiency levels of participants in the literature are defined either by their institutional status (Zhai & Feng, 2014; Ye, 2015; Chen, 2015; Ding & Xiao, 2016) or by the period of their learning instruction (Chen, 2015; Liu, 2017). Participants' proficiency levels vary among L2 Chinese studies, which echoes the L2 field (Wu & Ortega, 2013). As a result, it is problematic to compare the findings of the studies that have investigated the oral performance of L2 learners of Chinese. Concerning methodology, the majority of the existing studies are cross-sectional studies. However, there is an increasing trend to conduct longitudinal research in this area, most of which are case studies (e.g., Feng, 2018; Shi, 2002; Wu, 2017; Zhou, 2016).

Similar to other L2 studies, L2 Chinese studies have paid attention to the main factors which affect language development assessed by CAF measures. Contextual factors, in particular, study abroad as a learning context, have received increasing attention by L2 Chinese researchers (e.g., Du, 2013; Wright & Zhang, 2014). Moreover, some studies also investigate the impact of other factors on L2 speaking performance during study abroad, such as the quality of interaction (e.g., Diao, Donovan,

& Malone, 2018), and task-type (e.g., Wright, 2020). However, there are only limited empirical studies which investigate L2 Chinese speaking development during study abroad (e.g., Du, 2013; Wright & Zhang, 2014; Wright, 2018, 2020). Furthermore, no empirical studies have looked at the L2 Chinese speaking development during study abroad and the retention of SA effects during Formal Instruction back home. Measuring the effects of learning contexts (SA and Formal Instruction) on L2 Chinese speaking development has not been extensively applied. The effects of learning contexts on relationships between the CAF constructs and those between the sub-constructs, seldom ever studied in L2 Chinese oral performance, will be analysed in the current study.

1.3 The structure of the thesis

To measure the oral development of English-speaking learners of Chinese, in this study, Chapter 2 firstly reviews the existing L2 speaking studies concerning definitions and measures for assessing accuracy, complexity, and fluency in language performance. The two widely documented theories concerned with capturing the effect of task features and conditions on L2 learners' CAF production: Skehan's (1998, 2009a) Limited Attention model and Robinson's Cognition Hypothesis (2001a, 2001b, 2005) will be outlined. The interrelationships between CAF constructs and subconstructs within each construct predicted by the two models will also be provided. Considering the scarcity of research which investigates the effects of learning contexts on learners' oral CAF production, the effects of the main factors (i.e., learning contexts) on L2 oral development will also be explored. In particular, the effects of Study Abroad (SA) on L2 Chinese oral development will be reviewed. Chapter 3 presents the research questions that emerge from the literature review as well as the study's predictions. This chapter also outlines the study's methodology, including instruments and procedure. The data coding criteria as well as the CAF measures analysed in this research will also be presented. Chapter 4 reports the results of the study in terms of the effects of SA and Formal Instruction (FI) at home on L2 Chinese oral development. In Chapter 5, the development of the CAF measures during pre- and post-SA as well as during FI at home after SA will be discussed to investigate the impact of learning contexts on L2 Chinese speaking development. The correlations between CAF constructs and subconstructs within each construct of CAF will be interpreted to examine the trade-off effects hypothesis in language performance. Chapter 6 concludes the study, restates the main findings and offers some suggestions for future research. It also considers the limitations of the current research.

Chapter 2: Measuring Complexity, Accuracy, and Fluency

The notions of complexity, accuracy, and fluency (CAF) can consistently capture components of L2 oral performance (Housen & Kuiken, 2009; Housen, 2012; Kuiken, 2019). Therefore, the CAF components are frequently used to assess L2 learners' oral performance (e.g., Skehan, 2003; Robinson & Ellis, 2008). For this reason, CAF framework will be employed to assess the L2 speaking Chinese in this study.

This chapter begins with the definitions of CAF constructs and their applications (See Section 2.1) in L2 speech performance assessment. The purpose of this is to present the current state of L2 oral assessment. Due to Chinese-specific features, CAF measures used in L2 speaking Chinese studies (See Section 2.2), will next be addressed to provide the rationale for the CAF measures that are used in this research. The following section (See Section 2.3) reviews the theoretical hypotheses and empirical studies concerning how to identify correspondences between CAF constructs and between subconstructs with CAF. This has not been thoroughly examined in L2 Chinese oral performance. Finally, the main factors, in particular, learning contexts (study abroad and Formal instruction at home) which affect L2 oral performance identified in the literature will be reviewed (See Section 2.4). This section seeks to show where L2 speaking Chinese studies are lacking. Unlike other L2 Research, very few empirical studies have looked at L2 Chinese speaking development when studying abroad (e.g., Du, 2013; Liu, 2009; Wright & Zhang, 2014; Wright, 2018, 2020). These four sections lead to the research gap, which is the effect of learning contexts (study abroad and Formal instruction at home) on the speaking development of L2 Chinese learners as well as the relationship between CAF constructs and between subconstructs with CAF. These are the two areas where the study aims to contribute to.

2.1 Definitions and measurement of CAF constructs

In L2 research, the measures and indicators used to investigate L2 writing and speaking development vary greatly among studies. Apart from the holistic and subjective ratings, quantitative measures (frequencies, ratios, formulas) of general or specific linguistic properties of L2 production have been analysed in order to achieve more precise and objective accounts of an L2 learner's level within each dimension and sub-dimension of proficiency (Housen & Kuiken, 2009). Regarding CAF

measures, both general and specific measures have been analysed to investigate both written and spoken data to explore L2 development (Skehan, 2003; Robinson & Ellis, 2008; Housen & Kuiken, 2009). General measures can be applied to assess the data elicited by a wider variety of tasks, but they cannot fully capture minor differences that finer-grain analysis can. On the other hand, specific measurements can capture small differences in data related to a specific task or population, such as targeting accuracy or complexity in a specific area (e.g., the article system), or using more general measures (Skehan, 2003). However, using specific measures can limit generalisability (Vercellotti, 2012). The following section will review the general measures of CAF constructs that are used to investigate L2 speaking performance in particular.

2.1.1 Accuracy

Definitions of accuracy

Accuracy is considered to be the most straightforward and most internally consistent construct among the CAF triad (Housen & Kuiken, 2012; Norris & Ortega, 2009). There has been extensive debate in existing studies about the definition of accuracy. For instance, for some scholars, accuracy concerns how well language is produced in relation to the rule system of the target language (Skehan, 1996a). Later in Skehan's studies, accuracy relates to L2 learners' implicit language system, where accuracy is a learner's belief in norms, and refers to performance which is native-like through its rule-governed nature (Skehan, 1996b). Accuracy is also interpreted as L2 learners' capacity to avoid errors in production (e.g., Wolfe-Quintero et al., 1998; Skehan & Foster, 1999). However, others assert that accuracy is an explicit performance of the internal L2 system concerning L2 use and that L2 learners' strategy towards L2 use reveals their underlying cognitive coping mechanism (Ellis, 2003; Ellis & Barkhuizen, 2005).

Recent studies claim that accuracy refers to the extent to which an L2 learner's performance deviates from a norm (Housen et al., 2012), and the extent to which a person adheres to a set of rules (Norris & Ortega, 2009). This has been followed by subsequent research on defining accuracy as involving the extent to which a learner's language production aligns with target-language norms (Juan-Garau, 2014). Despite the lack of a unified definition, it appears that researchers have reached a consensus on one point, which refers back to Polio (2001) who claimed that linguistic accuracy is a broad term that is generally associated with the absence of errors, which may or may not include word choice,

spelling or punctuation errors. Based on this account, accuracy is also termed correctness, coping with deviations from the norm, which are normally characterised as errors (Housen & Kuiken, 2009).

Furthermore, another question has arisen: whether or not the criteria for evaluating and identifying errors should be tuned to prescriptive standard norms (as demonstrated by an ideal native speaker of the target language), or whether non-standard and non-native usages are acceptable (Housen & Kuiken, 2009). Concerning how to determine the criteria of evaluating accuracy, Housen et al. (2012) suggested interpreting accuracy as acceptability and appropriateness in CAF in a broader way. Therefore, it is clear that although accuracy is often argued to be conceptually simple, it is actually highly problematic in both its interpretation and application to assess L2 data (Housen et al., 2012).

Measuring accuracy

Despite the inconclusive arguments about its definition and evaluation criteria, the common element across these studies concerns the relationship of accuracy to the analysis of errors. In one early study, Bardovi-Harlig and Bofman (1989) analysed errors at three levels: syntactic, morphological, and lexical. Syntactic errors are those which relate to word order and sentence conjunctions; morphological errors consist of those in nominal morphology, verbal morphology, determiners, articles, and prepositions. Finally, lexical errors relate to idiomatic expressions and word choice. These three categories have been adopted by later studies. For example, Llanes and Muñoz (2009) examined L2 English learner's language oral gains during a short-term SA programme (3–4 weeks). Two quantitative measures were used: the ratio of error-free clauses and the average number of errors per clause. Learners' errors were classified into three types: morphological errors, syntactic errors, and lexical errors. The results did not show significant gains in oral accuracy. A recent study by Kafipour and Khojasteh (2011) compared the written performance of native English speakers with L2 English learners at three levels, morphological, syntactic, and semantic. The results revealed that L2 English learners made similar types of errors to native learners.

Among the three levels used to investigate accuracy, there is a dearth of morphological empirical studies (Juan-Garau, 2014). Operationally, in the studies on L2 performance assessment, accuracy has been normally measured by two types of general measures from a quantitative perspective (Vercellotti, 2012). The ratio of error-free measures have been analysed based on the production units

selected across studies, such as, the ratio of error-free clauses (e.g., Skehan & Foster, 1999; Wu, 2017; Vercellotti, 2017); the ratio of error-free Analysis of Speech Units (AS-units) (e.g., Tonkyn 2012; Ferrari, 2012; Wu, 2017); the ratio of error-free Terminable Units (T-units) (e.g., Zhai & Feng, 2014); the percentage of correct verb forms (e.g., Yuan & Ellis, 2003); the ratio of error-free Chinese sentences (e.g., Wang, 2002; Jin & Mak, 2013; Zhai, 2011; Ye, 2015); and the ratio of correct pronunciations (e.g., Zhai & Feng, 2014). The measures concerning the frequency of errors have also been investigated, such as, the number of errors per word (e.g., Takiguchi, 2004; Koizumi, 2005); the number of errors per clause (e.g., Llanes & Muñoz, 2009); the number of errors per AS-unit (e.g., Mora & Valls-Ferrer, 2012); errors per-T-unit (e.g., Bygate, 2001); lexical errors (e.g., Foster & Skehan, 1996); and errors per 100 words (e.g., Li, 2010; Mehnert, 1998).

For L2 Chinese studies, some L2 Chinese researchers have replaced T-units with Chinese sentences as the basic production unit due to the shortage of subordinate clauses in Chinese (e.g., Wang, 2002; Jin & Mak, 2013; Zhai, 2011; Ye, 2015). Moreover, some Chinese language-specific measures have been used, for example, the ratio of error initials/finals/tones (e.g., Chen, 2015a, 2015b), and tonal accuracy (Kim et al., 2015). Therefore, as illustrated, based on the segmentation production unit applied in different studies, accuracy has normally been measured by three types of errors (syntactic, morphological, and lexical) and two types of general measures (the ratio of error-free measures and the frequency of errors).

2.1.2 Fluency

Among the CAF triad, there is less agreement in the field of applied linguistics with regard to fluency and complexity compared to accuracy (Housen & Kuiken, 2009, 2012).

Definitions of fluency

Fluency, in general, is conceptualised either in a broad or a narrow sense (Lennon, 1990). In the broader sense, fluency concerns non-native speakers' overall language proficiency (Bosker et al., 2013). Conversely, fluency in the narrow sense, is a component of speaking proficiency. Some definitions have been proposed which are process-based (Alghizzi, 2017). For example, for some scholars, fluency is the speaker's ability to mobilise an interlanguage system to communicate in real-time (Skehan, 1996b). Conversely, for others, fluency relates to the underlying encoding process as the

production of language in real-time without undue pausing or hesitation (Ellis & Bakhuizen, 2005). Builded on previous definitions, fluency is considered as the ability to communicate one's intended meaning effortlessly, smoothly and it connects language use with no or little disruption (Tavakoli et al., 2016). Some definitions are interpreted by describing product-based performance from a qualitative perspective (Alghizzi, 2017). For instance, fluency is seen to be related to temporal features. Under this definition, fluency is regarded as producing speech at the tempo of native speakers, which is not impeded by silent pauses, hesitations, filled pauses ('ers' and 'erms'), self-corrections, repetitions, and false starts (Lennon, 1990). Therefore, the product- and process-based definitions differ among researchers, in this narrow sense (Kormos & Dénes 2004).

In recent studies, L2 fluency has been sub-divided into utterance fluency, cognitive fluency, and perceived fluency (Segalowitz, 2010, 2016). The difference between these three sub-categories has been distinguished in the following terms: utterance fluency relates to the features of utterances that reflect speakers' cognitive fluency; cognitive fluency is the efficiency of the operation of the underlying processes responsible for the production of utterances; and perceived fluency is the inferences listeners make about speakers' cognitive fluency based on their perceptions. With the three notions of fluency, utterance fluency can be objectively investigated by measuring the temporal aspects of a speech sample (De Jong et al., 2013). Moreover, several studies assert that the notion of utterance fluency, which is largely temporal, is also a construct with several aspects. For instance, Skehan (2003) further divided it into three components, namely, breakdown fluency (e.g., pause frequency), speed fluency (e.g., speech rate), and repair fluency (e.g., the frequency of false starts, repairs, reformulations) due to different speech features.

Considering the various definitions and interpretations of fluency, it is vital to specify the sub-components with specialised definitions when investigating the construct. This study investigates the domain of utterance fluency (Segalowitz, 2010, 2016), which further consists of three sub-components: breakdown fluency, speed fluency, and repair fluency (Tavakoli & Skehan 2005; Tavakoli, 2016).

Measuring utterance fluency

As indicated before, three distinct sub-constructs of utterance fluency have been identified (Skehan, 2003; Tovakoli & Skehan, 2005; Norris & Ortega, 2009): (1) breakdown fluency, which is

assessed by silence-related measures; (2) speed fluency, which can be captured by rate- and time-related measures; and (3) repair fluency, which is measured by self-correction measures (Norris & Ortega, 2009).

Breakdown fluency

Breakdown fluency refers to the ongoing flow of speech and can be measured by counting the number and length of filled and unfilled pause (De Jong et al., 2013). With regard to defining a pause, Riggenbach (1991) suggested that pauses shorter than 0.2 seconds should be considered as micro-pauses, and many subsequent studies have adopted this measure as the threshold of a pause (Kormos & Dénes, 2004). However, it is important to note that there are some other cut-off points for pause length in the literature, such as 0.25 seconds (Kormos & Dénes, 2004; De Jong et al., 2013, 2016), 0.28 seconds (Towell, 2002), 0.3 seconds (Raupach, 1980; Tonkyn, 2012), 0.4 seconds (Derwing et al., 2004; Tavakoli & Skehan, 2005), and even 3 seconds (Fulcher, 1996). The difference in length of a pause varies among studies and this issue deserves further discussion because setting the threshold for a pause is the starting point to investigate breakdown fluency. This will be discussed in more detail later in this chapter, and the cut-off point of a pause that is used in this study will also be provided (See section 2.2.2.2 Pause marking).

To measure breakdown fluency, the frequency and duration of pauses are typically analysed. The former has often been measured by the number of silent and filled pauses, while the latter is normally analysed using the mean duration of silent and filled pauses (De Jong, 2013, 2016). With regard to the reliability of the frequency and duration of pauses in assessing breakdown fluency, Bosker et al. (2013) analysed three representative breakdown measures: the number of silent pauses per second of spoken time, the number of filled pauses per second of spoken time and the mean length of silent pauses, to investigate the contributions of three fluency aspects (pauses, speed and repairs) to perceived fluency. They concluded that pause frequency is likely to be a more reliable predictor of L2 breakdown fluency than pause duration. However, this issue requires further investigation. Because pauses (silent and filled) can be complex as fluency measures (Wright, 2020). They are very likely to be affected by speakers' individual patterns of speech. For instance, a speaker may tend to pause or repair more in their L1 than others, but this is not due to any difficulties in producing L2 speech. Silent pauses may indicate that time is being used for speech planning rather than utterance planning, while

filled pauses may indicate successful strategies for holding a turn, particularly in dialogues, and are not always a clear indication of articulatory fluidity (de Jong, 2016; Tavakoli, 2011). This is why in this study the frequency and length of silent and filled pauses will be investigated, with the aim of exploring the reliability of breakdown fluency indicators. This is necessary because none of these indicators should be disregarded when a learners' fluency performance is assessed.

Speed fluency

Speed fluency concerns the speed with which speech is delivered (De Jong et al., 2013). In general, it is measured by speech rate (SR) and mean length of runs (MLR). Speech rate is calculated as the number of syllables produced each minute (including pause time). It is considered the best predictor of fluency (Kormos, 2006) as well as the most widely used indicator of speed fluency (Bosker et al., 2013, 2016b). Mean length of runs (MLR) is interpreted as a continuous stream of running speech (measured in words) not interrupted by disfluent pauses or hesitations, which reflects the length of language produced between two pause boundaries (Freed, 2000). This measure is considered to be a valuable predictor for measuring fluency (Tavakoli & Skeha, 2005).

Furthermore, the phonation-time ratio has also been used to analyse speed fluency. Specifically, this means the percentage of time spent speaking as a proportion of the time taken to produce the speech sample. It has also been found to be a good predictor of fluency (Kormos & Dénes, 2004; Valls-Ferrer & Mora, 2014). It is worth noting that articulation rate is typically calculated by the number of syllables divided by the amount of time (excluding pause times). This measure as a non-confounded measure is normally used in studies, which aims to measure specific aspect of fluency (Bosker et al., 2013). Because articulation rate does not mathematically relate to other measures of fluency, it can be regarded as a pure measure of speed. For other measures, such as speech rate, which is obtained by the number of syllables divided by total time including silences. Therefore, speech rate is numerically associated with the number and duration of pauses (Bosker et al., 2013; De Jong et al., 2013; De Jong, 2016a; Valls-Ferrer & Mora, 2014).

Repair fluency

Repair fluency is concerned with reformulation, replacement, false starts and repetitions of words and phrases (Tavakoli & Skehan, 2005). It is further argued to be how often speakers use false starts, make corrections, or produce repetitions (De Jong et al., 2013). To investigate repair fluency, the number of repetitions and repairs are generally analysed (e.g., Bosker et al., 2013; De Jong, 2013, 2016). This aspect will be used in this study.

Summary of utterance fluency measures

Concerning the measurement of fluency, it is suggested that it should be measured using these three main characteristics: speed, breakdown, and repair fluency (Tavakoli & Skehan, 2005; Tavakoli, 2016). Within these three main categories, the most analysed and reliable measures are as follows (Kormos, 2006; Kormos & Dénes, 2004; Bosker et al., 2013, 2016b).

1) Speed: articulation rate and speech rate; and mean length of runs;
2) Breakdown: pause duration and frequency;
3) Repair: the number of disfluencies per minute (i.e., repetitions, repairs, restarts); average pause time and length.

2.1.3 Complexity

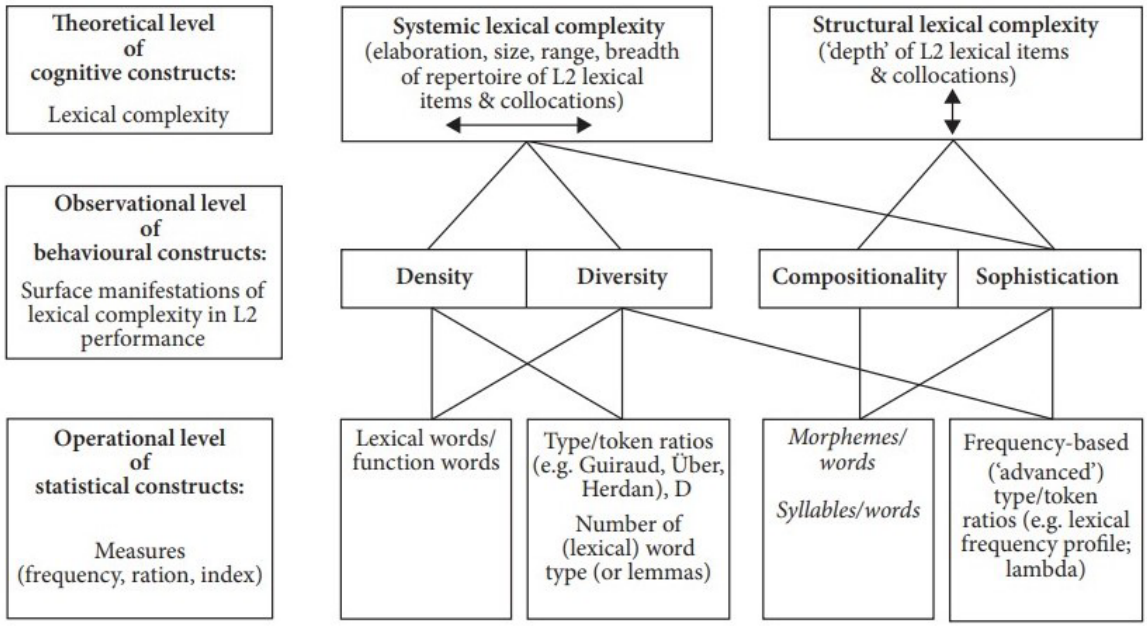
Complexity is considered to be the most controversial construct in the CAF triad (Housen & Kuiken, 2009, 2012). It has been generally interpreted as the use of more challenging and difficult language and the extent to which learners can produce elaborate language (Ellis & Barkhuizen, 2005). Complexity measures have been widely analysed in L2 studies despite the contradictory and mixed interpretations from different researchers using different criteria (Bulté & Housen, 2012). The following section will address the two most widely analysed components of linguistic complexity in L2 research: syntactic complexity and lexical complexity. The analysis will focus on these components because other dimensions of linguistic complexity (e.g., morphological complexity) have only been measured by a few studies (e.g., Yuan & Ellis, 2003; Ellis & Yuan, 2004; Verspoor et al., 2012). One of the reasons why only a few L2 studies have used morphological measures (i.e., inflectional, derivational) is the fact that English, the most frequently investigated L2, is a weakly inflected language (Bulté, 2013).

2.1.3.1 Lexical complexity

Definitions of lexical complexity

Lexical complexity has been called lexical variation (variety), lexical density, lexical sophistication (rareness), lexical richness, and lexical diversity across studies (e.g., Wolfe-Quintero et al., 1998; Bulté et al., 2008; Johansson, 2008; Koizumi, 2005; Yu, 2007, 2009). If there are no differentiations presented before a study's analysis, these different interpretations and terms might easily result in confusion and a lack of clarity (Yu, 2009). However, despite lexical complexity being a vital area of investigation in CAF, researchers have not reached a consensual definition of this construct (Bulté et al., 2008). This issue can potentially be solved by referring to Bulté and Housen (2012). In their study, they stated that lexical complexity can be examined at three different levels: the theoretical level (cognitive), the observational level (performance), and the operational level (quantitative). See Figure 1 below.

Figure 1. Lexical complexity at different levels (Bulté & Housen, 2012:28)



However, similar to syntactic complexity, lexical complexity is also defined at the observational level as a behavioural construct in only a few studies (Bulté et al., 2008; Norris & Ortega, 2009; Skehan, 2003; Ortega, 2003; Bulté et al., 2008; Norris & Ortega, 2009). Instead, the majority of L2

studies only define lexical complexity at the operational level, that is, as an operational-statistical construct (Bulté & Housen, 2012). Lexical complexity in L2 research is typically operationalised as a statistical construct, which is measured by quantitative measures (Bulté & Housen, 2012; Norris & Ortega, 2009). This study will use the operational level of lexical complexity following Bulté & Housen (2012).

Measuring lexical complexity

The most widely used quantitative lexical complexity measures can be subdivided into three different categories: density, diversity, and sophistication (Skehan, 2003; Bulté et al., 2008; Bulté, 2013). Lexical density refers to the amount of lexico-semantic information contained in a language sample. Lexical diversity is associated with the variety and range of lexical items used, and lexical sophistication with the intrinsic complexity of the individual lexical items (Bulté, 2013). The fourth category, the compositionality of lexical elements has been added by Bulté and Housen (2012). Specifically, this is the number of formal and semantic components of lexical items (e.g., phonemes, morphemes, denotations).

Bulté and Housen (2012) reviewed 40 L2 studies (published from 1996 to 2008) and concluded that lexical diversity is the most widely analysed among the four aspects (e.g., Daller et al., 2003; Koizumi, 2005; Kormos & Dénes, 2004; Malvern et al., 2004; Malvern & Richards, 2002; Tavakoli & Foster, 2008; Tajima, 2003; Yuan & Ellis, 2003). However, lexical density and sophistication have also been relatively well-analysed in L2 studies (e.g., Koizumi, 2005; Mehnert, 1998; Michel et al., 2007; Ortega, 1995; Robinson, 1995; Vermeer, 2000). However, compositionality has only been calculated in a few studies (e.g., Verspoor et al., 2008; Spoelman & Verspoor, 2010). This is probably due to the fact that compositionality is a relatively new aspect of lexical complexity that has only been introduced following Bulté and Housen (2012). Moreover, it can be argued that the mean length of words as a measure of lexical compositionality, especially when measured by looking at the number of morphemes per word, not only measures lexical complexity but also partly syntactic complexity (Bulté, 2013). At the observational level, lexical complexity can be measured statistically. This section will review the three main categories: density, diversity, and sophistication (See Table 1).

Lexical density measures the ratio of the number of lexical words. Lexical density is assessed by the number of lexical words divided by the total number of words, or by the total number of function words in a sample (Bulté & Housen, 2012; Polio, 2001; Wolfe-Quintero et al., 1998). Operationally, lexical density is obtained by the ratio of the lexical (or content) words to the grammatical (or function) words in a text (Bulté, 2013). Despite the issue of identifying which words are content words and which are function words (see Halliday, 2009), nouns, adjectives, lexical verbs (i.e., excluding auxiliary verbs) and certain adverbs are considered to be content words in general. Determinatives, conjunctions, pronouns, auxiliary verbs, and even prepositions are considered to be function words (Bulté, 2013). However, it is arguable that instead of a dichotomy, the opposition between lexical and functional words should be conceived of as a continuum. Prepositions, for instance, can be considered to have both lexical and functional characteristics (Bordet & Jamet, 2010). Moreover, it is controversial to link lexical density to L2 complexity in a straightforward manner. This is because, it has been found that there are significant differences between groups of learners using lexical density (e.g., Wolfe-Quintero et al., 1998; Norris & Ortega, 2009). Therefore, lexical density is not considered to be a good measure (Bulté, 2013) and consequently it has not been widely used (Alghizzi, 2017). In light of these controversial issues, lexical density will not be used in this study.

Lexical diversity is often measured by type-token ratios and the number of word types (Bulté and Housen, 2012). Type-token ratio (TTR) has been claimed to be the best-known measure (e.g., Malvern & Richards, 2000, 2002; Daller et al., 2003; Tavakoli & Foster, 2008). It is calculated by the number of different lexical items divided by the number of tokens (total number of words). However, it can be problematic in terms of lower TTRs which are automatically gained in longer texts (Bulté, 2013; Liu, 2017). Thus, TTR has often been considered to be an unsatisfactory measure in assessing lexical diversity by previous studies (e.g., Broeder et al., 1993; Vermeer, 2000; Malvern & Richards, 2002). To reduce the effect of text length on the measurement of lexical diversity, other alternative indicators have been put forward. For example, there are adaptations of TTR, such as mean segmental TTR (Yuan & Ellis, 2003) and Guiraud's Index (Koizumi, 2005). Other indicators have also been introduced, such as D score (Malvern & Richards, 2000, 2002). Based on a random sampling of words in a text, D is a calculation of the probability of repeated words in these random samples (Bulté, 2013). Among all of these indicators, Guiraud's Index is regarded as the most stable to analyse lexical diversity (van Hout & Vermeer, 1988) which outperforms TTR in assessing speech data (Vermeer, 2000).

Guiraud's Index is obtained by dividing the number of types by the square root of tokens (types / √ tokens).

Measures of lexical sophistication are generally calculated by comparing the words in a language sample with the words contained in previously established word lists, based on the relative frequency of the words (e.g., Laufer & Nation, 1995). Generally, lexical sophistication is assessed by frequency-based (advanced) type-token ratios (Bulté & Housen, 2012). Some issues arise from this assessment at the operational level. For instance, a consensus has to be reached between researchers on what are basic words and what are sophisticated words (Alghizzi, 2017). This leads to an equally important issue regarding the question of which corpora or criteria should be used to assign frequency degrees or to decide basic and advanced words (Bulté, 2013). Because frequency lists drawn from different corpora (e.g., books, newspapers, L2 textbooks, etc.) are very likely to lead to different classifications and significantly different results, this can help explain why measures of lexical sophistication have not been used as often in previous studies as lexical diversity measures (Bulté, 2013). Specifically, a review found that only two out of 40 relevant studies analysed lexical sophistication (Bulté & Housen, 2012). Despite the issues, however, different frequency lists have been used in L2 studies, for instance, the Lexical Frequency Profile (LFP) (Laufer & Nation, 1995), Poisson Distribution (Skehan, 2009), and word lists that are based on British National Corpus. The Lexical Frequency Profile (LFP) has been consistently used when analysing lexical sophistication. The LFP categorises the proportion of words at various degrees of frequency: the first 1,000 words, the second 1,000 words, The University Word List, and words that are not included in either of these lists. In other words, the LPF allows researchers to reveal the ratio of words learners use at various lexicon frequency layers (Alghizzi, 2017).

In conclusion, among the four categories of lexical complexity (density, diversity, sophistication, and compositionality) differentiated at the operational-statistical level (Bulté & Housen, 2012), lexical diversity and sophistication are the most (relative) reliable aspects despite certain remaining problems.

Table 1. Review of lexical complexity measures used in L2 studies

Lexical Complexity	**Measures**	**Studies**
a. Variety	Type token ratio	Daller et al. (2003)

	Guiraud index	Daller et al. (2003) Koizumi (2005)
	Advanced TTR	Daller et al. (2003)
	(Word types)² /words	Tajima (2003) Malvern et al. (2004)
	Mean segmental type-token ratio (MSTTR)	Yuan & Ellis (2003)
	Index of lexical diversity	Kormos and Dénes (2004)
	D score	Malvern and Richards (2002) Tavakoli and Foster (2008)
b. Density	Number of lexical words per word	Vermeer (2000) Koizumi (2005)
	Weighted lexical density: ([Number of sophisticated lexical words] + [Number of basic lexical words] x 0.5) / Number of words	Mehnert (1998) Koizumi (2005)
	Lexical words/function words	Ortega (1995) Robinson (1995)
	Lexical words/total words	Michel et al. (2007) Robinson (1995)
c. Sophistication	Less frequent words/total words	Gass (1999); Iwashita et al. (2008)
	The number of sophisticated word types per word	Daller et al. (2003) Koizumi (2005)
	Sophisticated word types per word types	Wolfe-Quintero et al. (1998)
	Sophisticated lexical words per lexical words	Wolfe-Quintero et al. (1998)
	The index Lambda	Skehan (2009a)
	Guiraud Advanced	Daller et al. (2003)

2.1.3.2 Syntactic complexity

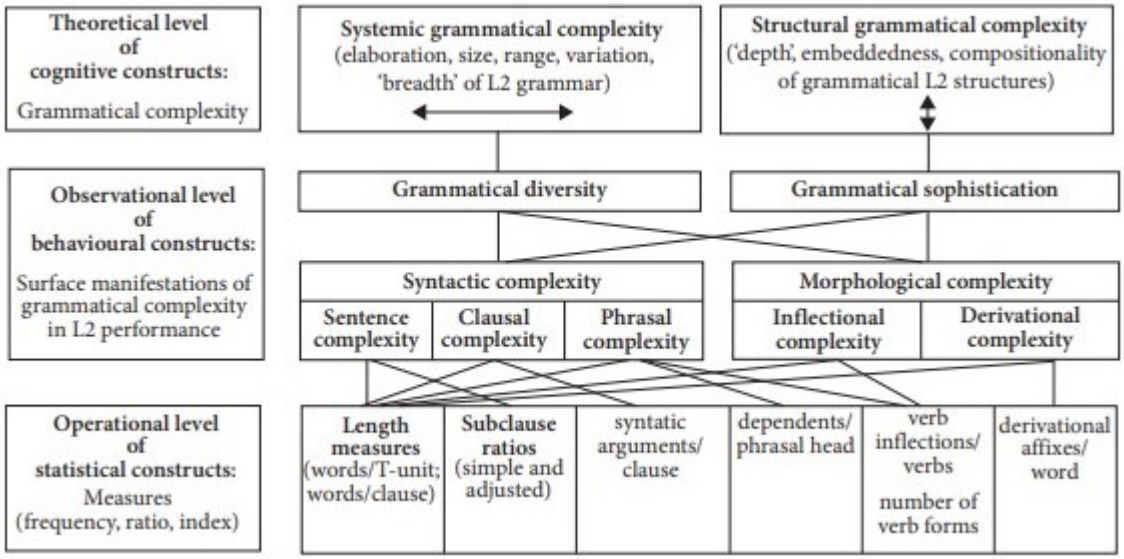
Definitions of syntactic complexity

Syntactic complexity has been interpreted differently in a number of studies. For example, Foster and Skehan (1996) interpreted syntactic complexity as progressively more elaborate language and a greater variety of syntactic patterns. Wolfe-Quintero et al. (1998) viewed syntactic complexity as a wide variety of both basic and sophisticated structures that are available and can be accessed quickly. Ortega (2003) interpreted the construct as the range of forms that surface in language production and the degree of sophistication of such forms. As indicated, the construct is associated with sophisticated and varied structures. However, a key question relates to how to determine the level of structure sophistication and the range of syntactic patterns (Alghizzi, 2017). Those interpretations towards the same linguistic feature illustrate the requirement for a theoretically motivated metric of linguistic complexity (Bulté & Housen, 2012).

In order to solve the issue of how to define syntactic complexity, Bulté and Housen (2012) proposed that the construct can be examined on three different levels, namely, the theoretical level, the

observational level, and the operational level (Figure 2). The first level is an abstract and hypothetical construct which is a part of the cognitive system concerning its number of components, the embeddedness of these components, and the nature of the correlations that exist among them. The second level is more concrete. The theoretical notions of complexity can be observed through language performance at different levels, for example, in the use of different strategies for combining and embedding clauses, applying different verb forms, and using a more common vocabulary. The third level relates to the analytical measures and instruments used to investigate the complexity of a language sample from a quantitative perspective. It is important to point out that to have meaningful and valid measures, as well as to make meaningful interpretations, the links between these three levels should be made explicit. For example, it is necessary to establish the meaning of syntactic complexity theoretically and how it manifests itself in actual language performance at the observational level. Also, it should be established how these manifestations can be quantified operationally. However, syntactic complexity has only been defined at the observational level as a behavioural construct by a few studies (Skehan, 2003; Ortega, 2003; Bulté et al., 2008; Norris & Ortega, 2009). In contrast, the majority of L2 studies only define syntactic complexity at the operational level (Bulté & Housen, 2012). Similar to the majority of L2 studies, this study will apply the operational level of syntactic complexity. The rationale for doing so is because quantifiable measures and indicators can be analysed at the operational level to investigate the syntactic complexity of learners' language performance.

Figure 2. Different levels of grammatical complexity (Bulté & Housen, 2012:27)



Measuring syntactic complexity

Norris and Ortega (2009) highlighted that syntactic complexity itself is multidimensional, and therefore must be measured multidimensionally. Moreover, three measurable sub-constructs in syntactic complexity have been identified: 1) complexity via subordination, which is measured via clauses; 2) overall or general complexity, which is assessed by length-based measures; and 3) subclausal complexity via phrasal elaboration, which is gauged by the mean length of a clause.

Among the measures noted above, length-based measures capture the mean length of a certain unit of analysis. They are normally calculated by dividing words (or morphemes) by a chosen production unit (Bulté & Housen, 2012; Norris & Ortega, 2009). Those length-based measures are interpreted as a global or generic metric of linguistic complexity (Norris & Ortega, 2009). This category of measures varies among studies mainly based on one of the three production units, namely, T-unit (Hunt, 1967), C-unit (Loban, 1976), and AS-unit (Foster et al., 2000) for oral speech segmentation. Specifically, syntactic measures are assessed by the length of T-unit (Hunt, 1965), the mean length of C-unit (Loban, 1976), and the mean number of words per AS-unit (e.g., Tavakoli & Foster 2008; Michel, Kuiken & Vedder, 2007; Jensen & Howard, 2014).

In L2 Chinese studies, considering different segmentation units, the length-based measures have been operationalised via two routes. One is the mean length of sentences (Shi, 2002; Wang, 2002; Ye, 2015; Zhu, 2009). The other is the mean length of AS-units (Chen, 2015a; Wu, 2017), which is calculated by the number of words divided by the number of AS-units, or the number of syllables of sentences (Zhu, 2009). Moreover, the mean length of correct AS-units has also been applied in recent studies, which is calculated by the total number of words divided by the total number of correct AS-units (Wu, 2017). It is worth noting that one of the add-on sub-constructs, namely, the frequency of certain sophisticated forms is operationalised at the syntactic level (Chen, 2015a).

Concerning complexity by subordination, this is normally measured by the clauses per C-unit/T-unit/AS-unit and subordinate clauses per total clauses (Norris & Ortega, 2009). In the field of L2 Chinese studies, it has been measured by the mean number of clauses per AS-unit (Chen, 2015a; Wu, 2017), and the ratio of clauses to AS-units (Wu, 2017). The third type, complexity via phrasal elaboration is measured by only one measure, namely, the mean length of the clause. However, this has rarely been analysed in L2 studies so far. Operationally, the most commonly applied syntactic measures are length-based measures of overall complexity (e.g., mean length of T-unit/AS-unit), and measures of subordination (e.g., number of clauses) (Kuiken et al., 2019).

However, these overall measures have increasingly become the object of criticism. This is because it is questionable whether the complexity of L2 development can be captured in terms of global length measures and subordination ratios. Therefore, other measures of syntactic complexity, which may reveal syntactic development at different levels of proficiency, should be applied (Kuiken, 2019). Based on this concept, Norris and Ortega (2009) argued that syntactic complexity can be distinguished at three levels in a hierarchical fashion and posited the developmental processing as follows: 1) coordination index is considered as having great predictive power when measuring syntactic complexity at the beginner level of L2 development; 2) subordination measures are valuable when measuring learners at intermediate and upper-intermediate levels; and 3) measure mean length of the clause, which is the only measure to date that assesses complexification at the subclausal or phrasal level. Mean length of clause is regarded as the measure with the most predictive power to examine L2 learners at an advanced level and has become an increasing topic of interest in the L2 field (Kuiken et al., 2019).

In summary, presenting the originality of CAF constructs and their development in other L2s serves as a foundation for understanding the state of research on assessing L2 learners of Chinese using CAF measures, which will be addressed in the following section.

2.2 CAF measures on L2 oral Chinese studies

This section begins by discussing the base units in assessing spoken data in previous studies (See Section 2.2.1), which leads to why AS-unit as the base unit is chosen for data analysis in this study. The CAF measures employed in L2 oral Chinese studies are then presented (see Section 2.2.2), with the intention of explaining what CAF measures should be chosen for this study.

2.2.1 Production units in the literature

To analyse spoken data, Foster et al. (2000), after identifying 87 studies in the L2 field, concluded that production units can be categorised into three types: semantic (e.g., proposition, communication unit (C-unit), intonational (e.g., tone unit, utterance), and syntactic (e.g., sentence, Terminable unit (T-unit), Analysis of Speech Unit (AS-unit)). Similarly, in L2 Chinese studies, C-units (i.e., Zhai, 2011), T-units (i.e., Zhai & Feng, 2014; Zhou, 2016), AS-units (i.e., Chen, 2015, 2020; Wu, 2017), and a Chinese sentence (i.e., Wang, 2002; Shi, 2002; Guo, 2007; Jin, & Mak, 2013; Ye, 2015) have been applied as the basic units to assess the oral data of L2 Chinese learners. Among these, three units: C-unit, T-unit and AS-unit have been used most widely.

A communication unit (C-unit), mainly as a semantic unit, has been defined as, utterances, for examples, words, phrases and sentences, grammatical and ungrammatical, which provide referential or pragmatic meaning (Pica et al., 1989). However, using these semantic based criteria is problematic because it is hard to process analysis with certainty (Foster et al., 2000). A T-unit has been defined as one main clause plus any subordinate clause or non-clausal structure that is attached to or embedded in it (Hunt, 1965, 1970), that is, an independent clause, accompanied by any associated dependent clauses (Larsen-Freeman, 2009). However, it has been revealed that this definition of a T-unit is not adequate in assessing speech data because learners do not always speak in full sentences as expected in written data (Foster et al., 2000; Luoma, 2004; Vercellotti, 2012).

As a modified version of a T-unit or a C-unit, an AS-unit has been defined as, "a single speaker's utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s) associated with either" (Foster et al., 2000:365). That is, an AS-unit is an independent clause, or an independent sub-clause unit which can be interpreted to be a full clause in the discourse, together with a subordinate clause including a finite or non-finite verb element and one other element at minimum. Mainly as a syntactic unit, the AS-unit has been considered as a valid unit to analyse oral data. Such a unit helps handle dysfluency features of spoken language data, such as false starts, repetitions, and corrections (Foster et al., 2000). Also, this unit allows intonation and pause information to be taken into consideration when coding oral data. Those clauses with finite verbs separated by pauses reaching and exceeding 500 milliseconds with a falling intonation are coded for separate AS-units, even if a subordinate conjunction (i.e., but, because) occurs (Foster, 2000; Vercellotti, 2012). Additionally, Foster (2000) offered three levels of application, which allow systematic exclusion of certain data, such as non-linguistic fillers, echoic responses, one-word minor utterances, and interlocutors' speech, for the purpose of coherent analysis. AS-units have been widely used because when analysing oral data, decisions on segmenting and coding have to be made (Foster, 2000; Koizumi, 2005; De Jong et al., 2016; Kahng, 2014; Tavakkoli et al., 2016; Wright & Tavakkoli, 2016; Vercellotti, 2012, 2019; Chen, 2015, 2020). Notwithstanding their shortcomings, AS-units are considered to be an accessible, clearly defined and easily analysed unit that is valid and reliable in assessing oral language data (Foster et al., 2000). This study will therefore take the AS-unit as the base unit for data analysis.

2.2.2 CAF measures in L2 Chinese speaking studies

To help provide the rationale for the CAF measures that are used in this study, in this section, the CAF measures that have been employed to analyse L2 spoken Chinese in previous research are reviewed. The measures and their calculations corresponding to the L2 Chinese studies are listed in the tables below.

2.2.2.1 Accuracy measures

Table 2. Accuracy measures in L2 Chinese speaking studies

Sub-constructs	No.	Measures/Calculation	L2 Chinese Studies
Phonetic accuracy	1	The number of correct syllables divided by the number of total syllables	Zhai and Feng, 2014

	2	The ratio of correct consonants	Chen 2015a, 2015b, 2020
	3	The ratio of correct vowels	
	4	The ratio of correct tones	
	5	Tonal accuracy (the number of wrong tones divided by the total number of words minus the number of filler words)	Kim et al., 2015
Lexical accuracy	6	The number of lexical errors of AS-units	Chen, 2015a
	7	The number of 'information bits' (IB)	Ye, 2015
	8	The ratio of lexical errors	Ding and Xiao, 2016
	9	The ratio of lexical error types	Ding and Xiao, 2016
Syntactic accuracy	10	The ratio of the correct sentence	Ye, 2015
	11	The ratio of error-free T-units (REFT)	Zhai and Feng, 2014
	12	The number of syntactic errors of AS-units	Chen, 2015a
	13	The ratio of correct sub-clauses	Wu, 2017
	14	The ratio of correct AS-units	Wu, 2017

In terms of defining and counting an error, this can be more problematic concerning accuracy rather than fluency and complexity. This is because accuracy can be impacted subjectively (Juan-Garau, 2014) in particular, for learners at both intermediate and advanced levels (Serrano, 2007). Existing L2 Chinese studies have explored accuracy at three levels (See Table 2): phonetic accuracy (i.e., Chen, 2015a, 2015b; Kim et al., 2015; Zhai & Feng, 2014), lexical accuracy (i.e., Chen, 2015a; Ding & Xiao, 2016; Ye, 2015; Zhai & Feng, 2014), and syntactic accuracy (i.e., Chen, 2015a; Ye, 2015; Wu, 2017; Zhai & Feng, 2014).

For phonetic accuracy, there are different calculations in the literature. For instance, phonetic accuracy has been obtained from the ratio of correct pronunciations calculated by the number of correct syllables divided by the number of total syllables (e.g., Zhai & Feng, 2014). Furthermore, considering a syllable as the basic unit of Chinese language (Liao, 2018), phonetic accuracy has been measured by three sub-categories, consonants, vowels, and tones. Under this approach, the ratio of correct consonants, vowels, and tones are calculated respectively (Chen, 2015a, 2015b). Acknowledging that learning tones is one of the most challenging aspects of Chinese acquisition, Kim et al. (2015) measured tonal accuracy as an important indicator to reveal learners' overall proficiency. Moreover, apart from these above three measures, the ratio of reading errors (i.e., character-recognising errors) has also been used to measure L2 Chinese learners' reading accuracy (Chen, 2015b).

Lexical accuracy has also been measured in L2 Chinese studies. For example, this has been done by assessing the number of lexical errors in AS-units (e.g., Chen, 2015a), and the ratio of lexical errors and the ratio of lexical error types (e.g., Ding & Xiao, 2016). The former was obtained by the

number of lexical errors divided by total number of lexical items; while the latter was calculated by the number of lexical errors in a type divided by the total number of lexical errors. Furthermore, twelve 'information bits' (IB) with universality were employed in Ye (2015), where the higher number of correct lexical items that related to any of the twelve IB, was considered to represent greater lexical accuracy. However, none of these indicators have been sufficiently exemplified in these studies.

To investigate syntactic accuracy, based on the different production units applied in different studies (e.g., T-unit, AS-unit, Chinese sentences), error and error-free ratio measures have been analysed. These include, the ratio of correct sentences (e.g., Ye, 2015), the ratio of error-free T-units (e.g., Zhai & Feng, 2014), the number of syntactic errors in AS-units (e.g., Chen, 2015a), and the ratio of correct sub-clauses and the ratio of correct AS-units (e.g., Wu, 2017).

2.2.2.2 Fluency measures

To examine the oral fluency of L2 learners of Chinese, existing studies (e.g., Chen, 2012; Ye, 2015; Zhai & Feng, 2014) have adopted the aforementioned measures (see section 2.1.2), such as Speech Rate (SR), Mean Length of Runs (MLR), number and length of pauses. Moreover, considering the particular features of Chinese, some measures have been modified in assessing L2 Chinese (Liao, 2018). For instance, instead of using the number of words, the number of syllables as an adjusted indicator was employed (e.g., Wang, 2002; Du, 2013; Ye, 2015).

The oral fluency of L2 learners of Chinese has been investigated mostly with three sub-categories of utterance fluency (speed fluency, breakdown fluency, and repair fluency). As outlined in Table 3 below, a large number of fluency measures have been analysed in L2 Chinese speaking studies.

Table 3. Fluency measures used in L2 Chinese speaking studies

	Fluency measures	**Calculation**	**L2 Chinese Studies**
1	Speech Rate	The total number of syllables / Time of utterance (in seconds)	Chen, 2012; Du, 2013； Kim et al., 2015； Guo, 2005; Zhai and Feng 2014; Ye 2015; Chen, 2017; Wang, 2018; Feng, 2018; Wright and Zhang, 2014； Zhang, 2001
		Total number of syllables/ Time of utterance (in seconds) ×60	Ding and Xiao, 2016; Wu, 2017;
		SR (Average Number of syllables without corrections, repetitions, and pauses)	Chen, 2020

2	Articulation rate (AR)	The total number of syllables /duration of utterance (without pause time)	Guo, 2005; Tavakoli, 2016; Wang, 2018; Wright, 2020
3	Phonation/ Time Ratio	Duration of articulation excluded pause divided by the total duration of utterance	Guo, 2017; Wang, 2018; Wright, 2018
4	Overall fluency	The number of valid tokens/ Time of utterance (in seconds)×60	Wen, 2006, 2010
		Number of valid syllables/ Time of utterance (in seconds)×60.	Ding and Xiao, 2016
5	The ratio of pruned length (RPL)	The number of pruned syllables divided by the total number of syllables	Guo, 2005; Zhai and Feng, 2014; Wang, 2018
6	The ratio of repairs	The number of repairs per minute	Chen, 2012; Chen, 2015b; Ding and Xiao, 2016; Fengyue, 2018
7	The ratio of repetitions	The number of repetitions per minute	Chen, 2012; Chen, 2015b; Ding and Xiao, 2016; Fengyue 2018
8	The ratio of repairs and repetitions	The number of re repairs and repetitions per minute	Zhang, 2001; Wu, 2017; Wang, 2018
9	Mean length of run (MLR)	The total number of syllables divided by the total number of silent pauses.	0.2 seconds (Chen, 2015); 0.3 seconds (Feng, 2018; Wang, 2018; Wu 2017); or 1 second (Zhai & Feng, 2014); Wright and Zhang, 2014; Wright, 2020; Chen, 2020
10	The Ratio of silent pauses	The number of silent pauses per minute	Chen, 2012; Zhai and Feng, 2014; Chen, 2015a, 2015b; Feng, 2018;
11	The Duration of silent pauses	The average length of silent pauses per minute	Chen, 2012; Zhai and Feng, 2014; Chen, 2015a, 2015b; Feng, 2018; Wright, 2020
12	The Ratio of filled pauses	The number of filled pauses per minute	Chen, 2012; Zhai and Feng, 2014; Chen, 2015a, 2015b; Feng, 2018;
13	The Duration of filled pauses	The average length of filled pauses per minute	Chen, 2012; Zhai and Feng, 2014; Chen 2015a, 2015b; Feng, 2018
14	The Ratio of pauses (including silent and filled pauses)	The number of pauses per minute	Zhai and Feng, 2014; Ye, 2015; Ding and Xiao, 2016; Wu, 2017; Wang, 2018
15	The Duration of pauses (including silent and filled pauses)	The average length of pauses per minute	Zhai and Feng, 2014; Ye, 2015; Ding and Xiao, 2016; Wu, 2017; Wang, 2018
17	Hesitation rate	Total number of repetitions, retracings, repairs and filled pauses/the total number of characters	Wright, 2018, 2020
18	Number of Pauses	Number of clause-internal pauses (250ms)	Wright, 2020

Pause marking

With regard to fluency measures, pause marking is an essentional starting point to determine pause-related indicators. Therefore, the studies on pauses in the field of L2 Chinese studies are reviewed in this section.

Any pause in oral productions, such as "eh/er" and "eh", are non-lexical fillers. Such non-lexical fillers are not recognised as words, and they contain little or no semantic information (Riggenbach,

1991). Instead, they are regarded as filled pauses (e.g., Chen, 2015; Ye, 2015; Ding & Xiao, 2016). Concerning the transcription (or exclusion) of the pauses, there are different standards for the cut-off point of pauses in the literature. They are as follows:

- Marking any pause reaching or exceeding 3 seconds which were excluded in assessing speech rate (Jin & Mak, 2013); while *unfilled pauses* reaching and exceeding 1 second were marked to calculate pause time (Jin & Mak, 2013).
- Marking any unnatural pauses meeting or exceeding 2 seconds between sentences (Zhang, 2001).
- Marking pauses meeting or exceeding 1 second (Zhai, 2011; Zhai & Feng, 2014).
- Marking any unnatural pauses exceeding 0.5 seconds within a sentence (Zhang, 2001).
- Marking silent pauses located between sentences reaching or exceeding 0.5 seconds, filled pauses as 0.3 seconds (Liu, 2019) following Riggenbach (1991).
- Marking any pause reaching or exceeding 0.3 seconds (Zhang & Wu, 2001; Guo, 2007; Ding & Xiao, 2016; Wu, 2017; Feng, 2018; Wang, 2018).
- Marking as filled or unfilled pauses each pause meeting or exceeding 0.2 seconds (Chen, 2012, 2015; Ye, 2015; Liu & Wu, 2016; Zhang, 2019).
- A silence equal to, or longer than, 250 milliseconds was considered as a silent pause (Kahng, 2014; Préfontain, 2013; De Jong, 2016a).
- Marking as unfilled pause each pause reaching or exceeding 0.75 milliseconds (Chen & Zhou, 2016).

As illustrated above, the most common standard for defining a pause, including silent and filled pauses, is 0.3 seconds, which follows Raupach (1980). In the literature, three thresholds have typically been suggested for dysfluency pauses, starting from 0.2 seconds (Riggenbach, 1991) to 0.3 seconds (Raupach, 1980; Valls-Ferrer & Mora, 2014), to 0.4 seconds (Tavakoli & Foster, 2008). It has also been suggested that 250-300 ms is the optimal cut-off point for measuring the number of pauses when investigating L2 proficiency (De Jong & Bosker, 2013). In this study, considering the low competency level of participants, I have adopted 0.3 seconds as the standard to mark unfilled/silent pauses and filled pauses as this is the most widely accepted duration (e.g., Feng, 2018; Guo, 2007; Ding & Xiao, 2016; Wu, 2017; Wang, 2018; Zhang, 2001).

Regarding the two sub-categories of pauses, there are different approaches in the literature. For instance, some studies combine the two types of pauses into one dysfluency feature (e.g., Guo, 2007; Zhai & Feng, 2014; Ding & Xiao, 2016; Wu, 2017; Wang, 2018). In contrast, in the majority of studies filled pauses and silent pauses are examined separately (Chen, 2012, 2015a, 2015b; Ye, 2015; Liu & Wu, 2016; Feng, 2017; Wright & Tavakkoli, 2016; Zhang, 2019).

To investigate repair fluency, some studies investigate repairs and repetitions as one indicator (e.g., Zhang, 2001; Liu & Wu, 2016; Zhang, 2019), but the number of repetitions and repairs are generally assessed separately (e.g., Chen, 2012, 2015b; Ding & Xiao, 2016; Feng, 2018). In terms of repetitions, different sub-categories have been analysed in previous L2 Chinese studies. For instance, repetitions of morphemes (e.g., Ding & Xiao, 2016; Wang, 2018), repetitions of words (e.g., Ding & Xiao, 2016; Wang, 2018), repetitions of phrases (e.g., Ding & Xiao, 2016; Wang, 2018), and repetitions of clauses (e.g., Wang, 2018) have been investigated. Moreover, the scope of repairs (or corrections) differs among studies. This aspect has been divided into different components in L2 Chinese studies:

a. Five components: reformulations, replacements, repetitions, hesitations, and false starts, were investigated (e.g., Zhang & Wu, 2001; Guo, 2007) following Foster and Skehan (1996).
b. Five components: removing hesitations, reformulations, replacements, repetitions, false starts, were investigated (Wu, 2017).
c. Repetitions, hesitations, false starts, corrections, excluded/ pruned syllables were included (Zhai & Feng, 2014).
d. Repairs of pronunciation, lexical items, sentences were analysed (Ding & Xiao, 2016).
e. Corrections including three phenomena: corrections of pronunciation, lexical items, grammar, and pragmatic errors; repetitions; reformulation after quitting unfinished utterances (Liu, 2019).

Moreover, repairs (repeated and reformulated expressions) have generally been considered as one sub-component of disfluency, and have been calculated as the aggregated total of filled pauses (um, er) and repairs (repeated and reformulated expressions) (e.g., Wright, 2013; Wright & Cong, 2014).

2.2.2.3 Complexity measures

To assess the oral complexity of L2 learners of Chinese, the two widely applied sub-constructs of complexity: lexical and syntactic complexity, have also been investigated in the literature. Reflecting the other L2 studies, to the best of my knowledge, no morphological measures have been analysed in L2 Chinese studies. This is probably due to the fact that Chinese language lacks morphological features.

Word segmentation and frequency in L2 Chinese speaking studies

Following Liu (2017), the THU Lexical Analyzer (THULAC online platform) for Chinese has been adopted in the present study for segmenting the transcription. The Hanyu Shuiping Cihui Dengji Dagang《汉语水平词汇等级大纲》 (HSCDD) known in English as the Outline of Vocabulary of Chinese Language Level (2001) has been widely used as a reference for the syntactic classification of units after segmentation. Based on the HSCDD, all the lexicons in each speech sample were divided into five different levels; namely, A, B, C, D-grade words and words which are not listed (甲级,乙级, 丙级,丁级 and 超纲). Regarding categorising participants' mastery of vocabulary, the HSCDD is largely used as a standard to differentiate different proficiency levels in oral performance (e.g., Chen, 2015; Ding & Xiao, 2016; Wu, 2016a; Zhou, 2016; Liu, 2017). For lexical items which are hard to categorise in speech samples, the Xiandai Hanyu Cidian 《现代汉语词典》(Modern Chinese Dictionary] is referred to as the optimal benchmark (e.g., Ding & Xiao, 2016; Wu, 2017). Furthermore, for assessing some specific words such as Liheci (detachable compound words), the Xiandai Hanyu Liheci Yongfa Cidian 《现代汉语离合词用法词典》(Dictionary of Usage of Detachable Words in Modern Chinese] is followed (e.g., Zhou, 2016).

Moreover, some studies (e.g., Wu, 2017) categorise words based on the comparatively new benchmark the Hanyu Guoji Jiaoyu Yong Yinjie Hanzi Cihui Dengji Huafen《汉语国际教育用音节汉字词汇等级划分》 (Classification of Syllables Characters and Lexical Items for Chinese International Education] (2010) (CSCLCIE). The New HSK (2012) is used generally as an official guide to

categorise L2 Chinese learners' produced lexical items at different proficiency levels, because the HSK is more practical and applicable than HSCDD due to its connection with the existing assessment system and because the HSK system is widely accepted as a formal standard to measure L2 Chinese learners' proficiency levels. Table 4 below compares the categorisation of words in the HSCDD (2001), the CSCLCIE (2010), and the New HSK (2012).

Table 4. Comparisons between HSCDD (2001), CSCLCIE (2010), New HSK (2012)

Chinese level vocabulary level outline (2001)	**Categories**	**The number of words**	
	A-grade	1,033	8,822
	B-grade	2,018	
	C-grade	2,022	
	D-grade	3,569	
New HSK (2012)	HSK 1	150	5,000
	HSK 2	150	
	HSK 3	300	
	HSK 4	600	
	HSK 5	1,300	
	HSK 6	2,500	
Classification of syllables characters and lexical items for Chinese International Education (2010)	*Level 1*	*2245*	11,902
	Level 2	*3211*	
	Level 3	*4175*	
	Level 3 appendix	*1461*	

Lexical complexity

Regarding lexical complexity in the L2 field, three main categories have been analysed: lexical density, lexical diversity, and lexical sophistication (Skehan, 2003; Bulté et al., 2008; Bulté, 2013). The last two categories have been analysed more frequently in L2 Chinese speaking studies (See Table 5)(e.g., Liu, 2017; Wu, 2017). This is probably due to the remaining issues of distinction between context words and functional words (Halliday, 2009) and of the reliability of the lexical density measures (See section 2.1.3.2).

Table 5. Lexical complexity measures in L2 Chinese speaking studies

Subconstructs	**No.**	**Measures**	**Calculation**	**L2 Chinese speaking studies**
Lexical diversity	1	Type-token ratio (TTR)	Type account / token account	Ye, 2015; Chen, 2015
	2	Guiraud's Index	Types / √ tokens (RTTR)	Chen and Li, 2016; Chen, 2020; Liu, 2017; Wu, 2017; Wright, 2018, 2020
	3	Transformations of TTR	Type ² / token	Ding and Xiao, 2016
	4	Corrected TTR (CTTR)	Types / √ 2 tokens (CTTR)	Liu, 2017

	5	D score		Wright and Zhang, 2014
Lexical sophistication	6	The ratio of beginning-level lexical items	The number of beginning-level lexical items/ total number of words	Chen, 2015; Ding and Xiao, 2016; Wu, 2016; Zhou, 2016; Liu, 2017; Wu, 2017
	7	The ratio of intermediate-level lexical items	The number of intermediate-level lexical items/ total number of words	
	8	The ratio of advanced-level lexical items	The number of advanced-level lexical items/ total number of words	
	9	The ratio of lexical items beyond the benchmark	The number of lexical items beyond/ total number of words	Wu, 2017

To measure lexical diversity, Type-Token Ratio (TTR) is the best-known indicator. The number of types is the total number of different words (word types); and the number of tokens is the total number of word forms in a speech sample (Veermeer, 2000). TTR is calculated by type account/token account. (Ye, 2015; Chen, 2015). However, the widely analysed TTR is problematic, because lower TTRs are automatically gained in longer texts (Chen & Li, 2016; Liu, 2017; Skehan, 2009; Wu, 2017). Other TTR-related indicators have also been analysed. For instance, Guiraud's Index (e.g., Chen & Li, 2016; Liu, 2017; Wu, 2017; Wright, 2020), Type ² / token (e.g., Wen, 2006; Ding & Xiao, 2016) and types / √ 2 tokens (CTTR) (e.g., Liu, 2017). Also, another lexical variety measure, the D score has been used but only relatively rarely (e.g., Wright & Zhang, 2014). The diversity index D (Malvern et al., 2004) serves as an index of lexical diversity. This index is a mathematical adaptation of the standard TTR that aims to reduce the intervening effects of text length and to provide an indication of the degree of words' repetition in a text (Bulté & Housen, 2014). However, none of these lexical diversity measures are satisfactory. In particular, TTR has been proved to be inadequate. However, Guiraud's Index seems to be adequate in giving a better indication of lexical richness (Vermeer, 2000), and can avoid the impact of longer texts when using TTR (Chen & Li, 2016; Liu, 2017).

Lexical sophistication is measured by frequency-based type-token ratios (Bulté & Housen, 2012). In L2 Chinese studies, two main benchmarks have been applied to categorise lexical items based on frequency. These are: the Hanyu shuiping cihui dengji dagang (HSCDD) (2001) (e.g., Chen, 2015; Ding & Xiao, 2016; Wu, 2016; Zhou, 2016; Liu, 2017) and the New HSK (2012) (e.g., Wu, 2017). Similarly, in this aspect, three predictors of the lexical sophistication in a hierarchal fashion have been based on both the HSCDD (2001) and the New HSK (2012), namely, the ratio of beginner, intermediate and advanced level lexical items. Furthermore, lexical items beyond the benchmark have

either been merged into an advanced level lexicon (e.g., Ding & Xiao, 2016) or as an individual predictor (e.g., Wu, 2017).

In conclusion, to analyse lexical complexity across existing L2 Chinese studies the most frequently applied indicators can be categorised into two types: 1) lexical diversity, which is normally analysed using Guiraud's Index (e.g., Chen & Li, 2016; Chen, 2020; Liu, 2017; Wu, 2017; Wright, 2018, 2020); 2) lexical sophistication, which is measured by the ratio of words at different levels. These two aspects of lexical complexity will be used in this study.

Syntactic complexity

Table 6. Syntactic complexity measures in L2 Chinese speaking studies

	No.	Measures	Calculation	L2 Chinese speaking studies
Syntactic complexity	1	The average length of As-units	The total number of lexical items /number of As-units	Chen, 2015a; Wu, 2017; Chen, 2020
	2	The length of the sentence	The number of syllables within valid sentences / the number of valid sentences	Ye, 2015
	3	The average length of correct AS-units	The total number of lexical items of error-free AS-units / the total number of error-free As-units	Wu, 2017
	4	Numbers of sub-clauses of As-units	The total number of sub-clauses / the number of As-units	Chen, 2015a, 2020; Wu, 2017
	5	Syntax levels of AS-units	Based on (Liu 1996)	Chen, 2015a
	6	Number of conjunctions	Not indicated	Chen, 2015a

On the basis of the base units applied in different studies, the mean length of utterance has been measured differently. For example, the length of the sentence (e.g., Ye, 2015) and the average length of AS-units (e.g., Chen, 2015a; Wu, 2017) have been used. Moreover, the length has been measured by the average length of correct AS-units (e.g., Wu, 2017). Assessing syntactic complexity via subordination has been analysed (e.g., Chen, 2015a, 2020; Wu, 2017) by analysing the number of sub-clauses of AS-units. Moreover, the syntax level of AS-units and the number of conjunctions have been analysed (e.g., Chen, 2015a). The former indicator is based on the criteria of the Hanyu Shuiping Dengji Biaozhun Yu Yufa Dengji Dagang (Chinese Proficiency Level Standard and Grammar Level Outline) (CPLSGLO) (Liu, 1996). Four syntactical levels are categorised in this book: A, B, C, and

D-grade (甲级、乙级、丙级、丁级). However, the average syntax level of AS-units will not be analysed in my study due to two reasons. First, the standard for categorising AS-units into different syntactic levels such as the CPLSGLO (Liu, 1996) is outdated and second, the participant's lexical complexity is based on the HSK (2012), and there might be minor deviations between the two benchmarks.

In conclusion, syntactic complexity has only been assessed by a few L2 Chinese speaking studies (e.g., Chen, 2015; Ye, 2015; Wu, 2017). Only a handful of measures have been analysed (see Table 6). This is very likely due to the lack of widely accepted benchmarks to categorise the syntactic levels of oral performance of L2 learners of Chinese. The only existing benchmark used is the CPLSGLO (Liu, 1996), which has no up to date digital version for researchers to use which makes processing a large amount of data impossible.

Summary of CAF measures analysed in L2 Chinese speaking studies

After reviewing the main L2 speaking studies, it can be concluded that these studies have often adopted measures used in other L2s to assess L2 Chinese fluency (Liao, 2018), such as speech rate (SR), number and length of pauses, number of false starts (e.g., Chen, 2012; Du, 2013; Guo, 2007; Jin & Mak, 2013; Wang, 2002; Ye, 2015; Zhai & Feng, 2014; Zhang, 2001). However, due to Chinese-specific features, some measures analysed in other L2s have had to be modified in L2 Chinese speaking studies (Liao, 2018). First, to examine L2 Chinese learners' pronunciation, the analysis unit for Chinese pronunciation should comprise Chinese syllabic and tonal features (Liao, 2018). Thus, a basic analysis unit for Chinese pronunciation has been redefined as a Chinese syllable (e.g., Jin & Mak, 2013; Wang, 2002). A syllable is equally considered as a morpheme, the smallest unit in spoken Chinese in most cases. Second, considering Chinese lexical measures, in measuring word tokens and types, systematic segmentation specifications are required because of the ambiguity of Chinese word boundaries (Jin & Mak, 2013). Third, regarding grammatical accuracy and complexity measures, the Chinese sentence as the unit has been used to replace the T-unit because there is a lack of subordinate clauses in Chinese (e.g., Jin & Mak, 2013; Wang, 2002; Ye, 2015; Zhai, 2011). Moreover, the AS-unit has also been widely used in this regard (e.g., Chen, 2015, 2020; Wu, 2017). Fourth, fluency measures

have also been adjusted to match the features of Chinese language. Specifically, the number of syllables has been used as an indicator of Chinese oral fluency, rather than the number of words, due to the challenges in Chinese word segmentation (e.g., Du, 2013; Wang, 2002; Ye, 2015). In conclusion, these Chinese-specific characteristics should be taken into account when using CAF measures in L2 speaking Chinese studies. Moreover, this revision of CAF measures investigating L2 speaking Chinese in previous research provides the rationale for the CAF measures selection in this research.

2.3 Relationships between complexity, accuracy, and fluency

Previous research has discovered that complexity, accuracy, and fluency are concerned with separate aspects of learners' language use (Foster & Skehan, 1996; Skehan, 1998b). For instance, Foster and Skehan (1996) distinguished three aspects of production: fluency, accuracy, and complexity. Both complexity and accuracy are concerned with form but have a significant distinction in emphasis. Specifically, complexity relates to the "restructuring" that arises as a result of the need to take risks, whereas accuracy reflects the learner's attempt to control existing resources and to avoid errors in a more conservative manner. Fluency reflects the primacy of meaning and the capacity to cope with real-time communication and it priorities idiom-based language to enable communication to proceed smoothly while avoiding using rule-based language. Wolfe-Quintero et al. (1998) stated that complexity is the scope of expanding or restructuring second language knowledge and accuracy is the conformity of second language knowledge to target language norms. This view links complexity and accuracy to L2 knowledge representation and the level of analysis of internalised linguistic information. Wolfe-Quintero et al (1998) regard fluency as related to linguistic L2 knowledge. Due to the fact that complexity, accuracy, and fluency are concerned with different aspects of a learner's production and knowledge, there are potential differences in correlation.

Researchers have considered if and how these constructs of language performance interact. This leads to another challenge concerning how to identify correspondences between these constructs, the factors that influence them, and how they are correlated (Housen et al., 2012). There have been studies on the interaction and correlation between the three constructs, as well as their connection to learner knowledge and their competitive, supportive, or (ir) relative, interrelationship (See Skehan, 1996, 1998; Larsen-Freeman, 2006; Skehan & Foster, 1997; Bygate, 2001; Robinson, 2001b; Yuan & Ellis, 2003; Michel, Kuiken, & Vedder, 2007; Norris & Ortega, 2009; de Bot, Lowie, & Verspoor,

2007; Housen & Kuiken, 2009; Ahmadian & Tavakoli, 2011; Ahmadian, 2011; Vercellotti, 2012, 2017, 2019).

This section firstly discusses the two most widely documented theoretical hypotheses which aim to account for the impact of task type and task conditions on performance, Skehan's Limited Attentional Capacity model (Skehan, 1998a; Skehan, 2009; Skehan & Foster, 2012) and the Cognition Hypothesis (Robinson, 2001, 2003, 2005, 2011). This is followed by a review of the studies which explore the correlations between CAF constructs and between their subconstructs and which reveal two major types of relationship: the trade-off effect (where a higher performance in one component corresponds to a lower performance in another) (e.g., Bygate, 2001; Skehan & Foster, 1996, 1997; Yuan & Ellis, 2003; Skehan, 2009) and connected improvement (e.g., Ahmadian, 2011; Ahmadian & Tavakoli, 2011; Vercellotti, 2012; Yuan & Ellis, 2003) during language performance. The review mainly focuses on studies of oral language performance. It will conclude with the main findings and potential problems revealed in the literature together with the prevailing trend in this area.

2.3.1 Two models on the relationship between CAF constructs

The two well-researched hypotheses concerning the relationship between CAF constructs present (relatively) competing arguments: Skehan's (1998, 2009a) Limited Attention Capacity model proposes competitive relationships in their interaction, such as the trade-off effect, whereas Robinson's (2001a, 2001b) Cognition Hypothesis claims supportive correlations and connected improvements.

Skehan's Limited Attentional Capacity model

The Limited Attentional Capacity model, which is also referred to as the Trade-off Hypothesis (Skehan, 2009a), starts from the initial assumption that there are attentional limitations on performance, associated with limited working memory size, and that the pressure on such limited resources will have implications for L2 production (Wang & Skehan, 2014). This is because learners deploy their limited capacities selectively to reflect whatever performance priorities they have or what the tasks and task conditions support (Foster & Skehan, 1999). As a result of the hypothesis, Skehan (1998) assumed that the three CAF dimensions are prone to compete for resource allocation in L2 task production. Therefore, focusing on one aspect of language performance is highly likely to make other dimensions

suffer. Furthermore, in some incidences, task characteristics and task conditions can prioritise new language and risk-taking, while on other occasions, they can predispose conservatism and error avoidance. Finally, at other times they can push learners to gain fluent control over aspects of the target language (Skehan, 2003b).

In other words, a trade-off effect on CAF constructs will occur depending on task features and/or conditions, preventing them from developing at the same time. This trade-off effect can occur (a) between meaning and linguistic aspects (form), causing learners to shift their focus onto fluency (increase) at the expense of complexity and accuracy (decrease); or (b) between meaning and only one linguistic aspect (form), resulting in fluency and accuracy gains or fluency and complexity gains; or (c) within linguistic aspects (form) themselves, potentially raising the prioritisation of accuracy and depleting the prioritisation of complexity and vice versa.

Among these three possible trade-offs, there is a particular tension between accuracy and complexity, which implies that simultaneously high levels of performance in these two components is unlikely (Skehan, 2009; Wang & Skehan, 2014). In contrast, fluency and accuracy, or fluency and complexity compete with each other to a far lesser extent (Skehan & Foster, 1997, 2001), that is, greater fluency might occur either with greater accuracy or greater complexity, but not both at the same time. In other words, fluency can increase with accuracy/complexity. The competitive relationships between CAF proposed by Skehan have been linked to the finding that each construct relates to different aspects of learners' language use. Specifically, both complexity and accuracy concern form, whereas fluency relates to meaning (Skehan, 1998b). Particular tension is very likely to occur during language performance within form, between control of form (accuracy) and interlanguage risk-taking (complexity). Furthermore, the competitive relationships between CAF constructs are related to the psycholinguistic processes of Levelt's (1989) model. According to this model, speaking is divided into three stages: conceptualisation (whose output is preverbal communication, and whose primary focus is the conceptual content and presentation of what will be uttered), formulation (which accepts the preverbal message and which then engages in processes of lemma selection and consequent syntax-building processes), and articulation (which converts the output of the formulater into actual speech). Both complexity and accuracy are linked to the stages of conceptualisation and formulation in Levelt's speech model, while fluency relates to formulation and articulation (Skehan 2009b; Wang & Skehan,

2014). In this regard, on the basis of Levelt's model, complexity and accuracy are closer to each other than fluency, which leads to particular tension during the language-speaking process.

The particular tension between complexity and accuracy is a basic accounts of the trade-off effects. The fundamental assumption is that as tasks become more difficult, the significance of attentional and memory constraints increases. However, this does not rule out the possibility of the trade-off effect. Another contribution of the Trade-off Hypothesis is its role in helping to explore the extent to which such limitations can be overcome by task characteristics and task conditions (Wang & Skehan, 2014). For instance, Skehan and Foster (2007) found that the effects of trade-offs can be the result of selective influences on different aspects of performance triggered by task characteristics. This means that actual performance depends on how the different combinations of independent variables interact to influence the language that is produced. Occasionally, complexity and accuracy can both be raised, due to the support of independent influences. Concerning the missing trade-off between complexity and accuracy under Skehan's prediction, it is the combination of task characteristics and conditions that will lead to the trade-off effect but occasionally complexity and accuracy will both be raised, such as via the familiarity of information (Skehan & Foster, 2012), and the degree of structure (Tavakoli & Skehan, 2005).

For instance, Tavakoli and Skehan (2005) conducted a study in which the level of structure was manipulated in several cartoon narrative retellings. In the study, Iranian learners of English were required to recount four cartoon series narratives that varied in degree of structure (operationalised as the number of pictures in the picture series whose order could be changed without compromising the story). All three measured performance areas (complexity, accuracy, and fluency) were elevated. This study showed that structure advantaged accuracy and fluency. But it also produced greater complexity, which was attributed to information integration. In this sense, because different characteristics support different performance areas, task characteristic manipulation overcomes trade-off limitations. Complexity and accuracy are therefore not driven forward by the same thing (task difficulty), but by two independent influences and task structure leads to greater accuracy, while information integration produces higher complexity.

This particular interactive influence, a joint increase of accuracy and complexity, may be difficult to achieve ordinarily, but the studies mentioned above show that it is possible. The fundamental assumption of the Trade-off Hypothesis (Skehan, 1998, 2009a) is that limited attention is a necessary starting point and trade-off research aims to examine how pedagogic goals can be met within such constraints, even if they are difficult to achieve.

Robinson's Cognition Hypothesis

In contrast to Skehan's limited resources, the Cognition Hypothesis assumes that attentional resources are multiple and noncompeting attentional pools that learners can access. It also suggests that learners can use multiple attention resources at the same time to pay attention to multiple aspects of language (Robinson, 2001b). Furthermore, the hypothesis proposes that tasks should be designed and sequenced on the basis of gradual increases in cognitive complexity (Wang & Skehan, 2014).

The cognition hypothesis has pedagogical implications for simple to complex task design and sequencing. The related framework categorises "task complexity" (cognitive factors), "task conditions" (interactive factors), and "task difficulty" (learner factors) to classify and sequence L2 pedagogic tasks (Robinson, 2011). Task conditions relate to interactive factors, such as participation and participant variables, while learner factors are defined in terms of the language learner rather than task features. For the purposes of the present study of exploring the effects of task features and conditions on learners' performance, task complexity factors are most relevant. As such, the other two areas will not be pursued here. In terms of task complexity, a theoretical distinction has been made between two categories of dimensions: resource-directing and resource-dispersing dimensions (Robinson, 2005). The task design can either direct resources (this does not hinder performance) or disperse resources (which does hinder performance) (Robinson & Gilabert, 2007).

The effects of task complexity, by increasing the cognitive demands of the task given to learners along two different dimensions, are different. Increasing task complexity along the resource-directing dimension (e.g., talking about more elements rather than just a few elements) can lead learners to map the increasing conceptual demands of tasks to language performance. The accuracy and complexity of adult L2 language production, therefore, can be facilitated, but fluency is likely to be negatively

affected (Robinson, 2003, 2011). This leads to L2 accuracy and L2 complexity often developing together when influenced by increasing task complexity, possibly, but not necessarily, at the cost of fluency. In other words, task complexity increases performance in each of these general areas, which contrasts with the default position of the Trade-off Hypothesis.

Increasing task complexity along the resource-directing dimension often leads to fluency to contrast with complexity and accuracy. Specifically, the resource-directing dimension includes increased cognitive and conceptual demands along with increasing task complexity that can be met by specific aspects of the linguistic system. Increasing task complexity along this dimension can potentially direct learners' attention and memory to the way that L2 structures and codes concepts, leading to language development. Increasing task complexity along this dimension during L2 performance is associated with some recapitulation of a sequence of conceptual development in childhood, which can be met by the use of specific aspects of the L2 that code these familiar adult concepts. It thus represents a natural order for sequencing the conceptual and linguistic demands of L2 pedagogical tasks (Robinson, 2005).

In contrast, increasing complexity along the resource-dispersing dimension, (e.g., through a lack of planning time, through multiple tasks, or through the need to use unfamiliar information), can lead the fluency, accuracy and complexity of production to be negatively affected. This is because increasing task complexity through resource-dispersing factors creates challenges for learners, which hinders their access to their existing repertoire of L2 knowledge (Robinson, 2005). Along with resource-dispersing factors, increasing task complexity divides attentional and memory resources from the features of linguistic code (Robinson, 2011) and does not direct learners to any particular aspects of the language code which can be used to meet the additional task demands (Wang & Skehan, 2014). Therefore, fluency, accuracy, and complexity are influenced negatively.

However, two main criticisms have been directed at the Cognition Hypothesis. First, some scholars argue that there is a lack of compelling evidence to support its predictions, especially when it comes to the aforementioned joint influence on accuracy and complexity. It is frequently the case that either accuracy or complexity is improved, but not both. Occasionally, in some studies, both complexity and accuracy have witnessed a joint increase (e.g., Foster & Skehan, 1999; Tavakoli & Skehan,

2005), but the increase is not sufficient to robustly support the Cognition Hypothesis. Another criticism concerns how much each component of task complexity actually contributes to higher levels of complexity. For example, planning, which is interpreted as resource-dispersing in the Cognition Hypothesis, is said to result in lower performance (if planning time is not available). However, some research on planning raises counter arguments to this interpretation. Specifically, researchers (e.g., Skehan, 2009c, Wang & Skehan, 2014) suggest that planning has different effects on different aspects of performance, with stronger effects on complexity and fluency and smaller, but less reliable, effects on accuracy.

The similarity and differences between the two models

After presenting an overview of the salient features of the Limited Attentional Capacity model and the Cognition Hypothesis, this section discusses the similarities and differences between them. The first similarity that it is important to note is both the Limited Attentional Capacity model and the Cognition Hypothesis aim to explore the same aspects (complexity, accuracy and fluency) of language development. Second, the two influential models agree on the idea that L2 learners make non-neutral decisions when completing tasks (Skehan & Foster, 2001) and that their performance is dependent on the task type and conditions. The two influential models are also both concerned with capturing the effect of task features and conditions on L2 learners' CAF performance (Skehan & Foster, 2001). The final major similarity is that the two models are both insightful in exploring how pedagogic goals can be achieved, from easy to complex in a sequence, by investigating task design and conditions.

However, although there are similarities between the two models, there are also differences. First and most importantly, the starting points of the two models are distinct. For instance, the Limited Attentional Capacity model assumes that L2 learners can only access limited attentional resources. However, in contrast, the Cognition Hypothesis explicitly rejects the notion of limited attentional resources and instead proposes that L2 learners can access an unlimited attentional pool. Moreover, another difference concerns the two dimensions of task complexity, where Skehan does not make a distinction between resource-directing and dispersing, which leads to the claim that accuracy, fluency, and complexity simultaneously decrease on complex tasks along any dimension (Robinson, 2011). Apart from the distinction of resource-directing/dispersing, the relationship between accuracy and

complexity is another important difference between the two models (Skehan & Foster, 2007). On the assumption of non-limited attentional resources, Robinson's model predicts greater complexity and accuracy when influenced by increasing task complexity (Robinson, 2011). Therefore, when accuracy and complexity are both increased, both Robinson and Skehan make predictions, but for different reasons. While task difficulty is the motivator for Robinson, according to Skehan, the motivator is not task difficulty, but rather the combination of task characteristics and task conditions (Tavakoli & Skehan, 2005). Finally, both models hypothesise how to determine the task complexity factor, but they have opposing viewpoints on how the manipulation and sequencing of the cognitive characteristics of tasks influence L2 participants' CAF and how their attention is deployed when executing their performance. The Limited Attentional Capacity model focuses on how tasks are implemented through language, cognition and performance conditions (Skehan, 1996b, 1998) and how learners' performance is affected as a result. However, the Cognition Hypothesis highlights the task features that influence the difficulty of the task, and how learners' performance is impacted as a result. From this perspective, the Limited Attentional Capacity model is very likely to be impacted by learners' proficiency levels, whereas the Cognition Hypothesis is more associated with language learners' development within a short time period.

In conclusion, considering the similarities and differences between the two models, it is clear that they both have merits and shortcomings. Furthermore, it is very unlikely that either of these two models can provide the whole picture of L2 learners' language development since language itself does not develop in a straightforward manner. Therefore, some other studies have sought support from Dynamic Systems Theory (de Bot, Lowie, & Verspoor, 2007; de Bot, 2008; Norris & Ortega, 2009; Polat & Kim, 2013) to explain L2 learners' language performance (e.g., Larsen-Freeman, 2006, 2009; Vercellotti, 2017, 2019).

2.3.2 Experimental studies on the relationship between CAF constructs

In terms of the relationship between CAF constructs, two major types of interaction, namely, the trade-off effect and connected improvement, have been proposed theoretically via the two well-documented models described above (see 2.3.1). These two types of interaction have also been investigated by empirical studies. The mainstream research explores the relationships between CAF constructs (e.g., Ahmadian & Tavakoli, 2011; Skehan & Foster, 1997; Foster, 2001a; Yuan & Ellis, 2003;

Vercellotti, 2012, 2017). However, a handful of studies have specifically aimed to explore the correlations between subdomains within each construct of CAF (Mora & Valls-Ferrer, 2012; Vercellotti, 2012, 2017), and, in particular, the domain of complexity (Bulté, 2013; Vercellotti, 2019; David et al., 2009; Larsen-Freeman, 2006).

The present research follows the second approach and aims to further explore the relationship between subdomains within each construct of CAF. Therefore, the next section presents a review of the studies which explore the trade-off effect and the connected improvement pattern between complexity, accuracy, and fluency in language performance, and, in particular, in oral performance. A review of the correlations between the subdimensions of each CAF construct will also be provided.

Studies on the trade-off effect and connected improvement between constructs

As Foster and Skehan (1996) discovered, both complexity and accuracy concern form, whereas fluency reflects the primacy of meaning. Skehan (1998a) further claimed that there is tension between meaning (measured as fluency) and form (either complexity or accuracy) in learners' language performance. Specifically, this view suggests that there is a particular tension within form (between accuracy and complexity) as well as a meaning-form tension (between fluency and complexity, or between fluency or accuracy) that occurs as a secondary tension during language performance (Wang & Skehan, 2014). Based on the assumption of limited mental resources, Skehan (2009) predicted a competitive relationship between CAF, which leads to the trade-off effect. These limitations to learners' language performance have been widely accepted by many researchers (e.g., Ahmadian & Tavakoli, 2011; Bygate, 2001; Crookes, 1989; Michel, Kuiken & Vedder, 2007; Skehan & Foster, 1996; Yuan & Ellis, 2003; Ellis & Barkhuizen, 2005).

In an early study, Skehan and Foster (1996) found a trade-off between accuracy and complexity when exploring the effect of planning during three oral tasks. On all measures, the planning group outperformed the non-planning group. Specifically, the planning group outperformed the non-planning group in accuracy but not in complexity during the narrative task, and the planning group outperformed the non-planning group in complexity but not in accuracy during decision-making tasks. During the personal information task, however, the planning group outperformed the non-planning group on all three measures. Skehan and Foster (1996) also discovered that planning led to greater accuracy on the

personal and narrative tasks, but not on the decision-making task; that on the narrative task, planning led to greater accuracy, but without evidence of complexity; and that on the decision-making task, planning led to greater accuracy, but without evidence of complexity. According to Skehan and Foster (1997), the trade-off between complexity and accuracy, within the form, was influenced by the pressure on learners' limited working memory which was caused by various types of planning. Yuan and Ellis (2003), when studying the effect of planning on oral language performance, found a similar trade-off effect between accuracy and fluency, a meaning-form tension, based on group score comparisons.

Similarly, Crookes (1989) reported greater complexity and lexical variety for tasks done under a planning time condition, but, interestingly, no greater accuracy, which suggested a trade-off between complexity and accuracy. In a study that looked at the effect of task repetition, Bygate (2001) found a similar effect between complexity and accuracy. Specifically, Bygate demonstrated how complexity and fluency (but not accuracy) improve together when learners repeat a task, suggesting a trade-off effect between accuracy and complexity. Interestingly, when testing the Cognition Hypothesis, Michel, Kuiken, and Vedder (2007) discovered that students who completed a more difficult task had increased accuracy but decreased fluency (due to the dialogue condition), with no significant effect on language complexity, implying a trade-off between accuracy and fluency. Yuan and Ellis (2003) also found a trade-off between accuracy and fluency, a meaning-form tension, with a careful online planning condition. A trade-off effect between accuracy and complexity, within form tension, with an online planning condition (OLP), were also reported. They concluded that the task design is very likely to direct learners' attention.

Likewise, due to the task design, Ahmadian and Tavakoli (2011) found higher accuracy and grammatical complexity at the expense of fluency. When describing a cartoon-based narrative, learners who were encouraged to undertake careful online planning (learners have ample time to plan their speech) had higher accuracy and grammatical complexity than students in the pressured online planning condition (learners are required to produce language in 6 minutes), who had higher fluency. As a result of this between-group study, strong performances in accuracy and grammatical complexity were found at the expense of fluency, indicating that there is a meaning-form tension, but not a tension within form and between control of form (accuracy) and interlanguage risk-taking (complexity).

As mentioned in these studies, a particular tension within the form (between accuracy and complexity), as well as a secondary tension-meaning (measured as fluency) form (accuracy and complexity) tension have been reported in line with Skehan's (1997, 1998) predictions. Moreover, Michel, Kuiken, and Vedder (2007) reported that the task seems to direct learners' attention. Certain tasks appear to relieve some of the tension on attentional resources, such as personal information tasks, which have higher accuracy and fluency, and pre-task planning, and this allows learners to produce language with more complexity and fluency (Skehan, 2009). This has led to studies which show a supportive relationship between CAF constructs.

In contrast to the trade-off effect, some researchers have found that the three constructs of CAF show a connected improvement pattern. For instance, Skehan and Foster (1996) examined the oral performance of pre-intermediate L2 English learners in three oral tasks (personal information, narrative, and decision-making), and found that their accuracy (proportion of error-free clauses), complexity (clauses/c-units), and fluency with planning in a personal information task all showed growth. This is very likely because complexity is promoted by planning, meanwhile, accuracy and fluency are enhanced by the information task due to information familiarity (Mora & Valls-Ferrer, 2012). Following Skehan's (2009c) argument, this connected improvement in all three performance areas has been attributed to different reasons triggered by the task design.

Ahmadian and Tavakoli (2011) investigated the effects of the simultaneous use of task repetition and careful online planning (operationalised as the provision of ample time for task performance) on the CAF of EFL learners. They investigated the oral performance of Iranian intermediate-level EFL learners when they described a cartoon-based narrative with the simultaneous use of careful online planning and task repetition conditions. They pointed out that this simultaneous use of careful online planning and task repetition positively impacted the learners' accuracy, complexity and fluency. Their results revealed that task repetition positively impacts complexity and fluency, while online planning advantages fluency. They concluded the lack of the expected trade-off effect (between accuracy and complexity) was influenced by separate aspects of the task design. This links to Skehan's (2009c) study which concluded that raised accuracy and complexity are not automatically attributable to the Cognition Hypothesis. Alternatively, complexity and accuracy are expected to come together through task manipulation and task conditions.

In the same year, when analysing the effects of massed task repetition (11 times) of an oral narrative task transferred to the performance of an interview task, Ahmadian (2011) reported that his repeated measures with students in two conditions (massed repetition and control) showed that the repetition group increased their complexity and fluency scores, particularly words/AS-unit, but not their accuracy scores, including error-free clauses. This supports Skehan's Trade-off Hypothesis between accuracy and fluency, implying that focussing on complexity and fluency restricts the capacity for accurately processing language. In this regard, the findings of Ahmadian's (2011) study are consistent with those of previous task repetition studies (Ahmadian & Tavakoli, 2011; Bygate, 2001), that is, that task repetition positively impacts participants' complexity of language on a different task. The connected improvement in both complexity and fluency impacted by task repetition is contrary to the meaning and form tension, but it supports the tension between meaning (measured as fluency) and form (either complexity or accuracy) (Skehan, 2009; Wang & Skehan, 2014).

Questioning the inevitability of the trade-off effect, Vercellotti (2012) analysed the oral performance of 66 English learners over 3-9 months with nine CAF measures. The analysis showed connected improvement between CAF constructs in language performance during topic-prompted monologues with one-minute planning time. Vercellotti (2012) suggested that instructed language performance growth occurs uniformly, rather than along individual paths. Similar results have been observed in her later studies when learners' oral performance was elicited in multiple topic-centred monologues. For instance, Vercellotti (2017) observed the oral performance of L2 learners of English over several months. Her within-individual correlation analysis also showed a connected-improvement pattern between CAF constructs, within which most of the correlations were weak because the CAF measures capture different aspects of language performance. It is worth noting that the data analysis in those two studies was based on individual learners' performance. Vercellotti (2017) concluded that individual development did not show CAF trade-offs, and that her data did not support the expected trade-off effect which is often based on group means.

Studies on the trade-off effect and connected improvement within each construct of CAF

Although the majority of studies have focussed on the correlations between CAF constructs, there is an increasing trend to investigate the correlations between subconstructs within/across each construct (e.g., Bulté, 2013; Mora & Valls-Ferrer, 2012; Spoelman & Verspoor, 2010; Vercellotti, 2012, 2017, 2019). However, this work has often only explored L2 writing development (e.g., Bulté, 2013; Spoelman & Verspoor, 2010). The competitive relationship in interaction (the trade-off effect) and supportive in correlation (connected improvement) revealed between CAF constructs have also been observed between subconstructs of each domain of CAF as well as between subconstructs across CAF.

Lexical complexity, encompassing lexical variety and sophistication, as a subdomain of complexity, has been revealed to have a mixed relationship, i.e., a competitive and a supportive relationship, both with syntactic complexity (with complexity), and accuracy (across domains of CAF). However, this largely depends on the nature of the measures. For instance, Yuan and Ellis (2003) reported a lexical variety and accuracy trade-off in the oral production of narratives, suggesting that when learners use more varied lexical items, more errors occur.

In terms of the relationships between lexical complexity measures and accuracy, on the basis of a review of his research, Skehan (2009a) stated that lexical sophistication competes with accuracy and syntactic complexity. Specifically, Skehan noted that less frequent words are associated with lower accuracy and mainly lower syntactic complexity, which reveals a trade-off between lexical sophistication and syntactic complexity, and between lexical sophistication and accuracy. However, as a subconstruct of lexical complexity, lexical diversity (measured by D) has been shown to be positively related to accuracy (measured by error-free clauses). Lexical diversity (D) is an indicator of the extent to which L2 speakers avoid the recycling of the same set of words. The less recycling of vocabulary during a language performance (higher lexical variety) is related to higher accuracy. This is very likely because speakers do not experience trouble (fewer errors corrections) during their utterances, which provides them more time to avoid the repetition of lexical items during speaking. Skehan (2009a) concluded that lexical variety (D) correlates negatively with complexity in the majority of cases. In other words, L2 speakers recycle vocabulary most (lower lexical variety) which enables them to achieve greater complexity. In general, trade-offs between lexical variety and complexity are revealed.

Within the construct of complexity, it has also been found that lexical complexity is negatively correlated with syntactic complexity via subordination suggesting a trade-off effect. Within the subdomain of lexical complexity, lexical diversity had very low, nearly non-existent, correlations with lexical sophistication revealing that they are independent of one another.

However, Vercellotti (2012) reported that lexical variety is positively correlated with AS-unit-level accuracy and clause-level accuracy showing a connected growing pattern between lexical complexity and accuracy, as well as lexical complexity and syntactic complexity. Vercellotti (2012) concluded that the joint improvement between lexical complexity and accuracy was very likely influenced by the data analysis (within individual correlation) based on individual learners' performance with a longitudinal design.

Similarly, a connected improvement between accuracy measures (the percentage of error-free AS-units and error-free clauses) was found by Vercellotti (2012). Furthermore, a joint improvement between fluency measures (mean length of pause, mean length of fluent run, and phonation time ratio) was also revealed in her study. No trade-offs were found between complexity measures, which encompass the subcategories of between syntactic measures (length of AS-unit and clauses per AS-unit), and between syntactic and lexical complexity measures (length of AS-unit and lexical variety) (Vercellotti, 2012). Moreover, within the subdomain of lexical complexity, a supportive relationship between grammatical complexity and lexical variety has also been noted (Vercellotti, 2017). Building on her earlier work, Vercellotti (2019) reported no trade-offs when investigating the oral development of syntactic complexity in the short topic-based monologue speech of 66 English L2 learners over three academic semesters. Instead, the results showed overall growth over time on both commonly used measures of syntactic complexity (e.g., length of AS-unit, clause length, subordination index) and more specific measures of structural complexity (e.g., syntactic variety and weight of complexity scores). Within the construct of complexity, a connected and supportive development between these complexity measures was reported, suggesting learners are not forced to prioritise certain aspects of language performance at the expense of others with increasing proficiency. This is consistent with the findings of her earlier studies (Vercellotti, 2012, 2017), suggesting that with increasing proficiency, cognitive resources are available for complexifying language performance. This longitudinal analysis revealed simultaneous growth in those measures, unlike cross-sectional data.

The available studies (Vercellotti, 2012, 2017, 2019; Spoelman & Verspoor, 2010) which have explored the subconstructs of complexity, have revealed a supportive relationship between complexity measures. Furthermore, David, Myles, Rogers, and Rule (2009) found lexical variety (Guiraud's Index) was significantly correlated with global grammatical complexity when aggregated across age groups, suggesting a connected improvement pattern. Vercellotti (2019) concluded that concerning the development of syntactic complexity, these results can only reflect the development of the complexity of ESL oral monologues elicited with topic prompts within a longitudinal design but that a different pattern is very likely with other task types. Furthermore, Vercellotti (2019) concluded that the lack of trade-offs within the construct of complexity, supports certain theories (e.g., Dynamic Systems Theory) that view language as a complex, interrelated system where development in one area does not necessarily hinder growth in another (de Bot, 2008), even among closely related subsystems, such as within syntactic complexity. Indeed, because of this, the increasing trend is to apply Dynamic Systems Theory to capture the effect of task features and conditions on L2 learners' CAF performance.

According to studies examining the relationships between complexity, accuracy, and fluency as well as between the subdimensions of each CAF in oral performance, trade-off effect and the connected improvement pattern occur both between and within each subdomain of CAF. Regarding the task-related elements, which are the determining factors impacting the correlations between CAF measurements, the details will be provided as follows.

Interim summary

Previous studies on L2 oral performance have revealed that CAF measures are sensitive to task type and conditions (Mora & Valls-Ferrer, 2012). In general, research on L2 learners' task performance has revealed that connected improvement among complexity, accuracy, and fluency are enhanced through task features and conditions. For instance, fluency and accuracy are improved through information familiarity, such as personal tasks, which leads to higher accuracy and fluency (e.g., Foster & Skehan, 1996). Concerning the degree of structure, structured tasks have been found to be more fluent and sometimes more accurate (e.g., Skehan & Foster, 1999; Tavakoli & Skehan, 2005). It has also been revealed that compared to monologic tasks, dialogic tasks lead to greater accuracy and complexity

(e.g., Tavakoli, 2016). In terms of the effects of planning on oral performance, pre-task planning consistently produces greater complexity and fluency (e.g., Skehan, 2009c). Similarly, task repetition has a significant effect on the fluency and complexity of learners' performances (e.g., Ahmadian, 2011; Ahmadian & Tavakoli, 2011; Bygate, 2001). Furthermore, concerning the interrelationship between CAF measures, data analysis (group means vs. individual growth curves), research design (cross-sectional vs. longitudinal), the nature of measures, and task design, including task type and conditions, can explain the major differences in findings among different research (Vercellotti, 2017).

These generalised findings on the effects of task type and conditions on oral performance have been revealed by previous studies. Therefore, when interpreting the results of studies on the trade-off effect as well as connected improvement among CAF measures, it is necessary to consider the separate influence from different variables concerning task type and conditions. For instance, Foster and Skehan (1997) reported accuracy and complexity are affected significantly on the personal task. However, planning produces significantly better results only on the narrative task for accuracy and the decision-making task for complexity. Clearly, both task types (personal task, narrative task, and decision-making task) and conditions (planning) have an impact on learners' oral performance. Planning has a significant impact on oral fluency, but the effects of planning on language accuracy and complexity are less clear.

However, there are some potential shortcomings in the previous CAF studies. For instance, the research pays most attention to the relationships between CAF constructs (e.g. Ahmadian, 2011; Ahmadian & Tavakoli, 2011; Bygate, 2001; Michel, Kuiken & Vedder, 2007; Mora & Valls-Ferrer, 2012; Yuan & Ellis, 2003; Vercellotti, 2012), while the interactions between subconstructs within each construct and/or subconstruct of CAF have been explored much less frequently (e.g., Mora & Valls-Ferrer, 2012; Vercellotti, 2017, 2019). Therefore, in an effort to help remedy this imbalance in the existing research, this study will investigate the correlations between subconstructs within CAF to explore whether the trade-off effect, as well as connected improvement between subconstructs within CAF, can contribute to the interpretation of task design and conditions on L2 learners' oral performance.

Moreover, there are unclear criteria to decide which indicator represents each construct of CAF. It is also challenging to numerically count the weight of CAF indicators used to analyse oral CAF

development. For instance, in terms of complexity measures, Guiraud's index, lexical word ratio, clause-to-AS-unit ratio, mean length of AS-unit were investigated in Mora & Valls-Ferrer (2012) and clauses/c-units and syntactic variety were used to represent complexity development in Skehan and Foster (1996). This inconsistency in the use of CAF measures is common across CAF studies and it leads to questions concerning the reliability of their findings. In order to generate more robust findings and to enable comparison between different studies, there is a need to use uniform indicators. This area requires attention in future research and is why the present study attempts to use different indicators to the best degree to capture the subdomains of fluency and complexity development.

Furthermore, the majority of studies have employed cross-sectional designs to explore the trade-off effect in learners' performance across different proficiency levels since data can be collected comparatively easily from participants within a short period (Ahmadian & Tavakoli, 2011; Bygate, 2001; Chen, 2015; Yuan & Ellis, 2003; Vercellotti, 2019). It is also important to note that the results generated regarding learners' performance are impacted by their proficiency levels and the duration of the examining period. To remedy these shortcomings, Norris and Ortega (2009) suggested exploring the interactions between the three CAF constructs with longitudinal studies (e.g., Mora & Valls-Ferrer, 2012; Vercellotti, 2012; 2017; 2019) so that the language learning process as the interactions between CAF constructs can be examined over time. Moreover, concerning the relationship between complexity, accuracy, and fluency in language performance, Skehan (2009c) concluded that the trade-off effect in explaining how attentional limitations constrain second language performance is only a starting point. Therefore, the Cognition Hypothesis and a focus on task design are needed to understand results concerning learners' second language performance when explored by complexity, accuracy, and fluency measures. Skehan (2009c) has called for CAF studies to explore and identify oral performance in more contexts and conditions.

Responding to this call, researchers have started to investigate the effects of learning contexts (Study Abroad and Formal Instruction at home) on relationships among CAF measures, which imply how different linguistic elements develop in relation to each other. This is one important limitation of SA research focusing on linguistic development. The majority of research has focused on how SA influences the development of individual linguistic elements in L2 oral performance. There are only

several studies have explored relationships among CAF measures after SA (e.g., Leonard & Shea, 2017; McManus et al., 2021; Mora and Valls-Ferrer, 2012). For example, Mora and Valls-Ferrer (2012) examined the effects of two learning contexts (Study Abroad and Formal Instruction) on the oral production of advanced-level Catalan-Spanish learners of English over a 2-year period (including a 3-month SA period). The speech samples were elicited through a guided interview without planning time. The results revealed that in terms of the relationships among CAF measures after SA, in general, correlations are relatively weak and largely nonsignificant. If there are significant correlations, the results suggested that more fluent learners also produce more accurate and complex language. Specifically, fluency (mean length of run) was revealed to improve in tandem with complexity via length (words per AS-unit) and subordination (clauses per AS-unit). Connected growth in fluency and accuracy has been reported as well in this study, and has revealed that errors per AS-unit are positively correlated with Pause Time Ratio (PTR) and pause frequency, which suggested that learners who produce more errors also pause more. The connected improvements between CAF measures after study abroad in this study, which is very likely arise because study abroad assists fluency (e.g., Collentine & Freed, 2004; Du, 2013). However, in contrast to the findings of Mora and Valls-Ferrer (2012), which showed that no relationships among fluency and lexis after SA, McManus et al. (2021) found significant and long-lasting (before, during and after SA) relationships between fluency and lexis. McManus et al. (2021) concluded that the difference of relationships between fluency and lexis is possibly because the different task types used in these two studies. In contrast to oral interviews used in Mora and Valls-Ferrer (2012), a monologic task was used to elicit the learners' speech in McManus et al. (2021). In terms of accuracy, McManus et al. (2021) found that accuracy correlated with fluency only before and after SA, but not during SA. This indicates that accuracy relationships among other CAF indicators were not stable, unlike the fluency-lexis relationships.

To contribute to the limited area of how different linguistic elements develop in relation to each other in the SA field, Leonard and Shea (2017) investigated the speaking development of advanced learners of Spanish at the beginning and end of a 3-month stay abroad. The analysis of CAF relationships revealed pre-SA relationships (positive correlations) between fluency, lexical complexity, and accuracy, though syntactic complexity did not correlate with any other CAF indicators. All of the CAF dimensions were significantly improved at post-SA. The individual CAF gain scores did not show any meaningful connections. Moreover, no significant connections between the various CAF gain scores

were discovered. Leonard and Shea (2017) suggested that CAF indicators may grow independently of one another in the short term (e.g., short-term fluency development is independent of accuracy), implying that CAF relationships arise slowly, driven by longer language use.

To investigate how different linguistic elements develop in relation to each other in the Formal Instruction at home context, Mora and Valls-Ferrer (2012) noted that complexity and accuracy measures did not correlate significantly and strongly with most of the fluency measures. This is a similar pattern of correlations as noted after SA in this study. However, the specific correlations between CAF were not reported in Mora and Valls-Ferrer's study, which makes it hard to interpret the results in a precise manner. Furthermore, concerning the effects of the two learning contexts (Study Abroad and Formal Instruction at home) on oral performance, Mora and Valls-Ferrer (2012) concluded that accuracy and fluency saw more gains during Study Abroad than during a period of Formal Instruction. They also found that complexity remained stable across the whole examining period despite improved fluency and accuracy.

In line with the trend of exploring the relationship between complexity, accuracy, and fluency in language performance in different learning contexts, this study will investigate the oral performance of English-speaking learners of Chinese in different learning contexts (Study Abroad and Formal Instruction) using a longitudinal design. The main aim is to see whether there are connected improvements or trade-offs in any areas of CAF and contribute to testing the attentional limitations proposed by Skehan (2009c). Moreover, the interrelationship between CAF constructs, similar to the majority of CAF studies, as well as the relationship between subconstructs within fluency and complexity, will also be explored. Therefore, the study attempts to contribute to an area that has not yet been investigated in great detail. This attempt will pay particular attention to the understudied field of L2 speaking performance in Chinese.

2.4 Main factors affecting L2 oral performance

Identifying factors which affect the synchronic manifestation and diachronic development of the CAF triad has become an issue of increasing attention in the L2 field (Housen et al., 2012). Typically, two categories of factors have been examined to explore their impact on L2 learning and L2 use:

internal and external linguistic factors (Housen et al., 2012). The former consists of linguistic features (e.g., patterns, rules, items), while the latter includes learner factors (e.g., age, motivation, aptitude); contextual factors (e.g., Formal Instruction At-Home vs. Study Abroad); task variables, (e.g., oral or written, monologic or dialogic) (e.g., Larsen-Freeman, 2009; Mora & Valls-Ferrer, 2012); type of pedagogical intervention (e.g., task-based teaching) (e.g., Larsen-Freeman, 2009; Norris & Ortega, 2009; Jesen & Howard, 2014); and the complexity of the task (e.g., Robinson, 2011). These factors are often investigated individually or together, as shown below in Table 7.

Table 7. Investigation of factors on oral performance in L2 studies

Factors	**Sub-categories**	**Studies**
Learning contexts	(Abroad/domestic)	Freed, Segalowitz, and Dewey (2004); Pérez-Vidal and Juan-Garau (2009); Segalowitz and Freed (2004); Collentine (20090; Jensen & Howard (2014); Du (2013); Liu (2009), Mora and Valls-Ferrer (2012); Wright and Zhang (2014); Wright (2020);
Experimental factors	Task features/types	Skehan (2009a); Skehan and Foster (1997); Larsen-Freeman (2009); Li et al. (2003); Liu (2017a); Wang (2011); Wang (2018)
	Planning	Yuan and Ellis (2003); Skehan (2009a);
	Task complexity	Ellis (2009); Ellis and Yuan (2004); Tavakoli and Skehan (2005); Robinson (2011); Robinson (2015)
	Proficiency levels	Serrano (2011)
	The relationship between objective and subjective analysis	Jin and Mak (2013); Wu (2017)
Learner factors	Identity	Chen (2020)
	Learning anxiety	Zhang (2001)

One of the external factors which affects L2 acquisition is context, which normally comprises three categories: Foreign Language classroom in a domestic setting, Study Abroad (SA), and Intensive Domestic Immersion context (IM) (Collentine, 2009). FL is often replaced by Formal Instruction At Home (FI) in other studies. Contextual factors, in particular, study abroad as a learning context has attracted much attention in L2 studies (e.g., Mora & Valls-Ferrer, 2012; Serrano, Llanes & Tragant, 2011; Valls-Ferrer & Mora, 2014; DeKeyser, 2014). However, SA has only received limited attention in L2 Chinese studies (e.g., Diao, Donovan & Malone, 2018; Du, 2013; Liu, 2009; Kim et al., 2015; Wright & Zhang, 2014; Wright, 2018, 2020). Moreover, because L2 language acquisition itself is multifactorial, some studies also investigate other factors during Study Abroad. For example, Diao, Donovan and Malone (2018) investigated the quality of interaction with host families on the L2 Chinese learners' oral gains during a one-semester SA period. During a recent study, Wright (2020) examined

task effects (rehearsed vs. spontaneous speech, in monologic and dialogic mode) on L2 Chinese oral fluency during a 10-month SA period.

Since only a handful of empirical studies have examined L2 Chinese speaking development during Study Abroad (e.g., Du, 2013; Liu, 2009; Wright & Zhang, 2014; Wright, 2018, 2020), to enrich this area, this study will explore the effect of Study Abroad on the speaking development of L2 Chinese learners. The next section will review the effect of learning contexts, in particular, Study Abroad on the oral performance of L2 learners of Chinese.

2.4.1 Study Abroad

Study Abroad (SA) research has grown extremely rapidly over the last two decades, a situation which has been stimulated by the growing global popularity of SA programmes (Yang, 2016), and large-scale projects such as SALA (Perez-Vidal, 2014) and LANGSNAP (Devlin, 2019). SA studies have often been conducted through SA-FI (Formal Instruction) comparisons (e.g., Collentine, 2004; Segalowitz, Freed, & Collentine, 2007) as well as through SA-FI-IM comparations (e.g., Segalowitz & Dewey 2004). However, these comparative studies have only been relatively scarce (Yang, 2016). As the first effort to synthesise SA research, Freed (1995a) pointed out that SA is beneficial in many aspects, but it might not be superior to FI classroom instruction for some aspects of linguistic development, such as morphosyntactic abilities. Support for SA's benefits came from Segalowitz and Freed (2004) who compared the oral gains from both FI and SA contexts on the oral performance of 40 native speakers of English studying Spanish over one semester. Their results showed that learners achieved significant gains in oral performance during SA.

However, some studies have suggested that SA and FI benefit different aspects of speech development (e.g., Juan-Garau & Pérez-Vidal, 2007; Trenchs-Parera, 2009). SA appears to benefit oral fluency and vocabulary skills more than at-home taught situations (e.g. Segalowitz and Freed 2004; Pizziconi, 2017), the reverse (FI) has frequently been observed for accuracy and syntactic complexity abilities (e.g. Isabelli, 2010; DeKeyser, 2017). For example, Juan-Garau and Pérez-Vidal (2007) aimed to explore the effects of both SA and FI on the oral performance of L3 university learners of English at an advanced level. Their data was collected through a role-play task and from four tests before and after the 3-month SA over 2 years. The results showed that the learners made significant oral fluency

gains during SA, while in the FI context, on the one hand, the frequency of errors, the length of clauses and sentences, and grammatical complexity were negatively affected but, on the other hand, subordinates and vocabulary improved significantly.

There have also been studies which explore SA benefits through SA-FI-IM comparisons. For example, Freed, Segalowitz and Dewey (2004) compared the oral fluency gains of 28 students of French in the three contexts, formal language classrooms in an FI (FI) context, an intensive summer immersion (IM) programme, and a Study Abroad (SA) setting. The results showed that the FI group made no significant gains, and the IM group made significant improvements. However, the SA group did not achieve many more gains than the other two groups. The results also showed that the learners enrolled in the intensive language programme outperformed SA learners in terms of L2 fluency.

Such mixed findings have led researchers to re-examine the impact of SA. The core of SA research lies in the validity of SA, and whether L2-immersive environments can be provided for learners (Yang, 2016). Some studies have found that SA might not always enable consistent improvements (Kinginger, 2011). Specifically, although SA is beneficial in many ways, it might not be superior to FI classroom instruction in some important aspects of linguistic development (e.g., morphosyntactic abilities) and for all levels of development (Collentine & Freed, 2004). It has also been revealed that oral proficiency gains during SA do not always outperform an FI instruction context (e.g., Juan-Garau & Pérez-Vidal, 2007; Cohen & Shively, 2007; Collentine, 2009; Pérez-Vidal, 2014). For instance, study-abroad learners also do not necessarily achieve greater language gains than their peers who study the target language in the FI context (Cohen & Shively, 2007; Collentine & Freed, 2004). As indicated, immersion might not always benefit L2 learners' performance. Likewise, studies have showed that SA groups make less progress in their fluency development, particularly in terms of hesitations and breakdowns (i.e., mean length of utterance, disfluency) (e.g., Freed, Segalowitz & Dewey, 2004; Wright, 2018, 2020; Wright & Zhang, 2014). These contradictory findings, as well as the main distinction between the SA and FI contexts, namely that SA offers immersive learning conditions which are missing from the FI context, have led researchers to explore SA in its own right (Sanz, 2014).

It is traditionally assumed that L2 learners' language development is aided by extensive access to the target language during SA (Paige et al., 2012; Dewey et al., 2014). Specifically, overall proficiency has been revealed to significantly improve (e.g., Segalowitz & Freed, 2004; Pérez-Vidal & Juan-Garau, 2011; LIanes & Serrano, 2017; Wright & Zhang, 2014) alongside particular aspects of learners' linguistic development, such as oral fluency in terms of general fluidity (i.e., greater output, less silences) (e.g., Collentine & Freed, 2004; Du, 2013; LIanes & Serrano, 2017; Mora & Valls-Ferrer, 2012; Pérez-Vidal & Juan-Garau, 2011; Serrano, LIanes & Tragant, 2011; Llanes & Muñoz, 2013; Pérez-Vidal et al., 2012). For instance, after reviewing several studies (Segalowitz & Freed, 2004; Lafford, 2004; Collentine, 2004) on the effects of learning contexts on L2 language acquisition, Collentine and Freed (2004) concluded that students achieve significant gains in oral fluency, in particular, in terms of ease and smoothness of speech, which is produced at more native-like speed as measured by temporal and hesitation phenomena as a result of SA. Another meta-analysis by Collentine (2009) which reviewed 13 SA studies revealed that fluency often benefits from SA experience, which is consistent with the conclusions drawn from the SA literature. This is also in line with the findings of studies which have examined L2 Chinese fluency development during SA (Du, 2013; Kim et al., 2015). For instance, Kim et al. (2015) analysed 24 L2 Chinese learners' development in regard to fluency (speech rate, frequency of filled and unfilled pauses, and mean pause length), during a 16-week SA programme in China. The students completed three questions based on a modified version of the Simulated Oral Proficiency Interview (SOPI) at the beginning and end of the programme. The study found that during the SA the learners' objective fluency measures (speech rate, frequency of filled and unfilled pauses, and mean pause length) improved. Generalized the findings of these previous SA research (e.g., Tullock & Ortega, 2017; Valls-Ferrer & Mora, 2014), it has been concluded that overall fluency increases during study abroad when speech becomes more rapid (speed), exhibits fewer and shorter pauses and hesitations (breakdown), and contains fewer self-repairs (repair).

Concerning the effects of SA on oral accuracy and complexity, there have been mixed findings. For instance, while some studies have shown accuracy to increase significantly during the SA context (e.g., Juan-Garau, 2014; Llanes & Muñoz, 2013; Mora & Valls-Ferrer 2012; Pérez-Vidal et al., 2012), others have found no statistical improvement (e.g., Serrano, LIanes & Tragant, 2011). The discrepancy between the findings is very likely to result from the length of the SA period. For instance, Serrano, LIanes and Tragant (2011) examined three groups of Spanish-speaking university students who were

exposed to English in three different contexts: FI intensive, FI semi-intensive and SA. The SA students' oral post-tests data were collected at two time points: 15 days and two months after the pre-test. The results revealed that the SA context was the most advantageous for oral development in terms of fluency and lexical complexity, but not accuracy. Therefore, learners find it difficult to improve their oral accuracy within a short SA period (i.e., 2 months).

Regarding lexical complexity, it seems there is a less clear benefit from SA. If there is any, this is generally attributed to the rich linguistic contact with native speakers that SA enables (Dewey, 2008). Some SA research showed that SA context does not guarantee greater lexical gains than FI context (Collentine, 2004). In contrast, vocabulary/lexical breadth development has been found to improve significantly because of an increased lexical repertoire formed during SA (e.g., Collentine & Freed, 2004; Dewey, 2008; Leonard & Shea, 2017; Kim et al., 2015). Applying CAF measures to assess lexical development, SA seems to be more advantageous for the development of oral production in terms of lexical diversity (measured by Guiraud's Index) (e.g., LIanes & Serrano, 2017; Juan-Garau & Pérez-Vidal, 2007; Mora & Valls-Ferrer, 2012). In contrast, some other research suggests that learners' lexical complexity does not significantly improve during SA (Pérez-Vidal & Juan-Garau, 2011; Pérez-Vidal et al., 2012; Serrano, LIanes & Tragant, 2011; Llanes & Muñoz, 2013; Pérez-Vidal et al., 2012; Wright, 2018, 2020). In terms of syntactic complexity, several previous studies have found that SA is very beneficial (e.g., Juan-Garau & Pérez-Vidal, 2007; Jensen & Howard, 2014; Pérez-Vidal & Juan-Garau, 2011; Llanes & Muñoz, 2013; Mora & Valls-Ferrer, 2012) concerning overall complexity which is normally measured by the length of units. However, some other research has found no statistical significance concerning the complexity by subordination that is measured by the clauses per unit (LIanes & Serrano, 2017; Serrano, LIanes and Tragant, 2011; Mora & Valls-Ferrer, 2012).

Though contradictory findings have been revealed regarding the effect of SA on oral performance, a consensus has been reached that not all of the aspects of oral performance gain significant improvements. If gains are made, they tend to occur in oral fluency and vocabulary rather than accuracy and syntactic complexity (e.g., Leonard & Shea, 2017; Mora & Valls-Ferrer, 2012; Segalowitz et al., 2004; Serrano, Llanes and Tragant, 2011; Valls-Ferrer & Mora, 2014; DeKeyser, 2014). For instance, in their longitudinal study which traced the oral performance of L2 advanced learners of English over 2 years, Mora and Valls-Ferrer (2012) explored the impact of learning contexts involving

formal instruction (FI) and 3-month study abroad (SA) on speech production. The speech data were elicited by an interactive oral role-play interview containing seven fixed questions at three collection times. The questions were considered to be at a similar complexity level to elicit similar speech output. 13 CAF measures were used to provide a comprehensive description of the participants' oral gains during the FI and SA period. The results showed significant improvements on all fluency measures except the dysfluency ratio. In contrast, accuracy was revealed to only improve marginally. In terms of complexity, there were no gains in both lexical and structural complexity. Similarly, Mora and Valls-Ferrer (2012, 2014) revealed that during the SA period, fluency showed robust improvement, complexity did not improve, while accuracy achieved modest improvement.

Furthermore, in terms of the retention of SA effects on oral gains, the findings concerning fluency and accuracy differ from each other (6 months to 3 years formal instruction after SA). Specifically, learners' oral fluency gains from a 3-month SA period have been suggested to revert to their previous levels after 6 months back home without instruction (e.g., Juan-Garau & Pérez-Vidal, 2006, 2007; Pérez-Vidal & Juan-Garau, 2005). However, other longitudinal studies (e.g. Howard 2009; Huensch & Tracy-Ventura 2017; Llanes 2012; Regan 2005) have in general found that linguistic gains made abroad are maintained over the course of a year when participants continue to receive formal instruction. Therefore, the impact of SA on oral gains have reached inconclusive results. To help interpret this situation, Juan-Garau (2014) examined the effects of both at-home FI and a 3-month SA period on oral accuracy development over 2.5 years. The data was collected from undergraduate advanced-level L3 learners of English taking an oral role-play task at four data collection times. The results showed that the majority of the learners benefited from the SA sojourn, while for those who didn't it was assumed that the 3-month period was insufficient to produce desired outcomes. Concerning the retention effects of SA, the data revealed that 15 months after the SA, the accuracy levels maintained their stability, showing that the SA impact was still observable more than a year later.

More recently, McManus et al. (2021) investigated advanced L2 French and L2 Spanish learners' CAF speech development over 21 months (including a 9-month SA). The oral performance on an oral picture-based narrative once before (at the end of year two of a 4-year degree), three times during (third year abroad), and twice after a nine-month stay abroad (within 9 months in FI context) were measured. Results in general showed study abroad benefited fluency (measured by speech rate and

mean length of run), lexis (measured by D scores), accuracy (measured by percentage of error-free ASUs and percentage of error-free clauses), and syntactic complexity (measured by Clauses/ASU and mean length of ASUs). Those gains made during SA for fluency, lexis and accuracy were maintained in general on return to the formal instructed context. Though small declines in fluency were noted following the return home. Lexical diversity during SA decreased marginally before increasing again after the return to the taught context. The results suggested that SA enables advanced learners, in particular, to proceduralize/automatize their existing linguistic knowledge through meaningful practice opportunities, as well as to generate new morphosyntactic knowledge and/or restructure existing morphosyntactic knowledge.

To contribute to SA research in understanding of the long-term evolution of foreign language proficiency, Huensch et al. (2019) looked at the retention/development of oral fluency and proficiency, and discovered that language contact/use and peak proficiency obtained were both critical factors in fluency and proficiency retention three years after formal instruction ceased (four years after study abroad). Speech samples were elicited in LANGSNAP (Mitchell et al., 2017) using picture-based narratives created to be as identical to each other as possible and administered in sequence approximately one year apart. Five fluency variables representing aspects of speed (measured by speech rate), breakdown (measured by the number of filled pauses per 100 words, and the number of silent pauses per 100 words), and repair fluency (measured by the number of repetitions per 100 words, and the number of corrections per 100 words).The results showed that speed and breakdown fluency indicators (e.g., speech rate, silent pauses) that considerably improved after study abroad were maintained four years later. In contrast, repair fluency indicators (e.g., repetitions, corrections) did not change significantly during study abroad or four years later. The study concluded that both language contact/use and proficiency attained are important variables in the long-term maintenance of overall proficiency. Tracy-Ventura et al. (2021) expanded on Huensch et al. (2019) by using a corpus-based approach to investigate the long-term evolution of lexical diversity post-SA and post-formal language instruction in two groups of participants, one L2 French and one L2 Spanish. It investigated to what extent lexical diversity (measured by D and Moving Average Type-Token Ratio (MATTR) changed four years after study abroad. The results showed that the group as a whole exhibited continued improvement at post-SA (three years after formal teaching ceased). The authors concluded that the increases in oral lexical

diversity over time is very likely due to increased automaticity in lexical access that is improved in online speech production with continued practice/use.

As indicated, SA research on the oral retention of SA effects after L2 language learners return to formal instruction at home context within a same cohort has grown in recent years. Some studies (e.g., Juan-Garau & Pérez-Vidal, 2006, 2007; Pérez-Vidal & Juan-Garau, 2005) revealed that learners' oral gains during SA period decrease after SA over the following year with no formal instruction. However, other studies (e.g., Howard 2009; Huensch et al., 2019; McManus et al., 2021; Tracy-Ventura et al, 2021) have in general found that linguistic gains made abroad are maintained over the following year to three years, during which participants continued to receive formal instruction/language contact. The difference among those studies concerning the FI maintenance is very likely attributed to internal factors (e.g., proficiency levels), and external factors (e.g., length of SA period (3 month vs.1-year SA), and availability of formal instruction /language contact (with/ without)) in FL at home context after SA. Among those factors, Tracy-Ventura et al. (2021) suggested that higher-level proficiency is a strong predictor of language retention after investigating L2 French and L2 Spanish oral gains using longitudinal data, which were collected before, during, and after their year abroad as part of the Languages and Social Networks Abroad Project - LANGSNAP (e.g., Mitchell et al., 2017; Huensch et al., 2019; McManus et al., 2021; Tracy-Ventura et al, 2021). In general, the limited SA research to date suggests that in terms of predicting the amount of FI maintenance, proficiency level is more important than the amount of exposure (input and time) (Huensch et al., 2019).

2.4.2 Other factors

To understand the effects of SA on L2 fluency more precisely, other variables in the methodologies used across previous studies should be taken into consideration (e.g., Wright, 2020). In terms of the key factors which influence SA benefits and outcomes, it is agreed that the SA research has produced mixed findings (Tullock & Ortega, 2017; Yang, 2016). This might result from the interaction between internal factors related to learners and external factors associated with the context. For instance, after reviewing 401 SA studies, meta-analysis by Tullock and Ortega (2017) concluded that two main explanations are considered to account for the lack of consistency that SA research emerged. First, what linguistic gains can be demonstrated are determined by what language characteristics chosen as a focus and what outcome measures are used. For instance, L2 language learners' participation

in meaning-based communication is likely to improve oral fluency but not necessarily morphosyntactic accuracy, especially when errors do not hinder understanding (Collentine, 2009). Second, SA benefits depend on certain factors, but which tend to vary across studies. Specifically, there are four major variables that are discussed across studies: the pre-departure proficiency of SA students (e.g., Llanes, 2011; Valls-Ferrer & Mora, 2014), individual differences (e.g., DeKeyser, 2014; Sanz, 2014), the age of participants (e.g., Llanes & Serrano, 2017), and the duration of the SA programme investigated (e.g., Llanes & Serrano, 2011; Serrano et al., 2016).

It has been claimed that for learners undertaking SA immersion, their initial proficiency levels should be taken into consideration, as this might influence the degree and type of activities learners engage in the SA context (e.g., Segalowitz & Freed, 2004; Colletine, 2009; Dewey, 2014). For instance, in their longitudinal study, Valls-Ferrer and Mora (2014) examined the impact of onset proficiency level and language contact profiles on the fluency development in Formal Instruction and three-month Study Abroad contexts. The data were collected through semi-guided interviews and an SA condition questionnaire from four tests at different time points over 30 months, which elicited speech output through seven fixed questions designed to be similar to real-life conversations. The results showed that lower initial fluency levels and greater language contact were related to fluency gains during SA. Similarly, Juan-Garau (2014) investigated a 3-month SA sojourn on the oral accuracy of 43 advanced-level L3 English learners. The results showed that students with lower pre-departure levels benefitted the most from SA. The study pointed out that one possible explanation for the fact not all the participants became more accurate related to the length of stay. Specifically, 3-months is likely to be insufficient to bring about the desired outcomes in all learners. Other researchers have also argued that one semester may not be enough for potential gains to be realised (e.g., Segalowitz & Freed, 2004).

Concerning the effects of onset proficiency level on learners' oral gains during study abroad, however, there are mixed findings. Leonard & Shea (2017) examined the speaking development of 39 advanced English learners acquiring Spanish in a pre and post-SA format (3-month). The results showed that participants experienced significant gains across complexity, fluency, and accuracy. These gains, however, were not spread uniformly across all dimensions or across all learners. Prior to study abroad, learners with higher levels of L2 linguistic expertise and faster L2 processing speed reported greater gains in accuracy and syntactic and lexical complexity. Some previous studies looked into the

question on what is the appropriate level for students planning to participate in a study abroad program. For example, Brecht, Davidson, and Ginsberg (1993) stated that Students with a high level of language proficiency were more likely to use the target language than students with a lower level of proficiency. Students' language proficiency must be sufficient to enable them to interact with other students (Bacon, 2002). According to Kubler (1997), the optimal time for most students to study in China is once they have attained the intermediate level, which is after they have completed two years of college Chinese or its equivalent. Good students have a firm command of basic vocabulary and grammar and are able to communicate in simple Chinese with the majority of Chinese speakers they encounter, allowing them to make the most of their study abroad experience. More recently, Li (2014) contributes to the effects of linguistic proficiency on the oral development of L2 Chinese during SA. Two groups (intermediate group and advanced group) of American learners of Chinese completed a Computerized Oral Discourse Completion Test at the beginning and toward the end of their 15-week SA sojourn. The results showed that no group reduced planning time, and only the advanced group improved their speech pace. In terms of the best time to go abroad, this study's findings imply that having around four semesters of formal instruction before going abroad is a better option for developing the ability to make requests in L2 Chinese.

To explore the effect of age on learners' oral development, Llanes and Serrano (2017) examined 197 L2 learners of English from three age-related groups: primary school, secondary school, and university in both FI and SA contexts. The instruments used were a written composition and an oral picture-cued narrative task with three tests. The learners' oral and written development was measured by fluency, lexical and syntactic complexity, and accuracy. The results revealed that when the learning context was excluded, older students surpassed younger students. However, when both learning context and age were taken into account, the results revealed that younger SA participants tended to do better than older SA participants regarding oral skills. Additionally, the results revealed that the SA context outweighed the FI context, in particular concerning oral skills across the majority of measures except syntactic complexity. This is similar to their previous research (Serrano, Llanes & Tragant, 2011), which reported that in contrast to the FI context, the SA group showed an increase in oral fluency and lexical complexity, but not in accuracy and syntactic complexity.

Some scholars have explored the effect of the duration of SA on L2 learners' oral proficiency as the length of SA sojourn varies across studies (e.g., Lara et al., 2015; Llanes & Serrano, 2011; Rees & Klapper, 2007; Serrano et al., 2016). There is no general agreement as to whether and to what extent SA residence can be considered as short- or long-term. Moreover, contradictory findings have been revealed concerning the duration of SA and its effect on L2 learners' linguistic gains. For instance, Llanes and Serrano (2011) investigated the effect of length of SA on the oral development of two groups of Spanish-speaking students learning English in the UK. Their data were collected from 46 participants by means of an oral narrative after a stay of two months (25 students) or three months (21 students). The learners' oral samples were analysed in terms of fluency, accuracy, lexical richness and complexity. The results suggested that there were no significant differences in the gains made by the two groups of learners. It concluded that an additional month abroad may not be long enough to produce significant differences in learners' second language development. Likewise, Lara et al. (2015) examined the oral performance of L2 learners of English during both a 3- and 6-month SA, the results surprisingly showed that not all of the participants' CAF measures who undertook a 6-month SA outperformed those who stayed for 3 months, e.g., oral accuracy. In terms of longer duration SA, there have been similar findings. For instance, Rees and Klapper (2007) undertook a longitudinal study designed to assess the progress made by UK foreign language undergraduate students during SA. Those who undertook a 12-month SA only showed a slightly greater gain (which failed to reach statistical significance) than those who undertook a 6-month SA sojourn. This revealed that longer stays do not necessarily lead to proportionately greater proficiency gains. Moreover, after reviewing 11 quantitative studies which compared the L2 linguistic gains of SA and FI learners, Yang (2016) concluded that short-term SA (from 11 weeks up to 13 weeks), is more effective than long-term SA (more than 14 weeks to up to 3.5 years) in terms of L2 linguistic development.

The contradictory findings outlined above have triggered researchers to examine the programme itself. For instance, Dewey et al. (2014) concluded that the pre-programme proficiency factor can be overcome by the impact of the programme itself, namely, the programme design. Indeed, a consensus on programme and programme intervention has been reached in recent studies. This is because it is at the core of whether the SA provides learners with an L2-immersive environment during the SA period (Yang, 2016). For instance, there are some SA studies which claim that setting clear

learning outcomes (Berg, 2009) as well as other extracurricular activities and opportunities are necessary to encourage learners to engage in the immersion community to maximise their achievements (Kinginger, 2011; Isabelli-García et al., 2018). Specifically, it is thought that when students study abroad, official language-partner platforms, social networks, and community learning can be initiated at the college level to promote their L2 Chinese learning and acquisition outside of the classroom setting. This has been agreed upon in some recent studies. Social networks in particular play an essential role in facilitating L2 learners' proficiency gains in the SA context. The types of relationships developed through communities, such as friendship, kinship, and participation in the workplace, facilitate proficiency gains (Paige et al., 2012; Dewey et al., 2012; Dewey et al., 2013; Du, 2013) and learners who lack community engagement during a SA sojourn can only be expected to make limited gains (Coleman & Chafer, 2010; Dewey, 2008; Mora & Valls-Ferrer, 2012).

Moreover, some studies have explored the impact of testing methods (i.e., task type) on oral development during SA (e.g., Wright, 2018; Wright, 2020). These studies have suggested that task loads should be taken into account because task load effects might override SA impact, in particular, the rehearsed/planned monologue task (Wright, 2018, 2020). For instance, Wright (2020) evaluated task types on L2 Chinese fluency development during a 10-month SA sojourn. Specifically, Wright compared the oral performance in 4 tasks with different task loads (rehearsed vs. spontaneous speech, in monologic and dialogic mode) between pre- and post-SA. The results revealed that significant differences between the rehearsed monologue and other tasks found pre-SA were generally not found after SA. This suggested that task load effects outweigh SA fluency. Moreover, performance in rehearsed speech tasks, in particular, the monologue, was better than in spontaneous tasks across most measures both pre- and post-SA. This can be explained by preparation time. Namely, pre-planning strongly elicits complexity and fluency but raises accuracy to a lesser extent (Skehan, 2009a; Tavakoli & Skehan, 2005; Mora & Valls-Ferrer, 2012).

In conclusion, the majority of SA research has focused on European languages. In particular, English and Spanish have become the two major target languages in this area (Yang, 2016). However, the effect of SA on L2 Mandarin has been under-explored and only a few empirical studies have been conducted in this area (e.g., Du, 2013; Kim et al. 2015; Wright & Zhang, 2014; Wright, 2018, 2020).

2.4.3 Previous work on L2 Chinese oral development

Existing L2 Chinese studies mostly focus on learners' written and oral development over a short-term period and adopt cross-sectional designs (e.g., Chen, 2012, 2015, 2020; Chen & Zhou, 2016; Huang, 1997; Li, 2010; Li et al., 2003; Liu, 2017; Guo, 2007; Wang, 2002; Ye, 2015; Zhai, 2011; Zhai & Feng, 2014; Zhang, 2000; Zhang & Wu, 2001; Zhou & Zhang, 2006). Only a few studies have explored L2 Chinese learners' oral development with a longitudinal design. Moreover, the majority of existing studies are written in Chinese and examine the oral and written development of L2 Chinese learners who are based in mainland China. Furthermore, hardly any studies which assess L2 Chinese learners' oral development have been written in English or conducted outside of China (Wu, 2008; Du, 2013; Jin & Mak, 2013; Jiang, 2013; Li, 2014; Liu, 2009; Wright & Zhang, 2014; Wright, 2018, 2020). Most of the studies conducted both inside and outside China have used participants enrolled on four-year Chinese undergraduate degree programmes.

The existing L2 Chinese studies have addressed speech performance from different perspectives, such as context, learners and testing methods. These studies have examine the impact of the following factors: 1) context: learning environment (Du, 2013); 2) learners: proficiency level (Chen, 2012, 2015a, 2015b; Ye, 2015); identity (Chen, 2020); and 3) testing methods: different task types, such as self-introduction, oral interview, topic discussion, picture-cued description, read aloud, and sentence repetition (Li et al., 2003; Wang, 2011; Wang, 2018); the effect of text types, such as orientational, narrative, descriptive, and argumentative (Liu, 2017a); temporal feature (short/long-term) (Ding & Xiao, 2016, Zhai & Feng, 2014, Shi, 2002, Zhou, 2016; Wu, 2017; Feng, 2018); and the relationship between objective and subjective analysis (Jin & Mak, 2013; Wu, 2017).

Apart from these different angles, there is one factor that has remained highly consistent across all the studies. Specifically, because nearly all L2 Chinese studies have been conducted when the learners are in the target language country, i.e., China, in this sense, all the learners have been in a SA context. Moreover, the majority of studies have adopted a cross-sectional design, whereas for those with a longitudinal approach, they have either assessed learners' development over a short period (Ding & Xiao, 2016; Zhai & Feng, 2014; Ye, 2015) or they have used case studies over a longer period

(Shi, 2002; Zhou, 2016; Wu, 2017; Feng, 2018). However, studies which investigate factors (i.e., formal instruction at-home context, the same cohort group, evolving proficiency levels) over a long-term period on English undergraduates of Chinese are very scarce (e.g., Wright, 2020).

Furthermore, most of the L2 studies which apply the CAF framework to assess the effects of study abroad on L2 learners' oral development focus on L2 English (Tullock & Ortega, 2017). While there is an increasing trend to measure the effect of study abroad on L2 spoken Chinese (e.g., Du, 2013; Kim et al, 2015; Wright, 2020; Wright & Zhang, 2014), the majority of research which has explored L2 Chinese learners' speaking development have used a cross-sectional design (Chen, 2012; Zhai & Feng, 2014; Wu, 2014; Ye, 2015; Ding & Xiao, 2016; Chen & Zhou, 2016; Liu, 2017). In terms of the proficiency levels of participants in the previous studies, they have either been defined by their institutional status (e.g., Zhai & Feng, 2014; Ye, 2015; Chen, 2015; Ding & Xiao, 2016; Zhou, 2016), by their period of instruction (Chen, 2015; Liu, 2017; Wu, 2017) or by instructed hours (Zhang, 2018). This is similar to other studies in the broader L2 field (Wu & Ortega, 2013).

Moreover, there is an increasing trend in using longitudinal design to measure L2 Chinese learners' oral development (Shi, 2002; Zhou, 2016; Wu, 2017; Wright, 2018, 2020). Case studies are often applied due to the time-consuming nature of that type of research (Shi, 2002; Zhou, 2016). It is necessary to examine the oral development of English-speaking learners of Chinese by employing CAF measures and a longitudinal design. This is particularly the case concerning the effects of study abroad on L2 Chinese Speakers of English as this area has only been investigated by a handful of empirical studies (e.g., Du, 2013; Kim et al., 2015; Wright, 2018, 2020; Wright & Zhang, 2014). Furthermore, to my best knowledge, there is no study investigating the effects of FI maintenance on oral development of L2 learners of Chinese after study abroad. Therefore, this study seeks to fill the gap by examining how learning contexts (study abroad and Formal instruction at home) affect the speaking development of L2 learners of Chinese.

Chapter 3: The study

The first part of this chapter will present the study's research objectives and questions. The second part focuses on the methodology. This latter part is divided into three sections which describe the participants, the procedures, and the Complexity, Accuracy and Fluency (CAF) measures adopted.

3.1 Research objectives

In the field of L2 language learning, most studies examine complexity, accuracy, and fluency as a general construct of L2 proficiency (Tavakoli, 2016). Regarding L2 Chinese language development, there has been increasing interest in applying CAF to assess L2 learners' performance (Chen, 2012; Zhai & Feng, 2014; Wu, 2014; Ye, 2015; Ding & Xiao, 2016; Chen & Zhou, 2016; Liu, 2017). Similar to other second language studies, the majority of CAF-related studies adopt cross-sectional designs rather than longitudinal designs (cf. Vercellotti, 2015). Furthermore, a lot of the studies with cross-sectional designs have used different proficiency groups to explore learners' development (Huang, 1997; Zhang, 2000; Zhang & Wu, 2001; Wang, 2002; Li et al., 2003; Zhou & Zhang, 2006; Guo, 2007; Li, 2010; Zhai, 2011; Chen, 2012; Xiao, 2013). In terms of the studies that have adopted a longitudinal design, most have only examined the oral development of learners of Chinese over a relatively short period, such as less than one academic semester or two to three months (Ding & Xiao, 2016; Zhai & Feng, 2014; Ye, 2015) and the longitudinal studies which have had a longer duration (i.e., one year or over) have nearly all been single-participant studies (e.g., Shi, 2002; Zhou, 2016; Wu, 2017; Feng, 2018). Moreover, there have only been a very few studies that explore the correlation among the CAF framework (e.g., Ye, 2015; Chen, 2015a; Wu, 2017). Instead, most studies have either investigated oral development as measured by one domain or subdomain of CAF (e.g., Chen, 2012; Zhai & Feng, 2014; Ding & Xiao, 2016; Feng, 2018; Liu, 2016; Wang, 2018; Wang & Jin, 2020), or speech development as measured by two domains of CAF (e.g., Chen, 2015b; Zhou, 2016).

Having reviewed the main studies on the oral development of L2 Chinese learners, we can conclude that there is virtually no existing research that tracks the longitudinal speech development of English-speaking learners of Chinese with evolving proficiency levels in terms of CAF constructs. Similarly, there are no studies that track how oral development is impacted by learning context (Formal

Instruction and Study Abroad), and also the relationship among the CAF constructs and sub-constructs across the full developmental trajectory (cf. Spoelman & Verspoor, 2010). As a result, it can be concluded that there is a dearth of research which examines what happens as proficiency grows in relation to the performance areas of complexity, accuracy, lexis, and fluency (Skehan, 2009). As noted above, the majority of L2 Chinese studies have adopted cross-sectional designs. The consequence is it is challenging to understand how group-level differences at varying proficiency levels compare to individual learner trajectories. Therefore, there is a need to use a longitudinal approach to focus on the developmental patterns of individual learners, which will complement the cross-sectional studies (Spoelman & Verspoor, 2010; Mora & Valls-Ferrer, 2012; Kuiken et al., 2019). Although there has been increasing interest in studying Mandarin (e.g., Han, 2014; Lu, 2017; Tao, 2016), until now few empirical studies have investigated effects of SA on L2 Mandarin speech development (e.g., Wright, 2018, 2020).

In summary, there is a need for a study with a semi-longitudinal design to explore the oral development of L2 learners of Chinese. To meet this gap, two crucial issues should be considered. Firstly, it is necessary to explore and capture the speech development of the same cohort of L2 Chinese undergraduates affected by two factors: learning context (i.e., FI and SA), and timescale (short-term and long-term within the examining period). Secondly, it is also necessary to explore when, how, and why different aspects of speech performance increase or decrease within the developmental trajectory. Given these considerations, this study focused on the specific development of students' oral competence with a semi-longitudinal design. The students who participated in the study were all enrolled on a four-year Chinese programme at the tertiary level of education in the context of Ireland. After presenting the essential features of the CAF framework (Skehan, 1989, 2009; De Graaff & Housen, 2009; Robinson, 2001b, 2005; Robinson & Gilabert, 2007; Housen et al., 2012) and its differing implementation in the field of Chinese L2 acquisition (e.g., Li, 2003; Zhai & Feng, 2014; Liao, 2018), this research analysed the data collected from curricular oral tests across 28-months to assess the oral development of L2 Chinese undergraduates in Ireland.

In the context of Ireland, teaching and learning Chinese as a second/foreign language in Ireland has received much attention. Chinese language teaching first appeared in the Irish educational system in 2006-2007, when two Confucius Institutes (CI), affiliated with the Chinese Ministry of Education,

were established in two respective universities, University College Cork (UCC) and University College Dublin (UCD) (Osborne et al., 2019). Both the Institute of Chinese Studies at UCD and the School of Asian Studies at UCC provided degree programs related to the Chinese language. Additionally, the main functions of CI is to provide teaching resources tailored to their respective institutions. Therefore, the Chinese language courses at UCC and UCD were mostly taught by the Chinese language instructors that CI offered.

In terms of assessing Chinese as an second/foreign language, one approach to defining successful L2 performance is to characterise it as complex, accurate and fluent. Thus, it is essential to comprehensively explore learners' performance and examine the relationship between different aspects of their performance (Tavakoli, 2016). A key goal of this study was to depict L2 Chinese learners' oral development as measured by CAF. In particular, the effect of learning context (FI and SA) on L2 Chinese oral performance was investigated. The other goal was to examine the correlations between subconstructs both within and between CAF. To explore the interdependency and multidimensionality of CAF constructs, it is necessary to explore more dimensions of all three related constructs of CAF within the same learner cohort (Jensen & Howard, 2014). It is also essential to measure complexity, lexis, accuracy, and fluency to capture the different facets of performance with a longitudinal study that can obtain a full view of developmental trajectories (e.g., Norris & Ortega, 2009; Spoeman & Verspoor, 2012; Vercellotti, 2012).

3.2 Participants

The data analysed in this study were derived from a corpus of oral production collected by the Department of Asian Studies, University College Cork. Specifically, the data were collected from 10 native English-speaking learners of Chinese who were undertaking an undergraduate degree in Commerce with Chinese Studies at University College Cork. No participants had contextual or long-term exposure to Mandarin before they studied at UCC. All of the students spent a full academic year abroad during year three, comprising approximately 10 months (from September to the following July), and were enrolled full-time at a university in Shanghai. To ensure confidentiality, this study will not reveal any information that could disclose their identities, such as the precise dates when data was collected.

The ten participants in this study were aged from 18 to 22 (See Table 8). There were more male participants (n=6) than females (n=4), which does not reflect a general trend in SA participation. The participants were examined with a semi-longitudinal design that included three data collection times. The data were collected in a Formal Instruction at home (FI) context and a 10-month Study Abroad context. The FI period was both before and after the SA period.

In terms of the participants' pre-SA proficiency level, they were all enrolled in the same university degree and received approximately 360 hours of instruction in the FI at home context. Furthermore, all the participants sat the HSK3 four months before the SA period to enable them to attain a one-year scholarship for the SA. However, three of them did not pass this exam (See Appendix D). Their HSK levels were considered as their onset proficiency levels and were either HSK2 (n=3) or HSK3 (n=7). When the participants reached their final attainment level one year after returning to the FI context, they had all either reached HSK3 or HSK 4 considering the instructional level that they were allocated based on their performance. These HSK levels are equivalent to B1 to B2 of CEFR, although the equivalency between these two standards remains controversial (Peng et al., 2020).

Table 8. The profile of participants

Participants No.	Age	Gender	Proficiency level (Pre-SA)
1	18	M	HSK3
2	18	F	HSK3
3	18	F	HSK3
4	22	M	HSK2
5	18	M	HSK3
6	18	F	HSK2
7	18	F	HSK3
8	19	M	HSK2
9	18	M	HSK3
10	18	M	HSK3

Note. The study will take into account the difference in HSK levels between subjects in the data analysis portion. As a result, data at the group level will be provided, as well as individual performance at each level (HSK2 and HSK3).

3.3 Methodology

This section describes the methodology used to investigate the effects of study abroad on the speaking development of 10 adult English-speaking learners of Chinese in Ireland. It details the instruments, the procedure for collecting data, and the study abroad period.

Table 9. Three stages of data collection

Session	Session 1		Session 2	Session 3
Date of oral tests	3-Dec		25-Oct	6-Apr
Year of students	Year-2	Year 3	Year-4	Year-4
Semester	1		1	2
Hours of Instruction	360 hours	540 hours	36 hours	108 hours
context	FI	SA	FI	FI

An overview of the three stages of data collection is presented in Table 9. Between session 2 (S2) and session 3 (S3), another two oral tests as continuous assessments, which were part of the undergraduate programme, were conducted. The four oral tests (session 2, session 3, and two continuous assessments between session 2 and session 3) that took place in the fourth year of the programme, at intervals of approximately 1 to 1.5 months, covered 144 hours of formal instruction for the whole academic year. Considering the research question on the effect of learning context on oral development, the two continuous assessments between session 2 and session 3 were not investigated. Instead, session 2, as the first post-SA session, and session 3, as the final attainment level during the FI at home context, were selected and analysed. Defined by their instructional status, and by feedback from the experienced instructors who lectured the participants during the programme, the learners at the pre-SA (S1) were classified as being at late beginner level. At post-SA (S2), they were at low intermediate level and at S3 they were high-intermediate level.

3.3.1 Instruments

The data collection points were named Session 1, Session 2, and Session 3, respectively (See Table 9). All three oral sessions were part of the assessment in the curriculum and took place over 28 months in the context of formal instruction in Ireland. It is also important to note that the students studied in China for a whole academic year during the third year of their undergraduate programme. During all three sessions during the FI at home period, there was not a particular focus on oral communication skills. Instead, the majority of the instruction consisted of traditional grammar teaching and practice, and learners were barely exposed to Chinese outside of the classroom. These three sessions for data collection were as follows:

Session 1: Students were tested at the end of the first semester of the second year of their course, prior to SA, thus enabling the study to investigate the gains in acquisition as a result of the 360 hours of formal instruction that they had completed by this point in their degree programme.

Session 2: Students were tested three months after they returned to the FI at home context after their 10-month SA. During the SA period, the host college offered 540 hours of Chinese classes to the participants over two terms covering 34 weeks. Therefore, the first test that the participants took after they returned home, namely, at session 2. This was when 5-week formal instruction (36 hours) was completed after SA in the instructional classroom setting in the year four.

Session 3: When learners had been back in the FI context for 6 months, S3 as the final attainment level within the year four was investigated to examine the learners' overall language learning progress. Learners received 144 hours of formal instruction in the year four.

S1 and S2 were considered as pre-SA post-SA test respectively. S2 and S3 were used to measure the effects of at-Home FI maintenance, and, to explore a possible decrease in the target language learning curve post-SA within an academic year of two semesters (with 6-month formal instruction).

In terms of textbooks, the students used the *New Practical Chinese Reader Textbook* (2nd Edition) 1, 2 and 4 in the first, second, and fourth years respectively, when the participants were in the FI context. In the first year of classroom-based Chinese language learning, all the students were required to undertake 168 hours of instructed learning, while in the second year and fourth year, 192 and 144 hours of classroom work were required. During the Formal Instruction period, there was no specific focus on training students' oral communication skills. A speaking class was not provided to the participants. These students only attended a 7-8 hour Chinese language lesson in which they were taught listening, speaking, reading, and writing skills. Also, they experienced limited exposure to Chinese outside the classroom.

3.3.2 Procedure and Ethics

Data analysed in this study were speech samples elicited by questions, which were part of the continuous oral assessments within the curriculum of the UCC Chinese and Commerce programme. A day before the students took the tests, they were able to access a larger number of fixed questions for preparation purposes. The questions were relative to the learning content when the learners were in the formal instruction context. However, none of the examination questions were revealed until they took the tests. During the oral tests, each participant was tested individually by their instructor. They were

all asked to produce free speech. The instructor mostly kept silent apart from when learners only managed limited production. In this scenario, the instructor asked students related questions to elicit their oral production. The participants in the FI context in this study were taught by the same Chinese language teacher who also served as the instructor during the three oral tests. The instructor in the FI context is the researcher, whom did not inform participants in advance before the three tests took place. This could enable data was produced in a naturally occurring manner without occuring participation bias. But the informed consent were sought after the three tests had been done.

3.3.3 The study abroad period

To fully understand the oral performance of the 10 participants during the 10-month SA, their HSK scores pre-SA (See Appendix D), their allocated class level, and their test scores from the Chinese language course are provided.

Concerning the proficiency level of the ten participants, they were allocated to two different language proficiency levels based on their performance on placement tests which they took after they arrived at the host college in Shanghai. During the whole academic year in the SA context, eight of the ten participants were allocated to one group with other language learners at Level A in term 1. In term 2 they had reached Level B. The other two participants were allocated to the Level B class in term 1, and they reached Level C in term 2.

The Chinese language courses offered to the participants at the different levels were as follows:

- Level A: Chinese Reading and Writing, Chinese Conversation, and Chinese Listening.
- Level B: Chinese Reading and Writing, Chinese Listening, Extensive Reading (B).
- Level C: Chinese Reading and Writing, Chinese Listening, Chinese Writing, Extensive Reading (C).

In terms of the instruction hours within the single academic year provided by the host college, there were 34 weeks of instruction. Each week there were approximately16 hours of language classes composed of 6.5 hours of reading and writing, 6.5 hours of listening, and 3.5 hours of conversation classes. Therefore, in total, the students received 128 hours of Chinese Reading and Writing, 128 hours of Chinese Listening and Speaking, and 64 hours of Extensive Reading during the study abroad

period. Other modules concerning commerce and Chinese society were taught in English. Furthermore, participants participated in organized social events such as 5-7 day cultural tours and 1-day foreign cultural exhibitions during the SA period, depending to the host university's curriculum and activities. The host university, on the other hand, did not provide any internships.

3.4 The CAF measures used in this study

After reviewing the main CAF measures used in L2 oral Chinese studies, fourteen measures were chosen to analyse the oral development of L2 learners of Chinese in this study.

Accuracy measures

In the field of L2 oral Chinese studies, accuracy can be analysed at three levels: phonetic (Chen, 2015a, 2015b; Kim et al., 2015; Zhai & Feng, 2014), lexical (Chen, 2015a; Ding & Xiao, 2016; Ye, 2015; Zhai & Feng, 2014), and syntactic (Chen, 2015a; Ye, 2015; Wu, 2017; Zhai & Feng, 2014).

As expected, the participants in this study produced nonnative-like pronunciations. The basic analysis unit for Chinese pronunciation is considered to be a Chinese syllable consisting of an initial, a final, and a tone (e.g., Liao, 2018; Jin & Mak, 2013; Wang, 2002). However, it is challenging to define tone error in a quantitative manner. This reflects an existing controversy concerning the criteria for identifying errors and evaluating accuracy, where Housen et al. (2012) argued that appropriateness and acceptability should be taken into account when accuracy is considered. The produced phonological variables should then be considered to be treated either as errors or alterations of the target language norm (Isabelli-García et al., 2018). Furthermore, normally phonetic accuracy is based on native speakers' judgment rather than software. Therefore, in this study, phonetic accuracy was not analysed.

Regarding syntactic accuracy, this relates to syntactic errors that need to be classified at different degrees. One proposal that has been made is to grade errors based on their level or the extent to which they compromise communication at a developmental level (cf. Pallotti, 2009). For instance, Kuiken and Vedder (2007) measured written L2 accuracy with a total number of errors per T-unit and categorised three degrees of errors. Under their categorisation first-degree errors are minor deviations in spelling, meaning, or grammatical form that do not interfere with the comprehensibility of the text,

second-degree errors are more serious, and third-degree errors make the text nearly incomprehensible. However, classifying syntactic errors of oral performance to different degrees requires dealing with the dysfluency features of oral performance. Moreover, there are no up to date benchmarks available to cope with the challenge of categorising the syntactic structures of Chinese language into different degrees except for the Chinese Proficiency Level Standard and Grammar Level Outline (Liu, 1996). Therefore, syntactic errors are challenging to classify operationally and consequently were not used in this study.

In terms of lexical errors, these are related to lexical expressions and word choice (Bardovi-Harlig & Bofman, 1989). Considering participants' proficiency levels and the duration of the examining period in the study, accuracy was not expected to increase. This prediction was made because accuracy is only expected to improve when English learners of Chinese are at the advanced stage (cf. Chen, 2015; Ye, 2015). Moreover, as this study was constrained by limitations to time and knowledge, it only analysed lexical accuracy. This was obtained by the ratio of error-free lexical items, calculated by one minus the ratio of the lexical errors, which in turn was calculated by the number of lexical errors divided by the total number of lexical items.

Fluency measures

To measure the three sub-categories of utterance fluency: speed fluency, and breakdown and repair fluency, the most commonly used indicators in each subcategory were employed. To assess speed fluency, Speech Rate (SR) and Mean Length of Runs (MLR) were used. SR was calculated by the total number of syllables (excluding filled pauses) divided by the time of utterances including pause time in seconds, multiplied by 60. This gave the produced syllables in one minute. MLR was calculated by the number of syllables divided by the number of silent pauses.

To measure breakdown fluency, the duration, and ratio of pauses including silent and filled pauses are typically examined separately (e.g., Chen, 2012; Zhai & Feng, 2014; Chen, 2015a; 2015b; Feng, 2018). Only a few studies have assessed pauses as a whole (Zhai & Feng, 2014; Ye, 2015; Ding & Xiao, 2016; Wu, 2017; Wang, 2018). Breakdown fluency is associated with the length and frequency of pauses (filled or unfilled), which is significant to assess fluency (cf. Wood, 2001).

In this study, the silent pauses and filled pauses were coded and calculated separately. There were five reasons for doing so:

1) Based on the mixed and indecisive criteria for the cut-off points for silent pauses and filled pauses, it is necessary to evaluate what are suitable criteria for the cut-off points for both filled and silent pauses in relation to learners' proficiency levels. Also, it is important to explore whether the cut-off points for filled pauses and silent pauses should be identical or if they should be analysed with different criteria considering learners' proficiency levels.
2) Silent and filled pauses have been differentiated in previous studies. Filled pauses are regarded as being highly idiomatic, and are associated with L1 use. Also, in L2 processing, fillers can be used as a communicative strategy to compensate for resource deficits (Préfontaine & Kormos, 2016). The number and duration of silent pauses are mathematically associated with speech rates, because the more or longer a speaker is likely to pause, the slower the speech rate will be (De Jong, 2016b).
3) Compared to other hesitation features, such as filled pauses and repairs, it has been suggested that unfilled pauses might be salient in determining the fluency level of non-native speakers (Riggenbach, 1991). Furthermore, a silent pause with a duration of around 250-300 milliseconds has been found to yield the highest correlation between silent pauses and L2 proficiency (De Jong & Bosker, 2013). To further explore the findings, it was necessary to analyse the filled and silent pauses individually to avoid diluting the statistical significance.
4) It has been pointed out that silent and filled pauses can be affected by speakers' individual speaking styles. For instance, some speakers may be more likely to pause more than others in their L1, not because of any problems in creating L2 speech (Wright, 2020). Silent pauses can be considered as an indicator of time used for the speech planning process, not utterance planning; while filled pauses can be used as a successful strategy for holding one's turn, which may not always indicate a lack of utterance fluidity (de Jong, 2016; Tavakoli, 2011).
5) Operationally, silent pauses can be automatically coded in PRAAT or CLAN once the standard has been set (Li, 2015; De Jong, 2016a), while in this study, filled pauses had

to be marked manually by the researcher herself using the audio editing software Audacity 2.3.2.

In conclusion, in fluency research, pausing, including silent and filled pauses, can be very complex (Wright, 2020). Nevertheless, pauses are included here for comparability with other studies. Therefore, two aspects were analysed in this study: the average length of pauses and the frequency of pauses. Specifically, four indicators were analysed: the average length of filled pause (ALFP), the average length of silent pause (ALSP), the number of filled pauses per 100 syllables (FP100), and the number of silent pauses per 100 syllables (SP100).

In this study, repetitions were considered as one sub-component of self-repairs. Following Kormos (2006), repetitions, false starts, and self-corrections have been merged into the one category of dysfluency, which was assessed as a whole. The number of repetitions and repairs per 100 syllables (RR100) were analysed.

Complexity measures

This study is concerned with linguistic complexity, which can be quantitatively analysed across three subdimensions: lexical, syntactic, and morphological complexity. Of these three, lexical and syntactic complexity have received the most attention in the existing L2 research. However, morphological complexity, in contrast, has rarely been measured (Bulté & Housen, 2014), especially in L2 Chinese studies. This situation has likely come about because there are no available measures that can be used to analyse the inflectional and derivational complexity of the Chinese language. This because Chinese language morphology lacks transparency (Du, 2013). Considering the focus of the present study, i.e., to analyse the oral development of English-speaking learners of Chinese, using these two widely analysed subdimensions: lexical and syntactic complexity measures, were adequate to achieve the study's goal. Therefore, these two subdimensions of complexity were analysed.

Lexical complexity

Lexical complexity can be analysed across four aspects at the operational level: lexical diversity, lexical sophistication, lexical density and lexical compositionality. Of these four, lexical diversity has been the most commonly measured dimension in the literature. Lexical density and sophistication have also been analysed in L2 research, however, lexical compositionality has been rarely assessed (cf. Bulté & Housen, 2012). Therefore, one of this study's main contributions is its focus on lexical diversity and sophistication.

Two measures of lexical complexity target these two related but distinct aspects of lexical complexity. Considering the evolving proficiency levels of the participants in this study, instead of applying type-token ratio, Guiraud's Index as an indicator of lexical diversity, was analysed. This approach follows existing L2 Chinese studies (e.g., Chen & Li, 2016; Liu, 2017; Wu, 2017). This approach intended to reduce the intervening effects of text length (Bulté & Housen, 2014). Lexical sophistication was measured by the ratio of words at different levels. Specifically, in this study, the new HSK (2012) was chosen as the corpus to categorise words at different levels instead of the *Hanyu Shuiping Cihui Dengji Dagang* (HSCDD) (2001). This is because the corpus should reflect the L2 input received by the learners as closely as possible (Bulté, 2013). With regard to lexical sophistication, operationally, the words categorised under HSK 1 and 2 are considered as beginner level words, the words categorised under HSK 3 and 4 are regarded as intermediate level, and the words categorised under HSK 5 and 6 and beyond are at the advanced level.

Syntactic complexity

It has been shown that different indicators reveal different stages of L2 development. For instance, phrasal complexity increases when learners are at more advanced stages, while syntactic complexity via subordination is mainly achieved when learners are at an intermediate stage of L2 development (cf. Norris & Ortega, 2009). Considering the proficiency levels of the participants in the study, syntactic complexity via subordination was analysed. Moreover, the length of syntactic complexity is considered to be the most sensitive measure to reveal L2 learners' attainment (cf. Bulté & Housen, 2012). Furthermore, the two most analysed syntactic complexity measures in previous studies are length and subordination (Kuiken et al, 2019). Therefore, these two indicators were also analysed in this study: 1) the number of syllables per AS-unit was calculated by the total number of syllables divided by the

number of AS-units; 2) the number of sub-clauses per AS-unit was calculated by the total number of clauses divided by the total number of AS-units (e.g., Chen, 2015; Wu, 2017).

3.5 Data transcription coding and trimming

The collected speeches from the three tests were transcribed by the researcher, who is a native Chinese speaker with several years of experience teaching Chinese as a second language. After transcribing the data, the researcher examined it for accuracy and categorized it into clauses and AS-units (Foster et al, 2000), which are sentence-length utterances designed for oral language. If there are any transcribed data showing that learners produced proper grammatical sentences during the three oral examinations, but they were not relevant to the themes, they would be considered invalid data. This section describes how the transcripted speech samples were trimmed for analysis. Following this, it sets out how the pruned data were coded for CAF measures and indicators.

3.5.1 Data trimming

Data trimming 1: editing out interlocutor speech

The speech data were collected under exam conditions. During the test, when participants had difficulties in producing utterances, the examiner asked them questions to elicit responses. In order to be consistent when comparing the data across the three sessions, the examiner's speech needed to be pruned. The standards for removing interlocutor speech used in this study followed existing studies (e.g., Du, 2013; Liu, 2017; Zhou, 2016; Wu, 2017; Feng, 2018). Elliptical question responses, imitative utterances, and single word Yes/No responses before calculating the mean length of utterance (MLU) were edited out (e.g., Johnston, 2001; Tomas & Dorofeeva, 2019; Wu, 2017). The pruned data were termed 'data Ⅰ' in this study. The examiners' utterances were edited out from the participants' utterances (E: examiners' utterances; P: participants' utterances) as follows:

a. Any one word utterance used to answer or confirm the examiners' questions were edited out. For example:

(1)
E:"所以很重要？"
E: suǒ-yǐ hěn zhòng-yào
E: so very important
E: So it's important?

P9: “对。”
P9: duì
P9: correct
P9: Yes.

The word “对”produced by the participant was pruned because it was only used to confirm the examiner’s question.

b. Negotiation to confirm pronunciation and meaning conveyed by both parties were excluded from the analysis. For example:

(2)
E：“对有些年轻人频繁跳槽的现象，你怎么看？”
E： Duì yǒu xiē nián-qīng rén pín-fán tiào-cáo de xiàn-xiàng ， nǐ zěn-me kàn？
E: to have some young person frequently de phenomenon, you how see?
E: What do you think of the phenomenon that some young people frequently change jobs?

P9: “频繁是什么意思？”
P9: Pín-fán shì shén-me yì-si ？
P9: frequently be what meaning
P9: What does this mean?

E: “frequently. ”

P9：“谢谢。对有些年轻人频繁跳槽的现象，……”
P9： xiè-xie。Duì yǒu xiē nián-qīng rén pín-fán tiào-cáo de xiàn-xiàng ， nǐ zěn-me kàn？ ……
P9: thanks to have some young person frequently de phenomenon, you how see……
P9: Thank you. What do you think of the phenomenon that some young people frequently change jobs? ……

For the example above, the oral production before the participant expressed their answer from “对有些年轻人频繁跳槽的现象，……” was edited out.

c. Proper nouns produced within speech samples were kept. But they were not calculated when obtaining the number of syllables and lexical items. For example:

(3)
P1: “在科克有 Cobh。 是有很多旅游，旅游的人去 Cobh 看海，Titanic 和也去 Blarney Castle, 去 Blarney Stone。”
P1: zài Kē-kè yǒu Cobh. shì yǒu hěn duō lǚ-yóu ， lǚ-yóu de rén qù Cobh kàn hǎi ， Titanic hé yě qù Blarney Castle, qù Blarney Stone。

P1: in Cork have Cobh be have many tourism tourism REL human go Cobh see sea Titanic and also go Blarney castle go Blarney stone.
P1: There are Cobh in cork. There are many tourists. Tourists go to Cobh to see the sea, Titanic and visit blarney castle and blarney stone.

d. English words produced in the speech samples were deleted. The whole clause including the English words were deleted as well. For example:

(4)
P9: “德国人不迟到，因为他们 organised. ”
P9: dé-guó rén bù chí-dào ， yīn-wéi tā-men organised
P9: Germany human not late because they organised
P9: The Germans are not late because they are organised.

The whole clause “因为他们 organised” was deleted due to the fact that it includes the English word “organised”.

The dataset

The ten participants’ data were used to assess the oral development of their L2 Chinese under the CAF framework. It is worth noting that the transcribed speech samples which were analysed in this study were the participants’ oral production after the pruning of interlocutor speech, as has been widely applied in the existing literature (e.g., Du, 2013; Liu, 2017; Zhou, 2016; Wu, 2017; Feng, 2018; Polat & Kim, 2014). An overview of the pruned data from this study’s participants after the data trimming process is presented in Table 10. 22,488 Hanzi made up the raw data from the participants. The interlocutor speech was removed (9.4%), leaving a total of 20,336 Hanzi with a duration of 181.8 minutes that were used for data analysis.

Table 10. Data per participant and the total

Participant No.	Oral Data		Raw Data
	Duration (minutes)	Number of Hanzi	Number of Hanzi
1	18.13	2251	3143
2	13.15	1406	2129
3	18.82	1711	1850
4	22.65	3901	3201
5	22.99	2681	3138
6	24.82	1941	2099
7	15.37	1540	1657
8	18.32	1980	2115
9	14.03	1607	1692
10	13.52	1318	1464
Total	**181.8**	**20,336**	**22,488**

Transcribing the pruned data

To analyse the oral data using the CAF measures, the oral samples needed to be transcribed and decoded. Firstly, the recorded speech samples were transcribed manually. The dysfluency features of the oral speech samples were marked in the written transcription. The transcription standards used are set in Table 11.

Table 11. The transcription standard

Transcription symbols	**Meaning**	**Illustration**
E	The examiners' utterances	**E:** 如果你的朋友借了你 10 欧，但是他 忘记了。
P	The participants' utterances	P: 一个星期或者两个星期。
/e/,/em/	Filled pauses	我住在/e/公寓/e//e/房租/e/是还可以。
{ }	Repetitions	我的床{床}旁边/e/ ，放着一个衣柜。
ǁ	Self-corrections	这样的人具有获得优厚的ǁ优*[xuo51]的<stop>。
~~	False starts	因为/e/他们/e/，我觉得他们总是想/e/这山望着那山高。
*	Unclear or vague syllables	墙上挂着一个 * [tʂɑŋ55]。

3.5.2 Coding accuracy

In terms of measuring accuracy, lexical errors were categorised into various types in this study, which is in line with previous studies (e.g., Ding & Xiao, 2016). The standards used to categorise and code lexical errors in this study is shown in Table 12. It is worth noting that English words used in the speech samples were not marked and counted as errors, but instead were excluded from the total number of syllables.

Table 12. Categorisation of lexical errors

Type	Subcategories	Examples
Lexical errors	a. Words are inconsistent with the topic addressed	他每个月总是付（花）很多钱买衣服，去饭店吃饭。
	b. Similar words or related words are confused	衣服（衣柜）对面放着一个书架。
	c. The keywords are missing	书架旁边放着一个桌子，也(放着)一把椅子。
	d. Pseudo words invented by participants	肺癌症，很多人有肺癌症 (肺癌)。
	e. Lexical items are redundant or omitted	我住地方是有点小，但是（很）干净。
	f. Lexical items do not match properly	我住的地方离科克大学有点儿近（远）。

Note. Lexical errors are underlined with "__". The words in the brackets are the words expected in the context.

Regarding the categorisation and calculation of lexical errors, the standards used in the study are detailed below.

a. If the same lexical error occurred more than once in the transcribed script from one participant collected from one test, it was considered as one error regardless of the frequency. For example:

(5) “我家有两个客厅，我家有一很大客厅，一很小客厅。”
wǒ jiā yǒu liǎng gè kè-tīng， wǒ jiā yǒu yī hěn dà kè-tīng， yī hěn xiǎo kè-tīng 。
There are two lexical errors: the classifier *ge* 个 and modifier *de* 的. Each of these two were omitted in the utterance twice, but each omission was only counted as one lexical error. The correct expression is as follows, within which the omissions are shown in ():

我家有两个客厅，我家有一 (个) 很大 (的) 客厅，一 (个) 很小（的）客厅。
wǒ jiā yǒu liǎng gè kè-tīng， wǒ jiā yǒu yī (gè)hěn dà （de ）kè tīng， yī （gè ）hěn xiǎo （de ）kè tīng 。
My house have two CL living room my house have one CL very big REL living room one CL very small REL living room
There are two living rooms in my house. There is a large living room and a small living room in my house.

b. Lexical errors that occurred in the speech script which were corrected by participants themselves were not considered as lexical errors but repairs. For example:

(6) “我家离很近，no， 我家离学校很近。”
wǒ jiā lí hěn jìn ， no， wǒ jiā lí xué-xiào hěn jìn 。
my house be-away-from very close my house be-away-from school very close
My house is close to， no, my house is close to the school.

c. Any unclear words due to participants’ poor articulation in the speech sample were labelled as “**” and excluded from the analysis. For example:

(7) “我书桌/e/，放着，放着**。”
wǒ shūzhuō, /e/，fàng zhe ，fàng zhe &&。
My table /e/ put ASP put ASP &&
There is ** on my desk.

3.5.3 Coding fluency

Three measures were marked and calculated to assess fluency: the number of silent pauses, the number of filled pauses, and the number of repairs and repetitions. In terms of pauses, the standard of

marking pauses (filled/ silent) that was used was the length of the pause should meet or exceed 0.3 seconds. Two types of pauses were marked and calculated in this study:

1. Unfilled pauses, also termed silent pauses (SP).
2. Filled pauses (FP), such as “um/em” “eh” and “er”. These are normally non-lexical and are not recognised as words as they contain little or no semantic information (Riggenbach, 1991).

Other filled pauses, such as ‘sound stretches’ and ‘lexical fillers’ such as “I mean” and “you know”, were recognised as words but convey little or no semantic information (Riggenbach, 1991) and thus were not examined in this study. For example:

a. FPs were uttered because of the difficulties in eliciting words in an utterance. e.g.,
“墙上挂着一个钟和一个，/e/，一张画儿。”
qiáng shàng guà zhe yī gè zhōng hé yī gè ，/e/，yī zhāng huàr 。
wall on hang zhe one CL and one CL, /e/, one CL painting
There is a clock and a, /e/, a painting on the wall.

The filled pause /e/ in this example was produced in order to hold the conversation until the speaker found the right classifier *zhang* to repair the original classifier *ge*.

b. FPs were used at the beginning or end of utterance because of the informal start. e.g.,
“/e/,我不住在家里，我住在公寓，房租是还可以。”
/e/, wǒ bú zhù zài jiā-lǐ ， wǒ zhù zài gōng-yù ， fáng-zū shì hái kě-yǐ 。
/e/, I not live at home-in I live at apartment rent be fairly okay
/e/, I don't live at home. I live in an apartment. The rent is OK.

The filled pause /e/ was used as an informal start to make time to elicit the utterance.

3.5.3.1 Pause marking

Concerning pause marking, two subsections will be presented as follows because two types of pauses were marked and calculated in this study.

Silent pauses

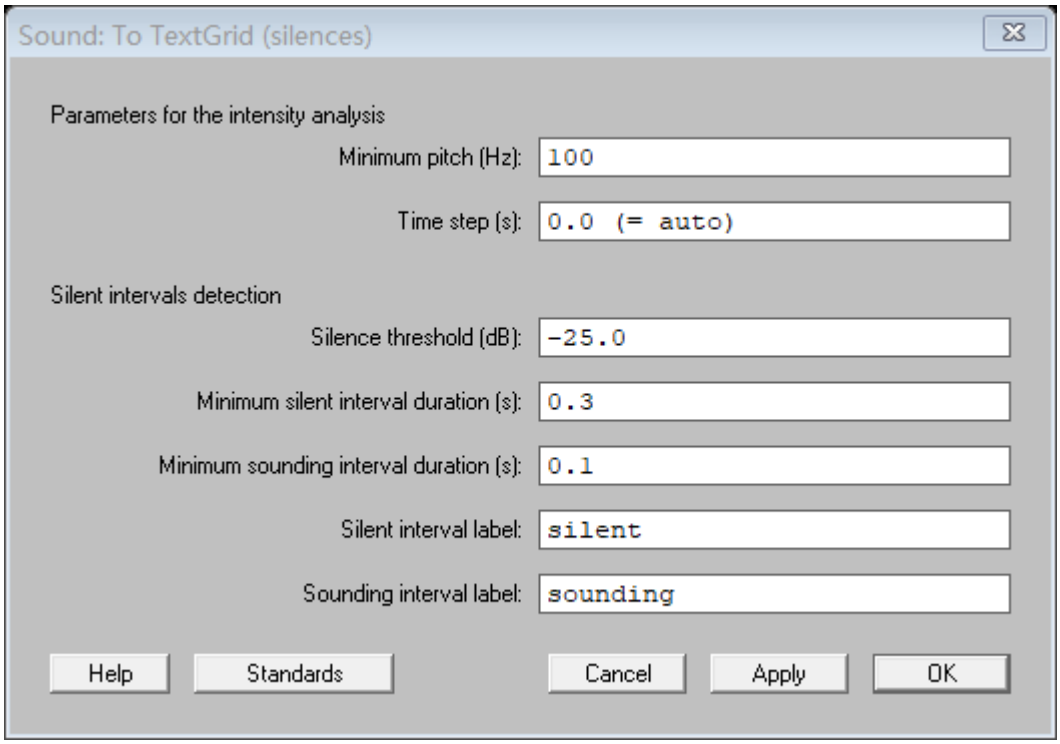
This section illustrates how the silent pauses were coded by calculating the number of silent pauses per 100 syllables and the average length of silent pauses. All the speech files that needed to have the examiners' speech edited out (see 3.5.1) were loaded into Praat in MP3 format. Then the files were edited as follows:

1. Under the “Praat objects” window, click “Open” and “read from file” in a sequence, and then choose the file that needs to be analysed.
2. Click “View & edit” and select the part that needs to be deleted. Then click “Edit-cut”. This results in the speech produced by examiners being cut.
3. The duration of the samples in seconds is available within the “sound” editing window.
4. Under the “Praat objects” window, the file can be saved in a .wav format.

However, due to the poor transcription quality of existing AI transcription software, all of the speech samples were transcribed by the researcher herself. Concerning calculating the duration and ratio of silent pauses, the recorded files were transferred into .mp3 files to be coded in Praat (See https://www.fon.hum.uva.nl/praat/). The procedure for using Praat is as follows:

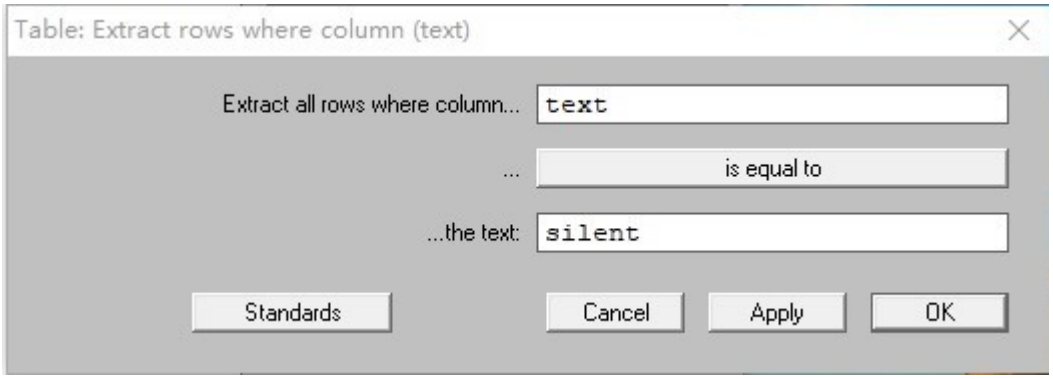
1. Under the “Praat objects” window, click “Open” and “Read from file”, and then choose the file that needs to be analysed.
2. Click “Annotate”, select “To textgrid (silence)”, and set “0.3” on the option for “Minimum sounding interval duration”. Doing so means that silent pauses which reach or exceed 0.3 seconds are coded (see Figure 3 below).

Figure 3. Annotation of silent pauses in Praat



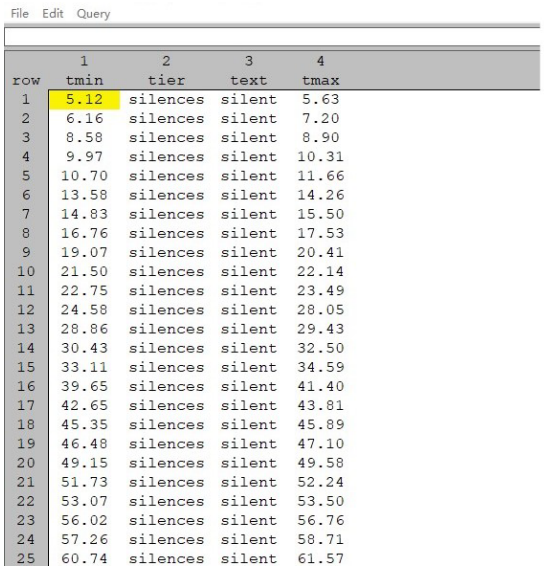
3. A “Textgrid” file is created after clicking “OK”.
4. Under the “Textgrid” file, click “Tabulate” and then select “Down to table”. In terms of the option “Time decimal”, in this study, “2” was selected in light of the set standard of 0.3 seconds.
5. Under the “Table” file, click “Extract”. In the “Extract rows where column (text)” window, input “text” in the “Extract all rows where column” option box, and input “silent” in the “the text” option box. Then click “OK” to finish this step (See Figure 4 below).

Figure 4. Extraction of silent pauses in Praat



6. A “Table silent-only” file is created. Afterwards click “View & Edit” and all of the silent pauses that reach or exceed 0.3 seconds from the speech samples are coded with starting and ending time points (see Figure 5 below).

Figure 5. Silent-only file created in Praat


File Edit Query

row	1 tmin	2 tier	3 text	4 tmax
1	5.12	silences	silent	5.63
2	6.16	silences	silent	7.20
3	8.58	silences	silent	8.90
4	9.97	silences	silent	10.31
5	10.70	silences	silent	11.66
6	13.58	silences	silent	14.26
7	14.83	silences	silent	15.50
8	16.76	silences	silent	17.53
9	19.07	silences	silent	20.41
10	21.50	silences	silent	22.14
11	22.75	silences	silent	23.49
12	24.58	silences	silent	28.05
13	28.86	silences	silent	29.43
14	30.43	silences	silent	32.50
15	33.11	silences	silent	34.59
16	39.65	silences	silent	41.40
17	42.65	silences	silent	43.81
18	45.35	silences	silent	45.89
19	46.48	silences	silent	47.10
20	49.15	silences	silent	49.58
21	51.73	silences	silent	52.24
22	53.07	silences	silent	53.50
23	56.02	silences	silent	56.76
24	57.26	silences	silent	58.71
25	60.74	silences	silent	61.57

7. Click “Save” and then select “Save as tab-separated file” to create a text-only file.
8. Now open the file with Microsoft Word, and copy-paste it into an Excel sheet. The total number of silent pauses and their starting and ending time points are now available.

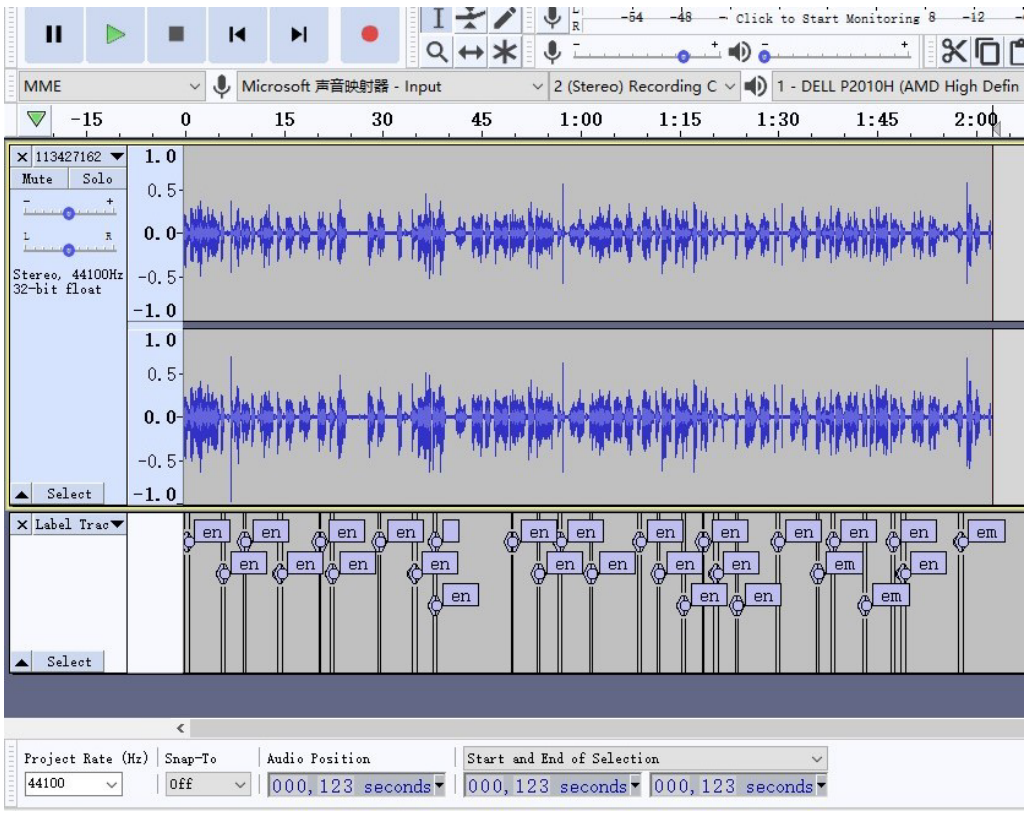
Filled pauses

To mark the filled pauses (FP), the pruned speech samples, after editing out interlocutor speech (see 3.7.1) in .wav format, were loaded into Audacity 2.3.2. The filled pauses at the beginning and end were marked in each participant’s transcript to calculate the average length of pauses. The reason for marking any pause at the beginning and end of each complete utterance is that the Average Length of Filled Pause (ALFP) is calculated by the total length of pauses divided by the total number of pauses, including pauses at the beginning and the end. See Figure 6 below for a speech sample coded with marked pauses. The procedure for using Audacity 2.3.2 to code filled pauses is as follows:

1. Click “File” and select “Open”, select the file that is to be coded.

2. Play the file and when there is a filled pause, locate and mark it manually.
3. Click “Edit”, “Labels” and “Add Label at Selection” in sequence. A new label is then created in a new label track underneath the audio track. Then type “en, ah, oh” on the label to annotate the selected filled pauses. See Figure 6 below for the results after the labelling of one speech sample.

Figure 6. Labelling filled pauses



4. To extract the number of filled pauses marked within the speech sample, click “Edit-labels”. A new window is then presented with the number of filled pauses. The start and end times of the marked filled pauses are also available, as shown in Figure 7.

Figure 7. The filled pauses with start and end times in one speech sample



	Track	Label	Start Time	End Time	Low Frequency	High Frequency
1	1 - Label Track	en	000,000 seconds	000,001 seconds	-----,-- Hz	-----,-- Hz
2	1 - Label Track	en	000,006 seconds	000,006 seconds	-----,-- Hz	-----,-- Hz
3	1 - Label Track	en	000,009 seconds	000,009 seconds	-----,-- Hz	-----,-- Hz
4	1 - Label Track	en	000,014 seconds	000,015 seconds	-----,-- Hz	-----,-- Hz
5	1 - Label Track	en	000,020 seconds	000,021 seconds	-----,-- Hz	-----,-- Hz
6	1 - Label Track	en	000,022 seconds	000,023 seconds	-----,-- Hz	-----,-- Hz
7	1 - Label Track	en	000,029 seconds	000,030 seconds	-----,-- Hz	-----,-- Hz
8	1 - Label Track	en	000,035 seconds	000,035 seconds	-----,-- Hz	-----,-- Hz
9	1 - Label Track		000,038 seconds	000,038 seconds	-----,-- Hz	-----,-- Hz
10	1 - Label Track	en	000,038 seconds	000,038 seconds	-----,-- Hz	-----,-- Hz
11	1 - Label Track	en	000,050 seconds	000,050 seconds	-----,-- Hz	-----,-- Hz
12	1 - Label Track	en	000,053 seconds	000,054 seconds	-----,-- Hz	-----,-- Hz
13	1 - Label Track	en	000,057 seconds	000,057 seconds	-----,-- Hz	-----,-- Hz
14	1 - Label Track	en	000,062 seconds	000,062 seconds	-----,-- Hz	-----,-- Hz

5. Finally, click “Export” and a label track in txt format is created, which can be saved. Then copy-paste the statistics into Excel, and the total duration of the filled pauses in seconds can be calculated. The ALFP can then be obtained via the total duration of pauses divided by the total number of filled pauses.

3.5.3.2 Repairs and repetitions

Concerning the indicator of repairs and repetitions used in this study, it was obtained by dividing the number of repetitions and repairs by the total syllables, multiplied by 100. This gives the number of repetitions and repairs per 100 syllables. Operationally, three types of repetitions were marked and calculated as shown in (﹏) below.

1) Repetitions of words，e.g.,

（8）“我的床，床旁边/e/，放着一个衣柜。”
wǒ de chuáng， chuáng páng-biān /e/， fàng zhe yī gè yī-guì 。
I REL bed, bed beside /e/, put ASP one CL wardrobe.
My bed, next to my bed /e/, there is a wardrobe.

Repetitions within a word，e.g.,

（9）“他们可能不会频繁地工作, 勤，勤奋地工作。”
tā-men kě-néng bú huì pín-fán de gōng-zuò , qín ， qín-fèn de gōng-zuò 。
they may not can frequently REL work, morpheme (qín) frequently REL work
They may not work frequently, work diligently.

2) Repetitions of phrases，e.g.,

（10）“我住地方，/e/有/e/，我住地方是/e/有点小，但是干净。”

wǒ zhù dì-fāng ， /e/ yǒu /e/，wǒ zhù dì-fāng shì /e/yǒu-diǎnr xiǎo ， dàn-shì gàn jìng 。

I live REL place / e / have / e / I live REL place / e/ a little small but clean.

Where I live, / e / there is / e /, where I live is / e/ a little small, but clean.

3) Repetitions of clauses，e.g.,

（11）“他们买很多东西，不，他们买很多东西，去城市里面看看常常。”

Tā-men mǎi hěn duō dōng-xi ， bú ， tā-men mǎi hěn duō dōng-xi ， qù chéng-shì lǐ-miàn kàn -kan cháng-chang 。

they buy very many things, no, they buy very many things, go city inside have-a-look often.

They buy a lot of things, no, they buy a lot of things, often go to the city centre to have a look.

In terms of repairs, three sub-categories were marked and calculated as shown in (__) below.

a. Repairs of prounciations，e.g.,

（12）“因为我很喜欢看电视，/e/，看电*[in11]，看电影。”

yīn-wéi wǒ hěn xǐ-huān kàn diàn -shì ，/e/，kàn diàn *[in11]，kàn diàn-yǐng 。

because I like watch TV /e /watching electricity [in11], watch movie

Because I like watching TV, /e /, watching TV * [in11], watching movies.

b. Repairs of words, e.g.,

（13）“墙上挂着一个钟和一个，一张画儿。”

qiáng shàng guà zhe yī gè zhōng hé yī gè ， yī zhāng huàr 。

wall on hang ASP one CL clock and one CL one CL painting

There is a clock, a, and a painting on the wall.

c. Repairs of phrases，e.g.,

（14）“我住的客厅，我的客厅，里面放着沙发。”

wǒ zhù de kè-tīng ， wǒ de kè-tīng ，lǐ-miàn fàng zhe shā-fā 。

I live REL living room I REL living room inside put ASP sofa

The living room I live, my living room, there is a sofa in it.

3.5.4 Coding Complexity

Following the majority of studies which assess the oral development of L2 learners of Chinese (e.g., Ye, 2015; Chen, 2015a, 2015b; Ding & Xiao, 2016; Liu, 2017; Wu, 2017), two sub-dimensions

of complexity were analysed: lexical complexity and syntactic complexity. To measure lexical complexity, lexical items in the transcribed speech needed to be segmented and categorised into different levels. To measure syntactic complexity, the number of AS-units and the number of clauses were coded. In previous section (2.2.2.3), word segmentation and frequency in L2 Chinese speaking studies are reviewed. The section then describes how lexical and syntactic complexity measures were coded in the transcribed speech samples in this study.

In this research, the New HSK (2012) was used as the benchmark to categorise produced lexical items. This is because HSK (2012) has been widely accepted as a suitable assessment tool. Operationally, the words categorised under HSK 1 and 2 were considered as beginner level words, the words categorised under HSK 3 and 4 were regarded as intermediate level ones, and the words categorised under HSK 5 and 6 and beyond were considered to be advanced level words. For the lexical items that are not included in the HSK (2012), the online official Contemporary Chinese Dictionary was used.

3.5.4.1 Coding lexical complexity

When measuring lexical complexity, and error rates, this study excluded dysfluency outputs which do not convey information, such as:

- informal starts to sentences
- repetitions
- corrections/repairs
- common fillers “eh/er” and “ehm”
- unfilled/silent pauses

To measure lexical complexity, the following need to be enumerated: word tokens and word types. Moreover, the number of HSK 1 to 6 words and those beyond the HSK (2012) were segmented and counted by the following four lexical complexity measures: Guiraud’s Index (types / √ tokens), the ratio of beginner, intermediate, and advanced level lexical items. The transcribed speech sample of one participant shown below in Figure 8 exemplifies how samples were segmented and categorised into different levels based on the HSK (2012) and the Contemporary Chinese Dictionary. Before the transcribed speech sample was segmented in THULAC - the online Chinese word segmentation

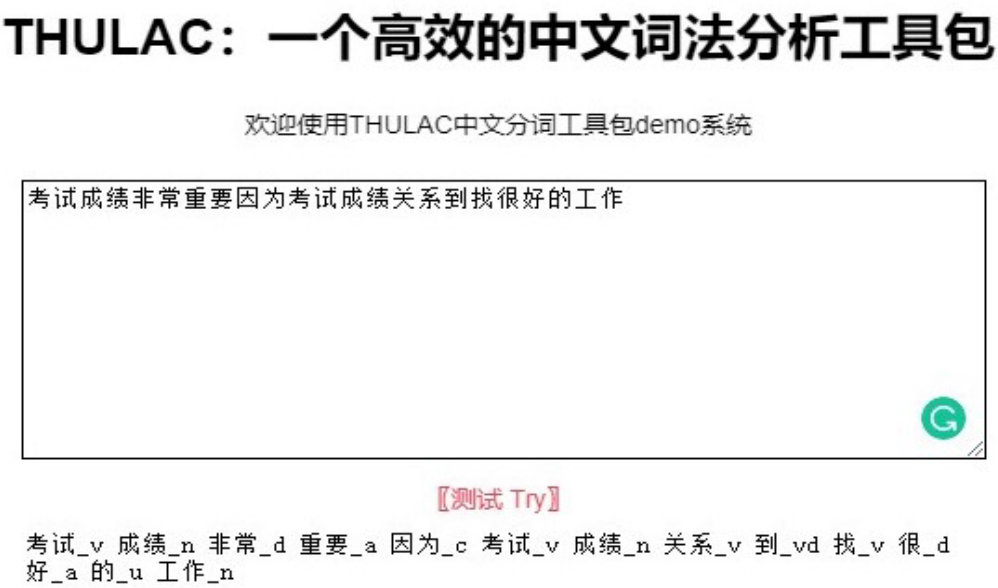
Toolkit (e.g., Liu, 2016), two kinds of syllables were pruned. First, all of the filled pauses which convey no meaning such as “en” “ah” and “oh” were pruned. Second, the syllables which were part of one lexical item and which cannot be regarded as one lexical item were edited out. Furthermore, those syllables which cannot be considered as individual words but belong to a word in the context were also pruned in the calculation of the number of words. For example,

(15) “年轻人频，频繁跳槽越来越流行。”
Nián-qīng rén pín， pín-fán tiào-cáo yuè-lá-iyuè liú-xíng.
Young person pin, frequent job-hopping more and more popular
Job hopping is becoming more and more popular among young people.

(16) “有的商店有免费的送货，买东，上网买东西是很方便。”
Yǒu-de shāng-diàn yǒu miǎn-fèi de sòng huò， mǎi morpheme (dōng)， shàng-wǎng mǎi dōng-xi shì hěn fāng-biàn.
Some stores have free REL delivery, buy east, surf-the-internet buy things be convenient
Some stores have free delivery, it's very convenient to buy things online.

After pruning the syllables of filled pauses and morphemes within a single word, transcribed speech samples were entered into the online THULAC platform for the first segmentation step. The speech text of each transcription was automatically segmented. The example below shows how THULAC was used for the first step of segmenting words in a transcribed script.

Figure 8. Using THULAC for the first step of segmenting words in a script



In terms of coding the number of tokens and types in a speech sample, Figure 8 shows an example of one segmented speech text using the THULAC online segmental tool. Based on the aforementioned definitions of types and tokens, there are 14 tokens (考试, 成绩,非常,重要,因为,考试,成绩,关系,到,找,很,好,的, 工作), and 12 types (考试, 成绩,非常,重要,因为,关系,到,找,很,好,的, 工作). However, some errors occur using THULAC. Regarding the errors automatically segmented using THULAC, for the words that are not on the HSK (2012) list, and the words which cannot be found in the Contemporary Chinese Dictionary, a second segmentation step and categorisation were required as shown in (“ ”) below.

1. Some phrases that were automatically categorised as lexical items by THULAC needed to be further segmented into lexicons on the basis of the HSK (2012) and the Contemporary Chinese Dictionary. For example, “不好”，“一个”，“看书 ”，“很多” .
2. Four-syllable idioms and Chinese sayings were automatically segmented into two or three sub-components and were labelled as one word, such as “月光族”, “这山望着那山高”， “更上一层楼”， “踏踏实实”.
3. Proper nouns including the names of countries, places, and festivals were also considered as words, such as “爱尔兰”， “淘宝”，”欧元”， “Instagram”.

4. Affixes such as “们” are not words. Such affixes and the words that they are attached to were categorised based on the level of the words. For example, “我们”,“他们”,“朋友们” were categorised under HSK 1.
5. Abbreviations were transformed into their original forms, such as “四年前”（四,年,以前）.

On the basis of the rules introduced above, the majority of the words and expressions in the participants’ speech samples could be found on the list of HSK 5000 lexical items. Those that could not be found on the HSK (2012) were regarded as lexical items beyond the HSK system. In this research, they were categorised as advanced level words.

3.5.4.2 Coding syntactic complexity

In this study, AS-units as the base unit were used to investigate the oral development of English-speaking learners of Chinese. Each transcript (n=10) was coded for syntactic complexity by enumerating the total number of syllables, the total number of AS-units, and the total number of clauses. The next section presents how the AS-units were coded. Based on data I, the examiners’ speech was pruned (see section 3.5.1), and self-corrections, false-starts, repetitions, reformulations were edited out to calculate the number of AS-units. This process was termed data trimming 2.

Data trimming 2: coding AS-units

To code AS-units, several rules had to be made. The process for doing so followed Foster et al. (2000), who claimed that it is essential to systematically exclude dysfluency features, such as false-starts and corrections, from the total word count when calculating the length of syntactic complexity. This measure of complexity was calculated by the words per unit. Hence, based on pruned data I, false-starts, repetitions and self-corrections/repairs were further edited out to calculate the number of AS-units and sub-clauses.

There are two reasons for editing out the dysfluency features of an utterance when coding AS-units. First, in an utterance, if there is a self-correction, the final version is counted as an AS-unit, and the previous versions are excluded (Foster et al., 2000). Therefore, repairs which do not affect the

coding of the number of AS-units can be edited out. Second, when coding AS-units, false-starts, repetitions, reformulations have been considered as being within one AS-unit in previous studies (e.g., Chen & Li, 2016; Wu, 2017). Therefore, editing out the false starts, repetitions, reformulations based on pruned data I (see section 3.5.1) would not affect the coding of the number of AS-units and clauses.

The examples below show how false starts, repetitions and repairs were excluded in the AS-units calculating process. The original version is in (_ _) and the pruned version for analysis is in (﹏).

a. False starts:
"因为他们, 我觉得他们总是想这山望着那山高。"
yīn-wèi tā-men , wǒ jué-de tā-men zǒng-shì xiǎng zhè shān wàng zhe nà shān gāo
because they I think they always think this mountain look-over ASP that mountain high
Because of them, I think they always think the grass is always greener on the other side.

The pruned version:我觉得他们总是想这山望着那山高。

b. Repetitions:
"我觉得频繁跳槽的，跳槽的人不太好。"
wǒ jué-de pín-fán tiào-cáo de， tiào-cáo de rén bú tài hǎo
I think frequently job-hopping REL job-hopping REL human not too good
I think people who change jobs frequently, change jobs frequently are not very good.

The pruned version:我觉得频繁跳槽的人不太好。

c. Repairs:
"墙上挂着一个钟和一个，一张画儿。"
qiáng shàng guà zhe yī gè zhōng hé yī gè， yī zhāng huàr
wall on hang ASP one CL clock and one CL one CL painting
There is a clock and a painting on the wall.

The pruned version:墙上挂着一个钟和一张画儿。

When calculating the length of AS-units, one-word utterances, English words, as well as incomplete clauses were excluded.

d. One-word utterances, e.g.， *yeah, Okay, uh, um* and *对*.

e. English expressions, e.g.，

“No, understandable. 我觉得是可以。”
No, understandable. wǒ jué-de shì kě yǐ .
No, understandable. I think be okay.
No, understandable. I think (it) is okay.

The pruned version: 我觉得是可以。

f. The As-units and sub-clauses within an As-unit that do not express a complete meaning in the context were edited out，e.g.，

“上网我跟我……，我常常看电影，看 Youtube.”
shàng wǎng wǒ gēn wǒ ……， wǒ cháng-chang kàn diàn-yǐng ， kàn Youtube.
on internet me with me... I often watch movies watch Youtube.
I surf the Internet with me... I often watch movies and Youtube.

The pruned version: 我常常看电影，看 Youtube.

The data after the coding of the AS-units and clauses were named ‘pruned data Ⅱ’. The working AS-unit used in this study was a single speaker’s utterances consisting of an independent unit or an independent unit together with a subordinate clause. This subordinate clause includes a predicate so that it can be interpreted to be a full clause in the discourse. Following Foster et. al. (2000), in data Ⅱ, AS-units in the transcripts were marked as follows: AS-unit boundaries were marked by [/], clause boundaries connecting two clauses within an AS-unit were marked by [::], as illustrated in (/ /) below.

(17) /我不住在家里/我住在公寓/房租是还可以/我住的地方离科克大学有点儿近/步行需要十五分钟/我住地方是有点儿小::但是干净/
/wǒ bú zhù zài jiā-lǐ /wǒ zhù zài gōng-yù /fáng-zū shì hái kě yǐ /wǒ zhù de dì-fāng lí kē-kè dà -xué yǒu-diǎnr jìn /bù-xíng xū-yào shí wǔ fèn-zhōng /wǒ zhù dì-fāng shì yǒu-diǎnr xiǎo ::dàn-shì gān-jìng /
/I not live at home / I live at apartment / rent be fairly OK / I live REL place be-away-from a Cork University a little close / walk need be five minute / I live place be a little small:: but clean/
/I don't live at home / I live in an apartment / the rent is OK / my place is a little close to University College Cork/ it takes 15 minutes to walk / my place is a little small:: but it's clean/

In conclusion, there were two types of data in this study, 'data I' were the data with the interlocutor speech edited out (see section 3.5.1) and 'data Ⅱ' were the data used for AS-units coding (see section 3.5.4.2).

3.5.5 Summary of CAF measures analysed in this study

The list of measures and indicators analysed in this study are shown in Table 13 below, which also shows the calculation of the 14 CAF indicators. Finally, the table also presents the data type on which the indicators were coded, i.e., data Ⅰ (see section 3.5.1) and data Ⅱ (see section 3.5.4.2).

Table 13. CAF measures analysed in this study

Constructs	Sub-constructs		Indicators	Calculation	Data type
Accuracy	Lexical accuracy		The ratio of error-free lexical items	1- The number of lexical errors/ the total number of lexical items(type)	Data Ⅰ
Fluency	Speed fluency		Speech Rate (SR)	The total number of syllables/ Time of utterance including pause time×60	Data Ⅰ
			Mean Length of Runs (MLR)	The total number of syllables /the total number of silent pauses(≥ 0.3 seconds)	Data Ⅰ
	Breakdown fluency		The average length of filled pause (ALFP)	The total length of filled pauses /the total number of filled pauses	Data Ⅰ
			The average length of silent pause (ALSP)	The total length of silent pauses /the total number of silent pauses	Data Ⅰ
			The number of filled pauses per 100 syllables (FP100)	The total number of filled pauses / the total number of syllables × 100	Data Ⅰ
			The number of silent pauses per 100 syllables (SP 100)	The total number of silent pauses / the total number of syllables × 100	Data Ⅰ
	Repair fluency		The number of repairs and repetitions per 100 syllables (RR100)	The number of repairs and repetitions/ the total number of syllables×100	Data Ⅰ
Complexity	Lexical complexity	Lexical diversity	Guiraud's Index	Types / √ tokens	Data Ⅰ
		Lexical sophistication	The ratio of beginner level words	The number of HSK 1& 2 words / the total number of words	Data Ⅰ
			The ratio of intermediate level words	The number of HSK 3 & 4 words /the total number of words	Data Ⅰ
			The ratio of advanced level words	The number of HSK 5 & 6 words and beyond / the total number of words	Data Ⅰ

	Syntac-tic com-plexity		The number of syllables per AS-unit	The total number of sylla-bles / the total number of AS-units	Data Ⅱ
			The number of sub-clauses per AS-unit	The total number of clauses / the total number of AS-units	Data Ⅱ

3.6 Research questions

To explore the oral development of L2 learners of Chinese under the CAF framework, this research measured the oral speech of English-speaking learners derived from topic-based oral tests that were conducted as part of a four-year undergraduate Chinese programme at University College Cork, Ireland. This research aimed to explore the oral development of English-speaking learners of Chinese by applying the CAF framework, and to discuss the relationships among the sub-components of CAF. The study compared the two widely applied competitive theories, the Trade-off Hypothesis (Skehan, 1998; Skehan & Foster, 1999) and the Cognition Hypothesis (Robinson, 2001, 2007). The study was orientated around two main research questions:

1. How do the complexity, accuracy, and fluency of the oral performance of instructed English-speaking L2 Chinese learners develop over a four-year college-level Chinese programme which includes a 10-month study abroad period?

Considering the two learning contexts (SA and FI at home), this first research question can be divided into two sub-questions:

1a) How do the complexity, accuracy, and fluency of the oral performance of instructed English-speaking learners of Chinese develop during pre- and post SA?
2b) How do the complexity, accuracy, and fluency of the oral performance of instructed English-speaking learners of Chinese develop during the FI at home context?

2. Are the general relationships between CAF constructs, and the relationships between the sub-constructs of complexity and fluency, competitive or supportive in the oral performance of learners during the examining period?

The second question aimed to explore the effects of learning context (SA and FI at home) on the interrelationships between CAF measures as well as between the different dimensions within complexity and fluency to investigate the oral performance of instructed English-speaking learners. This

research question can be subdivided into three separate categories, targeting specific relationships respectively:

2a) between complexity, accuracy and fluency, including (a) between complexity and accuracy, (b) between complexity and fluency, (c) between accuracy and fluency

2b) between subconstructs of complexity, including (a) different subdimensions of lexical complexity, (b) different subdimensions of syntactic complexity, (c) different subdimensions of lexical complexity and syntactic complexity.

2c) between subconstructs of fluency, including (a) different subdimensions of speed fluency, (b) different subdimensions of breakdown fluency, (c) different subdimensions of speed, breakdown and repair fluency.

Taking the Study Abroad experience into consideration, pre-SA session (S1) and post-SA session (S2) were explored to investigate the learners' oral development as well as the relationships between CAF measures. Moreover, to investigate the effect of Formal Instruction at home, the learners' final attainment level at the end of S3, i.e., 6 months' formal instruction after SA, was investigated. To compare the post-SA session (S2), the oral development as well as the relationship between CAF measures impacted by the FI at home maintenance factor have been provided.

Research question (1) aimed to reveal what changes, if any, occurred in the oral development of English-speaking learners of Chinese, in particular, in relation to the effects of different learning contexts (SA and FI), over a 28-month period, as measured by complexity, accuracy, and fluency indicators. In particular, it also aimed to analyse if as was expected there was a decrease in oral performance of English-speaking learners of Chinese during the FI at home context. Research question (2) sought to explore how the three dimensions of CAF correlated with each in terms of the oral development of English-speaking learners of Chinese and how this was impacted by the different learning contexts (SA and FI). This study also aimed to contribute to the debate concerning the Trade-off Hypothesis and the Cognition Hypothesis.

3.7 Predictions

Predictions for the learners' oral performance were made for two periods surrounding the 10-month Study Abroad: the pre- and post-SA periods (S1-S2) and the delayed FI at home maintenance

period (S2 & S3). On the basis of the literature review set out previously the following predictions were made in this study.

3.7.1 CAF development during pre-SA and post-SA (S1-S2)

Considering the benefits of SA on oral development, the study hypothesised that a limited improvement in fluency measures would occur, but that in terms of speech fluidity there would be greater growth. As for disfluency and repairs, only a small and non-significant reduction was expected. Accuracy was not expected to see radical change during the examining period. It has been suggested that there is no radical difference in accuracy between learners at beginner level and intermediate level. Moreover, in light of Ferrari's (2012) review of previous longitudinal and cross-sectional studies, L2 learners' accuracy is expected to increase in the long run (i.e., after three years). In terms of syntactic complexity, this study predicted mixed results. The length of syntactic complexity is expected to increase significantly after 3-9 months' SA according to previous SA research (Jensen & Howard, 2014; Mora & Valls-Ferrer, 2012). Syntactic complexity via subordination can also reveal statistically significant growth after 1-semester of SA (Llanes & Muñoz, 2013; Pérez-Vidal & Juan-Garau, 2011). However, some other studies (LIanes & Serrano, 2017; Mora & Valls-Ferrer, 2012; Pérez-Vidal et al., 2012; Serrano, Llanes and Tragant, 2011) have found that progress in subordination is unlikely after SA. Lexical development is expected to benefit highly from SA. Thus, for lexical sophistication, the ratio of beginner level words was expected to dominate the oral production of the learners across the 28-month period considering their proficiency level in this study. This is because intermediate and advanced-level lexical items are mainly expected to only be used when learners are at an advanced level according to Chen (2015a). In this study the ratio of intermediate and advanced level lexical items was predicted to increase radically after the 10-month SA period due to the exposure to the target language this would enable.

Moreover, considering the impact of pre-planning on the oral performance of the English-speaking learners of Chinese in this study, pre-planning was expected to strongly raise the learners' complexity and fluency, but to raise accuracy by a lesser extent according to previous studies (Skehan, 2009a). Therefore, significant improvement was expected in complexity and fluency, impacted by the task of topic prompted monologues with planning time. In particular, fluency and lexical variety were not expected to show statistically significant improvement during pre-SA and post-SA according to

Wright (2020). Instead, based on past studies (Freed, Segalowitz, & Dewey, 2004; Isabelli-Garca et al., 2018; Du, 2013), growth was only expected in speed fluidity.

3.7.2 CAF development during FI at home context (S2-S3)

In general, according to past research findings (Collentine, 2004; Freed, Segalowitz & Dewey 2004; Juan-Garau & Pérez-Vidal, 2006, 2007; Mora & Valls-Ferrer, 2012), the study hypothesised that when learners were in the FI at-home context, there would be a lack of significant improvement in their oral gains, because the L2 learners would not be able to access the large amounts of authentic input that they could during the SA period, except for subordinates and vocabulary. Learners in an FI context seem to concentrate on learning vocabulary and subordinating at the expense of accuracy and fluency. Also, SA gains are very likely to backslide 6 months after returning to an FI at home context according to previous research (e.g., Juan-Garau & Pérez-Vidal, 2006, 2007; Pérez-Vidal & Juan-Garau, 2005). This is very likely due to the lower proficiency levels (late beginner to high intermediate) of the participants in this study. Therefore, in this study, the complexity measures which were expected to benefit from the SA experience, such as syntactic complexity via length, and advanced-level lexical items were also expected to revert to the pre-SA level. A similar backslide was also expected in fluency, in particular, dysfluency phenomena, which are normally not trained explicitly in the classroom setting, such as silent pauses and self-repetitions based on past studies (Huensch et al., 2019; Juan-Garau & Pérez-Vidal, 2006, 2007; Pérez-Vidal & Juan-Garau, 2005; Trenchs-Parera, 2009; Tracy-Ventura et al, 2021).

3.7.3 Correlations between CAF measures during pre- and post-SA (S1-S2)

The predictions made in terms of the correlations between CAF constructs, as well as between subconstructs within each construct, relating to the study abroad factor are detailed below. It should be noted that because accuracy was measured by only one indicator in this research, there were consequently no predicted correlations in the domain of accuracy.

In terms of the relationships between CAF constructs used to measure the oral performance of English-speaking learners of Chinese on topic-based monologues both pre- and post-SA, this study expected to see the trade-off effect in line with Skehan's predictions (Skehan, 2009; Wang & Skehan,

2014). Specifically, this was expected between complexity and accuracy to a large extent, but between fluency and accuracy, or between fluency and complexity to a lesser extent. Moreover, because of the rehearsal (planning) monologue tasks that the students would need to complete during the study, complexity and fluency in language performance were expected to have a connected improvement pattern. This is because the pre-planning of tasks strongly raises complexity and fluency according to predictions by Skehan (2009, 2009a). Therefore, it was thought to be very likely that the English-speaking learners of Chinese would produce more fluent and more complex language at the cost of accuracy.

In terms of the correlations within the domain of complexity, the study hypothesised that a general trade-off effect would be apparent between lexical complexity and syntactic complexity in line with the findings of previous studies (Skehan, 2009a; Larsen-Freeman, 2006; Spoelman & Verspoor, 2010). Moreover, lexical diversity was not expected to benefit from the SA experience, but the SA was predicted to advantage syntactic complexity and advanced-level lexical items (see section 3.7.1). In particular, a correlation between lexical variety and global grammatical complexity was expected to be negative at lower proficiency levels because the use of varied lexical items may reduce the resources which are available to attend to grammar at lower proficiency levels based on the findings of Vervellotti (2012). Therefore, tension between lexical diversity and syntactic complexity is predicted.

Within the sub-domain of syntactic complexity, it was thought that there might be a supportive relationship between complexity by length and by subordination, because both were expected to benefit from the SA experience. Within the subconstruct of lexical complexity, the relationship between lexical diversity and lexical sophistication is very weak, and they are independent of one another for non-native speakers based on the findings of Skehan (2009a). A trade-off effect was expected between the three measures of lexical sophistication indicating the percentage of different level words because they are mutually dependent and together add up to 100%.

Concerning correlations within the domain of fluency, in general, no trade-off effect was expected after the SA. This is because most speed and breakdown fluency measures capture different subdomains of fluency in oral performance, revealing a missing of trade-off effects due to the fact that SA benefit fluency. Moreover, influenced by the monologue task with planning that the participants were required to undertake, fluency was expected to be enhanced by planning according to Skehan's

findings (Skehan, 2009c). Therefore, a general connected-improvement relationship between fluency measures within the fluency construct was expected.

3.7.4 Correlations between CAF measures during FI at home context (S2-S3)

The predicted correlations between CAF constructs, as well as between subconstructs within each construct of the CAF, during the FI at home context (S2-S3) are detailed below.

With regard to the relationships between the CAF constructs when investigating the oral performance of English learners of Chinese on topic-based monologues during the FI at home context, the study expected the trade-off effect and that this effect would be broadly similar during both the pre- and post-SA periods. Considering the effects of FI at home maintenance on oral performance (see section 3.7.2), SA gains were considered very likely to backslide after a 6-month period back home and that learners would concentrate on learning vocabulary and subordinating at the expense of accuracy and fluency during the FI at home context based on previous research (Juan-Garau & Pérez-Vidal, 2007; Huensch et al., 2019; Tracy-Ventura et al., 2021). Specifically, when the learners did not reach a higher-level proficiency. Therefore, a general improvement in vocabulary was expected at the cost of fluency and the other complexity measures (syntactic complexity and lexical sophistication), which were expected to be enhanced by the SA factor.

Within the domain of complexity, similarly to the pre- and post-SA period, during the FI at home context, a general trade-off effect between lexical complexity and syntactic complexity was also expected. However, in contrast to the benefit to syntactic complexity which was expected from the SA period, during the FI at home context, advanced-level lexical items (see section 3.7.1), vocabulary (lexical diversity) and syntactic complexity via subordination were expected to be given more attention. Therefore, learners were predicted to be very likely to concentrate on vocabulary development at the expense of syntactic complexity and display a decrease in advanced-level words which would suggest back sliding in lexical sophistication.

Among fluency measures, learners at FI seem to concentrate on learning vocabulary (lexical diversity) and subordinating at the expense of accuracy and fluency. Furthermore, the oral fluency

gains which were expected from the SA period, were predicted to revert to the pre-SA level after a 6-month period back home during FI at home context. Similar to the correlations between the fluency measures related to the SA factor, a joint decrease was expected with no trade-off effect during the FI at home context.

Chapter 4: Data Analysis and Results

In this chapter, the average data of the ten participants' performance measured by the 14 CAF indicators has been applied to track the oral development of a group of English-speaking undergraduates of Chinese over 28 months (including a 10-month study abroad period). To investigate the two learning contexts (Study Abroad and Formal Instruction at home) on the relationship between CAF components when analysing the oral development of English-speaking learners of Chinese, the first two sessions (S1-S2) were analysed to consider the study abroad factor. Subsequently, regarding the effects of the FI at-home maintenance period, correlations among CAF components between S2 and S3 have been examined. Finally, the individual performance from each level of the two levels of Chinese proficiency among the 10 participants at three tests during the 28 month period will be anaysed to provide a more comprehensive developmental trajectory of the participants in the study.

4.1 SA related results (S1-S2)

A paired-samples t-test was conducted to compare the correlations between CAF in the pre-SA (S1) and post-SA (S2) conditions. The 14 CAF measures, categorised into the three domains: complexity, accuracy, and fluency related to the study abroad factor, are presented below.

Table 14. SA effects on oral performance (S1-S2)

Constructs			**Indicators**	**Mean**		**t**	**p**	**Statistics**
				S1	**S2**			
Complexity	Syntactic complexity		Syllables per AS –unit	18.11	34.04	-6.03	***.001	increase
			Clauses per AS–unit	1.50	2.09	-4.93	***.001	increase
	Lexical complexity	Lexical variety	Guiraud's Index	4.79	4.72	0.23	0.41	no difference
		Lexical sophistication	Lexical_beginner	0.66	0.55	4.39	***.001	decrease
			Lexical_intermediate	0.18	0.20	-1.29	0.114	no difference
			Lexical_advanced	0.16	0.25	-7.67	***.001	increase
Accuracy	Lexical accuracy		Ratio of error-free lexical items	0.89	0.86	2.83	**.01	decrease
Fluency	Speed fluency		SR	91.33	113.20	-4.32	***.001	increase
			MLR	3.27	4.91	-4.34	***.001	increase
	Breakdown fluency		ALFP	0.49	0.56	-2.16	**.03	increase
			ALSP	0.89	0.81	1.08	0.155	no difference
			FP100	13.23	11.35	1.14	0.142	no difference

		SP100	32.00	21.69	5.53	***.001	decrease
	Repair fluency	RR100	3.13	2.79	0.66	0.262	no difference

Note. *p < .05; ** p < .01; *** p < .001

Speech rate (SR), mean length of runs (MLR), the average length of filled pause (ALFP), the average length of silent pause (ALSP), the number of filled pauses per 100 syllables (FP100), the number of silent pauses per 100 syllables (SP100), the number of repairs and repetitions (RR100).

Regarding the effect of the SA period on the oral performance of participants measured by the 14 CAF measures, the changes in performance from S1 to S2 can provide a picture of the correlations of CAF measures (see Table 14). Firstly in terms of complexity and the two syntactic complexity measures, the length of AS-units showed a statistically significant increase ($t=-6.034$, $p<0.001$) as did the subordination of AS-units ($t=-9.432$, $p<0.001$). For lexical complexity, Guiraud's Index, which was used to measure lexical diversity, did not show any statistically significant difference ($t=-9.432$, $p=0.41$). For lexical sophistication, beginner-level words decreased significantly ($t=4.387$, $p<0.001$), advanced-level words increased significantly ($t=-7.665$, $p<0.001$), while there was a non-significant change in intermediate-level words ($t=-1.291$, $p=0.114$). The only measure of accuracy, lexical accuracy, showed a statistically significant decrease ($t=2.827$, $p=0.01$). Within the fluency domain, in terms of speed fluency measures, SR showed a statistically significant increase ($t=-4.318$, $p<0.001$). Similarly, MLR showed a statistically significant increase ($t=-4.338$, $p<0.001$). For breakdown fluency, SP100 showed a statistically significant decrease ($t=5.25$, $p<0.001$). Meanwhile, ALFP displayed a statistically significant increase ($t=-2.157$, $p=0.03$). The other two breakdown fluency indicators both revealed no significant differences - ALSP ($t=1.078$, $p=0.155$), and FP100 ($t=1.137$, $p=0.142$). The repair fluency measure, RR100, did not show a statistically significant change ($t=0.662$, $p=0.262$).

4.2 FI at home maintenance results (S2 - S3)

A paired-samples t-test was also conducted to compare the correlations between CAF in the FI at home context. The first session after the SA (S2), and the learners' final attainment level (S3) were analysed. The 14 CAF measures related to the FI at home maintenance factor are presented below. The results (See Table 15) showed that at S2 (one month after the SA with formal instruction) compared to S3 (six months after the SA with formal instruction).

Table 15. FI at home effects on oral performance (S2 vs. S3)

Constructs			Indicators	Mean		t	p	Statistics
				S2	S3			
Complexity	Syntactic complexity		Syllables per AS –unit	34.04	24.27	5.70	***.001	decrease
			Clauses per AS–unit	2.09	2.12	-0.20	0.422	no difference
	Lexical complexity	Lexical variety	Guiraud's Index	4.72	8.11	-6.68	***.001	increase
		Lexical sophistication	Lexical_beginner	0.55	0.67	-5.20	***.001	increase
			Lexical_intermediate	0.20	0.13	3.48	**.003	decrease
			Lexical_advanced	0.25	0.21	2.68	**.013	decrease
Accuracy	Lexical accuracy		Ratio of error-free lexical items	0.86	0.86	-0.08	0.471	no difference
Fluency	Speed fluency		SR	113.20	115.13	-0.31	0.382	no difference
			MLR	4.91	4.12	3.16	**.006	decrease
	Breakdown fluency		ALFP	0.56	0.49	3.62	**.003	decrease
			ALSP	0.81	0.81	0.04	0.486	no difference
			FP100	11.35	11.26	0.09	0.465	no difference
			SP100	21.69	25.85	-3.30	**.005	increase
	Repair fluency		RR100	2.79	2.33	1.61	0.071	no difference

Note. *p < .05; ** p < .01; *** p < .001

Speech rate (SR), mean length of runs (MLR), the average length of filled pause (ALFP), the average length of silent pause (ALSP), the number of filled pauses per 100 syllables (FP100), the number of silent pauses per 100 syllables (SP100), the number of repairs and repetitions (RR100).

Regarding the effects of formal instruction at home on the oral performance of the participants measured by the 14 CAF measures, changes in performance between S2 and S3 can provide a picture of the correlations between the CAF measures. Firstly, in terms of complexity and the two syntactic complexity measures, the length of AS-units showed a statistically significant decrease (t=5.70, p<0.001), whereas, the subordination of AS-units revealed no significant differences (t=-0.202, p=0.422). For lexical complexity, Guiraud's Index, which was used to investigate lexical diversity, showed a statistically significant increase (t=-6.68, p<0.001). For lexical sophistication, beginner level words showed a statistically significant increase (t=-5.198, p<0.001), advanced level words decreased significantly (t=2.682, p=0. 013), as did intermediate level words (t=3.484, p=0.003). The only measure of accuracy, lexical accuracy, showed no significant change (t=-0.075, p=0.471). Within the fluency domain, in terms of speed fluency measures, SR showed no statistical change (t=-0.308, p=0.382), whereas, MLR showed a statistically significant decrease (t=3.16, p=0.006). For breakdown fluency, SP100 showed a significant increase (t=-3.30, p=0.005). However, ALFP saw a significant decrease

(t=3. 62, p=0.003). The other two breakdown fluency indicators both revealed no significant differences - ALSP (t=0.037, p=0.486), and FP100 (t=0.091, p=0.05). The repair fluency measure, RR100, did not show a statistically significant change (t=1.608, p=0.071).

4.3 Correlations between CAF constructs related to SA effects (S1-S2)

This section outlines the results of the correlations between subconstructs within complexity and fluency as well as the correlations between CAF constructs related to the effect of SA (S1-S2). The results can be explored by reviewing the data presented in Table 14.

As shown by the data, post-SA (S2), in general, within the complexity domain, the two syntactic complexity measures, complexity via length and subordination, had significantly improved compared to pre-SA (S1), suggesting a strong joint improvement after the SA. Meanwhile, for the indicators of lexical sophistication, at S2, advanced-level lexical items were significantly higher than S1 with a significant decrease in beginner level words, revealing significant growth in lexical sophistication. However, intermediate level words showed no significant difference between pre- and post-SA. In terms of Guiraud's Index, used to measure lexical diversity, no significant difference was observed between S1 and S2. Therefore, the results support a weak trade-off effect between subdomains of lexical complexity, in particular, lexical variety and syntactic complexity. Within the domain of lexical complexity, a trade-off effect was observed between lexical diversity and lexical sophistication.

A similar picture emerged for the fluency measures. Within the fluency domain, in general, at post-SA (S2), the results demonstrated a trade-off effect between speed and breakdown and repairs. The two speed fluency measures used in this study (SR and MLR) were both significantly higher in S2 than S1, suggesting a connected improvement. However, breakdown fluency measures did not yield a unified trend. Specifically, while ALFP at S2 was significantly higher than at S1, suggesting decreasing fluency, SP100, in contrast, at S2 was significantly lower than at S1, revealing increasing fluency. In this regard, the trade-off effect was evident within breakdown fluency. The other three breakdown fluency measures (ALSP, FP100, RR100) at S2 were not statistically different from S1. For repair fluency, no statistical difference was observed between S1 and S2. Broadly speaking, speed fluency

measures gained great improvement, but repair fluency measures did not. Breakdown fluency measures showed moderate improvement.

Concerning the relationship between complexity, accuracy, and fluency, the analysis does not provide a unified picture, suggesting a trade-off effect at post-SA (S2). The results partly support Skehan's prediction (1996) of within meaning tension between accuracy (control) and complexity (risk-taking) as well as a meaning-form tension between accuracy (form) and fluency (meaning). The majority of complexity measures, both syntactic complexity via length and subordination and lexical sophistication showed a joint improvement, whereas lexical variety, as measured by Guiraud's Index, did not change statistically. Within fluency, speed fluency measures (SR and MLR) revealed significant improvement. In contrast, within breakdown fluency, only ALFP and SP100 revealed a significant change. The other breakdown fluency measures (ALSP, FP100), as well as repair fluency, showed no difference. Lexical accuracy was observed to be significantly lower post-SA than pre-SA. This mixed picture will be discussed in terms of the trade-off effect that results from processing capacity limitations and task design.

4.4 Correlations between CAF related to FI at home maintenance (S2-S3)

To address the FI at home maintenance factor on the relationship between subconstructs within fluency and complexity as well as between CAF measures, comparisons were made between S2 (post-SA) and S3 (the final attainment level). The resultant data are shown in Table 15.

Generalising the data, at the final attainment level (S3), the results partially support the trade-off effect among complexity measures. Within the domain of lexical complexity, a trade-off effect was noted between lexical diversity and lexical sophistication. This was suggested by a significant increase of lexical diversity (Guiraud's Index) and a significant decrease of lexical sophistication indicators. Beginner level words were observed to be significantly higher than at S2, meanwhile, there was a significant decrease in both intermediate and advanced level lexical items. A trade-off effect was also revealed between lexical diversity (Guiraud's Index) and syntactic complexity, suggested by a significant decrease in complexity via length and the number of syllables per AS-unit. However, in terms of

syntactic complexity, the number of clauses per AS-unit showed no statistical difference between S2 and S3.

Similarly, the trade-off effect has been partially supported by fluency measures at S3, but with some mixed findings. Specifically, MLR, as a speed fluency measure, saw a significant decrease compared to S2; breakdown fluency measures did not yield a unified picture; ALFP was significantly lower than S2; whereas a significant increase in SP100 was revealed. The other fluency measures, SR to measure speed fluency, ALSP and FP100 to investigate breakdown fluency, as well as RR100 to measure repair fluency, did not show a statistical difference compared to S2.

In terms of the relationship between complexity, accuracy, and fluency, the data also do not reveal a unified picture, suggesting the trade-off effect at the final attainment level (S3). The results partly support Skehan's prediction (1996) of a meaning-form tension between complexity (form) and fluency (meaning). Regarding complexity measures, both syntactic complexity via length and lexical sophistication showed a joint decrease, while one measure of breakdown fluency, ALFP revealed a significant increase. However, the other fluency measures (SR, ALFP, FP100, and RR100) did not show a statistical difference between S2 and S3 except for a significant decrease in MLR to measure speed fluency.

4.5 Individual performance during the 28 month period (S1-S3)

The pre-SA proficiency levels of the subjects were either HSK3 (n=7) or HSK2 (n=3) (see Appendix D). During the 10-month SA, eight of the ten participants were placed in Level A language group in the first term. They had progressed to Level B in the second term. Similarly, the other two participants out of the ten participants were assigned to the Level B class in term 1, and in term 2, they advanced to Level C. Within the 28-month period (including 10-month SA), to give a more comprehensive speaking developmental pattern for the participants, each level will provide one participant's oral performance on three oral tests (S1, S2, S3). The oral performances of participant 9 (HSK3 at pre-SA, lower level during SA) and participant 4 (HSK2 at pre-SA, higher level during SA), who serve as representatives of the two levels, are provided as follows.

Regarding the effect of the SA period on the oral performance of participant 4, a clear and steady increase is generally exhibited by the 14 CAF measures. Firstly in terms of complexity and the two syntactic complexity measures, the length of AS-units showed a significant increase from S1(16.00) to S2(47.50); as did the subordination of AS-units (S1(1.33), S2(2.75)). For lexical complexity, Guiraud's Index, which was used to measure lexical diversity, showed a decrease from S1 (5.64) to S2 (3.95). For lexical sophistication, advanced-level words increased significantly (S1 (0.18), S2 (0.29)), while there was a non-significant change in intermediate-level words (S1 (0.20), S2 (0.18)) and in beginner-level words (S1 (0.62), S2 (0.53)). The only measure of accuracy, lexical accuracy, showed a slight decrease (S1 (0.90), S2 (0.86)). Within the fluency domain, in terms of speed fluency measures, SR showed a statistically significant increase (S1 (106.57), S2 (155.38)). Similarly, MLR showed a significant increase (S1 (3.69), S2 (8.14)). For breakdown fluency, silence-related indicators showed a similar pattern. SP100 showed a significant decrease ((27.08), S2 (12.28)); as did ALSP (S1 (0.95), S2 (0.53)). Meanwhile, ALFP displayed an increase ((0.28), S2 (0.49)); as did FP100 (S1 (0.52), S2 (4.21)). The repair fluency measure, RR100, did not show a moderate increase ((1.04), S2 (2.28)). In conclusion, participant 4 made significant progress in terms of syntactic complexity. Similar to this, there was a large rise in advanced-level lexical sophistication. However, lexical variety significantly decreased. Within Fluency, speed fluency benefited great after SA. Within the subdomain of breakdown fluency, silence-related indicators (ALSP and SP100) decreased significantly showing increasing fluency. In contrast, the number and length filled pauses increased significantly showing decreasing fluency.

Regarding the effects of formal instruction at home on the oral performance of participant 4 measured by the 14 CAF measures, changes in performance between S2 and S3 did not provide a clear pattern among the CAF measures. Firstly, in terms of complexity and the two syntactic complexity measures, the length of AS-units showed a significant decrease (S3 (29.98)), whereas, the subordination of AS-units revealed no significant differences (S3 (2.56)). For lexical complexity, Guiraud's Index, which was used to investigate lexical diversity, showed a significant increase (S3 (10.59)). For lexical sophistication, beginner level words showed a slight increase (S3 (0.62)), advanced level words decreased significantly (S3 (0.20)), while no change was shown in intermediate-level words (S3

(0.18)). The only measure of accuracy, lexical accuracy, showed no significant change (S3 (0.86)). Within the fluency domain, in terms of speed fluency measures, SR showed a significant increase (S3 (194.82)), whereas, MLR showed a slight decrease (S3 (7.95)). For breakdown fluency, SP100 did not show any change (S3 (12.58)). However, ALFP saw a moderate decrease (S3 (0.37)). The other two breakdown fluency indicators both revealed no significant differences - ALSP (S3 (0.53)), and FP100 (S3 (2.78). The repair fluency measure, RR100, showed a slight decrease (S3 (2.02)). When participant 4 returned to the FI context for six months, in general, the syntactic complexity decreased slightly as well as advanced-level words revealing lexical sophistication. In contrast, lexical diversity revealed significant improvement. Within fluency, speed fluency kept increasing in a moderate rate. The change occurred on the silence (ALSP and SP100) and filled pauses (ALFP and FP100) after SA were still observable at this stage.

Regarding the effect of the SA period on the oral performance of participant 9 did not exhibit a clear and steady increase measured by the 14 CAF measures as participant 4 did. Firstly in terms of complexity and the two syntactic complexity measures, the length of AS-units did not show a significant increase from S1 (20.08) to S2 (25.33); as did the subordination of AS-units (S1 (1.46), S2 (1.83)). For lexical complexity, Guiraud's Index, which was used to measure lexical diversity, showed a slight decrease from S1 (4.53) to S2 (4.34). For lexical sophistication, advanced-level words showed slight increase (S1 (0.18), S2 (0.25)), while there was a slight decrease in intermediate-level words (S1 (0.18), S2 (0.15)) and in beginner-level words (S1 (0.65), S2 (0.60)). The only measure of accuracy, lexical accuracy, showed a slight increase (S1 (0.86), S2 (0.88)). Within the fluency domain, in terms of speed fluency measures, SR showed a slight increase (S1 (112.36), S2 (127.93)). Similarly, MLR showed a moderate increase (S1 (3.84), S2 (5.63)). For breakdown fluency, silence-related indicators showed a similar pattern. SP100 showed a significant decrease ((26.05), S2 (17.76)); as did ALSP (S1 (0.85), S2 (0.69)). Meanwhile, ALFP displayed a slight decrease ((0.45), S2 (0.42)); and FP100 showed a moderate decrease (S1 (14.56), S2 (9.21)).The repair fluency measure, RR100, exhibited a moderate increase ((1.92), S2 (2.63)). In summary, after SA, participant 9 only achieved slight gains in terms of syntactic complexity. This slight increase also took place advanced-level lexicons measuring lexical sophistication as well as in lexical diversity. Within fluency, speed fluency achieved slight gains; within breakdown fluency, only the number of pauses (FP10，SP100) decreased moderately. The

other indicators of breakdown fluency (ALFP, ALSP) did not make great change; as did in repair fluency.

Regarding the effects of formal instruction at home on the oral performance of participant 9 measured by the 14 CAF measures, changes in performance between S2 and S3 did not show significant changes. Firstly, in terms of the two syntactic complexity measures, the length of AS-units showed a slight decrease (S3 (21.57)), whereas, the subordination of AS-units revealed a slight increase (S3 (2.04)). For lexical complexity, Guiraud's Index, which was used to investigate lexical diversity, showed a significant increase (S3 (8.07)). For lexical sophistication, beginner level words showed a slight increase (S3 (0.65)), advanced level words did not show significant change (S3 (0.26)), while intermediate-level words exhibited a moderate decrease (S3 (0.09)). The only measure of accuracy, lexical accuracy, showed a slight decrease (S3 (0.82)). Within the fluency domain, in terms of speed fluency measures, SR showed a slight decrease (S3 (118.68)), as did MLR (S3 (4.24)). For breakdown fluency, SP100 showed a moderate increase (S3 (23.59)).However, ALFP saw a slight increase (S3 (0.46)). The other two breakdown fluency indicators both revealed a similar pattern – ALSP showed a moderate increase (S3 (0.83)), and FP100 (S3 (9.88)) showed a slight increase. The repair fluency measure, RR100, showed a significant decrease (S3 (1.01)). When participant 9 returned to the FI context for six months, there was a minor decrease in terms of syntactic complexity. No significant change was observed in lexical complexity measures. In terms of fluency indicators, slight decreases also were observed in speed fluency as well as in repair fluency. Slight increases were found in breakdown fluency indicators.

Chapter 5: Discussion

This semi-longitudinal study used a corpus dataset to analyse the development of complexity, accuracy, and fluency (CAF) in oral performance. Specifically, this study examined the development of the oral language performance of 10 English-speaking adult learners of Chinese across 28 months (including a 10-month SA experience).

The development of the CAF constructs was examined pre- and post-SA, during the FI at home context. Explanations concerning the first question on the two learning contexts (SA and FI at home) on the oral performance, investigated by the CAF measures, will be discussed in sections 5.1 and 5.2. Interpretations concerning the second question on the effects of learning context (SA and FI at home) on cognitive allocation within and between the CAF measures will be presented in sections 5.3 and 5.4.

5.1 SA effects on oral performance (S1-S2)

Oral development measured by the 14 CAF measures was reported to explore the SA effects. The results are discussed below in terms of complexity, accuracy and fluency development during the pre- and post-SA periods.

5.1.1 Development of complexity

Syntactic complexity

Concerning syntactic complexity, the AS-unit length increased substantially after the 10-month SA period. The significant improvement after SA in this study is in line with previous studies (Jensen & Howard, 2014; Mora & Valls-Ferrer, 2012; Valls-Ferrer, 2010). This confirms that words per AS-unit as a measure of overall complexity in oral production clearly benefited from the SA period (Jensen & Howard, 2014; Juan-Garau & Pérez-Vidal, 2007; Mora & Valls-Ferrer, 2012).

The other indicator of syntactic complexity used in this research, the number of clauses per AS-unit to measure complexity via subordination, was also shown to benefit from the SA period. This is consistent with previous studies (Pérez-Vidal & Juan-Garau, 2011; Llanes & Muñoz, 2013) showing that the subordination of complexity achieved significant gains during SA. Additionally, the number of clauses per AS-unit has been proven to increase together with the proficiency level of L2 learners (Kuiken & Vedder, 2012). In this study, this also applied when learners were at lower intermediate level after the 10-month SA sojourn compared with when they were at upper beginner level pre-SA.

Lexical complexity

Concerning lexical complexity, lexical diversity as a subdomain measured by Guiraud's Index, did not reveal improvement during the SA period. This limited gain is in line with previous studies (Pérez-Vidal & Juan-Garau, 2011; Pérez-Vidal et al., 2012; Wright, 2018, 2020) and further demonstrates that Guiraud's Index does not exhibit a significant increase after a 3-month to 10-month SA period. This finding further supports the notion that the development of lexical diversity is constrained by learners' proficiency levels. Specifically, advanced level learners outperform those at the beginner and intermediate levels and there is no significant improvement when learners are at the beginner and intermediate levels (Chen, 2015a; Ding & Xiao, 2016; Ye, 2015). Thus, in this study, lexical diversity did not show great gains pre-SA when the learners were at the upper beginner level and post-SA when they were at the lower intermediate level.

Referring to Levelt's speaking model, Skehan (2009a) hypothesised that lexical diversity portrays the processing of the formulator, which accepts the preverbal message, and which then engages in lemma selection and consequent syntax-building processes. Lexical diversity is closely related to using less demanding words effectively when attention is available. It's as if a restricted number of words prime each other, and once available, they may be incorporated more easily, avoiding the need for more extensive, and disruptive lexical retrieval. This can explain why after the SA, lexical diversity did not reach a peak like the majority of the other indicators. Lexical diversity is more clearly a formulator factor, which is concerned with online (in operation), moment-by-moment decisions during speaking (Levelt, 1989; Skehan, 2009), and prioritises less-demanding words. This can ease the ten-

sion between the conceptualiser (whose output is the preverbal message, and which is essentially concerned with the conceptual content and packaging of what will be said) and the formulator, in particular, for non-native speakers at lower proficiency levels.

For lexical sophistication, after the 10-month SA, the beginner level words decreased significantly and there was a significant increase in advanced-level words. Constrained by being at the lower intermediate proficiency level post-SA, the participants' intermediate-level words did not show a statistical difference after the SA period. Meanwhile, the significant increase of advanced-level words after SA is largely attributable to the calculation method used in this research, which included the words in the HSK 5-6 level bracket and the words not included in the HSK. In particular, the study's data show that the participants acquired a large number of words beyond the HSK system during the SA. The significant increase of advanced-level words including non-HSK words, categorised as advanced level words, can be attributed to two key reasons. Firstly, during the SA sojourn, the participants accessed various types of input and more sophisticated words in their daily lives in the naturalistic environment, and their lexical repertoire consequently expanded. Indeed, it has been proven that vocabulary/lexical development can improve significantly because of an increased lexical repertoire during SA (Collentine, 2004; Milton & Meara, 1995; Jensen & Howard, 2014). Secondly, the textbooks that the learners used in the classroom setting in the college in China did not follow the HSK glossary. Therefore, during the formal instruction during SA, learners acquired a significant amount of non-HSK words. The HSK is widely used as a benchmark to assess lexical sophistication in existing Chinese studies, including this study. Therefore, advanced level words, including those words beyond the HSK system, increased significantly.

5.1.2 Development of accuracy

Accuracy was only measured with one sub-construct, namely lexical accuracy, and it saw a statistically significant decrease from pre- to post-SA. This is also in line with previous studies which showed that significant gains in oral accuracy after SA are not guaranteed (e.g., Mora & Valls-Ferrer, 2012 (3-month SA); Segalowitz et al., 2004 (a semester SA); Serrano, Llanes & Tragant, 2011(2-month SA); Valls-Ferrer & Mora, 2014 (3-month SA)), which very likely relates to the length of the SA period (2 to 3 months) in those studies. In this study, the learners did not make great gains in

accuracy after a 10-month SA period, as accuracy was constrained by participants' proficiency levels during pre- and post-SA, that is, upper beginner to lower intermediate levels respectively. The results of previous research (i.e., Chen, 2015; Ye, 2015; Zhai & Feng, 2014) reveal that oral accuracy, in particular, lexical accuracy develops when learners are at the advanced level. No significant improvement can be expected when learners are at the beginner and intermediate levels. In this sense, constrained by the proficiency level of the participants in the research, no significant improvement could be expected when the participants were between upper beginner and lower intermediate levels.

5.1.3 Development of fluency

For the fluency measures, the results showed a speed fluency improvement (SR and MLR) as well as breakdown fluency (SP100). Both speech rate (SR) and mean length of runs (MLR) saw a significant increase after SA, meanwhile, the number of silent pauses per 100 syllables (SP100) decreased significantly after the 10-month SA. In other words, oral fluency showed speed improvement with fewer silent pauses. This is exactly in line with previous studies which assert that SA benefits oral fluency, in particular, speed fluency (e.g., DeKeyser, 2014; Freed et al., 2004). After SA, learners are very likely to speak faster and they also produce longer speech runs and their speech becomes less hesitant, containing fewer pauses (Mora & Valls-Ferrer, 2012), in particular, silent pauses. The limited gains demonstrated in fluency breakdown and repair (i.e., ALSP, FP100, and RR100) were consistent with previous findings in that the participants did not show a significant decrease in dysfluency (i.e., filled pauses, mean length of pause, repairs, and repetitions) after the 10-month SA (Wright & Zhang, 2014; Wright, 2020). The results are in line with previous research (Collentine & Freed, 2004; Mora & Valls-Ferrer, 2012; Valls-Ferrer & Mora, 2014) showing that learners are very likely to speak faster and that they also produce longer speech runs and their speech becomes less hesitant, containing fewer pauses. It can be concluded that after a 10-month SA period, fluency achieved significant gains and showed higher speed and longer speech runs with fewer silent pauses. Small and non-significant reductions were also found in the disfluency (total number of filled pauses) and repairs in the oral performance of English-speaking learners of Chinese.

Among the fluency indicators, SR, MLR, and SP100 were found to simultaneously improve after the 10-month SA. Referring to Levelt's speaking model, speech rate encompasses the working

of the whole model, the conceptualiser, formulator and articulator, but that changes occur primarily in the articulator (Towell et al., 1996). The significant increase of speech rate after SA suggests that the entire speech production process had been restructured, and that proceduralisation had occurred in the articulator. As an indicator of automatisation in language performance (Skehan, 2009), mean length of run (MLR) has a conceptual connection with automatic speech production processing (Kahng, 2014), and has been suggested to be strongly associated with L2 fluency (e.g., Kormos & Denes, 2004; O'Brien et al., 2007; Segalowitz & Freed, 2004). The increase in MLR is mainly attributable to the proceduralisation of different kinds of knowledge, including procedural knowledge of syntax and of lexical phrases. This might suggest that increased proceduralisation in the formulator of Levelt's speech model indicates greater time for planning each utterance and it should therefore become evident with longer pauses or a greater number of pauses (Towell et al., 1996). In this study, the length of the filled pause (ALFP) was significantly higher after SA.

MLR is highly related to the application of prefabricated language units and formulaic language, which are considered to facilitate L2 oral fluency (Boers et al., 2006). This confirms the previous assertion (Levelt, 1989) that low-fluency speakers tend to use hesitations and non-lexical fillers to provide themselves with a longer period for processing. However, as discussed in other studies, for highly fluent speakers, whole clauses and chunks of words are often used to save time for processing, which leads to the extension of the length of runs (MLR) between pauses reflecting their increased fluency (Wood, 2006). This was further confirmed in this study by the qualitative analysis of participants' speech samples at S2. For example, chunks of words such as "个人简历 ([gèrén jiǎnlì], personal resume)", "越来越好 ([yuèláiyuèhǎo], become better and better)", "更上一层楼 ([gèng shàng yī céng lóu], strive for further improvement)", "这山望着那山高 ([zhè shān wàng zhe nà shān gāo], the grass is always greener on the other side of the fence)", "万事开头难 ([wàn shì kāi tóu nán], the first step is always the most difficult)" appeared in the speech samples.

Apart from SR and MLR, another indicator of fluency that achieved significant improvement after the 10-month SA exposure to the target language context was SP100. The significant decrease of SP100 indicated improved fluency after the SA experience. This supports the hypothesis that silent pauses are a salient feature that determine speakers' fluency levels and contribute to judgments of nonfluency (Riggenhach, 1991).

Interestingly, ALFP (the average length of the filled pause), as a breakdown fluency measure, saw a significant increase after SA. This can be interpreted from two perspectives. Firstly, the significant increase of ALFP after the 10-month SA, suggests decreasing fluency. This supports previous findings (Chen, 2012, 2015; Ye, 2015) that the average length of filled pauses does not see significant improvement when learners are at the beginner and intermediate levels. Constrained by the learners' proficiency level, i.e., lower intermediate level after SA, a significant regress was reasonable. Secondly, a significant increase of ALFP after SA might not be an indication of decreasing fluency. For instance, filled pauses can be used as a successful strategy for holding one's turn (Wright, 2020), and therefore may not be a clear indication of a lack of utterance fluidity (de Jong, 2016; Tavakoli, 2011).

Overall, with the exception of SP100 and ALFP, the dysfluency subconstructs, such as FP100 and repairs and repetitions, saw no statistical improvement from the SA experience. Levelt (1989) stated that speakers self-monitor their speech during the articulation stage with regards to any aspect of speech, such as content, syntax, choice of words, and phonological forms, and these are aspects which can be attended to simultaneously by native speakers. However, for L2 learners these processes are not yet automatised, which lead L2 speech to be more problematic (Kormos, 2006, 2011; Segalowitz, 2010). Gaps in linguistic knowledge or slow processing in accessing knowledge can impede the construction of accurate or sophisticated grammar and lexical items, resulting in reduced speech speed, hesitations, filled pauses, and repairs (Segalowitz, 2010, 2016; Skehan, 2003; Tavakoli, 2011). Therefore, it is very likely that certain errors or dysfluency features can be attended to, while others might be ignored. Also, it has been suggested that both repairs and pauses act as monitoring processes during speech production, where the former is an overt-monitoring process and the latter is a covert-monitoring process (Kormos, 2006; Tavakoli et al., 2016).

5.1.4 Summary of CAF development

In conclusion, in this study, the benefits of the SA period mainly appeared in terms of significant improvements in the constructs of complexity and fluency. Complexity measures, in particular, syntactic complexity (length and subordination) as well as lexical sophistication saw significant growth. In the domain of fluency, speed fluency (SR and MLR), as well as breakdown fluency (SP100),

improved significantly. In contrast, repair fluency and lexical variety, saw no significant differences between pre- and post-SA. However, accuracy showed a statistical decrease between pre- and post-SA. In particular, filled pauses (i.e., ALFP), as a successful strategy for holding one's turn are used in utterance. The significant increase of filled pauses may not indicate a clear decrease in fluency. The results imply that the improvement in oral fluency, with a significant improvement in speaking speed, could be associated with more sophisticated words and complex units at the sentence and clause levels with longer filled pauses, but with no significant decrease in dysfluency except for fewer silent pauses, all occurring with a more limited vocabulary and more lexical errors.

These results are in line with previous studies and suggest that a period of SA favours learners' oral fluency (e.g., Freed et al., 2004; Du, 2013; Mora & Valls-Ferrer, 2012; Trenchs-Parera, 2009; Valls-Ferrer, 2010; Wright & Zhang, 2014) as well as syntactic complexity gains (Jensen & Howard, 2014; Juan-Garau & Pérez-Vidal, 2007; Mora & Valls-Ferrer 2012; Pérez-Vidal & Juan-Garau, 2011). When learners produce speech at a higher speed there are longer speech runs and fewer pauses. However, this is not at the expense of learners producing complex language, as their vocabulary becomes more sophisticated and their syntax becomes more complex. Yet, the significant gains in complexity and limited gains in fluency are at the cost of accuracy, leading to more lexical errors.

5.2 The effects of FI at home maintenance on oral performance (S2-S3)

As shown by the results (see Section 4.2), each domain of CAF was examined individually to further explore how oral development was impacted by the FI at home context.

5.2.1 Development of complexity

Syntactic complexity

The AS-unit length decreased significantly from S2 to S3 when the learners had returned to the FI context for six months. Considering the SA effects, the length of complexity exhibited significant improvement from S1 to S2, and then returned to the pre-SA level six months after the learners had

returned to the FI context at S3. This may be because the FI context is very likely to exert a negative effect on the length of sentences (Juan-Garau & Pérez-Vidal, 2007). Also, this study's findings concerning the increase of overall complexity after SA and the decrease of overall complexity during the FI context reveal an unstable developmental pattern when learners are at the beginner and intermediate levels. This is consistent with Ye (2015), who concluded that the length of sentences develops at a slow rate, and only improves significantly when learners are at an advanced level. In this research, although the AS-unit length reached a peak at S2 because of the benefit of SA, it then returned to the pre-SA level as a result of being constrained by the students' intermediate proficiency level.

The number of sub-clauses per AS-unit saw no significant difference between S2 and S3, and showed a stable higher level six months after returning to the FI context. Therefore, it is suggested that the delayed effects of SA are still noticeable in terms of the syntactic complexity improvements of intermediate level learners for at least one academic year after returning to the FI at-home context. This confirms that FI, as a form-focused context (DeKeyser, 2007c, 2014; Sanz, 2014; McManus et al., 2021), allows learners greater opportunities to focus on subordination (the number of clauses) as a result of focus-on-form practice in the classroom. During the FI at home context, learners have no immediate pressure to communicate compared to when they are in the SA context, which is a naturalistic meaning-focused context with communicative pressures (Pérez-Vidal, 2015). This is relevant as the oral samples in this study were elicited by rehearsed topic-prompted monologues collected during oral tests, which did not involve communicative pressure.

Lexical complexity

The two sub-domains of lexical complexity saw a contrasting picture at S3. Guiraud's Index, which measures lexical diversity, increased significantly at S3 compared to S2 (post-SA). This depicts an increasing trajectory of lexical diversity when the participants were not exposed to the target language environment. This is in line with the finding of previous research (Juan-Garau & Pérez-Vidal, 2007; Segalowitz and Freed 2004; Pizziconi, 2017; Wu, 2017) that FI at home seems to benefit vocabulary learning, since as the amount of time spent learning Chinese increases, more and more diverse lexical items are acquired. However, the statistical increase in lexical diversity contradicted the finding of previous research (Chen, 2015a; Ye, 2015; Ding & Xiao, 2016) which claimed that no significant

improvement is evident among learners at the beginner and intermediate levels (Chen, 2015a; Ye, 2015; Ding & Xiao, 2016; Ye, 2015). The results of this study revealed that lexical diversity might not always be constrained by learners' proficiency levels and that other confounding factors (i.e., FI at home context) triggered its improvement and outperformed the effects of learners' proficiency level on lexical diversity.

Looking at lexical sophistication at S3, namely six months after the participants returned to the FI context with formal instruction, beginner level words increased significantly, returning to the same level as pre-SA (S1). Meanwhile, intermediate and advanced level words decreased significantly. The decrease of the intermediate and advanced level lexical items in this study was very likely associated with learners' retrieval of more varied lexical items during the FI at home context, and was constrained by the participants' proficiency levels (Ding & Xiao, 2014), i.e., at intermediate level. Referring to the Levelt model (1989), Skehan (2009a) hypothesised that lexical sophistication relates more to the conceptualiser stage and to the nature of preverbal message implications for lemma retrieval, whereas lexical diversity relates to the formulator stage, which is concerned with online, moment-by-moment decisions during speaking. For native speakers, the processing of the three stages occur in parallel and are incremental (De Bot, 2000). For non-native speakers, mental lexicons are not as rich and well-organised as they are for native speakers, especially when learners are at a lower proficiency level (i.e., intermediate level). Therefore, the conceptualiser-formulator connection is more problematic for lower proficiency learners, who cannot process in formation and in parallel as native speakers do. Therefore, a lower burden in the preverbal message at conceptualiser stage (lower lexical sophistication) allows L2 learners to retrieve more unusual lexical items (higher lexical diversity) at the formulator stage (Skehan, 2009a).

5.2.2 Development of accuracy

Lexical accuracy maintained a steady level at S3, and was the same as S2 (86%), that is, lexical accuracy remained at a very high level six months after the learners returned to the FI context. This is consistent with the finding of Juan-Garau (2014) who revealed that the effects of SA are still noticeable more than one academic year later with respect to oral accuracy improvement. The fact that no significant improvement in lexical accuracy was observed in this study was largely a result of the constraint

imposed by the participants' proficiency level, which was upper intermediate at this stage (S3). The results of pervious research (Chen, 2015; Ye, 2015; Zhai & Feng, 2014) found that oral lexical accuracy develops significantly when learners of Chinese are at the advanced level. Thus, no significant growth can be expected when learners are at the beginner and intermediate levels.

5.2.3 Development of fluency

For the seven fluency indicators used, a mixed set of results emerged. For instance, the speed fluency measure, speech rate (SR), showed no statistical change, suggesting a stable level had been maintained from S2. This implies that SR, as a reliable predictor of speed fluency (Kormos, 2006), preserves the effects of SA, which are still observable six months after learners have returned to the FI context. However, the other speed fluency measure, mean length of runs (MLR) showed a statistically significant decrease at S3.

Following the predictions of Towell (1996), when a greater time is used for planning each utterance, so it shows up as longer pauses or a greater number of pauses. An increased MLR suggests an increased proceduralisation in the formulator of Levelt's speech model. Thus, a decrease of MLR suggests a lack of proceduralisation of linguistic knowledge in the formulator of Levelt's speaking model. The decrease in MLR was attributable to less time being devoted to planning each utterance. A decrease in MLR might occur with shorter pauses or a fewer number of pauses. Thus it was very likely that the length of the filled pause (ALFP) would be significantly lower at this stage (S3).

The significant decrease of MLR observed in this study relates to SP100 being used to measure the silent pause frequency. Because MLR was obtained by the number of syllables divided by the number of silent pauses, which relates to the silent pause frequency (i.e., SP 100). At S3, when learners had returned to the FI context for six months, SP100 showed a significant increase revealing a significant decrease in their oral fluency gains in this aspect. This is in line with previous research (e.g., Juan-Garau & Pérez-Vidal, 2006, 2007; Pérez-Vidal & Juan-Garau, 2005) which revealed a backslide in oral fluency gains after learners had returned to the FI context for six months. Also, a significant increase of silent pauses indicated a slower speech planning process following the findings of previous

research (de Jong, 2016; Tavakoli, 2011). This suggests that silent pauses can indicate that learners use the pause time for speech planning.

However, the average length of the filled pause (ALFP), was found to be significantly lower after the learners had returned to the FI context for six months. In contrast to S2, this significant decrease of ALFP after the SA period might not suggest increasing fluency as it initially seems to. This is because the learners were very likely to revert to their pre-SA level after six months back home (e.g., Juan-Garau & Pérez-Vidal, 2006, 2007; Pérez-Vidal & Juan-Garau, 2005), and they did not produce more complex language with higher fluency as they benefited at S2 after SA (see Section 5.2.1). Therefore, filled pauses as a successful strategy for holding one's turn during an utterance (cf. Wright, 2020) were not necessary.

Except for the significant changes in MLR, SP100 and ALFP, other breakdown fluency indicators (ALSP, FP100) and repair fluency (RR100), remained stable between S2 and S3. These results revealed the same pattern for learners at post-SA (S2). In other words, they did not undergo any change during the whole experimental period, which suggests that no radical change was observed in terms of the effect of the learning contexts on learners' oral performance. These results are very likely to be associated with the learners' proficiency levels, revealing that these aspects of oral development produced by English-speaking learners of Chinese develop very late (Chen, 2012; 2015a). These dysfluency phenomena suggest that L2 learners self-monitor their speech during utterance. Self-monitoring involves checking the correctness and appropriateness of the produced output (Kormos, 2011). The results are consistent with the finding of previous studies (Mora & Valls-Ferrer, 2012; Valls-Ferrer, 2010), which suggest that the FI at home period has little impact on learners' fluency. In this study such an outcome was very likely because the length of the FI period (six months) was too short. In contrast to the SA period, the FI period did not provide learners with sufficient L2 practice to help them become more fluent, and, they were also constrained by their proficiency level (intermediate level). Moreover, the decrease in fluency during the FI period can be explained by the participants' lack of opportunities for oral production practice. As mentioned in the methodology (see Section 3.3.1), during the FI at home context, participants largely did not receive specific training on oral skills in a conventional classroom-setting. Instead, the focus was on traditional grammar teaching and writing skills.

5.2.4 Summary of CAF development

Overall, after six months back in the FI context after a 10-month SA sojourn, a statistically significant decrease was observed in the learners' overall complexity (length of AS-units) and lexical sophistication as well as in fluency (MLR and SP100). In contrast, a significant increase was observed in lexical diversity and there was no statistical difference in repair fluency and accuracy measures as well as complexity via subordination.

Learners' oral gains in complexity and fluency during the 10-month SA reverted to their previous levels six months after their return home. The results contribute to the finding that oral gains are expected to decrease six months after a return to the domestic setting without formal instruction (e.g., Juan-Garau & Pérez-Vidal, 2006, 2007; Pérez-Vidal & Juan-Garau, 2005). The analysis reveals that even with formal instruction, not only did overall complexity decrease significantly six months after returning to the domestic context, but that lexical sophistication as well as fluency (i.e., MLR and SP100) also decreased significantly. This is most likely due to the learners' lower intermediate proficiency level at the time of the study's post-SA. After their year abroad, lower-level proficiency may not be enough to keep their FI oral gains over the following year with formal instruction (cf. Tracy-Ventura et al., 2021). Furthermore, this outcome confirms the finding of Juan-Garau and Pérez-Vidal (2007) and shows that FI learners are very likely to concentrate on learning vocabulary and subordinating at the expense of complexity and accuracy. The FI period does not seem to provide learners with opportunities to practice their oral performance to achieve more gains in a similar way as the SA period, a finding which supports previous research (Mora & Valls-Ferrer, 2012). This finding very likely arose due to the length of the data collection period (six-month FI period) in this study, which was too short.

Moreover, learners are prone to achieve limited gains during the FI context (Freed, Segalowitz & Dewey, 2004). In contrast, the SA setting is clearly more beneficial than FI at home (LIanes & Serrano, 2017; Mora & Valls-Ferrer, 2012). Presumably, the SA context provides learners increased opportunities for meaningful L2 interaction compared to the limited access to authentic input and scarce opportunities for learners to engage in authentic conversations in a typical FI context (O'Donnell,

2004). Therefore, it can be concluded that there is an overall developmental pattern - fluency achieves larger gains during SA compared to FI at home.

5.3 Correlations between CAF sub-constructs related to SA effects (S1-S2)

This section firstly discusses the correlations between the subconstructs within complexity and fluency in the oral performance of English-speaking learners of Chinese in relation to the impact of the study abroad factor. This is followed by a discussion of the correlations between the CAF constructs. Finally, the correlations are analysed and interpreted.

5.3.1 Correlations between subconstructs within CAF after SA

Within the complexity domain, after the 10-month SA period, a strong joint improvement between lexical sophistication and syntactic complexity was clearly present, including a supportive relationship within syntactic complexity. Within the construct of syntactic complexity, connected growth was evident between word complexity (average sentence length in morphemes) and sentence complexity (average number of clauses per sentence). These two indicators have been proved to be connected and supportive (Spoelman & Verspoor, 2010; Vercellotti, 2012, 2017, 2019). The growth processes of word complexity and sentence complexity are compatible with each other (Spoelman & Verspoor, 2010). An increase in clauses per AS-unit increases the overall length of an AS-unit (Vercellotti, 2012). There was no evidence of a trade-off effect within the sub-constructs of syntactic complexity, which is consistent with previous research (Spoelman & Verspoor, 2010; Vercellotti, 2012).

Within the subconstruct of syntactic complexity, both the number of syllables per AS-unit, as an indication of overall or general complexity, and the number of clauses per AS-unit, as an indication of complexity via subordination, showed significant improvement after SA. These two commomly-used measures to investigate syntactic complexity at distinct linguistic levels (cf. Norris & Ortega, 2009) showed a joint increase. Each measure of syntactic complexity is useful for capturing language development. These results suggest that with increasing proficiency, more cognitive resources are available for complexifying language performance in these two aspects.

After the 10-month SA period, generalised from the results of the paired t-tests (S1 vs. S2), a weak trade-off effect was observed between lexical diversity and lexical sophistication. Lexical sophistication saw significant improvement, whereas lexical diversity had a non-significant difference compared to pre-SA. The development of lexical sophistication showed significant improvement after SA, especially advanced level words, including the non-HSK words that the learners acquired from their SA experience. However, the development of lexical diversity seemed to be related to the learners' proficiency level, supporting the findings of previous studies (Chen, 2015a; Ding & Xiao, 2016; Ye, 2015), that no significant improvement in lexical diversity can be expected when learners are at the beginner and intermediate levels. Moreover, the results can be understood in relation to Levelt's (1989) model of speaking. Lexical sophistication relates more to the conceptualiser stage of the model, whose output is the preverbal message. In contrast, lexical diversity is more closely related to the formulator stage, which accepts the preverbal message, and which then engages in processes of lemma selection and consequent syntax-building processes (Skehan, 2009a). For non-native speakers at lower proficiency levels, i.e., lower intermediate level after SA in this study, higher lexical sophistication (increased advanced level words) are more demanding in the conceptualiser stage. This leads to negative implications in the formulator stage in terms of the retrieval of unusual lexical items. Thus, lexical diversity, as an indication of using unusual words, seems to have been impaired. Consequently, less demanding words were very likely to be produced more effectively.

Within the subconstruct of lexical complexity, the analysis revealed a negative correlation between lexical items at the beginner level and advanced level, suggested by the significant increase in advanced level words and the significant decrease in beginner level words. It stands to reason that learners produce more intermediate and advanced level words (higher lexical sophistication) while using fewer beginner level words (higher lexical sophistication), which indicates improvements in both levels of words. This result can be attributed to the fact that beginner, intermediate and advanced level lexical items are related to each other in terms of how they are calculated. Constrained by the learners' proficiency level, beginner level words dominated the lexical items. Also, because of the benefit of the SA experience, advanced level words, especially the words that learners acquired from their SA experience, outnumbered intermediate level words (see the results concerning lexical complexity in Section 5.1). This shows that when learners utter more lexical items at a beginner level, a lower ratio of lexical items at the advanced level are produced in their speech after the10-month SA period.

Similar to previous studies, correlations were examined between fluency measures (Mora & Valls-Ferrer, 2012; Witton-Davies, 2014; Tovakoli et al., 2016). The results, in general, showed a significant correlation in two aspects, which are described next.

Firstly, speed fluency, as measured by speech rate (SR) and mean length of runs (MLR), showed significant improvement. SR, in this study, was obtained by syllables per minute and MLR was enumerated by the total number of syllables divided by the number of silent pauses which reached and exceeded 0.3 seconds. The significant increase of SR and MLR co-occured with a significant decrease in the number of silent pauses (SP100). This means that higher speed (higher fluency) and longer clusters of syllables between two pauses (higher fluency) co-existed with fewer silent pauses per 100 syllables (higher fluency), that is, the participants produced fewer silent pauses when producing longer utterances at a higher speed. Therefore, they improved their fluency in three ways. Furthermore, these results did not show a trade-off effect. There was a supportive relationship between speed fluency (i.e., SR, MLR) and breakdown fluency (i.e., SP100). This is consistent with the finding that the mean length of pause and mean length of the fluent run had weak to moderate negative correlations (Vercellotti, 2012) indicating a supportive relationship. This was very likely because of the benefit of the SA experience, and meant that learners were prone to speak faster and to produce longer runs with fewer silent pauses (Collentine & Freed, 2004; Mora & Valls-Ferrer, 2012; Valls-Ferrer & Mora, 2014).

Secondly, the analysis revealed a significant increase in ALFP, suggesting decreased fluency. This implies that longer utterances with higher speed containing fewer silent pauses (higher fluency) co-occurred with a longer length of filled pauses (lower fluency). These data support a trade-off effect between SR, MLR, SP100, and ALFP. Broadly, there was a tension between speed fluency and breakdown fluency in this regard. As indicated above, within the fluency domain, there was a competitive relationship between speed fluency (i.e., SR and MLR) and breakdown fluency (i.e., ALFP). For non-native speakers, filled pauses are used as a successful strategy for holding one's turn during an utterance (Wright, 2020), because low-fluency speakers tend to use hesitations and non-lexical fillers to provide themselves with a longer period for processing (Levelt, 1989).

5.3.2 Correlations between complexity, accuracy, and fluency after SA

Generalising the analysis, the improvement in the majority of the complexity measures (syntactic complexity via length and subordination) and fluency, especially speed fluency, came at the expense of lexical accuracy with a longer length of filled pauses. Broadly, there was a trade-off effect between fluency and accuracy, following a "natural" meaning (fluency) -form (accuracy) tension predicted by Skehan (1998a). Moreover, a secondary tension within form, between control of form (accuracy) and interlanguage risk-taking (complexity) (Skehan, 1998b), was also observed.

The tension between complexity and accuracy observed in this study is unsurprising since an increase in complexity at the word and sentence level statistically increases the chances that more errors will occur. It is clear that an increase in complexity corresponds to a decrease in accuracy. In other words, more complex language is less likely to be error-free. This follows Skehan's (2009) assumption of tension between control (accuracy) and risk-taking (complexity). This is also in line with the findings of previous studies (Chen, 2015a; Vercellotti, 2012, 2017) which indicated a negative relationship between accuracy and complexity. These results imply that after a period of SA, when learners undertake rehearsed topic-prompted tasks, they are very likely to structure their language in a more ambitious manner. This "cutting-edge" language with more complex syntax and sophisticated words places significant demands on their attentional resources, and goes beyond what they can comfortably control. Therefore, accuracy becomes less controlled, leading to more errors (Foster & Skehan, 1996).

Between fluency and accuracy, a tension was revealed by a significant increase in speed fluency (SR and MLR) and a significant decrease in lexical accuracy. In other words, more lexical errors appeared with longer speech runs at a higher speed. Following the argument of Foster and Skehan (1996), that accuracy is concerned with form, learners attempt to maintain control over available resources and avoid mistakes in a more conservative manner. Fluency reflects the primacy of meaning and the ability to communicate in real time. Fluency also prioritises idiom-based language over rule-based language to allow conversation to flow smoothly. These results support Skehan's (1998, 2009) theory that tension between focusing on meaning (fluency) and focusing on form (accuracy) should be expected and that it will lead to a trade-off effect. As a result of the benefit from SA experience, as

well as the effects of planning (Skehan, 2009c), learners seem to adequately produce idiom-based language to enable their utterances to proceed more smoothly (Foster & Skehan, 1996).

Complexity and fluency

Connected improvement was broadly observed between complexity and fluency. Within the domain of complexity, the most general complexity measure, length of AS-unit, calculated by the number of syllables per AS-unit, increased significantly (higher complexity) after SA. Similarly, the number of clauses per AS-unit achieved great gains (higher complexity). The same growth was also noted in lexical sophistication (higher complexity). Among fluency indicators, speed fluency (SR, MLR) saw a significant increase (higher fluency) with a significant decrease in SP100 (higher fluency). This joint increase in complexity and increased fluency are contrary to the trade-off effect. This is very likely attributable to two reasons. Specifically, SA experience advantages fluency (Freed et al., 2004; Mora & Valls-Ferrer, 2012) and complexity, especially syntactic complexity (Juan-Garau & Pérez-Vidal, 2007; Jensen & Howard, 2014; Pérez-Vidal & Juan-Garau, 2011; Llanes & Muñoz, 2013; Mora & Valls-Ferrer, 2012) and lexical sophistication (Collentine & Freed, 2004; Dewey, 2008; Kim et al., 2015). Moreover, because the learners undertook topic-centered tasks with planning time in this study, complexity and fluency were promoted by planning time in general (Skehan, 2001, 2009c).

Summary of relationships between CAF constructs and subconstructs within CAF (S1-S2)

In terms of the interrelationships between CAF constructs and subconstructs after SA, Table 15 (see Section 4.1) presents the results over the pre- and post-SA periods.

In terms of the relationships within each construct, a supportive relationship was found between speed and breakdown fluency in the domain of fluency. In the complexity construct, joint improvement was revealed between lexical sophistication and syntactic complexity. Within syntactic complexity, general complexity (length of AS-unit) was observed to be in a supportive relationship with complexity via subordination (clauses per AS-unit). Within speed fluency, the speech rate was positively correlated with longer fluent runs.

In terms of the interrelationships between CAF constructs, in general, these results indicate tensions between control (accuracy) and risk-taking (complexity), and between focusing on meaning (fluency) and form (accuracy). These findings are therefore in line with Skehan's hypothesis (Skehan, 2009; Wang & Skehan, 2014). However, for fluency and complexity, the analysis revealed that increased complexity (length, subordination, and lexical sophistication) correlated with improved fluency, with less silent pausing and longer fluent runs at a higher speed. These results showed that there was no trade-off effect between complexity and fluency, but instead that there was a connected improvement pattern. This is consistent with Skehan and Foster (2012), who concluded that simultaneous beneficial effects often happen with complexity and fluency, or with accuracy and fluency, but less frequently with complexity and accuracy when learners' performance on different task features or under different conditions are investigated with CAF measures. Moreover, SA experience very likely advantages both fluency and complexity (Mora & Valls-Ferrer, 2012).

In conclusion, this study investigated the effects of a 10-month study abroad sojourn on the oral performance of English-speaking learners of Chinese. The results show that the learners improved their complexity and fluency at the cost of accuracy. The widely accepted limited attentional resources which result in a trade-off effect in language performance were observed during the rehearsed topic-prompted monologue tasks in this study. In broad terms, these findings agree with Skehan's Trade-off Hypothesis, which predicts that raised levels in one performance area may deplete the attentional resources available for other areas and, as a result, that performance in those areas may decrease (Skehan, 2009c). However, this study has also revealed simultaneous improvements between complexity and fluency at the cost of accuracy. This also supports the findings of previous empirical studies (Vercellotti, 2012, 2017), which revealed no trade-off effect between complexity and fluency when investigating L2 English-speaking learners' performance on semi-spontaneous monologues with pre-planning time.

5.4 Correlations between CAF related to FI at home maintenance

This section firstly discusses the correlations within complexity and fluency in the oral performance of English-speaking learners of Chinese in relation to the impact of the FI at home factor. This is followed by a discussion of the correlations between the CAF constructs.

5.4.1 Correlations between subconstructs within CAF related to FI at home maintenance

Within the complexity construct, after the learners had returned to the FI at home context for six months, a strong joint decrease in lexical sophistication and syntactic complexity was observed. Meanwhile, a significant increase in lexical diversity was also present.

Among lexical sophistication measures, the analysis revealed a negative correlation between lexical items at the beginner level and the intermediate and advanced levels, indicated by a significant decrease in intermediate level and advanced level words with a significant increase in beginner level words. After learners had returned to the FI at home context for six months, they produced more beginner level words together with fewer intermediate level and advanced level words, which suggested lower lexical sophistication. This seeming tension between lexical items at the beginner level and the intermediate and advanced levels is because these three indicators accumulated into one. Moreover, limited by learners' proficiency level, beginner level words dominated the lexical items.

The strong trade-off effect between lexical sophistication and lexical diversity was revealed by the significant increase in lexical diversity together with a significant decrease in sophistication. These results suggested that the learners were prone to produce more varied words after returning to the FI at home context for six months after the 10-month SA period. Meanwhile, they tended to neglect more sophisticated words. Linking Levelt's speaking model to these learners' performance, the trade-off effect between lexical sophistication and lexical diversity could be expected. Specifically, lexical sophistication relates to the conceptualiser stage, and lexical diversity is an indication of formulator processing in Levelt's model (Skehan, 2009a). Considering their proficiency level (upper intermediate) at this stage, when learners focus on producing more unusual lexical items (higher lexical diversity) during speaking, the heavy lexical demands on formulator processing makes the conceptualiser–formulator connection problematic. Thus, only limited attention could be given during the conceptualiser stage, which led to poor performance in lexical sophistication.

Within the fluency domain, in general, a joint decrease was found between MLR and SP100, suggested by a significant decrease in MLR (lower fluency) together with a significant increase in

SP100 (lower fluency). In other words, this connected decrease was observed in both speed fluency and breakdown fluency. A tension within breakdown fluency was observed between ALFP and SP100, indicated by a statistical decrease in ALFP and a significant increase in SP100. However, considering the nature of ALFP as a successful strategy for holding one's turn during an utterance (Wright, 2020), the decrease of ALFP might not suggest an increase in fluency in a clear-cut manner. Therefore, this contrast between ALFP and SP100 is not necessarily a trade-off, but they both might imply a decrease in fluency. This decrease in fluency was largely caused by the significant increase in lexical diversity. Therefore, learners seem to concentrate on learning vocabulary and subordinating at the expense of accuracy and fluency (Juan-Garau & Pérez-Vidal, 2007) during the FI at home context. This suggests a trade-off effect between fluency and lexical diversity. Broadly speaking, a tension between fluency and complexity was observed. These results mirror Skehan and Foster's (1997, 2001) predictions concerning the competition between fluency (meaning) and complexity (interlanguage risk-taking). In particular, this risk-taking interlanguage operates in the subdomain of lexical complexity, that is, lexical diversity. This also supports the idea that the trade-off effect can be expected both between subconstructs as well as within/across each CAF construct.

5.4.2 Correlations between complexity, accuracy, and fluency related to FI at home maintenance

Concerning the correlations between the CAF measures, the results, in general, revealed that the improvement of lexical diversity came at the cost of syntactic complexity, lexical sophistication, and fluency development. Broadly, there was a trade-off effect between lexical complexity measures, as well as between vocabulary development and syntactic complexity and fluency. It seems that the retrieval of varied lexical items (higher lexical diversity) requires more silent pauses (lower fluency) at the cost of the length of syntactic complexity (lower syntactic complexity) and lower lexical sophistication. In other words, the need to retrieve from a larger lexical repertoire seems to have a cost concerning how more complex syntax, more sophisticated lexical items, and a smooth flow of speech are maintained as this retrieval creates processing demands and consumes attentional resources (Skehan, 2009). This supports the Trade-off Hypothesis.

This finding reinforces Skehan's (2009c) call for research to explore and identify performance in CAF in more contexts and under more conditions. It implies that the trade-off effect is not only

present in what Skehan (1998a) claimed to be a "natural" meaning-form (fluency vs. complexity/accuracy) tension as well as a secondary tension within the form (between accuracy and complexity) during language performance. This trade-off can be found between subconstructs of each CAF construct, and across subconstructs, and is determined by the task type and learning context investigated. In terms of this study's English-speaking learners of Chinese completing a topic-prompted monologue task during an FI at home context, a trade-off was evident between lexical variety and lexical sophistication in their oral performance. Learners make decisions on the priortisation of attentional resources during communication and learning, which leads to the allocation of attentional resources in one direction and limits their availability elsewhere (Skehan, 1996). In this study, learners apparently prioritised the retrieval of varied lexical items (higher lexical diversity) and this resulted in lower performance in other areas due to their limited attentional resources. This was because learners are very likely to learn more and acquire more diverse lexical items when the amount of time spent learning Chinese is increased during the FI at home context (Wu, 2017) with formal instruction.

5.5 Evaluating CAF indicators

Based on the results of this study in investigating the oral development of English-speaking learners of Chinese from upper beginner to upper intermediate levels over 28 months (including a 10-month Study Abroad period), this section provides some suggestions for using CAF indicators to measure oral development.

Among complexity measures, in the subdomain of syntactic complexity, length-based measures (i.e., the number of syllables per AS-unit) is a sensitive indicator to capture overall complexity. Complexity via subordination (i.e., the number of clauses per AS-unit) has been revealed to support the argument of Norris and Ortega (2009) who claimed that subordination measures are valuable when measuring learners at intermediate and upper intermediate levels. This indicator is stable when learners are at the intermediate level. Between lexical complexity measures, Guiraud's Index, used to measure lexical diversity, has been observed to be the most stable indicator in assessing speech data. This supports Vermeer (2000), who emphasised the stability of the indicator, in particular, when learners are at the upper beginner and lower intermediate levels. To investigate the lexical sophistication of learners of Chinese, beginner and advanced-level lexical items are more sensitive than intermediate level words. Specifically, the lexical items that are beyond current widely-used benchmarks (i.e., the HSK) deserve

more attention in coding and categorisation. The observations of this study contradict the conclusion of previous studies (Chen, 2015; Ye, 2015) that syntactic complexity and lexical variety develop very slowly, and only learners at the advanced level outperform those at the beginner level. The results of this study showed that indicators of lexical variety and syntactic complexity can be used to capture oral development when learners are at the beginner and intermediate levels.

In terms of accuracy measures, previous research (i.e., Chen, 2015a) concluded that lexical accuracy develops significantly when English-speaking learners of Chinese are at the advanced-level, compared to when learners are at the beginner and intermediate levels. The development of lexical accuracy is impacted by learners' proficiency levels. The results of this study revealed that when learners were at the intermediate level, their lexical accuracy did not reveal a linear pattern of decrease with the increase of their proficiency level. Instead, it decreased when more complex lexical items and syntactic structures were retrieved together at a higher speed and with fewer silent pauses. These results further support the finding of previous L2 Chinese studies (Chen, 2015a; Ding & Xiao, 2016) which revealed an unstable developmental pattern, featuring both progress and regression, when learners are at the beginner and intermediate levels. Therefore, lexical accuracy might not be a reliable indicator when assessing learners at the beginner and intermediate levels.

Among fluency measures, this study sought to observe the differing weights of various indicators in capturing the learners' fluency development. Speed fluency, captured by rate-related and time-related measures, is the most sensitive domain; repair fluency, measured by self-correction measures (i.e., RR100) is the least sensitive domain; whereas breakdown fluency, measured by counting the number and length of filled and unfilled pauses, has an intermediate level of sensitivity. Among breakdown fluency indicators, the frequency of silent pauses (i.e., SP100) and the length of filled pauses (i.e., ALFP) are likely to be more reliable predictors of L2 breakdown fluency than other indicators (i.e., FP100, ALSP). They are sensitive to oral development at least when learners are at the upper beginner to upper intermediate levels. This is different from the finding of Bosker et al. (2013) who found that pause frequency is likely to be a more reliable predictor of L2 breakdown fluency than pause duration. In this study, both pause frequency and pause duration in their subdomains could predict L2 breakdown fluency.

Chapter 6: Conclusion

This chapter summarises the results of this study. It also outlines some pedagogical implications, research limitations, and recommendations for L2 Chinese oral development for future research.

6.1 Summary of findings

The purpose of this study was to contribute to the literature on second language oral development by assessing English-speaking learners of Chinese in two contexts: Study Abroad (SA) and Formal Instruction at home (FI), which have rarely been investigated in an Irish context. For this investigation, two research questions were generated. The first research question sought to explore how the oral CAF development of the same cohort of instructed English-speaking learners of Chinese was affected by two learning contexts (SA and FI) semi-longitudinally. The second research question was designed to investigate the interrelationships between the CAF constructs and the sub-constructs within CAF in terms of how they were impacted by the two learning contexts. The two widely documented models (Limited Attentional Capacity and the Cognition Hypothesis) were proposed to explain how L2 learners' attention is deployed in oral performance and how this is impacted by task design (e.g., different tasks and contexts). The Limited Attentional Capacity Model is based on the assumption that L2 learners only have limited attentional resources, which leads to the three dimensions of CAF to compete for resource allocation in L2 task production. In particular, the model envisions a tension between complexity and accuracy. However, it has been acknowledged that such a trade-off effect can be overcome by task characteristics and task conditions (Wang & Skehan, 2014) due to selective influences on different aspects of performance triggered by task characteristics (Skehan & Foster, 2007). In contrast, predicated on the assumption of non-limited attentional resources, Robinson's Cognition Hypothesis predicts greater complexity and accuracy when influenced by increasing task complexity (Robinson, 2011). Therefore, when accuracy and complexity both increase, both Robinson and Skehan make predictions, but for different reasons. For Robinson, task difficulty is the motivator. However, the motivator according to Skehan is not task difficulty, but rather is the combination of task characteristics and task conditions (Tavakoli & Skehan, 2005).

To answer the two research questions proposed, oral data were collected from 10 English-speaking university students majoring in Chinese, each of whom experienced SA and FI contexts across a 28-month period (including a 10-month SA period). Their oral performance during three oral tests was measured using 14 CAF indicators. The effects of two learning contexts (SA and FI) on oral performance, encompassing pre- and post-SA and FI at home contexts, were explored by paired-samples t-tests. Generalised from the results of paired-samples t-tests, the interrelationships between the CAF constructs and within their sub-constructs have been revealed.

The main findings of the paired-samples t-tests for the first research question on oral development related to the study abroad and formal instruction at home periods are summarised below.

1) During the pre- and post-SA periods, within the CAF constructs to measure oral performance, syntactic complexity (length and subordination) and lexical sophistication benefited significantly from the SA period. However, fluency only saw limited gains, but learners did produce longer fluent runs at a higher speed and fewer silent pauses. In contrast, accuracy decreased significantly, and the learners made more lexical errors after SA. This finding can be interpreted as showing that SA experience benefits complexity and fluency at the expense of accuracy. These findings are consistent with earlier research, which indicates that SA increases learners' oral fluency (e.g., Freed et al., 2004; Du, 2013; Mora & Valls-Ferrer, 2012; Trenchs-Parera, 2009; Valls-Ferrer, 2010; Wright & Zhang, 2014) as well as syntactic complexity (Jensen & Howard, 2014; Juan-Garau & Pérez-Vidal, 2007; Mora & Valls-Ferrer 2012; Pérez-Vidal & Juan-Garau, 2011). Moreover, the task type has to be taken into consideration when interpreting learners' oral performance as measured by CAF. This study's analysis relates to the rehearsed topic-centered monologue task with planning that the participants undertook. It has shown that their complexity and fluency were enhanced by pre-task planning (cf. Skehan, 2001, 2009c).

2) Six months after returning to the Formal Instruction at-home (FI) context, complexity measures, the length of complexity, and lexical sophistication exhibited a significant decrease. Similarly, fluency, in particular, speed fluency and breakdown fluency, were observed to significantly decrease. In contrast, lexical diversity saw significant improvement. However, no statistically significant difference was revealed in general regarding the Formal Instruction at-home effects on repair fluency, accuracy, and complexity via subordination. Learners in this study during the FI context were prone

to focus on learning vocabulary and subordination at the cost of fluency. This finding mirrored that of Juan-Garau & Pérez-Vidal (2007) showing that FI learners are very likely to concentrate on learning vocabulary and subordinating at the expense of complexity and accuracy. However, accuracy remained stable when the learners were in the FI period. This is because, lexical accuracy does not necessarily decrease when learners focus on learning vocabulary.

Additionaly, the finding expands the assumption that oral gains are expected to decrease six months after returning to a domestic setting without formal instruction (cf. Juan-Garau & Pérez-Vidal, 2006, 2007; Pérez-Vidal & Juan-Garau, 2005). The analysis showed that learners' oral gains decreased (i.e., speed fluency, breakdown fluency, lexical sophistication, overall complexity) except for vocabulary development (i.e., lexical diversity) after six months of returning to the domestic setting with formal instruction. FI gains were much smaller than SA gains, another finding that is in line with previous research (LIanes & Serrano 2017; Mora & Valls-Ferrer, 2012). In this research, a statistically significant decrease occurred in the length of complexity and lexical sophistication as well as fluency six months after returning to the FI context with formal instruction.

Moreover, the generalised results of paired-samples t-tests for the second research question concerning the relationships between CAF constructs, and between subconstructs within CAF, and the impact of study abroad and formal instruction at home showed that:

1) The trade-off effect occurred between certain CAF constructs after SA, in particular between accuracy and complexity, and accuracy and fluency. These results confirm the Trade-off Hypothesis and that tension exists between control (accuracy) and risk-taking (complexity), and between focusing on meaning (fluency) and form (accuracy) (Skehan, 1998, 2009; Wang & Skehan, 2014). Task characteristics and learning contexts have been discussed to interpret the results because the different task and contextual characteristics supported different performance areas (Skehan & Foster, 2012). In terms of the learning context, study abroad favoured oral gains, especially fluency in terms of speed and silent pauses, syntactic complexity (length and subordination), and lexical sophistication. These findings broadly support Skehan's Trade-off Hypothesis, which postulates that increased performance in one area may drain the attentional resources available for other areas, leading to a potential decline in

those areas' performance (Skehan, 2009c). This study has also shown that accuracy suffers when complexity and fluency improve simultaneously. This is consistent with earlier empirical studies (Vercellotti, 2012, 2017), which looked at L2 English-speaking learners' performance on semi-spontaneous monologues with pre-planning time and found no evidence of a trade-off between complexity and fluency.

2) Concerning the correlations between CAF constructs and between sub-constructs within CAF during the FI at home context, the study has revealed a tension within the complexity domain between lexical diversity and syntactic complexity via length, as well as between lexical diversity and fluency. This implies that vocabulary development comes at the cost of syntactic complexity and fluency in the FI context (Juan-Garau & Pérez-Vidal, 2007). The results support the Trade-off Hypothesis that, tensions can be found between subconstructs within each CAF construct, especially in the subconstructs of complexity in this study. The prioritisation of attentional resources is determined by task type and learning context. This conclusion supports Skehan's (2009c) demand for research to investigate and characterize CAF performance in more contexts and under more circumstances. Six months after returning to the Formal Instruction at-home (FI) context, the English-speaking learners of Chinese in this study prioritised their attention on the retrieval of varied lexical items which resulted in lower performance in other areas, in particular, syntactic complexity via length, and speed fluency. This expands the finding of previous research that when students spend more time learning Chinese in a formal instruction setting at home, they are very likely to acquire a wider variety of lexical items (Wu, 2017). In terms of task design, planning seems not to have improved fluency and complexity as occurred during the SA period, or at least this was the case concerning the interrelationships among CAF measures in rehearsed topic promoted monologues in this study.

In conclusion, in this study, the trade-off effect was evident during the oral performance of English-speaking learners of Chinese. The trade-off effect was present not only between CAF constructs after SA (e.g., between complexity and accuracy) but also between subdomains within CAF during the FI at home context (e.g., between lexical sophistication and syntactic complexity). This extends research that hasn't yet been completely applied on how learning contexts affect oral performance in L2 Chinese. However, some connected improvement occurred between CAF constructs, such as the joint improvement in complexity and fluency after SA. Likewise, connected improvement was

also observed within certain subdomains of CAF. For instance, a simultaneous decrease was observed in lexical sophistication and syntactic complexity via length six months after the learners returned to the FI context. Therefore, the trade-off effect was clearly important in the learners' oral development, while the exact pattern of the results that the learners achieved can be explained in relation to learning context (study abroad and formal instruction at home) and task design (rehearsed topic promoted monologues). Additionally, the sensitivity of CAF indicators has to be taken into consideration when measuring the oral performance of L2 learners of Chinese to capture reliable developmental patterns. These results contribute to research on the impact of learning contexts on the oral performance of English-speaking Chinese learners by examining how attentional resources are prioritized across CAF dimensions and subconstructs within CAF.

6.2 Teaching and learning implications

This research provides several pedagogical implications for English-speaking learners of Chinese and Chinese language teachers at college level in an Irish context, for joint programmes in the target language country, as well as for L2 Chinese oral assessment.

Firstly, the challenging aspects (i.e., pauses, repairs and repetitions, lexical variety) of the oral performance of adult English-speaking learners of Chinese revealed by the CAF measures should be given more attention during teaching and practice within a teaching curricula. For example, to equip learners to have a better engagement during study abroad, oral class should be added as a transition between SA and FI contexts. To improve the oral performance of learners, oral fluency should be accorded more attention. Specifically, improvements in oral fluency can be achieved when dysfluency features (i.e., pauses, repairs and repetitions) are reduced. However, learners at the beginner and intermediate levels are very likely to see non-significant change in lexcal repairs and repetitions. Dysfluency phenomena might result from vocabulary retrieval during speech. Those words generating dysfluency may be old words that learners should be familiar with, or new words that learners take risks to use. For the former type of vocabulary, opportunities such as repetition practice in the classroom setting can be provided. For the latter type of words, learners should be encouraged to take risks to use new words in the classroom.

In terms of improving lexical diversity during speaking, Chinese language teachers should be encouraged to develop students' habits in using different words when expressing similar meanings in explicit instruction. For example, students should be required to use synonyms (e.g., 漂亮 (piàoliàng) vs. 好看(hǎo kàn)) and antonyms (e.g., 不贵(búguì) vs. 便宜(piányi) when expressing similar meanings in a task. Moreover, some exercises can be used to make students practice varied words during speaking in classroom teaching. Based on an assigned topic, students should be encouraged to utter the longest expressions that they are able to make with more varied words to express their thoughts.

Concerning the use of intermediate level lexical items, these seem to often be neglected in L2 learners' speech. Based on the HSK system, intermediate level lexical items have been embedded in the teaching syllabus, but in this study the participants' data revealed that their use of them was unsatisfactory. Chinese language teachers should be aware of this potential problem in using intermediate level lexical items and class activities should be designed to train learners to acquire and practice intermediate-level words within the classroom setting in a systematic manner in the FI at home context. Acquiring intermediate level words is very likely to help students to more easily scaffold to the advanced level. This approach would help students to use intermediate level words when they reach intermediate proficiency level. Also, acquiring more intermediate level words might help learners to produce more varied words in their oral production.

Secondly, this research has revealed a general pattern of trade-offs between CAF constructs and between the subconstructs of complexity in the oral performance of English-speaking learners of Chinese in SA and FI at home contexts during rehearsed topic prompted monologue tasks. This finding can help Chinese language teachers to gain a better understanding of the pattern of oral development of English-speaking learners of Chinese that are triggered by different learning contexts and task types. Tasks with different characteristics and conditions might determine learners achieve certain goals of the three different CAF aspects (complexity, accuracy, and fluency) of oral performance.

Thirdly, in terms of oral assessment, instead of a holistic approach, to achieve a more rounded picture of the oral performance of L2 learners of Chinese, the CAF measures should be analysed separately. However, it is important to recognise that CAF is mainly used for research purposes, rather than for language assessment in schools. Yet, learners' oral performance can undoubtedly be evaluated

in a more objective fashion rather than being rated subjectively by their Chinese language teachers as is generally the case now. To achieve this, learners' oral performance should be measured in a concrete manner rather than being evaluated in a vague way. To this end, CAF research might be used to help Chinese language teachers to adjust and design their teaching plans and to train students to achieve improvements in specific areas.

6.3 Limitations and suggestions for future work

Although this study has achieved its main goals of assessing the oral development of 10 English-speaking learners of Chinese over 28 months (including a 10-months of Study Abroad sojourn) during a bachelor's degree programme, there are nonetheless some limitations relating to the study's research design and participants that should be noted. Also, there are some further limitations concerning the learning contexts applied and the CAF measures employed.

The first limitation that must be recognised concerns the study's research design. In particular, the data used in the study were derived from a limited corpus collected during examination conditions. As a result, this potentially constrains the findings' generalisability to reveal the oral development of English-speaking learners of Chinese, as the study's findings are only robust in terms of the oral development of English-speaking learners of Chinese during exam conditions. Furthermore, the data were analysed at the group level, which means that the differences among individuals were overlooked, and while group data can describe a process, it has no validity at the level of the individual (Larsen-Freeman, 2009). Moreover, the relatively small sample size used in this study was partially due to the semi-longitudinal design of the research. The continuous assessment required each participant's commitment across a period of 28 months. Within this time period, the participants undertook a 10-month study abroad sojourn. Given this, there were many internal and external factors that might have impacted the oral production of the participants. For example, these include group internal (learners) factors (i.e., learning difficulties, motivations) and external (environmental) factors (i.e., teaching quality, teachers' competency, and environment). Issuing questionnaires to the participants to explore such internal and external factors during the experimental period, and in particular, during the study abroad period, would have been one way of attempting to assess the impact of these factors. However, neither a questionnaire nor a Language Contact Profile (cf. Freed et al., 2004) were issued to the participants.

This inhibited a deeper exploration of individual variations which may have affected the learners' performance. For example, issues such as the learners' motivation levels, their attendance record, length of exposure to the target language, and their social networks when they studied abroad were not considered. In particular, language profiles during the 10-month immersion experience would have allowed deeper interpretation of the data collected after the participants returned to the domestic context. Future experimental studies are recommended to explore learners' language exposure inside and outside of the classroom in the target country. This would provide a more complete insight into individuals' development (e.g., Wright & Zhang, 2014).

The second major limitation arises from the CAF measures applied in this research. Specifically, in this study, 14 CAF measures were assessed to explore the participants' CAF development and their interrelationships with oral performance. However, the questions of what weight should be attributed to the different CAF indicators, how to pinpoint the level of sub-indicators and how to measure the distance between them all deserve further attention. Also, some specific measures could be assessed to explore the development of each domain of CAF in relation to the unique characteristics of Chinese language (Wang, 2018; Feng, 2018). For example, in terms of phonetic accuracy, phonetic errors can be sub-divided into three types: consonant, vowel, and tone errors (e.g., Chen, 2015).

In terms of the research design, in future studies, the oral development of English-speaking learners of Chinese needs to be examined using experimental designs which can avoid the first limitation which emerged in this study. Moreover, this study only focussed on the oral development of English-speaking learners of Chinese in the context of limited time and commitment. The improvement in learners' written development and the relationship between written and oral skills can therefore be investigated in future to provide more detailed insights into L2 Chinese learners' language development. Furthermore, the relationship between subjective ratings and CAF measures in evaluating learners' performance can also be analysed to explore the validity of the CAF measures. Finally, in this study, the participants' speech samples were graded by the examiners during the oral tests in line with the standard approach found in most L2 studies on oral performance. However, although this approach is commonly employed, we presently lack a detailed understanding concerning which features of oral performance trigger perceptions of oral gain by such judges (Tonkyn, 2012). Given this, in the future, the relationship between the distinguishing features of oral performance of L2 learners of Chinese

measured by CAF measures and the scores awarded by examiners using holistic scoring undoubtedly deserves more attention (e.g., Ortega, 2003; Tonkyn, 2012; Jin & Mak, 2013; Zhai, 2011).

References

Ahmadian, M. J., & Tavakoli, M., (2011). The effects of simultaneous use of careful online planning and task repetition on accuracy, complexity, and fluency in EFL learners' oral production. *Language Teaching Research*, 35-59.

Ahmadian, M. J., (2011). The effect of 'massed' task repetitions on complexity, accuracy and fluency: Does it transfer to a new task? *Language Learning Journal*, 1-12.

Alghizzi, T. M. (2017). *Complexity, accuracy, and fluency (CAF) development in L2 writing: the effects of proficiency level, learning environment, text type, and time among Saudi EFL learners* (Doctoral dissertation, University College Cork)

Bacon, M. S. (2002). Learning the rules: Language development and cultural adjustment during study abroad. *Foreign Language Annals*, 35, 637-646.

Bassetti, B. (2007). Effects of Hanyu pinyin on the pronunciation of learners of Chinese as a Foreign Language. In A. Guder, X. Jiang and Y. Wan (Eds.), *The Cognition, Learning and Teaching of Chinese Characters* (pp. 156-179). Beijing: Beijing Language and Culture University Press.

Berg, M. V. (2009). Intervening in student learning abroad: a research-based inquiry. *Intercultural education*, 20 (sup1), S15-S27.

Bosker, H. R., Pinget, A. F., Quené, H., Sanders, T., & De Jong, N. H. (2013). What makes speech sound fluent? The contributions of pauses, speed and repairs. *Language Testing*, 30(2), 159-175.

Bowden, H.W., (2016). Assessing second-language oral proficiency for research. *Studies in Second Language Acquisition,* 38(4), pp. 647-675.

Brecht, R., Davidson, D., & Ginsberg, R. (1993). Predictors of foreign language gain during study abroad. Washington, DC: National Foreign Language Center.

Bulté, B. (2013). *The development of complexity in second language acquisition-A dynamic systems approach* (Doctoral dissertation, Free University of Brussels)

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA, 23-46.

Bulté, B., & Housen, A. (2018). Conceptualizing and measuring syntactic diversity. *International Journal of Applied Linguistics*, 28(1), 147-164.

Bulté, B., & Housen, A. (2018). Syntactic complexity in L2 writing: Individual pathways and emerging group trends. *International Journal of Applied Linguistics*, 28(1), 147-164.

Bulté, B., Housen, A., Pierrard, M., & Van Daele, S. (2008). Investigating lexical proficiency development over time–the case of Dutch-speaking learners of French in Brussels. *Journal of French Language Studies*, 18(3), 277-298.

Bygate, M., (2001). Effects of task repetition on the structure and control of oral language. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching Pedagogic Tasks: Second Language Learning, Teaching and Testing* (pp. 23-48). Essex: Pearson Education Limited.

Cameron, L., & Larsen-Freeman, D. (2007). Complex systems and applied linguistics. *International journal of applied linguistics,* 17(2), 226-239.

Chen M, & Li Y. J. (2016). A study on the complexity of Korean native speakers' spoken Chinese. *Language and text application*, 4

Chen M, & Zhou Q. (2016). A study on Chinese reading fluency of Korean native speakers. *Chinese teaching and research*, 2

Chen M. (2020). The impact of identity on the oral complexity, accuracy and fluency of Chinese second language learners. *Language Teaching and Research*, (1), 5.

Chen, M. (2012). Meiguo liuxuesheng Hanyu kouyu chanchu de liulixing yanjiu [Chinese oral fluency of CSL learners of American English speakers]. *Yuyan Jiaoxue yu Yanjiu [Language Teaching and Linguistic Studies]*, 2, 17–24.

Chen, M., (2015). An experiment study of Reading Fluency and Accuracy in Chinese as Second Language Acquisition. *Research on Chinese Applied Linguistics*. (1), 123-138.

Chen, Y. F. & Hsin, S. C. (2010). The development of TCSL teacher training in Taiwan. T*eaching and learning Chinese in global contexts: CFL worldwide*, 166-178.

Chinese Proficiency Test Center of Beijing Language and Culture University, (2000), Vocabulary outline of Chinese proficiency test, Chinese 8000 word dictionary, Beijing Language and Culture University Press.

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in second language acquisition*, 227-248.

Collentine, J. (2009). 13 Study Abroad Research: Findings, Implications, and Future Directions. *The handbook of language teaching*, 218.

Collentine, J., & Freed, B. F. (2004). Learning context and its effects on second language acquisition: Introduction. *Studies in second language acquisition*, 26(2), 153-171.

Collentine, J., (2004). 'The effects of learning contexts on morphosyntactic and lexical development.' *Studies in Second Language Acquisition* 26 (2), 227-248

Daller, H., Van Hout, R., & Treffers-Daller, J. (2003). The lexical richness in the spontaneous speech of bilinguals. *Applied linguistics*, 24(2), 197-222.

David, A., Myles, F., Rogers, V., & Rule, S. (2009). Lexical development in instructed L2 learners of French: Is there a relationship with morphosyntactic development. In B. Richards, M. H. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller, V*ocabulary Studies in First and Second Language Acquisition: The Interface Between Theory and Application* (pp. 147-163). Hampshire: Palgrave Macmillian.

De Bot, K. (2000). A bilingual production model: Levelt's "speaking" model adapted. *The bilingualism reader,* 420-442.

De Graaff, R., & Housen, A. (2009). 38 Investigating the Effects and Effectiveness of L2 Instruction. *The handbook of language teaching*, 726.

De Jong, N. H. (2016a). Predicting pauses in L1 and L2 speech: The effects of utterance boundaries and word frequency. *International Review of Applied Linguistics in Language Teaching*, 54(2), 113-132.

De Jong, N. H. (2018). Fluency in second language testing: Insights from different disciplines. *Language Assessment Quarterly*, 15(3), 237-254.

De Jong, N. H., & Bosker, H. R. (2013). Choosing a threshold for silent pauses to measure second language fluency. In R. Eklund (Ed.), *Proceedings of the 6th Workshop on Disfluency in Spontaneous Speech (DiSS)* (pp. 17-20).

De Jong, N. H., & Mora, J. C. (2019). Does having good articulatory skills lead to more fluent speech in first and second languages? *Studies in Second Language Acquisition,* 41(1), 227-239.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J. H. (2015). Second language fluency: Speaking style or proficiency? Correcting measures of second language fluency for first language behavior. *Applied Psycholinguistics*, 36(2), 223-243.

De Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2013). Linguistic skills and speaking fluency in a second language. *Applied Psycholinguistics*, 34(5), 893-916.

De Jong, N., & Perfetti, C. A., (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. *Language Learning*, 1-36.

DeKeyser, R. M. (2007). Study abroad as foreign language practice. *Practice in a second language: Perspectives from applied linguistics and cognitive psychology,* 208-226.

DeKeyser, R. M. (2005). What makes learning second-language grammar difficult? *A review of issues. Language learning*, 55(S1), 1-25.

DeKeyser, R. (2017). Knowledge and skill in ISLA. In *The Routledge handbook of instructed second language acquisition* (pp. 15-32). Routledge.

DeKeyser, R. M. (2014). Research on language development during study abroad. *Language acquisition in study abroad and formal instruction contexts*, 313-325.

Del Río, C., Juan-Garau, M., & Pérez-Vidal, C. (2018). Teachers' assessment of perceived foreign accent and comprehensibility in adolescent EFL oral production in Study Abroad and Formal Instruction contexts: A mixed-method study. *Learning context effects*, 181.

Derwing, M. and J. Rossiter. (2003). 'The effects of pronunciation instruction on the accuracy, fluency, and complexity of L2 accented speech.' Applied Language Learning 13: 1–18.

Derwing, T. M., Rossiter, M. J., Munro, M. J., & Thomson, R. I. (2004). Second language fluency: Judgments on different tasks. *Language learning*, 54(4), 655-679.

Devlin, A. M. (2019). The interaction between duration of study abroad, diversity of loci of learning and sociopragmatic variation patterns: A comparative study. *Journal of Pragmatics,* 146, 121-136.

Dewey, D. P. (2004). A comparison of reading development by learners of Japanese in intensive *domestic* immersion and study abroad contexts. *Studies in Second Language Acquisition*, 303-327.

Dewey, D. P. (2008). Japanese vocabulary acquisition by learners in three contexts. *Frontiers: The Interdisciplinary Journal of Study Abroad*, 15, 127-148.

Dewey, D. P., Belnap, R. K., & Hillstrom, R. (2013). Social network development, language use, and language acquisition during study abroad: Arabic language learners' perspectives. *Frontiers: The interdisciplinary journal of study abroad*, 22(1), 84-110.

Dewey, D. P., Bown, J., & Eggett, D. (2012). Japanese language proficiency, social networking, and language use during study abroad: Learners' perspectives. *Canadian Modern Language Review*, 68(2), 111-137.

Dewey, D. P., Bown, J., Baker, W., Martinsen, R. A., Gold, C., & Eggett, D. (2014). Language use in six study abroad programs: An exploratory analysis of possible predictors. *Language Learnin*g, 64(1), 36-71.

Di Silvio, F., Diao, W., & Donovan, A. (2016). The development of L2 fluency during study abroad: A cross-language study. *The Modern Language Journal*, 100(3), 610-624.

Diao, W., Donovan, A., & Malone, M. (2018). Oral language development among Mandarin learners in Chinese homestays. *Study Abroad Research in Second Language Acquisition and International Education*, 3(1), 32-57

Ding, A. Q. & Xiao, X. (2016). A study on the oral lexical development of elementary Italian learners' Chinese. *Beijing: Chinese Teaching In The World* (2) 239-252.

Du, H. (2013). The development of Chinese fluency during study abroad in China. *The Modern Language Journal*, 97(1), 131-143.

DuFon, M. A., & Churchill, E. (Eds.). (2006). *Language learners in study abroad contexts* (Vol. 15). Multilingual Matters.

Ellis, N. C. & Robinson, P. (2008). An introduction to cognitive linguistics, second language acquisition and language instruction. In Robinson, P., & Ellis, N. C. (Eds.). *Handbook of cognitive linguistics and second language acquisition*. (PP.3-24). Routledge.

Ellis, R. (2009). *Implicit and explicit knowledge in second language learning, testing and teaching* (Vol. 42). Multilingual Matters.

Ellis, R. (Ed.). (2005). *Planning and task performance in a second language* (Vol. 11). John Benjamins Publishing.

Ellis, R., & Barkhuizen, G., (2005). *Analysing Learner Language*. New York: Oxford University

Ellis, R., (2003). *Task-based Language Learning and Teaching*. Oxford University Press.

Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. *Applied linguistics*, 27(3), 464-491.

Felker, E. R., Klockmann, H. E., & De Jong, N. H. (2019). How conceptualizing influences fluency in first and second language speech production. *Applied Psycholinguistics*, 40(1), 111-136.

Feng, Y. (2018). *A case study on the diachronic development of oral Chinese fluency of primary Korean students,* (Master's thesis, Jinan University).

Fortune, T. W., & Ju, Z. (2017). Assessing and exploring the oral proficiency of young Mandarin immersion learners. *Annual Review of Applied Linguistics*, 37, 264-287.

Foster, P. (2009). Lexical diversity and native-like selection: The bonus of studying abroad. *In Vocabulary studies in first and second language acquisition* (pp. 91-106). Palgrave Macmillan, London.

Foster, P., & Skehan, P., (1996). The influence of planning on performance in task-based learning. *Studies in Second Language Acquisition*, 18 (3), 299-324.

Foster, P., & Wigglesworth, G. (2016). Capturing accuracy in second language performance: The case for a weighted clause ratio. *Annual Review of Applied Linguistics*, 36, 98.

Foster, P., (1996). Doing the task better: How planning time influences students' performance. In J. Willis and D. Willis (eds), *Challenge and Change in Language Teaching*. London: Heinemann.

Foster, P., Tonkyn, A., & Wigglesworth, G., (2000). Measuring spoken language: A unit for all reasons. *Applied Linguistics,* 21, 354-375.

Fotos, S., (1993). Consciousness-raising and noticing through a focus on form: Grammar task performance versus formal instruction. *Applied Linguistic*, 14 (4), 385-407.

Freed, B. F. (1995). What makes us think that students who study abroad become fluent? *Second language acquisition in a study abroad context*, 9, 123-148.

Freed, B. F. (1995a). Language learning and study abroad. *Second language acquisition in a study abroad context*, 3, 34.

Freed, B. F. (1998). An overview of issues and research in language learning in a study abroad setting. *Frontiers: The interdisciplinary journal of study abroad*, 4(1), 31-60.

Freed, B. F., Dewey, D. P., Segalowitz, N., & Halter, R. (2004). The language contact profile. Studies in Second Language Acquisition, 26(2), 349-356.

Freed, B. F., Segalowitz, N., & Dewey, D. P. (2004). Context of learning and second language fluency in French: Comparing regular classroom, study abroad, and intensive domestic immersion programs. *Studies in Second Language Acquisition,* 26, 275-301.

Gabbianelli, G., & Formica, A. (2017). Difficulties and expectations of first level Chinese second language learners. *In Explorations into Chinese as a Second Language* (pp. 183-206). Springer, Cham.

Gaillard, S., & Tremblay, A. (2016). Linguistic proficiency assessment in second language acquisition research: The elicited imitation task. *Language Learning*, 66(2), 419-447.

Gass, S., A. Mackey, M. Fernandez and M. Alvarez-Torres. (1999). The effects of task repetition on linguistic output. *Language Learning*, 49: 549–80

Gass, S., Mackey, A., Alvarez-Torres, M. J., & Fernández-García, M. (1999). The effects of task repetition on linguistic output. *Language Learning*, 49(4), 549-581.

Gilabert, R., (2005). *Task Complexity and L2 Narrative Oral Production*. Unpublished Ph.D.-thesis, Universitat de Barcelona, Spain.

Goldman-Eisler, F. (1972). Pauses, clauses, sentences. *Language and speech*, 15(2), 103-113.

Grey, S., Cox, J. G., Serafini, E. J., & Sanz, C. (2015). The role of individual differences in the study abroad context: Cognitive capacity and language development during short-term intensive language exposure. *The Modern Language Journa*l, 99(1), 137-157.

Griffith, W., & Lim, H. Y. (2012). Performance-based assessment: rubrics, web 2.0 tools and language competencies. *Mextesol Journal*, 36(1).

Guo, X. (2007). Hanyu zuowei di'er yuyan de kouyu liulixing lianghua ceping [Quantifiable measures of speaking fluency in Chinese as a second language]. *Xiangfan Shifan Xueyuan Xuebao (Shehui Kexue Ban) [Journal of Xiangtan Normal University (Social Science Edition)]*, 29(4), 91–94.

Gürbüz, N. (2017). Understanding fluency and disfluency in non-native speakers' conversational English. *Educational Sciences: Theory & Practice*, 17(6).

Hammerly, H. (1991). *Fluency and Accuracy: Toward Balance in Language Teaching and Learning*. Multilingual Matters 73.

Hamp-Lyons, L., & Henning, G. (1991). Communicative writing profiles: An investigation of the transferability of a multiple-traits coring instrument across ESL writing assessment contexts. *Language Learning*, 41, 337–373

Heaton, J. B. (1966). *Composition through pictures*. Longman Group United Kingdom.

Homburg, T. J. (1984). Holistic evaluation of ESL compositions: Can it be validated objectively?. *TESOL quarterly*, 18(1), 87-107.

Housen, A., & Kuiken, F., (2009). Complexity, accuracy and fluency in second language acquisition. *Applied Linguistics*, 30(4), 461-473.

Housen, A., Kuiken, F., & Vedder, I. (Eds.). (2012). *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (Vol. 32). John Benjamins Publishing.

Huang Huijian (1997). Assessing communicative behavior by picture-cued reading tasks. *Modern foreign languages*, (2), 43-48

Huensch, A., Tracy-Ventura, N., Bridges, J., & Medina, J. A. C. (2019). Variables affecting the maintenance of L2 proficiency and fluency four years post-study abroad. *Study Abroad Research in Second Language Acquisition and International Education,* 4(1), 96-125.

Hulstijn, J. H. (2011). Language proficiency in native and nonnative speakers: An agenda for research and suggestions for second-language assessment. *Language Assessment Quarterly*, 8, 229–249.

Hunt, K. W. (1965). Grammatical Structures Written at Three Grade Levels. NCTE Research Report No. 3.Huench, A., Tracy–Ventura, N., Bridges, J., & Cuesta Medina, J. A. (2019). Variables affecting the maintenance of L2 proficiency and fluency four years post-study abroad. *Study Abroad Research in Second Language Acquisition and International Education*, 4, 96–125.

Isabelli-García, Christina, Jennifer Bown, John L. Plews, and Dan P. Dewey. (2018). "Language learning and study abroad." *Language Teaching* 51, (4), 439-484.

Isabelli-García, C. (2010). Acquisition of Spanish gender agreement in two learning contexts: Study abroad and at home. *Foreign language annals,* 43(2), 289-303.

Iwashita, N., Brown, A., McNamara, T., & O'Hagan, S. (2008). Assessed levels of second language speaking proficiency: How distinct? *Applied linguistics*, 29(1), 24-49.

Jensen, J. M. (2015). *Study abroad and complexity, accuracy and fluency (CAF) development: a longitudinal investigation of French and Chinese Learners of L2 English* (Doctoral dissertation, University College Cork).

Jiang, W. (2013). Measurements of development in L2 written production: The case of L2 Chinese. *Applied Linguistics*, 34(1), 1-24.

Jin, T., & Mak, B. (2013). Distinguishing features in scoring L2 Chinese speaking performance: How do they work? *Language Testing*, 30(1), 23-47.

Jin, T., Mak, B., & Zhou, P. (2012). Confidence scoring of speaking performance: How does fuzziness become exact? *Language Testing*, 29, 43–65.

Johnston, J. R. (2001). An alternate MLU calculation: Magnitude and variability of effects. *Journal of Speech Language and Hearing Research*, 44(1), 156-164.

Juan-Garau, M. (2014). Oral accuracy growth after formal instruction and study abroad. *Language acquisition in study abroad and formal instruction contexts*, 87-111.

Juan-Garau, M., & Pérez-Vidal, C. (2007). The effect of context and contact on oral performance in students who go on a stay abroad. *Vigo International Journal of Applied Linguistics,* (4), 117-134.

Kafipour, R., & Khojasteh, L. (2011). The Study of Morphological, Syntactic, and Semantic Errors Made by Native Speakers of Persian and English Children. *Studies in Literature and Language,* 3(3), 109-114.

Kahng, J. (2014). Exploring utterance and cognitive fluency of L1 and L2 English speakers: Temporal measures and stimulated recall. *Language Learning*, 64(4), 809-854.

Kim, J., Dewey, D. P., Baker-Smemoe, W., Ring, S., Westover, A., & Eggett, D. L. (2015). L2 development during study abroad in China. *System*, 55, 123-133.

Koizumi, R. (2005). Speaking performance measures of fluency, accuracy, syntactic complexity, and lexical complexity. *JABAET (Japan-Britain Association for English Teaching) Journal*, 9, 5-33.

Kormos, J. (2011). Speech production and the Cognition Hypothesis. *Second language task complexity: Researching the cognition hypothesis of language learning and performance*, 2, 39-60.

Kormos, J., & Dénes, M. (2004). Exploring measures and perceptions of fluency in the speech of second language learners. *System*, 32(2), 145-164.

Kubler, C. C. (1997). Study abroad as an integral part of the Chinese language curriculum. J*ournal of the Chinese Language Teachers Association*, 32(3), 15-30.

Kuiken, F., & Vedder, I. (2007). Task complexity, task characteristics and measures of linguistic performance. *Complexity, accuracy and fluency in second language use, learning and teaching*, 113-126.

Kuiken, F., & Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA*, 143-170.

Lafford, B. A. (2004). The effect of the context of learning on the use of communication strategies by learners of Spanish as a second language. *Studies in Second Language Acquisition*, 201-225.

Lara, R., Mora, J. C., & Pérez-Vidal, C. (2015). How long is long enough? L2 English development through study abroad programmes varying in duration. I*nnovation in Language Learning and Teaching,* 9(1), 46-57.

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.

Larsen-Freeman, D., & Cameron, L. (2008). *Complex systems and applied linguistics.* Oxford: Oxford University Press

Larsen-Freeman, D., (2010). Complex, dynamic systems: A new transdisciplinary theme for applied linguistics? *Language Teaching*, 45 (2), 202-214.

Lennon, P. (1990). Investigating fluency in EFL: A quantitative approach. *Language learning*, 40(3), 387-417.

Leonard, K. R., & Shea, C. E. (2017). L2 speaking development during study abroad: Fluency, accuracy, complexity, and underlying cognitive factors. *The Modern Language Journal*, 101(1), 179-193.

Levelt, W. J. (1983). Monitoring and self-repair in speech. *Cognition*, 14(1), 41-104.

Levelt, W. J. (1984). Spontaneous self-repairs in speech: Processes and representations. *In Proceedings of the tenth international congress of phonetic science* (pp. 105-117). Dordrecht Foris.

Li, S. (2014). The effects of different levels of linguistic proficiency on the development of L2 Chinese request production during study abroad. *System*, 45, 103-116.

Li, X. (2011). Survey on Chinese Language and Culture Learning & Teaching Program in Irish Secondary Schools. *The Guide of Science & Education*, 2011(2), 195-196.

Li, X., & Li, J. (2014). Hanyu kouyu kaoshi (SCT) de xiaodu fenxi [Validity analysis of Spoken Chinese Test]. *Shijie Hanyu Jiaoxue [Chinese Teaching in the World]*, 28, 103–112.

Liao, J. (2018). Acquisition and assessment of L2 Chinese speaking. *The Routledge Handbook of Chinese Second Language Acquisition*, 234-260.

Liu Y. L., (1996), *Chinese Proficiency Level Standard and Grammar Level Outline*, higher education press.

Liu Y., & Wu X. Y. (2016). A study on the oral fluency of learners of Chinese. *Chinese teaching and research*, (4), 32-41.

Liu Y. (2019). A discussion on the causes of oral non fluency in Chinese as a second language. *Language teaching and research*, (3), 9.

Liu, J. (2009). Assessing students' language proficiency: A new model of study abroad program in China. *Journal of Studies in International Education,* 14(5), 528–544.

Liu, W. (2013). A Contrastive Study on Irish Transition Year (TY) Education and Chinese Secondary Education. *Journal of Tianjin Normal University (Elementary Education Edition),* 14(3), 61-64.

Liu, Y. (2017a). The relationship between text types and task difficulty in Chinese L2 teaching. *Journal of Chinese language education,* 30, 1-15.

Liu, Y. (2017b). Effect of task types on lexical complexity in L2 Chinese speaking performance. *Journal of Chinese teaching in the World,* 31(2), 253-269.

Llanes, À., & Muñoz, C. (2009). A short stay abroad: Does it make a difference? *System,* 37(3), 353-365.

Llanes, À., & Muñoz, C. (2013). Age effects in a study abroad context: Children and adults studying abroad and at home. *Language Learning,* 63(1), 63-90.

Llanes, À., & Serrano Serrano, R. (2011). Length of stay and study abroad: Language gains in two versus three months abroad. *Revista Española de Lingüística Aplicada*, 2011, núm. 24, p. 95-110.

Llanes, À., & Serrano, R. (2017). The effectiveness of classroom instruction 'at home'versus study abroad for learners of English as a foreign language attending primary school, secondary school and university. *The Language Learning Journal*, 45(4), 434-446.

Lu, J., & Zhao, Y. (2011). Teaching Chinese as a foreign language in China: A profile. *Teaching and learning Chinese in global contexts,* 117-130.

Lu, Y., & Song, L. (2017). European benchmarking Chinese language: Defining the competences in the written language. *In Teaching and Learning Chinese in Higher Education* (pp. 13-34). Routledge.

Lynch, T. and J. Maclean. (2001). Effects of immediate task repetition on learners' performance. In M. Bygate, P. Skehan and M. Swain (eds) *Research Pedagogic Tasks: Second Language Learning, teaching and testing.* 99–118. London: Longman.

Malvern, D., & Richards, B. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing,* 19(1), 85-104.

Malvern, D., & Richards, B., (1997). A new measure of lexical diversity. In A. Ryan, & A. Wray, *Evolving Models of Language.* (pp. 58-71). Multilingual Matters.

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development.* New York: Palgrave Macmillan.

Mao, S., Ye, J. (Eds.). (2002). *Testing pronunciation in teaching Chinese as a second language*. Beijing: China Social Sciences Press.

Mason, B., & Krashen, S. (2019). Hypothesis: A class supplying rich comprehensible input is more effective and efficient than "Immersion." *IBU Journal of Educational Research and Practice*, 7, 83-89.

McKee, G., Malvern, D., & Richards, B. (2000). Measuring vocabulary diversity using dedicated software. *Literary and linguistic computing,* 15(3), 323-338.

McManus, K., Mitchell, R., & Tracy-Ventura, N. (2021). A longitudinal study of advanced learners' linguistic development before, during, and after study abroad. *Applied Linguistics,* 42(1), 136-163.

Meara, P., & Bell, H., (2001). P_Lex: A simple and effective way of describing the lexical characteristics of short L2 text. *Prospect,* 5-19.

Mehnert U. (1998). The effects of different lengths of time for planning on second language performance. *Studies in Second Language Acquisition,* 20: 52-83.

Metruk, R. (2019) Assessing spoken proficiency: holistic and analytic ways of scoring. *Language, Literature and Culture in Education,* 28.

Michel, M. (2017). Complexity, accuracy and fluency in L2 production. *The Routledge handbook of instructed second language acquisition,* 50, 68.

Michel, M. C., Kuiken, F., & Vedder, I. (2012). Task complexity and interaction :(Combined) effects on task-based performance in Dutch as a second language. *EUROSLA yearbook*, 12(1), 164-190.

Michel, M. C., Kuiken, F., & Vedder, I. 2007. The influence of complexity in monologic versus dialogic tasks in Dutch L2. *IRAL,* 241-259.

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL-International Journal of Applied Linguistics,* 107(1), 17-34.

Ministry of education of the people's Republic of China, National Languages Committe. (2010) *Classification of syllables and Chinese characters for international Chinese education.* Beijing Language and Culture University Press.

Mitchell, R., Tracy-Ventura, N., & McManus, K. (2017). *Anglophone students abroad: Identity, social relationships and language learning.* Routledge.

Mizera, G. J., (2006). *Working memory and L2 fluency.* (Doctoral dissertation, University of Pittsburgh).

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy, and complexity in formal instruction and study abroad learning contexts. *Tesol Quarterly,* 46(4), 610-641.

Mora, J.C., (2006). Age effects on oral fluency development. In C. Munoz (Ed.) *Age and the rate of foreign langue learning* (pp. 65-88). Clevedon: Multilingual Matters.

Mozgalina, A. (2015). *Applying an argument-based approach for validating language proficiency assessments in second language acquisition research: The elicited imitation test for Russian* (Doctoral dissertation, Georgetown University).

Mulvaney, D. (2015). Exploring the complexity and dynamics of the willingness to communicate in English during group interaction. *In International Conference on Language and Communication 2015 proceedings* (pp. 50-74).

Muñoz, C. (2006). The effects of age on foreign language learning: The BAF project. *Age and the rate of foreign language learning,* 19, 1-40.

Namaziandost, E., & Ahmadi, S. (2019). The assessment of oral proficiency through holistic and analytic techniques of scoring: A comparative study. *Applied Linguistics Research Journal,* 3(2), 70-82.

National Hanban (2012). *New Chinese Proficiency Test (HSK) vocabulary* (Revised Version)，Available on http://www.chinesetest.cn/godownload.do

Németh N, and J. Kormos. (2001). Pragmatic aspects of task-performance: The case of argumentation. *Language Teaching Research,* 5: 213-240.

Nihalani, N. K. (1981). The Quest for the L2 Index of Development1. *RELC Journal,* 12(2), 50-56.

Norris, J., & Ortega, L. (2003). Defining and measuring SLA. *The handbook of second language acquisition,* 716-761.

Norris, J.M., & Ortega, L. (2009). Towards an organic approach to investigating CAF in instructed SLA: The case of complexity. *Applied Linguistics,* 30(4), 555–578.

O'Donnell, K. (2004). *Student perceptions of language learning in two contexts: At home and study abroad* (Doctoral dissertation, University of Pittsburgh).

Ong, J., & Zhang, L. J. (2010). Effects of task complexity on the fluency and lexical complexity in EFL students' argumentative writing. *Journal of Second Language Writing,* 19(4), 218-233.

Ortega, L. (1995). The effect of planning in oral narratives by adult learners of Spanish (Research note No. 15). Honolulu, HI: University of Hawaii. *Second Language Teaching and Curriculum Center*.

Ortega, L., & Bynes, H. (2008). *Longitudinal studies and advanced L2 capacities.* New York: Routledge.

Ortega, L., (1999). Planning and focus on form in L2 oral performance. *Studies in Second Language Acquisition,* 21(1), 109-148.

Ortega, L., (2003). 'Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing.' *Applied Linguistics,* 24: 492-518.

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002). An investigation of elicited imitation tasks in crosslinguistic SLA research. *In the Second Language Research Forum, Toronto.*

Orton, J. (2010). The current state of Chinese language education in Australian schools. *Education Services,* Australia.

Orton, J. (2011). Educating Chinese language teachers–Some fundamentals. *Teaching and learning Chinese in global contexts: CFL worldwide,* 151-164.

Orton, J. (2013). Developing Chinese oral skills: A research base for practice. *Research in Chinese as a second language,* 9-32.

Osborne, C., Zhang, Q., & Xia, Y. (2019). The past and present of Chinese language teaching in Ireland. *Chinese Language Teaching Methodology and Technology*, 2(1), 32.

Paige, R. M., Berg, M. V., & Lou, K. H. (2012). *Student Learning Abroad: What our Students Are Learning, What they're Not, and What We Can Do About It.* Stylus Publishing, LLC.

Pan, M. (2016). Rating Scale Formulation. In *Nonverbal Delivery in Speaking Assessment* (pp. 159-179). Springer, Singapore.

Pérez-Vidal, C. & Juan-Garau, M. 2011. The effect of context and input conditions on oral and written development: A study abroad perspective. *International Review of Applied Linguistics in Language Learning,* 49(2): 175–185.

Pérez-Vidal, C. (2014). Study abroad and formal instruction contrasted. *Language acquisition in study abroad and formal instruction contexts,* 13, 17-58.

Pérez-Vidal, C., & Juan-Garau, M. (2011). The effect of context and input conditions on oral and written development: A study abroad perspective, 49(2), 157-185. Available on https://doi.org/10.1515/iral.2011.008

Pérez-Vidal, C., Juan-Garau, M., & Mora, J. C. (2011). The effects of formal instruction and study abroad contexts on foreign language development: the SALA project. *Implicit and explicit conditions, processes and knowledge in SLA and bilingualism,* 115-138.

Pérez-Vidal, C., Juan-Garau, M., Mora, J. C., & Valls-Ferrer, M. (2012). Oral and written development in formal instruction and study abroad: Differential effects of learning context. *Intensive exposure experiences in second language learning,* 65, 213.

Pizziconi, B. (2017). Japanese vocabulary development in and beyond study abroad: the timing of the year abroad in a language degree curriculum. *The Language Learning Journal,* 45(2), 133-152.

Polat, B & Kim Y. (2014). Dynamics of Complexity and Accuracy: A Longitudinal Case Study of Advanced Untutored Development, *Applied Linguistics,* (2) 184–207.

Préfontaine, Y. (2013). Perceptions of French fluency in second language speech production. *Canadian modern language review,* 69(3), 324-348.

Préfontaine, Y., & Kormos, J. (2016). A qualitative analysis of perceptions of fluency in second language French. *International Review of Applied Linguistics in Language Teaching,* 54(2), 151-169.

Rees, J., & Klapper, J. (2007). Analysing and evaluating the linguistic benefit of residence abroad for UK foreign language students. *Assessment & Evaluation in Higher Education,* 32(3), 331-353.

Riggenbach, H. (1991). Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse processes,* 14(4), 423-441.

Robinson, P. (1995). Task complexity and second language narrative discourse. *Language learning,* 45(1), 99-140.

Robinson, P. (2003). The cognitive hypothesis, task design, and adult task-based language learning. *The University of Hawai'I Second Langauge Studies Paper,* 21 (2).

Robinson, P. (2005). Cognitive complexity and task sequencing: Studies in a componential framework for second language task design. *IRAL-International Review of Applied Linguistics in Language Teaching,* 43(1), 1-32.

Robinson, P. (2015). The Cognition Hypothesis, second language task demands, and the SSARC model of pedagogic task sequencing. *Domains and directions in the development of TBLT,* 8, 87-121.

Robinson, P. (Ed.). (2011). *Second language task complexity: Researching the cognition hypothesis of language learning and performance (Vol. 2)*. John Benjamins Publishing.

Robinson, P. (2001a). Task complexity, cognitive resources, and syllabus design: A triadic framework for examining task influences on SLA. In P. Robinson, *Cognition and Second Language Instruction* (pp. 287-318). Cambridge: Cambridge University Press.

Robinson, P. (2001b). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics,* 27-57.

Robinson, P., & Gilabert, R., (2007). Task complexity, the cognition hypothesis and second language learning and performance. *IRAL,* 161-176.

Robinson, P., (2003). Attention and memory during SLA. In C. J. Doughty, & M. H. Long (Eds.), *The Handbook of Second Language Acquisition* (pp. 631-678). Malden, MA: Blackwell Publishing.

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, 533–554.

Rossiter, M. J. (2009). Perceptions of L2 fluency by native and non-native speakers of English. *Canadian Modern Language Review,* 65(3), 395-412.

Sample, E., & Michel, M. (2014). An exploratory study into trade-off effects of complexity, accuracy, and fluency on young learners' oral task repetition. *TESL Canada Journal,* 23-23.

Sato, M. (2014). Exploring the construct of interactional oral fluency: Second Language Acquisition and Language Testing approaches. *System,* 45, 79-91.

Segalowitz, N. (2016). Second language fluency and its underlying cognitive and social determinants. *International Review of Applied Linguistics in Language Teaching,* 54(2), 79-95.

Segalowitz, N. (2000). Automaticity and attentional skill in fluent performance. In H. Riggenbach (Ed.) *perspective on fluency* (pp25-42). Anna Arbor: The University of Michigan Press.

Segalowitz, N., & Freed, B. F. (2004). Context, contact, and cognition in oral fluency acquisition: Learning Spanish in at home and study abroad contexts. *Studies in second language acquisition,* 173-199.

Serrano, R., Llanes, À., & Tragant, E. (2011). Analyzing the effect of context of second language learning: Domestic intensive and semi-intensive courses vs. study abroad in Europe. *System,* 39(2), 133-143.

Shen, H. H. (2018). Chinese as a second language reading: lexical access and text comprehension. In *The Routledge Handbook of Chinese Second Language Acquisition* (pp. 134-150). Routledge.

Shen, H. H., & Jiang, X. (2013). Character reading fluency, word segmentation accuracy, and reading comprehension in L2 Chinese, *Reading in Less Commonly Taught Languages*, 1-25

Shi, J. (2002). A case study of the acquisition of Chinese sentence patterns by Korean learners. *Chinese Teaching in the World,* 4, 34–42.

Shum, M. S. K., Tsung, L., & Gao, F. (2010). Teaching and learning (through) Putonghua: From the perspective of Hong Kong teachers. *Teaching and Learning Chinese in Global Contexts: CFL Worldwide,* 45.

Skehan P. and P. Foster. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research,* 1: 185-211.

Skehan, P. (1998). Task-based instruction. *Annual review of applied linguistics,* 18, 268-286.

Skehan, P. (2009a). Lexical performance by native and non-native speakers on language-learning tasks. In *Vocabulary studies in first and second language acquisition* (pp. 107-124). Palgrave Macmillan, London.

Skehan, P. (2009b). Models of speaking and the assessment of second language proficiency. *Issues in second language proficiency,* 203-215.

Skehan, P. (2009c). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied linguistics,* 30(4), 510-532.

Skehan, P. (1996b). A framework for the implementation of task-based instruction. *Applied Linguistics,* 17(1), 38–62.

Skehan, P., & Foster, P. (2012). Complexity, accuracy, fluency and lexis in task-based performance. *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA,* 32, 199

Skehan, P., & Foster, P., (2001). Cognition and tasks. In P. Robinson (Ed.), *Cognition and second language instruction* (pp. 183–205). New York: Cambridge University Press.

Skehan, P., (2003). 'Task-based instruction.' *Language Teaching,* 36: 1-14.

Skehan, P., (2014). Limited attentional capacity, second language performance, and task-based pedagogy. *Processing perspectives on task performance,* 211-260.

Skehan, P., and Foster, P., (1999). The Influence of Task Structure and Processing Conditions on Narrative Retellings. *Language Learning,* 49, 93-120.

Skehan, P., Foster, P., & Shum, S. (2016). Ladders and snakes in second language fluency. *International Review of Applied Linguistics in Language Teaching,* 54(2), 97-111.

Skehan, P., Willis, E. J., & Willis, D. (1996). Second language acquisition research and task-based instruction. *Readings in Methodology,* 13.

Skehan, P. (1998a). Task-based Instruction. *Annual Review of Applied Linguistics,* 18. 268 – 286.

Spada, N., Shiu, J. L. J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning,* 65(3), 723-751.

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics,* 31(4), 532–553

Sun Xiaoming (2009). Research on the development model of productive vocabulary of learners of Chinese. Journal of research on education for ethnic minorities, (4), 121-124

Tajima, M. (2003). *The effects of planning on oral performance of Japanese as a foreign language.* (Doctoral dissertation, Purdue University)

Tavakoli, P. (2016). Fluency in monologic and dialogic task performance: Challenges in defining and measuring L2 fluency. *International Review of Applied Linguistics in Language Teaching,* 54(2), 133-150

Tavakoli, P. and Skehan, P. (2005). 'Strategic planning, task structure, and performance testing' in R. Ellis (ed.): *Planning and Task Performance in a Second Language*. Amsterdam: Benjamins.

Tavakoli, P., & Foster, P. (2008). Task design and second language performance: The effect of narrative type on learner output. *Language Learning,* 58(2), 439-473.

Taylor, L., & Galaczi, E. (2011). Scoring validity. *Examining speaking: Research and practice in assessing second language speaking,* 30, 171-233.

Tomas, E., & Dorofeeva, S. (2019). Mean Length of Utterance and Other Quantitative Measures of Spontaneous Speech in Russian-Speaking Children. *Journal of Speech, Language, and Hearing Research,* 62(12), 4483-4496.

Tonkyn, A., Housen, A., Kuiken, F., & Vedder, I. (2012). Measuring and perceiving changes in oral complexity, accuracy and fluency. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA.* Amsterdam: Benjamins, 221-245.

Towell, R. (2002). Relative degrees of fluency: A comparative case study of advanced learners of French. IRAL, *International Review of Applied Linguistics in Language Teaching,* 40(2), 117.

Towell, R. and J.-M. Dewaele. (2005). 'The role of psycholinguistic factors in the development of fluency amongst advanced learners of French.' in Dewaele, J.-M. (ed.): *Focus on French as a Foreign Language: Multidisciplinary Approaches.* Clevedon: Multilingual Matters.

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied linguistics,* 17(1), 84-119.

Tracy-Ventura, N., Huensch, A., & Mitchell, R. (2021). Understanding the long-term evolution of L2 lexical diversity: The contribution of a longitudinal learner corpus. *Learner Corpus Research Meets Second Language Acquisition,* 148.

Trenchs-Parera, M. (2009). Effects of formal instruction and a stay abroad on the acquisition of native-like oral fluency. *Canadian Modern Language Review,* 65(3), 365-393.

Třísková, H. (2017). Acquiring and Teaching Chinese Pronunciation. In *Explorations into Chinese as a Second Language* (pp. 3-30). Springer, Cham.

Tse, S. K., & Tan, W. X. (2011). Catering for Primary School Pupils with Different Chinese Language Proficiencies in Singapore through Differentiated Curricula and Instructional Materials. *Teaching and Learning Chinese in Global Contexts: CFL Worldwide,* 29.

Tseng, M. F. (2019). Creating a Task-Based Language Course in Mandarin Chinese. In *The Routledge Handbook of Chinese Language Teaching* (pp. 118-133). Routledge.

Tsung, L., & Cruickshank, K. (2010). Minority education for exclusion or access: Teaching Chinese as a second language in Xinjiang Uyghur Autonomous Region. *Learning and teaching Chinese in global contexts: Multimodality and literacy in the new media age,* 97-115.

Valls Ferrer, M. (2010). *Language acquisition during a stay abroad period following formal instruction: temporal effects on oral fluency development* (Doctoral dissertation, Universitat Pompeu Fabra).

Valls-Ferrer, M., & Mora, J. C. (2014). L2 fluency development in formal instruction and study abroad. *Language acquisition in study abroad and formal instruction contexts,* 111-136.

Van Geert, P. (2008). The dynamic systems approach in the study of L1 and L2 acquisition: An introduction. *The Modern Language Journal,* 92(2), 179-199.

Vercellotti, M. L. (2012). *Complexity, accuracy, and fluency as properties of language performance: The development of multiple subsystems over time and in relation to each other* (Doctoral dissertation, University of Pittsburgh).

Vercellotti, M. L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics,* 38(1), 90-111.

Vercellotti, M. L. (2019). Finding variation: assessing the development of syntactic complexity in ESL Speech. *International Journal of Applied Linguistics,* 29(2), 233-247.

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language testing,* 17(1), 65-83.

Verspoor, M., De Bot, K., & Lowie, W. (2011). *A dynamic approach to second language development: Methods and techniques.* Amsterdam: John Benjamins

Verspoor, M., Lowie, W., & Van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal,* 92(2), 214-231.

Wagner, R. K., Puranik, C. S., Foorman, B., Foster, E., Wilson, L. G., Tschinkel, E., & Kantor, P. T. (2011). Modeling the development of written language. *Reading and writing,* 24(2), 203-220.

Wang, X. H. (2010). *Analysis of productive vocabulary of primary Chinese learners* (Master's thesis, Shanghai Normal University).

Wang, X. Z, & Jin, X.Y. (2020). A Study on non-fluent oral filled pauses of Chinese Second Language Learners. J*ournal of Northeast Normal University (PHILOSOPHY AND SOCIAL SCIENCES EDITION)*, (2), 11.

Wang, X. Z. (2018). A study on oral fluency of Chinese as a second language. (Doctoral dissertation, Northeast Normal University).

Wang, J. (2002). Sanlei kouyu kaoshi tixing de pingfen yanjiu [A study on the scoring of three types of oral test items]. *Shijie Hanyu Jiaoxue [Chinese Teaching in the World],* 16, 63–77.

Wang, J. (2005). Yuyan nengli ziwo pingjia de xiaodu yanjiu [Validation of self-assessment of second language ability]. *Yuyan Jiaoxue yu Yanjiu [Language Teaching and Linguistic Studies]*, 5, 60–68.

Wang, J. (2011). Chuji Hanyu kouyu ceyan tixing yanjiu [A study on test item types of beginning Chinese oral test]. *Kaoshi Yanjiu [Examinations Research]*, 28(5), 67–76.

Wang, Z., & Skehan, P. (2014). Structure, lexis, and time perspective Influences on task performance. In P. Skehan (ed.), *Processing perspectives on task performance,* 155-185, Amsterdam, the Netherlands: Benjamins.

Wigglesworth, G., (1998). The effect of planning time on second language test discourse. In A. J. Kunnan (Ed.), *Validation in language assessment* (pp. 91-110). Mahwah, NJ: Lawrence Erl-baum Associates.

Wolfe-Quintero, K., S. Inagaki and H.-Y. Kim. (1998). *Second Language Development in Writing: Measures of Fluency, Accuracy, and Complexity.* Honolulu, HI: University of Hawai'i, Second Language Teaching and Curriculum Centre.

Wood, D. (2001). In search of fluency: What is it and how can we teach it? *Canadian Modern Language Review,* 57(4), 573-589.

Wood, D. (2006). The uses and functions of formulaic sequences in second language speech: An exploration of the foundations of fluency. *The Canadian Modern Language Review,* 63(1), 13-22.

Wright, C. (2018). Effects of time and task on L2 Mandarin Chinese language development during study abroad. In C. Sanz & A. Morales-Front (Eds.), *The Routledge handbook of study abroad research and practice* (pp. 166–180). New York: Routledge.

Wright, C. (2020). Effects of task type on L2 Mandarin fluency development. *Journal of Second Language Studies,* 3(2), 157-179.

Wright, C., & Cong, Z. (2014). Examining the effects of Study Abroad on L2 Chinese development among UK university learners. *Newcastle and Northumbria Working Papers in Linguistics,* 20, 67-83.

Wright, C., & Tavakoli, P. (2016). New directions and developments in defining, analyzing and measuring L2 speech fluency. *International Review of Applied Linguistics in Language Teaching,* 54(2), 73-77.

Wu, J. F. (2016). A study on the development of lexical richness in Chinese writing by native English speakers. *World Chinese teaching,* 30 (1), 129-142

Wu, X. Y. (2017). *A case study on the development of Chinese Proficiency of a Second-language Learner* (Master's thesis, Jinan University)

Wu, C. H. (2008). Filled pauses in L2 Chinese: A comparison of native and non-native speakers. In *Proceedings of the 20th North American Conference on Chinese Linguistics* (NACCL-20) (Vol. 1, pp. 213-227). Columbus, Ohio: The Ohio State University.

Wu, Q. (2017). *A Survey Report on the Chinese-Teaching in College and University in County of Dublin in Ireland*. (Master's Thesis, Chongqing Normal University, China).

Wu, S. L., & Ortega, L. (2013). Measuring global oral proficiency in SLA research: A new elicited imitation test of L2 Chinese. *Foreign Language Annals,* 46(4), 680-704.

Xiao, X. M. (2017). *A study of oral reading fluency on Chinese as a second language.* (Master's Thesis, Beijing Language and Culture University, China).

Xiaoqi, L., & Jinghua, L. (2014). Validity Analysis of Spoken Chinese Test [J]. *Chinese Teaching in the World,* 1.

Xu, K. (2014). Yanyu shengcheng shijiao xia liuxuesheng kouyu zhong jieci shiyong de kaocha yu fenxi [An analysis of prepositions in foreign students' oral Chinese: From language production perspective]. *Huawen Jiaoxue yu Yanjiu [TCSOL Studies],* 56(4), 32–38.

Yan, X., Maeda, Y., Lv, J., & Ginther, A. (2016). Elicited imitation as a measure of second language proficiency: A narrative review and meta-analysis. *Language Testing,* 33(4), 497-528.

Yang, J. S. (2016). The effectiveness of study-abroad on second language learning: A meta-analysis. *Canadian Modern Language Review,* 72(1), 66-94.

Yang, J., & Medwell, J. (2017). Learners' and teachers' beliefs about learning tones and pinyin. In *Explorations into Chinese as a Second Language* (pp. 141-163). Springer, Cham.

Ye, W. (2015). Yingyu muyuzhe hanyu kouyu shuiping fazhan yanjiu [The oral Chinese language development of native English speakers]. *Nanjing Shifan Daxue Wenxueyuan Xuebao [Journal of School of Chinese Language and Culture Nanjing Normal University],* 4, 170–174.

Yu, G. (2009). Lexical diversity in writing and speaking task performances. *Applied linguistics,* 31(2), 236-259.

Yu, Q. (2016). *Defining and Assessing Chinese Syntactic Complexity via TC-Units* (Doctoral dissertation, University of Hawaii at Manoa]

Yuan, F. (2010). Impacts of task conditions on learners' output in L2 Chinese narrative writing. *Journal of the Chinese Language Teachers Association,* 45(1), 67-88.

Yuan, F. and R. Ellis. (2003). 'The effects of pre-task planning and on-line planning on fluency, complexity and accuracy in L2 oral production.' *Applied Linguistics,* 24: 1-27.

Zhai, Yan (2012). Evaluation criteria for oral Chinese performance test. *Chinese teaching and research, January,* 2012, 44-52

Zhai, Y. (2011). Kouyu liulixing zhuguan biaozhun de keguanhua yanjiu [A study on the subjective criteria of speaking fluency]. *Yuyan Jiaoxue yu Yanjiu [Language Teaching and Linguistic Studies],* 5, 79–86.

Zhai, Y., & Feng, H. (2014). Jiyu "kantu shuohua" renwu de Hanyu xuexizhe kouyu liulixing fazhan yanjiu [Study of Chinese learners' speaking fluency development with picture description activity]. *Huawen Jiaoxue yu Yanjiu [TCSOL Studies],* 56(4), 1–7.

Zhang, Li (2001). Relation between foreign students' learning anxiety and their fluency of spoken Chinese. *Applied Linguistics,* (3), 44-49

Zhang, L. P. (2002). *Theory and practice of Chinese proficiency test.* Northern Taiwan: Normal University Bookstore.

Zhang, L. P. and Chen F. Y. (2005). Classification of Chinese vocabulary. *Proceedings of the 6th International Symposium on Chinese vocabulary semantics.*

Zhang, L. P. (2007). Development of Chinese language proficiency test (top Huayu). 2007 [trends and prospects of foreign language proficiency test] *International Symposium.* Taipei: National Chengchi University.

Zhang, P. (2019). *The effect of pre task preparation time on oral fluency and accuracy of intermediate Chinese learners* (Master's thesis, Shanghai Foreign Studies University)

Zhang W. Z. (2000). A qualitative study on the development of second language oral fluency. *Modern foreign languages,* (3), 273-282

Zhang, W. Z., & Wu, X. D (2001). A quantitative study on the development of second language oral fluency. *Modern foreign languages,* 24 (4), 341-351

Zhang, C., & Wang, H. (2018). The Development of Chinese Language Education in Ireland: Issues and Prospects. TEANGA, *the Journal of the Irish Association for Applied Linguistics,* 25, 34-51.

Zhou, A. J. & Zhang, C. (2006). Application of Cool Edit Pro software in oral English fluency measurement. *Foreign language audio visual teaching,* (2), 67-70

Zhou, J. (2016). *A case study on the diachronic development of Korean students' oral Chinese accuracy and complexity* (Master's thesis, Jinan University).

Zhou, M. (2010). Globalization and language order: Teaching Chinese as foreign language in the United States. *Teaching and learning Chinese in a global context: Multimodality and literacy in the new media age,* 131-150.

Zhou, Y. (2012). *Willingness to communicate in learning Mandarin as a foreign and heritage language* (Doctoral dissertation, University of Hawaii at Manoa).

Zhou, Y., & Wu, S. L. (2009). *Development and pilot of a Mandarin L2 elicited imitation task*. Unpublished manuscript, University of Hawai ‘i at Manoa, Honolulu.

Zhu, S. (2009). *A study on the dynamic oral text of Korean learners of Chinese* (in Chinese).（Master's thesis, Beijing Language and Culture University).

Appendixes

Appendix A: Normality of 14 CAF measures

Table 16. Normality distribution for each CAF measure

Con-structs	Measures	Tests of Normality	Signifi-cance
Fluency	Speech rate	Kolmogorov-Smirnov	.053
	Mean Length of Runs	Kolmogorov-Smirnov	.000
	The_average_length_of_filled_pause(ALFP)	Kolmogorov-Smirnov	.200
	The_average_length_of_silent_pause(ALSP)	Kolmogorov-Smirnov	.200*
	The_number_of_filled_pauses_per_100_syllables(FP100)	Shapiro-Wilk	.005
	The_number_of_silent_pauses_per_100-syllables(SP100)	Kolmogorov-Smirnov	.179
	The_number_of_repairs_and_repetitions_per_100_syllables(RR100)	Kolmogorov-Smirnov	.200*
Accuracy	Lexical_accuracy	Kolmogorov-Smirnov	.090
Complexity	Guiraud's_index	Kolmogorov-Smirnov	.200*
	Lexical_beginning	Kolmogorov-Smirnov	.179
	Lexical_intermediate	Kolmogorov-Smirnov	.200*
	Lexical_advanced	Kolmogorov-Smirnov	.200*
	The_number_of_syllables_per_AS_unit	Kolmogorov-Smirnov	.062
	The_number_of_sub_clauses_per_AS_unit	Kolmogorov-Smirnov	.073

Note. *p < .05.

Appendix B: Scores per participant

Table 17. Scores for each measure per participant

Constructs	Indicators	S1	S2	S3	Participant_No.
Accuracy	*Lexical_accuracy*	0.90	0.86	0.92	1
Fluency	*SR*	100.47	125.40	128.38	1
Fluency	*MLR*	3.47	4.30	4.55	1
Fluency	*ALFP*	0.36	0.50	0.41	1
Fluency	*ALSP*	0.87	1.15	0.95	1
Fluency	*FP100*	24.85	11.63	7.38	1
Fluency	*SP100*	28.83	23.26	21.98	1
Fluency	*RR100*	3.37	5.04	3.92	1
Complexity	*Guiraud's Index*	4.96	4.09	8.58	1
Complexity	*Lexcial_Beginning*	0.59	0.52	0.65	1
Complexity	*Lexical_Intermediate*	0.20	0.16	0.15	1
Complexity	*Lexical_advanced*	0.21	0.32	0.21	1
Complexity	*Syntactic_syllables*	16.30	36.86	30.33	1
Complexity	*Syntactic_subclause*	1.50	1.71	2.52	1
Accuracy	*Lexical_accuracy*	0.93	0.87	0.89	2
Fluency	*SR*	90.16	119.96	105.85	2
Fluency	*MLR*	3.18	5.79	3.70	2
Fluency	*ALFP*	0.47	0.56	0.41	2
Fluency	*ALSP*	0.97	0.74	0.72	2
Fluency	*FP100*	8.30	4.94	9.51	2
Fluency	*SP100*	31.41	17.28	26.99	2
Fluency	*RR100*	3.25	0.41	1.80	2
Complexity	*Guiraud's Index*	5.32	5.94	7.28	2
Complexity	*Lexcial_Beginning*	0.68	0.59	0.72	2
Complexity	*Lexical_Intermediate*	0.19	0.20	0.09	2
Complexity	*Lexical_advanced*	0.13	0.21	0.19	2
Complexity	*Syntactic_syllables*	19.79	24.30	21.61	2
Complexity	*Syntactic_subclause*	1.86	2.00	2.00	2
Accuracy	*Lexical_accuracy*	0.89	0.86	0.86	3
Fluency	*SR*	80.00	97.38	92.34	3
Fluency	*MLR*	2.56	4.12	3.45	3
Fluency	*ALFP*	0.51	0.52	0.49	3
Fluency	*ALSP*	0.80	0.54	0.71	3
Fluency	*FP100*	26.14	27.50	24.22	3
Fluency	*SP100*	39.00	24.29	29.02	3
Fluency	*RR100*	4.98	3.21	2.30	3
Complexity	*Guiraud's Index*	6.02	5.59	8.73	3
Complexity	*Lexcial_Beginning*	0.75	0.50	0.74	3

Complexity	***Lexical_Intermediate***	0.13	0.23	0.10	3
Complexity	***Lexical_advanced***	0.13	0.27	0.16	3
Complexity	***Syntactic_syllables***	18.54	31.11	19.96	3
Complexity	***Syntactic_subclause***	1.62	2.22	1.83	3
Accuracy	***Lexical_accuracy***	0.90	0.86	0.86	4
Fluency	***SR***	106.57	155.38	194.82	4
Fluency	***MLR***	3.69	8.14	7.95	4
Fluency	***ALFP***	0.28	0.49	0.37	4
Fluency	***ALSP***	0.95	0.53	0.53	4
Fluency	***FP100***	0.52	4.21	2.78	4
Fluency	***SP100***	27.08	12.28	12.58	4
Fluency	***RR100***	1.04	2.28	2.02	4
Complexity	***Guiraud's Index***	5.64	3.95	10.59	4
Complexity	***Lexcial_Beginning***	0.62	0.53	0.62	4
Complexity	***Lexical_Intermediate***	0.20	0.18	0.18	4
Complexity	***Lexical_advanced***	0.18	0.29	0.20	4
Complexity	***Syntactic_syllables***	16.00	47.50	29.98	4
Complexity	***Syntactic_subclause***	1.33	2.75	2.56	4
Accuracy	***Lexical_accuracy***	0.87	0.82	0.89	5
Fluency	***SR***	100.82	136.71	110.69	5
Fluency	***MLR***	4.39	5.31	3.74	5
Fluency	***ALFP***	0.72	0.64	0.61	5
Fluency	***ALSP***	0.87	0.82	0.87	5
Fluency	***FP100***	4.85	4.84	9.50	5
Fluency	***SP100***	22.78	18.82	26.73	5
Fluency	***RR100***	3.80	3.23	2.42	5
Complexity	***Guiraud's Index***	4.08	3.48	8.06	5
Complexity	***Lexcial_Beginning***	0.64	0.64	0.67	5
Complexity	***Lexical_Intermediate***	0.19	0.16	0.11	5
Complexity	***Lexical_advanced***	0.17	0.20	0.22	5
Complexity	***Syntactic_syllables***	15.80	41.33	27.00	5
Complexity	***Syntactic_subclause***	1.33	2.11	2.04	5
Accuracy	***Lexical_accuracy***	0.79	0.84	0.71	6
Fluency	***SR***	60.56	68.91	89.06	6
Fluency	***MLR***	2.37	3.22	3.28	6
Fluency	***ALFP***	0.65	0.68	0.62	6
Fluency	***ALSP***	0.88	1.07	0.89	6
Fluency	***FP100***	21.58	15.15	11.59	6
Fluency	***SP100***	42.23	31.06	30.47	6
Fluency	***RR100***	5.57	3.28	3.22	6
Complexity	***Guiraud's Index***	4.87	5.01	6.24	6
Complexity	***Lexcial_Beginning***	0.62	0.58	0.71	6
Complexity	***Lexical_Intermediate***	0.22	0.23	0.11	6
Complexity	***Lexical_advanced***	0.16	0.19	0.18	6
Complexity	***Syntactic_syllables***	23.94	36.00	20.26	6
Complexity	***Syntactic_subclause***	1.50	2.27	1.91	6

Accuracy	***Lexical_accuracy***	0.96	0.90	0.84	7
Fluency	***SR***	103.62	107.55	93.07	7
Fluency	***MLR***	3.41	4.84	3.38	7
Fluency	***ALFP***	0.52	0.54	0.49	7
Fluency	***ALSP***	0.68	0.65	0.78	7
Fluency	***FP100***	11.11	12.55	14.29	7
Fluency	***SP100***	29.33	20.66	29.63	7
Fluency	***RR100***	1.33	2.58	1.59	7
Complexity	***Guiraud's Index***	4.02	4.80	8.16	7
Complexity	***Lexcial_Beginning***	0.63	0.50	0.58	7
Complexity	***Lexical_Intermediate***	0.18	0.22	0.20	7
Complexity	***Lexical_advanced***	0.20	0.28	0.23	7
Complexity	***Syntactic_syllables***	18.75	33.88	27.00	7
Complexity	***Syntactic_subclause***	1.33	2.13	1.93	7
Accuracy	***Lexical_accuracy***	0.94	0.91	0.96	8
Fluency	***SR***	64.68	100.52	119.08	8
Fluency	***MLR***	2.22	4.07	3.50	8
Fluency	***ALFP***	0.49	0.57	0.44	8
Fluency	***ALSP***	1.13	0.82	0.82	8
Fluency	***FP100***	6.29	9.00	8.62	8
Fluency	***SP100***	45.14	24.57	28.57	8
Fluency	***RR100***	3.43	2.08	2.71	8
Complexity	***Guiraud's Index***	4.44	6.09	8.84	8
Complexity	***Lexcial_Beginning***	0.70	0.56	0.61	8
Complexity	***Lexical_Intermediate***	0.16	0.21	0.19	8
Complexity	***Lexical_advanced***	0.14	0.23	0.20	8
Complexity	***Syntactic_syllables***	15.91	32.11	27.07	8
Complexity	***Syntactic_subclause***	1.55	1.78	2.33	8
Accuracy	***Lexical_accuracy***	0.86	0.88	0.82	9
Fluency	***SR***	112.36	127.93	118.68	9
Fluency	***MLR***	3.84	5.63	4.24	9
Fluency	***ALFP***	0.45	0.42	0.46	9
Fluency	***ALSP***	0.85	0.69	0.83	9
Fluency	***FP100***	14.56	9.21	9.88	9
Fluency	***SP100***	26.05	17.76	23.59	9
Fluency	***RR100***	1.92	2.63	1.01	9
Complexity	***Guiraud's Index***	4.53	4.34	8.07	9
Complexity	***Lexcial_Beginning***	0.65	0.60	0.65	9
Complexity	***Lexical_Intermediate***	0.18	0.15	0.09	9
Complexity	***Lexical_advanced***	0.18	0.25	0.26	9
Complexity	***Syntactic_syllables***	20.08	25.33	21.57	9
Complexity	***Syntactic_subclause***	1.46	1.83	2.04	9
Accuracy	***Lexical_accuracy***	0.89	0.80	0.86	10
Fluency	***SR***	94.03	92.31	99.36	10
Fluency	***MLR***	3.56	3.71	3.45	10
Fluency	***ALFP***	0.48	0.65	0.62	10

Fluency	***ALSP***	0.91	1.08	1.03	10
Fluency	***FP100***	14.06	14.45	14.80	10
Fluency	***SP100***	28.13	26.95	28.95	10
Fluency	***RR100***	2.60	3.13	2.30	10
Complexity	***Guiraud's Index***	4.03	3.94	6.60	10
Complexity	***Lexcial_Beginning***	0.70	0.50	0.71	10
Complexity	***Lexical_Intermediate***	0.18	0.28	0.08	10
Complexity	***Lexical_advanced***	0.11	0.22	0.21	10
Complexity	***Syntactic_syllables***	16.00	32.00	17.88	10
Complexity	***Syntactic_subclause***	1.50	2.13	2.00	10

Appendix C: Markings per participant

Table 18. Markings per participant rated by the course's teachers at S1, S2 and S3

Student ID	**Session 1**		**Session 2**		**Session 3**	
	Oral (3 Dec)	Written (26March)	Oral (15 Oct)	Written(1 Nov)	Oral (4Ari)	Written (3Apr)
1	69	57	60	70	70	73
2	67	68	60	62	56	69
3	42	49	53	51	55	48
4	63	45	62	68	70	64
5	72	49	60	63	56	63
6	41	44	46	54	51	50
7	61	72	48	59	53	57
8	51	33	65	62	64	68
9	57	56	50	47	56	50
10	45	25	42	37	51	42

Appendix D: HSK Scores per participant

Table 19. HSK Scores per participant

Participants No.	**HSK3** （ 28 March before study abroad）			
	Listening **100**	**Reading** 100	**Writing** 100	**Total** 300
1	90	61	75	226
2	88	77	63	228
3	68	51	67	186
4	73	44	55	172
5	83	47	55	185
6	63	47	59	169
7	90	74	83	247
8	68	47	47	162
9	80	77	71	228
10	88	61	51	200
Note. The version of HSK the participants took was "HSK 2.0", which was released in 2010 and have been changed since July 2021. The passing score of HSK3 ("HSK 2.0") is 180.				

Appendix E: Tasks used in S1, S2 and S3

Sessions	Topic promoted tasks
S1	请介绍一下，你住在哪里，你住得地方怎么样？ Please tell us, where do you live and how is your place?
S2	对有些年轻人频繁跳槽的现象，你怎么看？ What do you think of the phenomenon that some young people frequently change jobs?
S3	你去过中国的哪些名胜古迹？最喜欢哪儿？ Which places of interest have you been to in China? Where is your favorite?