PRE-POST CHANGE IN L2 ORAL FLUENCY: THE LEXICO-SYNTAX OF LARGE FLUENCY GAINERS

by

David Crouch

A Dissertation

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

Department of English

West Lafayette, Indiana

May 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. April Ginther, Chair

Department of English

Dr. Margie Berns

Department of English

Dr. Tony Silva

Department of English

Dr. Harris Bras

Department of English

Approved by:

Dr. S. Dorsey Armstrong

Dedicated to my loving wife Shina (a.k.a. Hannah), who has loved, encouraged, and supported me throughout my studies.

ACKNOWLEDGMENTS

I would like to thank Dr. April Ginther, my kind, patient, and understanding advisor, without whom, this dissertation would not have been possible. A big thank you to my dissertation committee members, who have given me many helpful comments throughout this process. Also, thank you, Jie (Wendy) Gao, for cross-checking my results.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

The theory underlying L2 oral fluency has focused on cognitive processes, particularly proceduralization (Anderson, 1983; Levelt, 1989, 1999) and linguistic constructs, especially vocabulary and grammar (Segalowitz, 2010). Towell, Hawkins, and Bazergui (1996) argued that development of formulaic language enables automatic speech production. However, no research has studied the longitudinal development of L2 oral fluency concurrently with any of the following lexical variables: lexical frequency profile, formulaic language use, and MTLD (a measure of lexical diversity). The purpose of the present study is to clarify the process by which L2 oral fluency, syntax, and vocabulary develop concurrently.

Data analysis involved three sequential phases: oral fluency analysis, lexico-syntactic analysis, and discourse analysis. Oral fluency measures were calculated using the transcribed oral test responses of 100 L1-Chinese EAP learners at the beginning and end of a required two-course EAP language and culture sequence at Purdue University. The task completed was a computer-administered, two-minute argumentative speaking task. This study included eight oral fluency measures: speech rate, mean length of speech run, articulation rate, phonation time ratio, mean length of silent pause, mean length of filled pause, silent pause frequency, and filled pause frequency. For the ten participants who made the largest percentage-wise oral fluency gains (in terms of the oral fluency variable associated with the largest effect size of gains), oral transcripts were analyzed to compute descriptive statistics for the three lexical variables mentioned above and three syntactic variables: coordinate clause ratio, dependent clause ratio, and words per T-unit.

Results indicated significant change in all oral fluency measures, except mean length of silent pause and mean length of filled pause. The largest gains were made in mean length of speech run. Of the linguistic variables, the largest longitudinal change was associated with coordinate clause ratio. Discourse analysis of the transcripts of large fluency gainers' pre-post responses suggested that large fluency gainers used coordinate clauses to build more sophisticated discourse models in the post-test response than they did in the pre-test response. Implications for L2 oral fluency theory, EAP pedagogy, and L2 oral assessment are discussed.

CHAPTER 1. INTRODUCTION

1.1 Background

The ability to speak fluently is an important aspect of L2 proficiency. Along with grammatical accuracy, vocabulary knowledge, and other core language skills, oral fluency is important for conveying a spoken message comprehensibly and intelligibly. This is likely because listeners can more easily process an oral message that is delivered at an optimal rate (Clark, 2002) and with minimal pausing, especially mid-clause (Goldman-Eisler, 1968). Second language learners tend to experience a tradeoff effect between accuracy and fluency (Skehan, 1998), increasing one at the expense of the other.

A great deal of research has examined oral fluency from the perspective of the listener. Research into listener perception of fluency has found that some temporal measures of oral fluency correlate strongly and positively with listener perception of oral fluency (Derwing & Munro, 2004; Kormos & Dénes, 2004; Rossiter, 2009). High stakes second language testing research also reflects the importance of certain aspects of oral fluency (Ginther, Dimova, & Yang, 2010; Iwashita, Brown, & McNamara, & O'Hagan, 2008).

The theory underlying the development of L1 oral fluency has focused on psycholinguistic processes. Lexico-syntactic formulation and phonological articulation are the two linguistic constructs that have received the most attention in the strand of oral fluency (Levelt, 1989, 1999; Segalowitz, 2010; Towell, Hawkins, & Bazergui, 1996;). Psycholinguists have borrowed the concepts of working memory, proceduralization, and automaticity from mainstream psychology in an effort to explain how L1 learners improve their efficiency at formulating and articulating oral language (Anderson, 1983; Mohle & Raupach, 1987).

Some researchers have studied the longitudinal development of L2 oral fluency in terms of temporal measures of fluency as well as qualitative aspects of linguistic development. This strand of research has added immensely to our understanding of L2 oral fluency development. Most prominently, Towell et al. (1996) attributed oral fluency gains in L2- learners of French mostly to increased use of formulaic language. In contrast, Collentine (2004) provided some evidence from an experimental study that the use of more complex syntax facilitated pre-post oral fluency gains in L2-Spanish.

1.2 Organization

There are two main parts to this study. The first part will consist of a longitudinal analysis of the oral fluency development of 100 L1 Chinese university students over a period of two semesters. The purpose of the longitudinal analysis is to examine the development of fluency over time. The present study, to a large extent, provides evidence consistent with Levelt's Speech Production Model and Towell et al.'s (1996) findings, which are discussed in more detail below.

The second part of the study relates to the nature of the formulator in the Speech Production Model. Towell et al (1996) argued that development in the formulator was characterized by greater syntactic complexity in the responses of L2 learners after one year of immersion in the target language. However, they did not investigate the role of vocabulary development in the Speech Production Model. The present study fills that gap.

In the longitudinal analysis of oral fluency development, I will calculate descriptive statistics of the sample for each of the oral fluency variables at time 1 and time 2, respectively. Second, I will conduct a paired sample t-test to determine whether the pre-post differences observed in the sample were statistically significant. Using this pre-post research design, I will measure eight temporal measures of oral fluency.

Phase two of the present study was an exploratory analysis of the longitudinal development of lexical ability and syntactic complexity in the large fluency gainers. After identifying the oral fluency measure associated with the largest effect size change in the group of 100 examinees, the ten examinees ("largest fluency gainers" henceforth) who exhibited the largest percentage-wise gains with regard to that measure were identified. Then, the transcribed pre-test and post-test responses were analyzed in terms of three measures of lexical ability and three measures of syntactic complexity. Finally, the discourse of one large fluency gainer is discussed as an exemplar of how L2 discourse, lexico-syntax, and oral fluency develop together in the responses of large fluency gainers.

1.3 Research Questions

Several studies have focused on longitudinal development of L2 oral fluency in a study abroad setting. All such studies have found that L2 learners improve their oral fluency after

spending an extended period of time studying the L2 abroad (Collentine, 2004; Kim, Dewey, Baker-Smemoe, Ring, Westover, & Eggert, 2015; Huensch & Tracy-Ventura, 2017; Mora & Valls-Ferrer, 2012; Segalowitz & Freed, 2004; Towell et al., 1996;). Guided by the L2 Speech Production Model (De Bot, 1992; Levelt, 1989, 1999), some research on the longitudinal development of L2 oral fluency has analyzed the use of vocabulary (Collentine, 2004; Kim et al., 2015; Mora & Valls-Ferrer, 2012; Towell et al., 1996) and syntax (Collentine, 2004; Valls-Ferrer, 2012).

While these studies have yielded useful findings, the strand of longitudinal L2 oral fluency research could better represent the study abroad population in three key respects. First, all of these studies included study abroad participants whose primary purpose for studying abroad was to learn the target language. Such studies do not capture the large population of international students who study an L2 as an academic language, enabling them to take courses in a non-language major. According to Statista (2019), of the 1,078,822 international university students in the US in the 2017-8 academic year, the largest nationality group (363,341) was from China. Moreover, a majority of these Chinese students studied non-language disciplines like (in descending order of frequency): Business/Management (20.7%), Engineering (19%), and Math/Computer Science (17.2%), as opposed to "Intensive English" (2.1%). Second, only one study (Valls-Ferrer, 2012) analyzed the development of oral English as an L2, and no study of which the author is aware has included L1-Chinese participants. Third, as Segalowitz (2010) noted, most of the existing studies of L2 longitudinal development have small sample sizes. When using inferential statistics to analyze change in multiple variables over time, it is best to use a large sample size.

Fortunately, there is no shortage of L1-Chinese students studying in undergraduate programs at 4-year universities in the US, especially in the STEM fields. In fact, the large number of such students presents unique challenges for this student population and the universities that seek to help them integrate culturally and linguistically into the North American academic community.

The most significant challenge facing the L1-Chinese population at large North American universities is that with so many Chinese classmates, the temptation to socialize primarily with other Chinese students may overwhelm their motivation to integrate socially. This temptation

can preclude the kind of intensive oral English practice that Bybee (2008) argued is necessary for reaching advanced levels of oral fluency, lexical ability, and syntactic complexity.

Oral fluency studies have particularly focused on lexical and syntactic variables. This is probably because vocabulary and syntax figure prominently in the influential L2 Speech Production Model (De Bot, 1992; Levelt, 1989, 1999). While all of the studies cited above showed that studying abroad led to significant gains in L2 oral fluency, the results regarding gains in L2 vocabulary and syntax have been mixed. For example, Towell et al. (1996) attributed fluency gains mostly to acquisition of formulaic language, while Collentine (2004) found that study abroad participants made greater gains in syntax than in lexical ability. Moreover, the study abroad participants in Mora & Valls-Ferrer (2012) increased their lexical richness and length of AS-units without increasing the number of clauses per AS-unit.

Of course, L2 lexical ability and syntactic complexity are each multi-faceted constructs. For example, lexical ability encompasses the ability to use a wide range of words, including words at different frequency levels, as well as a variety of multi-word formulaic sequences. Similarly, syntax can be made more complex by adding subordinate clauses, adding coordinate clauses, or including more words in each production unit. Hence, including measures that gauge multiple aspects of lexical and syntactic ability would clarify the nature of the lexico-syntactic reorganization that takes place over time in L2 speakers who make large gains in oral fluency.

The present study used a pre-post longitudinal research design to measure the L2 development of 100 L1-Chinese examinees over the course of a two-semester language and culture course sequence. All participants were first year undergraduate students at a large, public STEM university in the US. Each participant took a computer-administered, semi-direct English language proficiency test, from which the responses to the "Express Your Opinion" free response item were analyzed for oral fluency, lexical ability, and syntactic complexity. The following research questions were investigated. The research questions of the present study were as follows:

1. How does the L2-English oral fluency of university-level L1-Chinese test-takers change over the course of two semesters of language and culture study?
2. How does the L2-English oral lexical ability of the ten largest fluency gainers change over the course of two semesters of language and culture study?

3. How does the L2-English oral syntactic complexity of the ten largest fluency gainers change over the course of two semesters of language and culture study?

CHAPTER 2. LITERATURE REVIEW

2.1 Introduction

The research questions of this dissertation involve temporal measures of fluency, longitudinal development of oral fluency, formulaic sequences, lexical frequency profiles, and a lexical diversity measure. Hence, the literature review discusses the research related to these topics. The literature review discusses the literature that is relevant to the present study in four sections. Each section ends with a summary of the literature reviewed and analysis of the literature as it relates to the present study.

Section 2.2 traces the theoretical development of oral fluency research from its beginning in the 1950's to the development of cognitive models of speech production in the 1980's. Section 2.2 starts with Goldman-Eisler's (1958a, 1958b) pioneering work on pausing, then moves on to Pawley & Syder's (1983) connection between fluency and vocabulary. After that, Anderson's (1983) introduction of the Adaptive Control of Thought (ACT) model is discussed, before explaining how Levelt (1989) adapted Anderson's ACT model to create the Speech Production Model.

Section 2.3 discusses research on oral fluency in the field of ESL and SLA. This section starts with a brief explanation of Lennon's distinction between fluency in the "broad sense" and fluency in the "narrow sense" (p. 389), as well as his research applying "temporal measures of fluency" (p. 392) to the longitudinal development of L2 oral fluency. Then, section 2.3 summarizes the longitudinal studies of L2 oral fluency development, including Towell et al.'s (1996) application of the Levelt Speech Production Model (1989) to L2 learners. The section ends with a review of cross-sectional studies involving temporal measures of fluency.

Section 2.4 covers the theory underlying formulaic sequences, lexical frequency, and their relationship to oral English proficiency in general and oral fluency in particular. This section starts with Wray's (2002) influential treatment of formulaic sequences. Then, it summarizes Laufer and Nation's (1995) Lexical Frequency Profile (LFP) and other research on the relationship between lexical measures and L2 oral English proficiency, as well as L2 oral fluency. Finally, it summarizes the methodology behind the Lexical Frequency Profile (Laufer & Nation, 1995) and the Spoken Academic Formulas list (Simpson and Vlach-Ellis, 2010).

2.2 Theoretical Background

Goldman-Eisler (1958a, 1958b) conducted some of the earliest widely-known research on L1 fluency and hesitation phenomena in the field of clinical psychology. She conducted a series of many experiments involving spontaneous speech in the 1950's and 1960's in London. She further influenced the development of oral fluency research by pioneering the use of special tools for measuring pause length and speech rate. It is also noteworthy that she started the practice of studying speech sounds and pausing in speech separately.

She argued that pausing in spontaneous speech was primarily the result of "freedom of choice" (p. 97) at points of "uncertainty" (p. 96) in the flow of speech. She described these points as locations where the previous speech content and linguistic structure provide the least "constraining" (p. 97) influence on the next word to be uttered. She argued that this freedom of choice requires planning and selection among various alternatives, which necessarily takes time. To state the hypothesis in other words, previous discourse constrains or limits the acceptable possibilities of the following discourse. This constraining influence, she hypothesized, allows the speaker to choose quickly among the different linguistic alternatives and continue speaking without pausing.

She tested this hypothesis by using word-guessing procedures. In one experiment (Goldman-Eisler, 1958a), the L1 English subjects (N=8) were given a sentence that was originally uttered spontaneously in a conversation, and they were asked to guess the first and each consecutive word in the sentence. Each sentence contained words uttered either fluently, before a pause, or after a pause. Two variables were included in the study: one quantitative variable and one categorical variable. The quantitative variable was transition probability of words to be guessed. The transition probability of words was calculated as the ratio of the number of times that participants guessed the word correctly and the total number of guesses. In other words, very easy to guess words had transition possibilities of near one, and impossible to guess words had transition possibilities of zero.

The categorical variable involved pausing. The words to be guessed were categorized based on the fluency with which they were uttered in the original recorded conversation from which the sentences were excerpted. The words were categorized as either uttered fluently, with no preceding pause (>.25 sec), uttered disfluently, with a preceding pause, or uttered fluently directly preceding a pause. Goldman-Eisler calculated a Chi-square statistic, which tests whether

categorical variables in a population are related to each other, for two categorical variables: words uttered after a pause, words uttered fluently, and words with transition probabilities of zero.

The results supported the original hypothesis that transition probability was influenced by whether a word was uttered fluently or after a pause in the original speech. The results showed that most of the words with transition probabilities of zero were originally uttered after a pause, while the vast majority of the words uttered fluently were "predictable at various levels of probability" (p. 100).

In a follow-up experiment (Goldman-Eisler, 1958b), subjects (N=15) read and completed the same spontaneously produced sentences from the experiment just mentioned. From each sentence, one word was deleted, and participants had to fill in the correct word. In the study just reviewed, some of the words to be filled in were determined to have high transition probabilities, while some were determined to have low transition probabilities. Three quantitative variables were included in this study for each sentence: speech rate, mean length of silent pause (before providing the missing word), and percentage of guesses that were correct. The author performed a one-way ANOVA with words of high versus low transition probability as the two factor levels and speech rate, mean length of silent pause, and percentage of guesses that were correct as the dependent variables.

Results showed that the subjects guessed less accurately for the same words that had originally been uttered after long pauses. Moreover, participants read these sentences slower and with longer pauses.

Goldman-Eisler concluded that those words before which the original speaker paused longest and that the subjects had the most trouble guessing were the words that contained the most information value. In line with her original hypothesis, she argued that these words were the ones that were least constrained by the preceding language and thus presented the speaker with the most freedom of choice.

To summarize, Goldman-Eisler presented the idea of freedom of choice as having a negative influence on L1 fluency. She described the constraining influence of preceding discourse as having a positive influence L1 oral fluency, reasoning that it limited the possibilities available at a point of high uncertainty in speech. However, she did not explain the linguistic or cognitive mechanisms underlying this constraining influence of discourse. Furthermore, she did

not explore the possibility that having a wide variety of well-organized lexical alternatives to choose from at a speech juncture of great uncertainty could facilitate continuance of speech. Pawley and Syder's (1983) later discussed the facilitative possibilities of freedom of choice.

Pawley and Syder (1983) discussed the connection between "nativelike selection" (p. 191) and "nativelike fluency" (p. 191). What interested Pawley and Syder (1983) the most was how the native speaker "selects a sentence that is natural and idiomatic from among the range of grammatically correct paraphrases, many of which are non-nativelike or highly marked usages" (p. 191). They gave a few examples of "nativelike" sentences and their non-nativelike alternatives. They considered those marked with a '*' below "non-nativelike".

1a. (at a party) "I'm so glad you could bring Harry!" (p. 195)

1b. "That you could bring Harry gladdens me so." (*) (p. 196)

2a. (telling the time) "It's twenty to six." (p. 197)

2b. "It's six less twenty." (*) (p. 197)

By "nativelike fluency" (p. 191) Pawley & Syder meant simply that native speakers can produce "fluent stretches of spontaneous connected discourse" (p. 191).

The authors criticized Chomsky's (1957) syntactic theory without discarding it. Pawley and Syder did not consider the Chomskyan notion of generative grammar and infinite linguistic creativity to be incorrect, but rather only part of the story of language production. They argued that linguistic innovation makes up part of linguistic competence, but innovation is far too cognitively demanding to be the dominant force in linguistic performance. They asserted that native speakers must have access to a large repertoire of memorized material in order to allocate processing capacity to the innovation--linguistic and otherwise-- that is necessary to fulfill the communicative demands of highly demanding situations.

The primary means by which Pawley and Syder argued that native speakers allocate processing capacity efficiently in highly demanding communicative situations are what they called "lexicalized sentence stems" (p. 191). They defined the unit as "a unit of clause length or longer whose grammatical form and lexical content is wholly or largely fixed" (p. 191). A few examples of such units are "It's on the tip of my tongue"; "That's easier said than done"; and "It's quarter past two" (p. 206-7).

They considered the cultural meaning of such units to be fixed for native speakers. In other words, every adult native speaker in a particular speech community knows the cultural

meaning behind the unit immediately upon hearing it. Moreover, they argued that lexicalized sentence stems represent a large proportion of the discourse of mature native speakers and that the typical native speaker knows hundreds of thousands of such units, even though they may only know a few thousand "single morpheme lexical items" (p. 210).

Again, Pawley and Syder (1983) were careful to tread lightly on the Chomskyan consensus of their time. They sought to integrate their lexical theory into the dominant syntactic paradigm. They did so primarily on grounds of processing capacity. Their attempt can best be summarized in the following quote: "What may be an economical or efficient way of organizing knowledge-in-principle may not be efficient for the demands of ordinary language use" (p. 218). They made a persuasive case that in a theory of grammar, complex lexical knowledge can work along with productive syntactic rules to add automaticity to linguistic creativity. Pawley & Syder (1983) recognized the importance of lexical knowledge that is complex, appropriate to certain communicative contexts, and well organized. However, it took others to expound the cognitive mechanisms by which such sophisticated knowledge develops.

Writing at the same time as Pawley & Syder (1983), a cognitive psychologist (Anderson, 1983) proposed an influential model of cognition (See figure 2.1 below) called "Adaptive Control of Thought" (ACT). This theory was ambitious in its attempt to explain all complex cognitive skills, including language, as manifestations of the same set of principles. This theory represented a break with the Chomskyan generative paradigm, which held language as a privileged cognitive process distinct from other cognitive processes. Anderson's notion of language as following the same set of principles as all other cognitive processes aligned him with the functional linguists (Firth, 1957; Halliday, 1975). More specifically, Anderson viewed language as a cognitive process that develops as the user performs communicative tasks and interacts with the world, which was consistent with the functionalist view of language as a system of meaning potential that evolves as language users function in the social world.

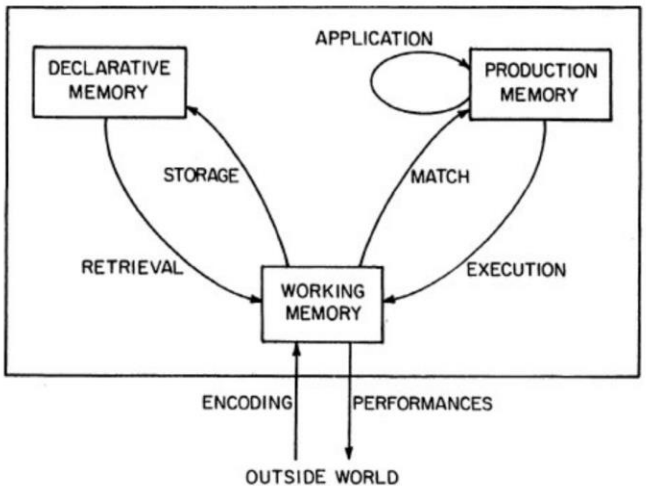

Figure 2.1. Anderson's (1983) ACT Model

One of Anderson's principles was that learning involves two types of knowledge: "declarative" and "procedural" (p. 19-20). Declarative knowledge (which is the form that all knowledge takes when it comes into being) is knowledge of facts, and procedural knowledge is knowledge of how to perform certain tasks. Declarative knowledge, which is stored in the form of chunks of information, is knowledge that X is true. Procedural knowledge takes the form of a series of "condition-action pairs" (p. 6). These condition-action pairs are If-Then statements that include a condition: if X is true, then do Y. Anderson called such sets of procedures "productions" (p. 19). These productions are stored in long-term memory, which is the system's principal storehouse of both declarative and procedural knowledge. Productions can be retrieved from long-term memory by working memory, which is the part of long-term memory that is active at any particular moment.

A simple example from daily life will clarify how the ACT Model works. Years ago, when a driver traveled to a new location in a large city, she had to first look at a map. The address of the new place and its location on the map are declarative knowledge. Driving to the location for the first time requires conscious attention, and the driver may get lost if she loses focus. The driver may even need to pull over at some point to look at the map. However, after traveling to the same location several times, she learns shortcuts and better ways to navigate traffic. These strategies are stored in her memory as procedural knowledge. In other words, she

reorganizes information in her long-term memory in such a way that the task becomes simpler, and she is able to perform the driving task more efficiently.

In this model, learning takes place through use and storage of knowledge in long-term memory. According to ACT, declarative and procedural knowledge work together with memory and the outside world. As information from the outside world is received, matched to patterns in existing knowledge, and used to perform tasks, memory units are created, and connections between memory units are strengthened. Automaticity of processing and thus more efficient task performance is the result of strengthening of connections.

One distinction between declarative and procedural knowledge is particularly relevant to L2 oral fluency. That distinction is that the use of declarative memory requires attention, while the use of procedural memory is more automatic, requiring little to no attention. Hence, conversion of declarative knowledge to procedural knowledge is the primary means by which individuals become more efficient at performing tasks. Returning to the discussion of L2 oral fluency, the means by which the L2 speaker converts declarative memory to procedural memory remains an open question. In other words, psycholinguists still have much to learn about how L2 speakers reorganize L2 knowledge in a way that it can be used more efficiently to produce L2 speech. However, one model in particular has provided a great deal of clarity on this question.

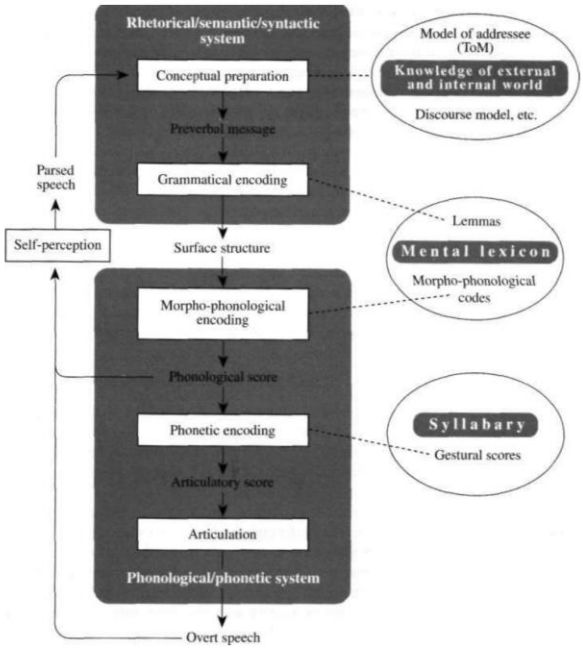

Figure 2.2. L2 Speech Production Model (De Bot, 1992; Levelt, 1989, 1999)

Levelt (1989, 1999) applied Anderson's work to the problem of L1 speech production, creating the influential Speech Production Model, which De Bot (1992) later adapted to L2 speech production. The adapted L2-model, which Levelt (1999) accepted, became the dominant theoretical model guiding research on L2 speech production, including the strand of L2 oral fluency. The model, which is also known as "the blueprint for the speaker as information processor" (p. 9), borrowed some of Anderson's (1983) ideas, including two kinds of knowledge: "declarative knowledge" and "procedural knowledge" (p. 10-11). In the model, the two different kinds of knowledge are processed in four phases (see figure 2.2 above): the “conceptual preparation” (p. 87), the formulation of the “pre-verbal message” (p. 87), “grammatical encoding” (p. 88), "articulation" (p. 88), and the “self-perception” (p. 88).

Before speech takes place, the speaker forms a communicative purpose based on declarative knowledge of the situation at hand and procedural knowledge of what language should be expressed. The formulator converts the communicative purpose into language structures and applies declarative knowledge of what language structures mean and procedural knowledge of how to use language structures to express the desired meaning. As can be seen in figure 2.2, in the grammatical encoding phase, the speaker retrieves lexical chunks that are stored in the long-term memory. In grammatical encoding, the speaker then forms utterances according to the syntactic rules that can be stored as declarative knowledge or procedural knowledge, depending on the speaker's current state of language proficiency.

In articulation, coordinates the phonological apparatuses necessary for producing the speech sounds needed to express the original communicative intent. As the speaker produces speech, the self-perception system processes what is being said, always comparing it to the original communicative intent. The Levelt Model is important because it effectively encapsulates the complex processes and sub-processes of speech production. This model also provided testable hypotheses about how oral fluency works.

This section laid out some of the fundamental theoretical concepts underlying oral fluency. Goldman-Eisler (1958a, 1958b) connected hesitation phenomena to structural constraints inherent in language and the informational content of words. She theorized that when foregoing language in an utterance constrains the speaker's freedom to choose the next word, the speaker speaks fluently. When she has the freedom to choose and the need to choose words of great informational content she requires time to do so, and the result is hesitation.

Pawley & Syder (1983) took up the idea of choice (they called it "selection") and connected it to the native speaker's lexical knowledge beyond the word level. Their solution to the problem of choice, posed by Goldman-Eisler, was a large, well-organized storehouse of "lexicalized sentence stems" to express the particular cultural idea that is appropriate to any given communicative context. They recognized the human processing limitations and the difficulty of Chomsky's (1957) generative grammar in dealing effectively with those limitations. In response they offered a reasonable supplement to productive syntactic rules.

Anderson (1983) offered a general model of cognition that explained how complex lexical knowledge like that described by Pawley & Syder (1983) could develop. Anderson's ACT model has been invaluable to the theory behind oral fluency. The ACT model provided a framework that enabled linguists and applied linguists to think in terms of processing, working memory, declarative and procedural knowledge, and the limitations thereof. This development led to some of the more sophisticated research into language acquisition in general and second language acquisition in particular, which is discussed more in the next section.

Levelt (1989, 1999) adapted Anderson's ACT model to the specific problem of speech production. His model provided testable hypotheses and raised questions about the relationship between abstract knowledge and observable language production. This did not resolve the question of whether oral fluency development is more closely associated with conceptual, syntactic or lexical knowledge. However, this model formalized the cognitive foundations of oral fluency development.

2.3 L2 Oral Fluency

The oral fluency of speakers matters a great deal in terms of listener perception of oral proficiency. We see this intuitively in our daily interactions with others. When a conversation partner or a public speaker exhibits frequent pauses, slow delivery, and unnecessary repetitions, restarts, and self-corrections, we quickly lose patience with the amount of effort necessary to decipher a speaker's meaning. L2 speakers are no exception to this rule. Communication breakdown is often the result of disfluent speech. Of course, we do not need to rely on intuition to know that L2 oral fluency is important. L2 oral fluency-related constructs are well defined, and there is a large body of research that supports the relationship between objective measures of L2 oral fluency, listener perception of L2 oral fluency, and objective measurement of L2 oral

proficiency. Section 2.3 presents some definitions related to L2 oral fluency and discusses at length some of the most influential empirical research on L2 oral fluency.

2.3.1 Definitions of Fluency from ESL/SLA

Lennon (1990) defined fluency as "an impression on the listener's part that the psycholinguistic processes of speech planning and speech production are functioning easily and efficiently" (p. 391). Rossiter, Derwing, Manimtim, and Thomson (2010) defined fluency as "flow, continuity, automaticity, or smoothness of speech" (p. 584). In terms of L2 pedagogy, Nation (2007) applied fluency to all four skills (reading, writing, listening and speaking). He described fluency practice as repetitive speed and automaticity exercises in which the language learner processes and comprehends or formulates and produces already known language under time pressure.

Lennon (1990) was one of the first scholars to study oral fluency systematically in the EFL context. He drew an important distinction between the two senses of oral fluency in the EFL context: "the broad sense" (p. 389) and the "narrow sense" (p. 389-90). He viewed fluency in the broad sense as synonymous with oral proficiency. On the other hand, he described the narrow sense as one objective criterion among many that experienced second language teachers and raters consider when assessing spoken language. Lennon recognized that fluency played an important role in listener perception of L2 oral proficiency. However, noting the rather impressionistic nature of fluency, he lamented that there were no commonly used quantitative measures for gauging oral fluency. To fill this gap, he proposed and popularized the use of "temporal measures" (p. 399). Lennon explained the purpose of temporal measures thus:

> The development of such a set of measures that might function as benchmarks for oral proficiency would help expand our understanding of the ingredients of fluency and could also serve an important diagnostic function for teachers and learners. (p. 399)

It should be noted that most of these quantitative measures involve time, which is why they are called "temporal measures". In fact, temporal measures of oral fluency have become an integral part of oral fluency measurement because time is an essential element of listener expectations. Of course, expectations play an important role in language in general.

Oller (1974) argued "that our ability to anticipate elements in a sequence is the foundation of all language skills" (p. 444). He meant by this that humans learn highly complex systems of meaning that are encoded in language, and expectations of the order in which linguistic elements are to be presented play an immense role in the success of a particular communicative act. He gave a few examples in support of his argument that sequence is essential to language. For example, he noted that we expect sounds to be articulated in a particular order to produce spoken words; we expect words to be arranged in a specific sequence to form grammatical sentences; and we expect sentences to be placed in a specific order to form a coherent paragraph.

Clark and Tree (2002) applied similar principles of language use expectations to oral fluency in conversation. They argued that listeners can more easily maintain their attention to an utterance and identify expressions when they come "at the expected moment" (p. 8). Moreover, in their view, speakers aim to provide the ideal delivery out of a desire to meet the listener's timing expectations. Clark & Tree (2002) defined the ideal delivery as "the way they (speakers) would have wanted to produce it (the utterance) if they had no problems" (p. 7).

They went on to make the case, based on their research on conversation analysis (CA), that various aspects of language that are intended to delay delivery can signal to the listener the speaker's intent to delay. They gave the examples of filled pauses "uh" and "um", the former, according to their analysis, signaling a short delay and the latter a longer delay. Another example of signaling delay that they gave was the use of non-reduced vowels in function words. For example, in the utterance "I would have to go down to *the*-- film school and talk to some of the people there" (p. 8), one of their participants used the high front unrounded vowel, /i/, in the pronunciation of the definite article *the* (rhyming with *see*) to signal a delay before the noun *film school*. Vowel elongation is just one of many ways that speakers can signal their intentions.

The point is that shared expectations of timing, delivery, and signaling between the speaker and the listener shape perception of oral fluency. Moreover, aspects of fluency and disfluency related to those shared expectations are evident in temporal measures of oral fluency. To cite an example given above, using filled pauses takes time without adding meaningful syllables, and thus reduces speech rate. To cite the other example given above, signaling delay by failing to reduce vowels in function words directly reduces articulation rate by increasing the amount of time it takes the speaker to articulate those function words. Sensitivity to common

fluency-related features, as well as linguistic features, is what makes temporal measures such good measures of language proficiency, as is discussed in more detail later in section 2.3.4.

Another important point that is discussed in more detail in section 2.3.4 is the importance of studying fluency as a multi-faceted construct. Skehan (2009) developed an influential framework for categorizing temporal measures. He proposed that the skill of oral fluency be broken down into three separate components: "speed, breakdown, and repair" (p. 512-3). Speed relates to the amount of language produced within a particular time period (within speech runs or within a response overall); breakdown refers to pausing (silent, filled, mid-clause, etc.); and repair has to do with the act of editing one's utterances in real time (repetitions, restarts, and self-corrections). Of course, while the different components of oral fluency can be separated in the realm of theory, in practice they are highly inter-related constructs.

This inability to disentangle the different aspects of complexity, accuracy, and fluency (CAF) led Norris and Ortega (2009) to advocate an "organic approach" (p. 557) to the study of CAF. They argued that CAF constructs are dynamic systems, each made up of multiple sub-systems that develop over time in inter-related ways that may not be completely predictable. In light of this claim, they asserted that researchers should measure multiple aspects of CAF constructs in order to provide empirical evidence that the theoretical claims made about second language development are supported. Norris and Ortega (2009) also argued that measurement should take into consideration the fact that some CAF measures are redundant and hence measurement should endeavor to minimize this redundancy.

In keeping with this organic approach to CAF development, the present dissertation examined the longitudinal development of multiple components of L2 oral fluency, which are represented by eight different temporal measures of oral fluency. While these measures overlap to some extent, there is a body of research that suggests that they do represent different components or sub-systems of a larger developing system. Furthermore, findings that the different components of the sub-system, as represented by pre-post differences in some or all of these eight measures, change at different rates over the same time period provide evidence that some of the theoretical claims related to oral fluency apply to L2 learners.

Norris and Ortega (2009) also emphasized the importance of "task specifications, behavior elicitation, and learning context" (p. 557-8) in the measurement of CAF. They deemed all of these factors essential to measuring effectively and drawing appropriate conclusions. This

point is quite relevant to the present dissertation, which includes oral fluency measures that have been studied in many other studies. That said, further investigation is warranted because an organic approach to the study of CAF development does not take for granted that empirical findings supporting theoretical claims apply to L2 learners regardless of the learners' L1, L2, learning context, and purpose for learning. Neither does an organic approach assume that one set of empirical findings supports theoretical claims without regard to task specifications or behavior elicitation methods. This is important because the present study used different task specifications than other studies to elicit oral responses from L2 learners (more on this point in Chapter 3: Methodology).

Because L2 oral fluency is a multi-faceted construct, many different oral fluency measures have been studied in relation to longitudinal development and listener perception thereof. The following table presents a detailed but non-exhaustive list of temporal measures of oral fluency that have been used over the past three decades to study L2 oral fluency. It is broken down based on the different aspects of oral fluency described by Skehan (2009): speed, breakdown, and repair. Note that some variables are composite measures, meaning that they include more than one aspect of oral fluency.

Table 2.1. Oral Fluency Variables Over the Past Three Decades

Category	Variables	Description	References
Amount	amount of speech	number of words in a response	(Kormos & Dénes, 2004; Riggenbach, 1991)
Speed	Articulation Rate (AR)	number of syllables divided by speech time (response time minus filled pause time and silent pause time)	(Ginther et al., 2010; Kormos & Dénes, 2004; Towell et al., 1996; Ushigusa, 2008; Van Gelderen, 1994)
Speed	Mean Syllable Duration (MSD)	phonation time divided by number of syllables	(De Jong et al., 2013; Ginther et al., 2010; Huensch & Tracy-Ventura, 2017)
Speed, Breakdown	Speech Rate (SR)	number of syllables divided by response time	(Derwing et al., 2004; Ginther et al., 2010; Goldman-Eisler, 1968; ; Huensch & Tracy-Ventura, 2017; Iwashita et al., 2008; Kormos & Dénes, 2004; Leaper & Riazi, 2014; Lennon, 1990[1]; Riggenbach, 1991; Towell et al., 1996; Ushigusa, 2008; Van Gelderen, 1994)
Speed, Breakdown	Mean Length of Hesitation-free Run	number of syllables divided by number of speech runs without silent pauses of more than .40 seconds	(Segalowitz & Freed, 2004)
Speed, Breakdown	Mean Length of Filler-free Run	number of syllables divided by number of speech runs without a filled pause	(Segalowitz & Freed, 2004)
Speed, Breakdown	Longest filler-free run	longest speech run (in syllables) without a filled pause	Segalowitz & Freed (2004)
Speed, Breakdown, Repair	Pruned Syllable Rate	number of syllables, not including "self-corrections, self-repetitions, false starts, non-lexical filled pauses, and asides" (Derwing et al, 2004, p. 665) divided by response time in seconds	(Derwing et al., 2004; Lennon, 1990; Rossiter, 2009)

[1] Lennon (1990) and Riggenbach (1991) used words instead of syllables as the production unit.

Table 2.1 continued

Category	Variables	Description	References
Density, Fluidity	Phonation time ratio (PTR)	phonation time (response time excluding silent and filled pause time) divided by response time	(Ginther et al., 2010; Kormos & Dénes, 2004; Towell et al., 1996; Ushigusa, 2008)
Density, Fluidity	Mean Length of (Speech) Run	number of syllables divided by number of speech runs	(Derwing et al., 2004; Hasselgren, 2002; Iwashita et al., 2008; Kormos & Dénes, 2004 ; Lennon, 1990; Towell et al., 1996; Ushigusa, 2008)
Density, Fluidity	Mean Length of Utterance	number of words per speaking turn (in dialogic tasks)	
Breakdown	Number of Silent Pauses	number of periods of silence of at least .25 seconds[2]	(Ginther et al., 2010; Goldman-Eisler, 1968)
Breakdown	Silent Pause Ratio	"silent pause time as a decimal percent of total response time" (p. 387)	(Ginther et al., 2010)
Breakdown	Pause Ratio	number of silent pauses of one second or more divided by speaking time	(Leaper & Riazi, 2014)
Breakdown	Silent Pause Time	"total time in seconds of all silent pauses in a given speech sample" (p. 387)	(Ginther et al., 2010)
Breakdown	Number of Filled Pauses	number of filled pauses	(Ginther et al., 2010)
Breakdown	Filled Pause Time	"total time in seconds of all filled pauses in a given speech sample." (p. 387)	(Ginther et al., 2010)

[2] In all studies except Riggenbach (1991) and Towell et al. (1996), the silent pause length threshold was set at .25 seconds. The former study used separate thresholds for separate pause length variables, while the latter used a threshold of .28 seconds, citing technical constraints.

Table 2.1 continued

Category	Variables	Description	References
Breakdown	Mean Silent Pause Length (MSP)	silent pause time divided by number of silent pauses	(Bosker et al, 2012; Goldman-Eisler, 1968; Kormos & Dénes, 2004; Towell et al., 1996)
Breakdown	Mean Filled Pause Length	filled pause time divided by number of filled pauses	(Ginther et al., 2010)
Breakdown	Silent Pauses per Second (SPS)	number of silent pauses divided by response time in seconds	(Derwing et al, 2004; Huensch & Tracy-Ventura, 2017; Kormos & Dénes[3], 2004; Rossiter, 2009)
Breakdown	Silent Pauses per 100 word	100 times the number of silent pauses, divided by number of words in the response	(De Jong et al, 2013)
Breakdown	Filled Pauses per 100 word	100 times the number of filled pauses, divided by number of words in the response	(De Jong et al., 2013)
Breakdown	Silent Pauses per Second Spoken	number of silent pauses divided by phonation time (response time excluding silent and filled pause time)	(Bosker et al., 2012)
Breakdown	Filled Pauses per Second Spoken	number of filled pauses divided by phonation time (response time excluding silent and filled pause time)	(Bosker et al., 2012)
Breakdown	Filled Pauses Per Minute	number of filled pauses divided by response time in minutes	(Kormos & Dénes, 2004)
Breakdown	Micro-pauses	number of silent pauses "of .2 seconds or less" (p. 426)	(Riggenbach, 1991)
Breakdown	Hesitations	number of silent pauses "of .3-.4 seconds" (p. 426)	(Riggenbach, 1991)

[3] Kormos & Dénes (2004) used minutes as the unit of time.

Table 2.1 continued

Category	Variables	Description	References
Breakdown, Location	Mean Silent Pause Duration between AS-Units	silent pause time between AS-Units divided by number of silent pauses between AS-Units	(Huensch & Tracy-Ventura, 2017)
Breakdown	Unfilled Pauses	number of silent pauses "of .5 seconds or greater" (p. 426)	(Riggenbach, 1991)
Breakdown	Filled Pauses	number of "voiced fillers, which do not normally contribute additional lexical information" (p. 426)	(Riggenbach, 1991)
Repair	Disfluencies per minute	number of "repetitions, restarts, and repairs" (p. 152) divided by response time in minutes	(Kormos & Dénes, 2004)
Repair	Repetitions per T-Unit	number of repeated words divided by number of T-Units (Hunt, 1970)	(Lennon, 1990)
Repair	retraced restarts	number of "reformulations in which part of the original utterance is repeated" (p. 427)	(Riggenbach, 1991)
Repair	unretraced restarts	number of "reformulations in which the original utterance is rejected (= false start)" (p. 427)	(Riggenbach, 1991)
Repair	Repetitions per 100 word	100 times the number of repetitions, divided by number of words in the response	(De Jong et al., 2013)

Table 2.1 continued

Category	Variables	Description	References
Repair	Self-corrections per 100 word	100 times the number of self-corrections, divided by number of words in the response	(De Jong et al., 2013)
Repair	Repetitions per Second	number of repetitions divided by response time in seconds	(Derwing et al, 2004; Huensch & Tracy-Ventura, 2017)
Repair	Restarts per Second	number of restarts divided by response time in seconds	(Huensch & Tracy-Ventura, 2017)
Repair	Maze Ratio	number of unnecessary repetitions, false starts and self-corrections divided by number of words	(Leaper & Riazi, 2014)
Repair	Corrections per Second Spoken	number of corrections divided by phonation time (response time excluding silent and filled pause time)	(Bosker et al., 2012)

2.3.2 Longitudinal Change in L2 Oral Fluency

Lennon (1990) conducted a pre-post analysis of oral fluency development in the speech of L1 German L2 learners of English (N=4) spending five months at a university in England. He collected data by means of a picture series narration task, which he administered twice, five months apart. He measured 12 different variables, which are characterized below according to Skehan's (2009) categorization scheme. These variables included three speed variables (mean length of speech run, words per minute, pruned words per minute), three repair variables (repetitions per T-unit, self-corrections per T-unit, & percentage of repeated and self-corrected words), six breakdown variables (filled pauses per T-unit, silent pause ratio, filled pause ratio, percentage of T-units followed by a pause, percentage of total pause time at all T-unit boundaries, and mean pause time at T-unit boundaries). Lennon conducted a one-tailed paired sample t test to test whether the participants' fluency changed over time. Of these twelve variables, three variables showed a significant pre-post change in the oral responses of the participants. Pruned words per minute (a speed variable) increased, while filled pauses/T-unit and percentage of T-units followed by a pause (two breakdown variables) decreased. This study suggested that L2 speakers' speed of delivery (adjusted for disfluency occurrences) can increase over a period of time as short as five months, and L2 speakers' pausing frequency can decrease over the same time period.

Taking a more theoretical approach, Towell, Hawkins, and Bazergui (1996) studied the longitudinal development of second language French oral fluency. To test a hypothesis based on Levelt's (1989) Speech Production Model, the authors used a quantitative and discourse analysis approach to study the pre-post longitudinal change in the L2 fluency of university level L1 English learners (N=12) of L2 French over a year spent studying abroad in France. The speech-elicitation task was a narrative retelling of a story. Each subject watched a short movie and then summarized the plot of the movie in French. They conducted the same procedures at time 1 (T1) in their L1 (English). The authors used paired sample t tests to compare each individual's L2 temporal fluency at time 2 (T2) to that at T1 and to measure each speaker's L1 fluency. They measured five temporal measures of fluency: speech rate, phonation/time ratio, articulation rate, mean length of run, and mean length of silent pause.

They used quantitative analysis of temporal fluency measures to test their hypothesis regarding second language oral fluency development. They based their theory on Anderson's

(1983) ACT Model (discussed in the previous section) and Levelt's (1989) Speech Production Model (discussed in the previous section). They based their research design on the assumption that increased automaticity of speech production is the result of "the conversion of declarative knowledge into procedural knowledge" (p. 90) at some phase in the Levelt Model. They were primarily concerned with determining at which phase in the model the most proceduralization takes place: the conceptualizer, the formulator, or the articulator.

More specifically, Towell et al. (1996) hypothesized that if a speaker's speech rate increased without mean length of silent pause increasing, this could be considered evidence of proceduralization of speech processes. Furthermore, if the subjects' mean length of speech run lengthened more than their articulation rate increased, then it could be assumed that more proceduralization took place in the formulator (the lexico-grammatical encoding phase); if the results turned out the other way around, then the implication would be that the articulation phase underwent a greater degree of proceduralization.

The first finding was that, unsurprisingly, the subjects' L1 fluency was higher than their L2 fluency at time one. Furthermore, their L2 fluency did not increase to the level of their L1 fluency, even after one year immersed in the target language. The longitudinal findings of the L2 phase of the study were that speech rate, mean length of speech run, articulation rate, and phonation time ratio were all associated with statistically significant pre-post increases. However, the change in mean length of silent pause was not statistically significant. Therefore, their hypothesis that longitudinal development of oral fluency is characterized by proceduralization of the speech processes was supported by the findings. Evidence of this was the fact that participants increased their speech rate without spending more time in utterance planning during pauses. The results showed that most of the increase in speech rate could be attributed to an increase in mean length of speech run, which was consistent with more proceduralization taking place in the lexico-grammatical encoding phase of Levelt's (1989, 1999) Speech Production Model.

In the the qualitative phase of the study, Towell et al (1996) analyzed the 12 subjects' responses in a pre-post comparison for lexical and syntactic complexity, following Nattinger and Decarrico's (1992) coding scheme. Having found that the subjects' improvement in mean length of speech run was the largest contributor to their oral fluency gains, they focused on the responses of subjects who improved their mean length of speech run considerably. They did so in

hopes of finding the linguistic source of proceduralization. They found that those subjects who increased their mean length of speech run the most did so by increasing their use of collocations as well as syntactic complexity.

Segalowitz and Freed (2004) extended this idea that L2 oral fluency development is related to development in other linguistic skills, namely, grammatical and lexical ability. They conducted an experimental study of longitudinal development of L2 Spanish oral fluency and various other linguistic variables. The researchers used a pre-post experimental research design, collecting data over the course of a 13 week semester from two groups of L1 English students (N=40). The control group (n=18) studied Spanish in a traditional university classroom setting, while the experimental group (n=22) studied Spanish at a university in Spain. The researchers collected oral data by administering the ACTFL/ETS Oral Proficiency Interview (OPI).

Segalowitz and Freed (2004) analyzed the OPI data for four oral fluency variables, and Collentine (2004) analyzed the same data for grammatical ability and lexical ability. To measure oral fluency, grammatical ability, and lexical ability, the authors extracted two two-minute portions of student speech from the same two points in the timeline of each interview at time 1 and time 2. The oral fluency variables included speech rate, mean length of hesitation-free runs, mean length of filler-free runs, and longest filler free run. As stated above, Collentine analyzed the OPI data for grammatical ability and lexical ability and Segalowitz and Freed(2004) cited it from another study (Collentine, 2004), which included a more fine-grained analysis. The oral fluency analysis consisted of "analysis of variance with the between group factors being Context (At Home, Study Abroad) and Oral Gain (Gain, No gain)" (p. 11). Collentine (2004) based grammatical ability on "17 measures of morphological, syntactic, and morphosyntactic structures at pre-test and post-test" (p. 6). First, they calculated accuracy percentages for each student on each grammatical structure at T1 and T2. Then, they calculated pre-post accuracy percentage gains for each student. Next, a discriminant analysis was performed to determine whether the accuracy percentage gains distinguished at a statistically significant level between the two groups (study abroad and stay at home). Lexical ability was operationalized by counting the number of unique words in seven different parts of speech (in the OPI oral data). Next, pre-post gains in this lexical variable were calculated for each participant. Then, a discriminant analysis was

conducted, similar to the one described above for grammatical ability. The authors also administered the SAT II Spanish subject test (not including the listening section) to both groups at T1 and T2.

Results showed that the experimental (study abroad) group made statistically significant gains with regard to three of the four oral fluency variables, while the control (stay at home) group showed no significant gains in any oral fluency variable. The experimental group's largest oral fluency gains were associated with speech rate, and their gains in mean length of filler-free runs and longest fluent run were also statistically significant. For grammatical accuracy, the at home group made larger gains than the study abroad group in 5 of the 17 variables but the discriminant analysis indicated that the gains for the other 12 grammatical accuracy variables did not distinguish between the two groups at a statistically significant level. For lexical ability, the gain in number of unique words distinguished between the two groups at a statistically significant level for only one part of speech category (adjectives); the at home group made more gains in number of unique adjectives than the study abroad group.

Other findings from this study are notable. First, a statistically significant majority of the members of the experimental group achieved a higher score on the Oral Proficiency Interview, while few students in the control group increased their score. On the other hand, the control group made gains in their score on the SAT II Spanish test, which mostly assesses grammar and vocabulary, while the experimental group made no significant gains in SAT II Spanish test scores. The authors concluded that, based on the findings, immersion in the target language tends to improve oral English proficiency and oral fluency but not grammatical and lexical knowledge more than traditional classroom language study.

However, fine-grained analysis of the grammatical ability development in this study was relevant to the Speech Production Model. First, the discriminant analysis indicated that the study abroad group increased its coordinate clause count significantly more than the at home group did. Collentine (2004) argued "that the increase in the production of coordinate clauses is most likely an artifact of the fact that the (study abroad group) increased its fluency during the treatment period, producing more words per segment" (p. 240). Also relevant to syntactic development is the fact that both the at home group and the study abroad group increased their subordinate clause counts, but the increases did not distinguish between the two groups. Since both groups increased their subordination by indistinguishable magnitudes, while the group that

made the largest fluency gains increased their coordination by a larger magnitude, these findings suggest that for students at this proficiency level, coordination tends to facilitate oral fluency development more than subordination does.

Huensch and Tracy-Ventura (2017) extended the study of L2 oral fluency development beyond the study abroad period to examine the nature of fluency development before the study abroad period and attrition after returning home from a study abroad experience. They examined the longitudinal development of L2 Spanish oral fluency in a group of L1 English learners (N=26) over a 21-month period, starting six months before a cohort of students studied abroad in Spain and ending six months after their return. The participants, who were all Spanish majors, were continuously enrolled in traditional Spanish language classes for the two-year period before studying abroad. There was no control group in this study. Data were collected via a picture series narration task. Data collection occurred once six months before arrival, again upon arrival, twice at three-month intervals during the study abroad period, and twice at three-month intervals after returning home. The authors purposefully included dependent variables from all three categories of oral fluency variables, based on Skehan's (2009) categorization scheme (presented in order below): three speed variables, four breakdown variables, and two repair variables. The dependent variables were the following: mean syllable duration, speech rate, mean length of speech run, silent pauses/second, filled pauses/second, mean silent pause duration within AS-units, mean silent pause duration between AS-units, restarts/second and repetitions/second. An AS-unit is defined by Foster, Tonkyn and Wigglesworth (2000) as "a single speaker's utterance consisting of an independent clause or sub-clausal unit, together with any subordinate clause(s) associated with it" (p.365). This measure was created as an alternative to the T-unit in spoken discourse to better suit the nature of spoken discourse.

The authors used the Friedman test to determine the size and statistical significance of oral fluency differences between time points. The Friedman test, which is the non-parametric alternative to the repeated measures ANOVA test, is a statistical test that is used to measure differences when variables of a related sample are measured more than twice. In other words, it is useful when measurements are taken from the same participants more than twice, and the statistical assumptions of repeated measures ANOVA are not met.

Results showed that the participants initially improved their articulation a great deal, represented by a large, statistically significant decrease in mean syllable duration during the pre-

study abroad period. They also increased their overall fluency, represented by speech rate in the pre-study abroad period. However, they did not increase their lexico-syntactic formulation, represented by a non-statistically significant change in mean length of speech run in the pre-study abroad period. During the study abroad period, all three speed variables, mean syllable duration, speech rate, and mean length of speech run were associated with statistically significant gains[4], with mean length of speech run associated with a smaller, more gradual gain than the other two speed variables. Furthermore, the participants retained their elevated speech rate and mean length of speech run, as evidenced by the fact that these variables did not decrease significantly between the times that the cohort arrived home and six months after return. Mean syllable duration, on the other hand, did increase significantly upon return, suggesting that improvement in articulation speed was not long-lasting. In contrast to mean length of speech run, gains in no other variable were retained after students returned home. This finding suggests that mean length of speech run represents a broad construct involving powerful underlying constructs.

Mora and Valls-Ferrer (2012) studied the L2 oral fluency, complexity, and accuracy gains made by 40 L2 English learners at three different time points. In addition to gains made over six months of focused English instruction at their home university in Spain, the authors reported gains made studying abroad in the United Kingdom. The authors also compared the L2 learners' CAF measures to those of L1 English speakers. The oral fluency measures included speech rate, mean length of speech run, articulation rate, phonation time ratio, disfluency ratio, pause frequency, and pause time ratio. The accuracy measures were error-free AS unit % and errors per AS unit. The complexity measures included a measure of lexical richness (Guirad's index), a measure of lexical density (lexical word ratio), clauses per AS unit, and mean length of AS unit.

Results showed that the learners made improvements (progressing towards the L1 English speakers) with regard to each and every fluency and accuracy measure, and they made gains on the lexical richness measure and mean length of AS unit. Furthermore, across all measures in which participants made gains, they made more gains during the study abroad period, even though it was half as long as the stay at home period. Furthermore, L2 speakers

[4] It should be noted that for mean syllable duration a decrease is considered a "gain" because this variable is negatively correlated with speed fluency.

differed significantly from L1 speakers at all time points with regard to all dependent variables except mean length of AS-units at T3.

The only study examining the longitudinal development of L2-Chinese was Kim et al. (2015). In this study, data were collected via an oral proficiency interview from 22 L1-English learners of Chinese twice, once before and once after the study abroad semester. Dependent variables fell into three categories: oral fluency, tonal accuracy, vocabulary, and task fulfillment. Oral fluency variables included speech rate, filled pauses per minute, unfilled pauses per minute, mean length of unfilled pauses, and a holistic fluency rating conducted by trained L1-Chinese raters using a ten-point fluency scale. The researchers measured tonal accuracy by counting the number of syllabaries whose tones were correct and divided by total number of syllabaries. There were three vocabulary variables: number of types, number of tokens, and type/token ratio. The same trained raters used a ten-point scale to rate each participant on whether they answered each oral interview question.

Results showed that participants exhibited statistically significant increases on speech rate and fluency rating, while showing significant decreases in all pausing variables. Moreover, they significantly increased their tonal accuracy and the number of unique words (types), but they significantly decreased their type-token ratio. Finally, they improved significantly on their task-fulfillment rating. The most interesting finding here relates to vocabulary; even as participants produced a wider range of vocabulary, they also tended to reuse the same words.

I will now interpret the results of this study with reference to the Speech Production Model. The fact that mean syllable duration (which is negatively correlated with articulatory efficiency) initially decreased the most suggests that mean syllable duration (and by extension, its reciprocal: articulation rate) represents a fairly narrow mechanical construct: efficiency of the functioning of the articulatory organs in the mouth. It makes sense that this skill develops quickly with increased practice and atrophies quickly when practice is less frequent. The fact that mean length of speech run developed more slowly also makes sense when we consider the nature of the formulator, for which Towell et al. (1996) used mean length of speech run as a proxy. Development in the formulator is a more complex phenomenon, entailing proceduralization of lexical knowledge, syntactic knowledge, and probably the integration of the two. Returning to the results, the initial decrease in mean silent pause duration within AS units may partially reflect the improvement in articulation just discussed. It is also consistent with an increase in lexical

knowledge. To clarify, as learners learned more new vocabulary in the pre-study abroad period, they paused less within each syntactic unit to retrieve a word. The fact that learners did not increase their mean length of speech run until the study abroad period began is consistent with the process of putting their newly learned vocabulary into use in an immersion setting. Proceduralization within the formulator probably takes place at the intersection of vocabulary and syntax, a process for which a threshold level of vocabulary is necessary but not sufficient. Integrating second language lexical knowledge with syntax is likely to require time spent immersed in the L2. This may be why the participants in Huensch and Tracy-Ventura (2017) did not increase their mean length of run in the at home phase of the study, and the at home group in Segalowitz & Freed (2004) did not increase their mean length of run, while the study abroad group in Segalowitz and Freed (2004) did.

Moving on to breakdown fluency, duration of pausing between AS-units may represent functioning of the conceptualizer of the Speech Production Model; with practice, participants can more quickly think of something else to say when they finish a thought, but as Skehan (2009) found, L1 speakers often pause to plan utterances between AS-units, so it is probably unrelated to language proficiency.

2.3.3 Cross-Sectional Oral Fluency Studies

Riggenbach (1991) conducted one of the earliest cross-sectional exploratory studies of L2 oral fluency to seek out the characteristics of L2 speech perceived fluent versus that perceived as non-fluent. This study included quantitative and qualitative analysis, but only the quantitative analysis will be discussed. The data analyzed were the recorded, naturally occurring conversations of L1 Chinese (N=6) participants: three perceived as very fluent and three very non-fluent. For their ESL classes the participants recorded their own conversations with L1 English interlocutors. No topic was assigned for the conversations. The participants' fluency level was holistically assessed by ESL teachers (N=12), which resulted in the very fluent/very non-fluent groups. Fluency occurrences that were traditionally associated with disfluency were identified and examined in context to determine what if any legitimate communicative function they had. Fluency occurrences included "micro-pauses", "hesitations", "pauses", "unfilled pauses", "filled pauses", "retraced restarts", and "unretraced restarts", "rate of speech", and "amount of speech" (p. 426-7). To compare the frequency of occurrences of each of the

variables, Riggenbach used a Mann-Whitney U/Wilcoxon Rank Sum[5], which is a nonparametric alternative to the independent t-test.

The quantitative analysis of these fluency occurrences showed that only two variables were associated with statistically significant differences between the highly fluent and highly non-fluent group: rate of speech and number of unfilled pauses. The fluent group exhibited a higher rate of speech and fewer unfilled pauses than the non-fluent group.

Riggenbach's quantitative findings had one very important implication related to the coding of oral fluency variables. Her practice of placing pauses of different length in separate pausing categories that amounted to different variables was not ideal. She found that, of these pausing variables, only the number of pauses that were .5 seconds or longer distinguished fluent L2 speakers from non-fluent L2 speakers. The arbitrary temporal cutoff points for pausing variables probably diminished the power of each pausing variable to distinguish proficiency levels. For the same reason, making fine distinctions between different kinds of restarts may have also diminished the power of each restart variable to distinguish the fluent from the non-fluent speakers. Most future studies of L2 oral fluency that included pausing variables would choose a single cutoff (usually .25 seconds) for silent pauses, including all silent pauses that equaled or exceeded this cutoff and excluding all that were shorter than the cutoff.

One study that applied such a silent pausing cutoff was Kormos and Dénes (2004). This study compared L1 English (N=3) and L2-English teachers' (N=3) perceptions of the oral fluency of L2 English learners. The study consisted of quantitative analysis of fluency measures and qualitative analysis of rater comments on L2 oral responses. The teachers, all of whom except one of the L1 English teachers were experienced language testers, rated the fluency of Hungarian EFL students (N=16). Half of the students were advanced and half lower intermediate (grouped based on language course placement). The participants were prompted to individually make up a story based on a cartoon strip. Each participant spoke for two to three minutes, and his or her speech was recorded and transcribed. The data were analyzed quantitatively based on ten different measures (p. 151-2): speech rate, articulation rate, phonation-time ratio, mean length of

[5] The Mann-Whitney U/Wilcoxon Rank Sum test is used to determine if two randomly selected samples chosen from independent populations differ significantly with regard to some variable. The Mann-Whitney U/Wilcoxon Rank Sum test is often chosen over the independent sample t-test because the former has no requirements with regard to the distribution of the sample.

speech run, number of silent pauses per minute, number of filled pauses per minute, mean length of pauses, number of disfluencies per minute, number of stressed words per minute, ratio of words to stressed words, D value (lexical variety), and *number of words*. Each teacher rated the test-takers' oral English fluency using a holistic scale and then wrote comments about what aspects of language they considered in their rating decisions. The L1 English and L2 English teacher ratings were considered separately, and Spearman rank order[6] correlations were used to calculate the correlation between ratings given and each of the ten measures listed above.

Results were as follows. Of the ten measures mentioned above, the following exhibited statistically significant correlations with ratings: speech rate, phonation-time ratio, mean length of run, mean length of pauses, number of stressed words per minute, D value (lexical variety), and number of words. Temporal measures speech rate and mean length of run correlated strongest and positively with fluency ratings. The quantitative results showed remarkable agreement between the L1 English and L2 English groups. All of the just-mentioned measures were statistically significant for both L1 English and L2 English teachers as distinct groups, and all measures deemed statistically significant by one group were statistically significant for the other group.

When the results of the quantitative findings discussed above and the qualitative analysis of the teachers' written comments were compared, some discrepancies stood out. The most striking finding was that number of stressed words per minute, despite not being mentioned in the teachers' written comments, was as highly correlated, or more so, with ratings than any other measure. Furthermore, both L1 English and L2 English teachers reported having weighed pausing as very important, but only mean length of pause was statistically significant, while no other pause-related measure was statistically significant. Of course, it should be noted that pausing time and frequency of pausing affect speech rate and mean length of speech run, so it is possible that the raters weighted pausing heavily as they claimed to have done, but this heavy weighting did not show up in all of the pausing variables. Finally, the L2 English teachers mentioned naturalness of speech as very important. It could be that stress and length of silent pause were what they meant by "naturalness" (p. 153).

[6] Spearman rank order correlation is a non-parametric alternative to the Pearson Product-moment correlation. Spearman is the more suitable test when one of the variables is in ordinal scale, which the holistic score is in this case.

Another study involving rater perception of fluency, Derwing, Rossiter, and Munro (2004) studied the question of whether temporal measures of fluency could distinguish oral English proficiency levels well among low proficiency test-takers across 3 different tasks. The study also investigated whether oral fluency measures are more closely associated with accentedness or comprehensibility. L1-Chinese advanced beginner ESL students (N=20) participated in two different speaking tasks: picture series narration and monologue. Raters (N=28) without formal training rated 20 samples of speech recorded from the foregoing tasks. The researchers explained the task procedures to the raters and asked them to rate the speech samples on a holistic scale based only on "flow and smoothness" (p. 664) for the fluency ratings. The raters also used a holistic scale to rate comprehensibility and accentedness. The samples were next analyzed to calculate the following temporal measures: mean length of run, speech rate, self-repetitions, pauses, and pruned syllable rate. All of these but mean length of run were reported per second.

Next, the researchers noted that the two measures that differed the most between the two tasks were pauses/sec and pruned syllables/sec. These two measures were then regressed on mean fluency ratings. The resulting model explained well over half of the variation in mean fluency ratings for each of the two tasks. The authors also reported that pruned syllables per second was the strongest predictor of fluency ratings, and repetitions per second was the weakest predictor. Furthermore, the fluency ratings were more strongly and positively correlated with the comprehensibility ratings than with the accentedness ratings.

Derwing et al. (2004) provided evidence for a few important oral fluency propositions. First, temporal measures of fluency are strong predictors of L2 oral fluency ratings, even when untrained raters do the ratings. This is important because it indicates that temporal measures gauge a construct that is noticeable even to untrained raters. Second, this proposition is true for the same L2 speakers completing two different tasks. Finally, the fact that fluency ratings correlate higher with comprehensibility than they do with accentedness has interesting implications for the Speech Production Model. The Speech Production Model claimed that vocabulary and syntax play an important role in oral fluency. Assuming that this claim is true, comprehensibility should be closely associated with oral fluency. The reason is that comprehensibility involves the integration of vocabulary and grammar to form sentences that make sense to the listener. Accentedness has more to do with sound substitutions than it does

with oral fluency. Hence, the correlation analysis is consistent with the Speech Production Model.

Taking this strand a step further, Rossiter (2009) studied the extent to which the perceptions of L2 oral fluency according to three groups of raters of different backgrounds were associated with each other. L1 English language experts (N=6), L1 English non-language experts (N=15), and advanced L2 English users (N=15) rated the oral responses of ESL learners (N=24) for "temporal fluency" (p. 401). The study also examined the longitudinal development of the ESL learners' oral fluency, which was measured by the pre-post change in their fluency ratings over 10 weeks of ESL instruction. The study involved quantitative and qualitative analysis of L2 oral responses. The researcher explained to the raters that they were to take such factors as pausing, speaking pace, self-correction, and a few other fluency measures that could be quantified in terms of ratios with a unit of time in the denominator. The speaking task was a story-telling task that involved describing a picture sequence, which they did twice, with ten weeks between performances. The raters rated each speaker using a holistic scale. As raters listened, they were to make notes of their opinions of the speakers' temporal fluency. The recordings were transcribed so that they could be analyzed quantitatively and qualitatively. The following oral fluency variables were measured: pauses per second and pruned syllables per second. Two methods of statistical analysis were used: Pearson correlation and repeated measures ANOVA. Pearson correlations between fluency ratings and temporal measures of fluency were computed for each rater group. Then, Pearson correlations were computed for inter-rater reliability within and across groups. One way repeated measures ANOVA was run with rater group as the factor and fluency ratings as the dependent variable to see if the ESL learners' fluency ratings increased over the ten-week period. Qualitative analysis of the data was done by counting the number of rater mentions of different oral proficiency criteria.

The findings showed that fluency ratings for all three rater groups correlated strongly and positively with the pruned syllable ratio and strongly and negatively with the pause ratio at both times one and two. Furthermore, the Pearson correlation analysis of fluency ratings indicated a high level of agreement within and across rater groups. This finding suggests that raters of different L1 backgrounds and different levels of linguistic expertise can perceive temporal measures of fluency and rate L2 learners reliably based on those measures. Another finding was that there was no statistically significant increase in fluency ratings, based on the ratings of any

of the three rating groups. This finding suggests that ten weeks is too short of a time for second language oral fluency, or possibly raters' perception of it, to increase.

Qualitative analysis of the oral data provided a couple of interesting findings. The authors analyzed mentions of fluency features related to the temporal aspects of fluency that the participant raters were instructed to base their ratings on separately from non-temporal criteria like pronunciation, grammar, and vocabulary. The single temporal criterion that was mentioned by far the most (almost half of all comments) was pausing.

This study had important implications for oral fluency research. First and foremost, as previous studies showed, temporal measures of oral fluency are perceptible to trained (Kormos & Dénes, 2004) and untrained raters (Derwing et al., 2004). Rossiter (2009) extended this strand of research by showing that L1-English speakers who are language experts, L1 English speakers who are non-language experts, and advanced L2 English users could all perceive the same temporal measures and rate based on them with a high level of agreement. Finally, the fact that almost half of the rater comments related to pausing suggests that pausing, which affects measurement of composite measures like speech rate, pruned syllable rate, and mean length of speech run plays a large role in perception of oral fluency as well as the measurement of temporal measures. Hence, the way we measure pausing matters a great deal.

Bosker, Pinget, Quené, Sanders, & De Jong (2012) extended the study of untrained rater perception of L2 oral fluency by examining rater perception of speed, breakdown, and repair separately. In other words, Bosker et al (2012) studied the relationship between various oral fluency measures for L2 speakers of Dutch (N=30) and the fluency ratings of untrained L1 Dutch speaking raters (N=80). The study included L2 Dutch speakers (N=38) of various L1 backgrounds. Each speaker performed three different speaking tasks: (1) "simple, formal, descriptive"; (2) "simple, argumentative, descriptive"; (3) and "complex, formal, argumentative" (p. 164), from which one twenty second portion was extracted from the middle of each recorded response.

The 80 raters were divided into four groups of 20 raters, each group participating in one experiment (four experiments in total). Raters in each group rated responses based on different criteria. In experiment one, raters were to rate "overall fluency" (p. 165) using an analytical scale comprised of the following three criteria: "(1) the use of silent and filled pauses, (2) the speed of delivery of the speech and (3) the use of hesitations and/or corrections" (p. 166). In experiment

one, Pearson correlations and multiple linear regression were used for statistical analysis. More specifically, bivariate Pearson correlations were calculated between each fluency measure and fluency ratings in experiment one only. In experiments two-four, multiple linear regression was used with temporal measures of fluency as the predictors and fluency ratings as the dependent variable.

In experiments two-four, all procedures were the same as in experiment 1, except the rating criteria were to be different, and the word "fluency" was not used in the rating instructions. In experiment two, raters were to rate based on "silent and filled pauses" (p. 166), in experiment 3, "speed of delivery", and in experiment 4, "the use of "repetitions and corrections" (p. 169).

The researchers analyzed the data based on six different temporal measures of oral fluency, representing speed, breakdown, and repair. The speed variable was mean length of syllable (also known as mean syllable duration); the breakdown variables were number of silent pauses/second spoken, number of filled pauses/second spoken, and mean length of silent pause. Repair was represented by number of repetitions/second spoken and number of corrections/second spoken. Mean length of syllable and mean length of silent pause were log-normalized. The raters were told to only consider three oral fluency-related phenomena when rating: 1) use of pauses; 2) speech rate; and 3) the use of repetitions and self-corrections.

In experiment 1, results showed that the speed variable (mean length of syllable) correlated the strongest and negatively with fluency ratings. Moreover, mean length of pause and number of silent pauses correlated moderately and negatively with fluency ratings. The correlations between each of the repair variables, number of repetitions/second spoken and number of corrections/second spoken, and fluency ratings were the weakest and negative. These findings indicate that when untrained raters rate based on speed, breakdown, and repair at the same time they are most influenced by speed and least influenced by repair.

The researchers then used the oral fluency measures to build a multiple linear regression model, with fluency ratings as the dependent variable. The regression analysis showed that there was not much difference between the percentage of the variation in fluency ratings explained by the two breakdown variables combined and the single speed variable in these two separate regression models; both explained a little over half of the variation in fluency ratings. Moreover, repair variables explained very little additional variation in fluency ratings, after including breakdown and speed variables. This finding suggests that when considered separately, speed

variables and breakdown variables are roughly equal in their power to explain variation in untrained fluency ratings, and both speed and breakdown have more explanatory value than repair variables.

For experiments 2-4, multiple linear regression models were also used. For experiment 2, breakdown variables were regressed on the raters' fluency ratings, which were supposed to be based on silent and filled pausing. Similarly, for experiment 3, mean length of syllable was regressed on the fluency ratings (based on speed). Finally, in experiment 4, the two repair variables were regressed on fluency ratings (this time, based on repetitions and corrections). The model containing the three breakdown variables explained the the vast majority of the variance in its dependent variable. The explanatory power of the other two regression models (the one containing mean length of syllable and the one containing the two breakdown variables) were very close to the same and below that of the breakdown model. The findings of experiments two-four suggest that when untrained raters focus on one aspect of fluency alone they are influenced most by breakdown and about equally by speed and repair.

The findings of of all four experiments in Bosker et al. (2012) considered together suggest that to the untrained ear, speed and breakdown are the most noticeable aspects of second language oral fluency, and repair is the least noticeable, especially when an untrained rater is listening for all three aspects of fluency. Hence, a research design that aims to most parsimoniously represent second language oral fluency will include at least one speed variable, at least one breakdown variable, or a variable that serves as a composite of the two, but including a repair variable is not totally necessary.

The next two studies reviewed went beyond the notion of rater perception of L2 oral fluency as a purely academic concern to the importance of L2 oral fluency in distinguishing holistic score points on high stakes tests. Iwashita, Brown, McNamara, and O'Hagan (2008) analyzed seven features of test-taker speech at five different proficiency levels for five different TOEFL iBT tasks to determine which measures best distinguished proficiency levels . However, only the results for fluency and vocabulary will be discussed. Eight different performances for each task at each of five levels (N=200) were rated by two trained TOEFL raters. Five different oral fluency dependent variables were measured: number of filled pauses, number of silent pauses, repair, total pause time, speech rate, and mean length of run. The first three of these were measured per minute, and speech rate was measured per second. Moreover, two vocabulary

dependent variables were measured: tokens per minute (a measure of word production) and types per minute (a measure of vocabulary range). The authors ran 2x2 ANOVA with task and level as the two factors to determine which of the measures distinguished levels.

Results showed that, for fluency, in descending order of effect size, speech rate, total pause time, and number of silent pauses showed "a clear relationship with proficiency level" (p. 41), meaning that they increased or decreased in a step-wise manner across proficiency levels. The effect size for speech rate was double that of the fluency variable with the next largest effect size. Moreover, the two vocabulary variables increased in step-wise fashion as score level increased. The effect size for tokens per minute was medium sized and slightly larger than that for types per minute. To summarize, more proficient speakers spoke faster, engaged in silent pausing less frequently, spent a lower proportion of their total time pausing, and used a broader range of vocabulary.

This study contributed some findings that are relevant to the theory and measurement of oral fluency. In keeping with the Speech Production Model, this study provided some evidence that some aspects of oral fluency develop as vocabulary range broadens. An oral fluency measurement insight that can be drawn from the findings of this study is that speech rate (a composite measure of oral fluency) distinguishes proficiency levels much better than either number of silent pauses or total pause time. This can be explained by the fact that speech rate measures pausing as well as speed, while number of silent pauses and total pause time only measure breakdown.

Ginther, Dimova, and Yang (2010) used a different research design to study the relationship between temporal measures of oral fluency and high-stakes holistic speaking test scores. They collected holistic ratings of test-taker performance on an argumentative task on an oral English proficiency test for international teaching assistants. Responses (N=125) were rated by two trained raters and collected for the two largest non-L1-English groups who took the test: Chinese (N=75) and Hindi (N=50). Responses for L1 English test-takers (N=25) were collected and analyzed but not rated. The rating scale was a seven-point scale, with all L1 English test-takers receiving a seven automatically. The study included 15 temporal measures of fluency: total response time, phonation time, phonation-time ratio, number of syllables, speech rate, articulation rate, mean length of speech run, silent pause time, filled pause time, number of filled pauses, number of silent pauses, mean filled pause length, silent pause ratio, and filled pause

ratio. The authors then conducted Spearman correlation analysis to see which measures correlated highest with holistic OEPT scores. Next they computed descriptive statistics according to L1 and score point; for example, Chinese 4's or Hindi 6's. Finally, they computed 99% confidence intervals[7] for each L1 score point group and compared intervals within each L1 group to see what patterns could be found.

Results showed that (in descending order) speech rate, mean syllables per run, articulation rate, and silent pause ratio correlated highest with holistic OEPT scores. The confidence interval analysis showed that within each L1, only speech rate, mean sllyables per run, and phonation-time ratio had non-overlapping 99% confidence intervals between score levels. While non-overlapping confidence intervals only occurred for adjacent score points in one comparison (Chinese 3's versus Chinese 4's), when comparing non-adjacent levels, several score level comparisons within L1 groups exhibited non-overlapping 99% confidence intervals. Even non-adjacent differences provide strong evidence of the power of a particular measure to distinguish proficiency levels because raters of this test were trained to take multiple aspects of oral English proficiency into consideration (pronunciation, verb use, vocabulary range, etc.). The authors concluded that the results contributed to the validity argument of the OEPT and supported the use of speech rate, mean length of speech run, and articulation rate in automated scoring of oral English proficiency.

Ginther et al. (2010) provided compelling evidence of the strong relationships between temporal measures of oral fluency and oral English proficiency. The strong bivariate correlations of speech rate and mean length of speech run respectively with oral English proficiency holistic score, along with the confidence interval analysis, showed yet again that composite measures of oral fluency, which encompass speed and breakdown, distinguish proficiency levels well. The authors made a persuasive argument that the power of these measures is derived from the fact that they represent core language skills, like vocabulary and syntax. The next study reviewed in the present literature review would provide some evidence for this argument that L2 oral fluency measures represent linguistic knowledge.

De Jong, Steinel, Florijn, Schoonen, and Hulstijn (2013) studied the relationships between L2 oral fluency, linguistic knowledge, and processing ability. The participants were

[7] Showing that the fluency measures of L2 speakers at different score levels have non-overlapping 99% confidence intervals represents strong evidence that students scoring at those different levels exhibit different levels of fluency.

adult L2 learners of Dutch (N=179). Each participant completed eight different monologic speaking tasks, which were balanced along the dimensions of complexity, formality, and discourse type. Trained undergraduate raters (N=3) rated L2 performances. Each participant was audio-recorded for two minutes and the data was transcribed and analyzed. The following repair fluency measures were included in the study: silent pauses, self-corrections, filled pauses, and repetitions each normed to a common base of 100 words. A speed variable and a breakdown variable were also included: mean syllable duration and mean pause duration, respectively.

Several linguistic knowledge variables were included in the study. Fill in the blank test scores were used as the measure of vocabulary knowledge. Scores on a variety of objective grammar tasks were used as the measure of grammatical knowledge. The measure of pronunciation was essentially the accuracy rate of pronunciation of monosyllabic words in a word list. Accuracy rate of word stress when pronouncing multi-syllabic words was used as a metric of word stress. Intonation was rated by the three trained raters by means of a sentence-reading task. L2 lexical retrieval ability was measured by means of a timed picture-naming task in which the time it took the participant to name a shown picture represented the participant's lexical retrieval ability. Articulation latency was measured by using the same timed picture-naming task, only this time the participant was asked to wait for an audio-visual cue before naming the picture. The time between the cue and the response was counted. Speed of completion of a timed grammatical transformation sentence completion task was used to measure ability to construct sentences.

Results showed that when measures of linguistic knowledge were tested for correlation with fluency measures, mean syllable duration stood out as the fluency measure with the highest magnitude bivariate correlations. The strongest positive bivariate correlation was between sentence construction speed and mean syllable duration. There were moderate and negative bivariate correlations between mean syllable duration and vocabulary knowledge and between mean syllable duration and grammatical ability. There was also a moderate and negative correlation between mean syllable duration and pronunciation quality.

This study provided some evidence that vocabulary and grammar knowledge are associated with L2 oral fluency. Since mean syllable duration is negatively correlated with oral fluency, the fact that this variable correlated moderately and negatively with vocabulary and grammatical ability suggests that knowing more words and grammar structures facilitates

articulation. This finding is not entirely consistent with the Speech Production Model, which associates grammar and vocabulary with the formulator, as opposed to the articulator. However, articulation is known to correlate strongly with mean length of run, which the Speech Production Model associated with grammar and vocabulary. One puzzling finding of De Jong et al. (2013) is that mean syllable duration correlated so strongly and positively with sentence building speed. One would think that possession of the knowledge to build sentences would facilitate articulation; however, a strong positive correlation with mean syllable duration suggests that the opposite is true.

The cross-sectional research reviewed up to this point has focused on temporal measures of fluency and their relationship to L2 fluency or oral proficiency, but as Norris and Ortega (2009) argued, the speech elicitation task plays an important part in CAF research. With this fact in mind, Leaper and Riazi (2014) examined the effects of speaking topic on complexity, accuracy, and fluency of second language learners. Given the purpose of the present literature review, I will discuss only the research design and results related to the fluency research question. The research design involved comparison of temporal measures of fluency across speaking topics. They had two purposes: first, to find the quantitative differences in fluency measures among test-taker groups who completed oral tasks with different topics, and second, to identify the interactive features of the discourse that could possibly elucidate why there were quantitative differences in fluency. Data were collected from 141 Japanese EFL university students who took group oral proficiency exams, which involved group discussion of four different prompts: mobile technology, the outdoors, family matters, and singles issues. The fluency measures included were speech rate, maze ratio, and pause ratio. The Kruskal-Wallis test, which is a non-parametric alternative to ANOVA, was used to test whether fluency measures differed between prompts.

Results showed that temporal measures of fluency differed across speaking prompts. The responses for the singles prompt exhibited a higher pause ratio than the other prompts due to the greater number of silent pauses, and responses to the family prompt had a higher maze ratio (e.g., more unnecessary repetitions, restarts, and self-corrections) than the mobile and outdoor prompts. The responses to the outdoors prompt exhibited a higher maze ratio than the family and singles topics.

Qualitative analysis of the discourse suggested that when test-takers were discussing topics that they had experience with they could speak quite fluently even if they did not have much to say about them. On the other hand, the family and singles topics required discussion of personal topics which test-takers may have felt uncomfortable talking about or required more thought. This may explain the higher frequency of pausing in the singles and family topics. Clearly the qualitative findings helped explain the quantitative results. The findings of this study suggest that researchers cannot take for granted that different prompts elicit the same level of L2 oral fluency. Hence, if more than one prompt is used in a study, the prompts should be shown to elicit oral responses with comparable levels of oral fluency.

This section discussed some of the more sophisticated empirical research related to L2 oral fluency. Section 2.3.1 provided some definitions of fluency from instructed SLA. Section 2.3.2 summarized four L2 longitudinal oral fluency studies. Finally, 2.3.3 provided a detailed review of some of the more sophisticated cross-sectional studies including oral fluency variables.

Norris and Ortega's (2009) organic approach to CAF research, which was discussed in section 2.3.1, has some important implications for the present study. First, it is necessary to study multiple variables in order to capture different aspects of the same construct. CAF development involves change in multiple related sub-systems at once; therefore, it is important to compare the direction and magnitude of change in different variables over the same period.

Second, the characteristics of the task that is used to elicit language are quite important in drawing valid conclusions from results obtained. For example, if different participants were given tasks with different prompts, did each prompt elicit a comparable level of oral fluency? This concern will be revisited later in this section summary and in the methods section.

Third, the language learning context and purpose for learning affect the interpretations that can be drawn from the results. For example, learners studying abroad for a semester are very different from TOEFL-screened learners in an EAP language support program at the beginning of a rigorous, four-year academic program. The former can be expected to have a lower proficiency level and more integrative purpose for learning. On the other hand, the latter would probably tend to have a higher proficiency level and more instrumental purpose for learning.

There are three points that I would like to make about the longitudinal studies summarized in section 2.3.3. First, they all involved participants on short term study abroad stays. Second, from their temporal measures of oral fluency, it is quite obvious that their

language proficiency was not particularly high. The present study looked at how the oral fluency of a group of TOEFL-screened, EAP students studying in a four-year program in the US developed over their first two semesters of college. Third, mean length of speech run and speech rate stand out as variables of interest. These two variables were associated with statistically significant increases in three (Huensch & Tracy-Ventura, 2017; Segalowitz & Freed, 2004; Towell et al., 1996) of the four longitudinal studies reviewed. These findings considered together suggest that these two variables are particularly good indices of longitudinal development in L2 oral fluency. Furthermore, the gains in these two variables reported by Huensch and Tracy-Ventura (2017) were retained even after the study abroad period in which the participants took part. This finding suggests that these two variables represent a fairly broad range of psycholinguistic knowledge.

I would like to make three points about the cross-sectional studies discussed in section 2.3.4. First and foremost, L2 oral fluency is important enough to continue studying in increasingly specialized research designs. Temporal measures of oral fluency distinguish between proficiency levels (Ginther et al, 2010; Iwashita et al, 2008). Moreover, they are closely associated with listener perception of fluency (Derwing et al., 2004; Kormos & Dénes, 2004; Rossiter, 2009). Secondly, as evidenced by the studies reviewed, both speed variables (especially speech rate and mean length of speech run) and a variety of breakdown variables are effective indicators of oral proficiency; therefore, both types of variables should be included in any oral fluency study that intends to provide an accurate picture of a group of learners' oral fluency. The third point relates to speaking topic. Leaper and Riazi (2014) provided some evidence that speaking topic can affect the oral fluency of test-taker responses. Hence, a rigorous research design that includes oral responses elicited from multiple speaking prompts should test the comparability of those speaking prompts with regard to oral fluency. In other words, each prompt should elicit responses that are no more or less fluent than any other prompt. If different prompts elicit a different level of oral fluency, then pre-post differences may be attributed at least partially to differences in the difficulty of the prompt, as opposed to being solely the result of oral fluency gains.

2.4 Vocabulary and Oral Fluency

Vocabulary plays a role in the efficient conversion of the speaker's message into spoken words (Levelt, 1989, 1999). However, there is not much direct empirical research to back up this theoretical argument. There is some evidence that vocabulary knowledge is associated with oral proficiency overall (Lu, 2012; You, 2014). There is also some evidence from Ushigusa (2008) and Hasselgren (2002) that use of formulaic language is associated with more fluent L2 speech. However, it still not known whether use of words at different word frequency levels is associated with L2 oral fluency measures. Section 2.4 reviews some of the research related to L2 oral fluency and vocabulary.

2.4.1 Formulaic Language

According to Wray (2002), a formulaic sequence is

> a sequence, continuous or discontinuous of words or other elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (p. 9)

She argued that formulaic sequences are essential for first language and second language acquisition. Agreeing with Pawley and Syder (1983) and taking issue with Chomsky (1957), she asserted that the human tendency toward linguistic innovation is overstated, or perhaps over-rated, and that in most situations language users depend heavily on language that they have heard and used many times before. As evidence against the Chomskyan view on innovation, she noted that if innovation were the dominant force in language performance, then synonyms like "large" and "great" would appear before the word "number" with frequencies that are not significantly different, but this is not the case. In performance data, "large" appears much more often than "great" before "number".

Similar to Pawley and Syder (1983), the reason that Wray gives for this tendency toward formulaicity is cognitive. She explains that humans find themselves in a wide variety of cognitively demanding situations. Thus, given the limitations on human cognitive processing capacity, it is simply easier to fall back on familiar, unanalyzed chunks of language. This frees

up working memory to attend to other aspects of language and human behavior like linguistic creativity.

Schmitt (2004) provided more evidence, arguing "There is plenty of evidence to suggest that formulaic sequences are typically stored and processed as unitary wholes, even if this is not true in every case." (p. 4) In support of this claim, he cited multiple studies that have suggested that formulaic sequences are "spoken more fluently (than creatively generated strings), with a coherent intonation contour" (p. 5). Schmitt also reasoned that since formulaic sequences are so often used for specific functions, like apologizing and requesting, this must be due to processing efficiency.

Wray also discussed what she saw as some serious problems involved in the process of identifying formulaic sequences. First of all, she pointed out that arbitrariness could not be removed from the process. While computer frequency searches can eliminate some types of human error, a human must set the frequency threshold above which a string of words is considered frequent enough to be considered a sequence. Setting this level requires some amount of arbitrary judgment. Of course, the same problem exists with any cutoff used in research. For example, as was discussed before, Riggenbach (1991) set several different silent pause length cutoffs, most of which resulted in variables that did not distinguish proficiency levels. Others (Ginther et al., 2010; Goldman-Eisler, 1968; Kormos & Dénes, 2004) set the silent pause threshold at .25 seconds, with more favorable results.

Wray (2002) discussed another source of arbitrariness involving the determination of formulaic sequence boundaries. She gave the example of the simple two-word string: “thank you” (p. 28). She questioned whether this is a cut and dry two-word formulaic sequence or if it is actually just a part of other longer sequences, like “thank you very much” or “thank you goodbye” (p. 28). She argued that if the researcher were to make this distinction based on nothing more than intuition, then such a procedure would at least partially defeat the purpose of conducting computer frequency searches to begin with.

Wray also lamented the lack of agreement on the definition of a formulaic sequence or even whether 'formulaic sequence' is the appropriate word for the linguistic phenomenon itself. She named dozens of alternative names that are used for this phenomenon and discussed at length over a dozen different "fundamental features" (p. 44) of the phenomenon. These include "internal structure, form, irregularity, variability, collocation, function, meaning, idiom and

metaphor, pragmatic meaning, and provenance" (p. 47-59), to which one might add the "lexicalized sentence stem" (p. 191) of Pawley and Syder (1983).

Wray argued that having so many fundamental features of formulaic sequences poses problems. First, if a string must exhibit all of these features in order to make the formulaic sequence list, then the list will necessarily be very short. If only one or a few of these features are necessary, then the list will be overly inclusive and thus lack practical value. Given that word lists have become important in vocabulary research, below is a brief review of some of the literature related to academic vocabulary lists.

2.4.2 Academic Vocabulary Lists

In agreement with Wray (2002), Simpson-Vlach and Ellis (2010) argued that lists of commonly co-occurring words meant for L2 pedagogical purposes should be based on more than just frequency. Simpson-Vlach and Ellis (2010) reasoned that some words, like "and of the" (p. 490) often co-occur without being "psycholinguistically salient" (p. 490), while some words, like "on the other hand" are both frequently co-occurring and psycholinguistically salient. They claimed that groups of words that meet one or both criteria should be identified and taught to L2 learners.

Hence, the authors compiled a list of words that frequently collocate in academic texts and are psycholinguistically salient. They called this list the "academic formula list". To compile this list, they created two academic corpora (each slightly over two million words): one with spoken texts and the other with written texts. Each corpus consisted of sub-corpora representing a cross-section of academic disciplines, and the spoken corpus containing a "non-departmental" (p. 493) corpus. The authors then searched these two corpora for collocations comprised of three, four, and five words each, which appeared more than ten times per million words in the corpus.

The formula list was compiled by comparing collocation frequencies between academic and non-academic corpora. To compare the academic frequency with the frequency in non-academic texts, two non-academic English corpora were chosen: one with written texts and one with spoken conversational texts. The list of "academic formulae" was compiled by carrying out a log likelihood statistical procedure comparing the collocations that occurred significantly more frequently in the academic corpora than in their respective non-academic counterparts. The formula list consisted of three lists of collocations that more frequent in: 1) academic writing, 2)

academic speaking, and 3) both academic writing and speaking (the core list). To ensure that words on the lists provided broad coverage, one of the criteria for inclusion on the list was that each word met the frequency requirement in three of the four academic sub-corpora of each mode (writing and speaking). The present dissertation used the academic spoken formula list (along with formulaic language lists from four other studies) to measure frequency of formulaic sequences in lexical analysis of oral responses.

A more established approach to vocabulary word lists is based on the frequency of individual words. The lexical frequency profile (LFP) of Laufer and Nation (1995) consists of the most frequent words that appear in written academic discourse. The LFP originally consisted of four levels: the most frequent 1000 words (K1), the second most frequent 1000 words (K2), the University Word List, and off-list words, which are specialized words that appear on none of the 3 lists. The University Word List was developed by Xue and Nation (1984), and it includes 800 words that appeared frequently in a corpus of 303,000 academic lectures, journals, and exam papers from 27 different disciplines.

Laufer and Nation (1995) tested the validity of the LFP on three groups of students (N=65) who were deemed to be at different English proficiency levels. A group of university level L2 English learners in New Zealand (n=22) of a variety of L1 backgrounds was taking an EAP class and was considered low intermediate based on a placement test. A group of EFL Israeli, college-level learners was deemed to be at a higher proficiency level as demonstrated by Cambridge Certification in the case of the lower Israeli group (N=20) and another Israeli group (N=23), having completed two semesters of course work in the English language and literature department, was considered to be higher still. In other words, the study included three different proficiency levels. Each student wrote two in-class argumentative essays in the same week and took a standardized vocabulary levels test involving a word completion task. The lexical frequency profile of each essay was calculated. In other words, the word count of each essay was computed, and each word was classified in one of four lexical frequency categories: K1, K2, University Word List, and off list (not in any of the three lists). The lexical frequency profile for each student included the percentage of the words in the essay that fall into each of the four lexical frequency categories. Two methods of statistical analysis were used in this study. First, there was a correlation analysis involving lexical frequency percentages (e.g., 80% K1, 6% K2, etc.) and the results of the vocabulary levels test. Second, there was a one-way ANOVA with

proficiency level group as the factor and lexical frequency percentages as the dependent variables.

ANOVA results showed that the lexical frequency profile distinguished between L2 writers of different proficiency levels. There were significant differences between proficiency levels in frequency of use of K1 words (higher level students used fewer), UWL words (higher level students used more), and off-list words (higher level students used more). Moreover, off-list words distinguished between adjacent levels most consistently. Similarly, frequency of K1, UWL, and off-list words each correlated in the expected direction with student scores on the vocabulary levels test. Vocabulary levels test scores correlated strongly and positively with frequency of use of UWL and off-list words and strongly and negatively with frequency of use of K1 words. However, K2 frequency did not significantly correlate with the vocabulary levels test scores. The findings of this study suggest that lexical frequency profile in L2 writing distinguishes proficiency levels and is correlated strongly with receptive L2 vocabulary knowledge; however, this was an L2 writing study, not an L2 speaking study. The relationship between lexical frequency and L2 oral fluency remains unclear.

2.4.3 Formulaic Language and Oral Fluency

Hasselgren (2002) went beyond the well-documented finding that temporal measures of fluency distinguish L2 proficiency levels. She used corpora to attempt to describe the formulaic language of fluently spoken English. She studied filled pauses quantitatively and qualitatively by means of a corpus of L2 oral English test-taker data made up of responses to "a three-task communicative oral test" (p. 104), which is administered in Norway to test-takers in pairs and graded on a six-point rating scale. The study analyzed spoken corpora of L1 English (n=18) and L1 Norwegian (n=43) oral test takers to determine how L1 English speakers used a category of formulaic language called “small words” (p. 150) differently from highly proficient L2 speakers and low proficiency L2 speakers. The L1 Norwegian responses were divided into two groups: those scoring five and above on the test (n=19) and those scoring four and below (n=24). The fluency measures included in the study were number of filled pauses and mean length of utterance (in words). The vocabulary measure included was frequency of "small word" use. The only small words that were included were short words that often appeared in the corpus data and

facilitated a continual flow of speech without adding semantically to the utterance. For example, 'well', 'I see', 'oh', and 'sort of' (p. 151).

Moreover, filled pauses and small words were categorized based on position: "turn-initial", "turn internal", and "turn-final"" (p. 152-3). Descriptive statistics of the results were provided.

Besides showing that mean length of utterance distinguished well between L2 proficiency levels and between L1 groups, the results suggest that particular uses of small words may facilitate improvement in fluency. Results showed filled pause frequency within turns was the only filled pause frequency location that distinguished well among the groups, with L1 speakers using significantly more than both L2 groups and the more proficient L2 group using more than the less proficient group.

Hasselgren's qualitative data analysis focused on how test-takers, who it should be remembered were testing in pairs, used small words to signal their communicative intent. For example, two common occurrences were use of small words to "take (or) hold... the (speaking) turn" (p. 160). This analysis involved an exhaustive corpus analysis of each small word in context by two L1 coders, who ended up reaching consensus through discussion on each and every case. Both the transcripts and the audio recordings were analyzed.

The analysis found that use of small words to take and hold a speaking turn both discriminated well among the different groups. The L1 group used small words more often than both L2 groups to take and hold a turn, and the more proficient L2 group used small words for both functions more often than the less proficient group. Another signaling function that discriminated well among groups was use of small words like "sort of" and "kind of" (p. 164-5) to "hedge". This finding is important because it showed that the connection between vocabulary and fluency has an effect on the ability of L1 and L2 English speakers to assert themselves in communicative situations. In other words, more proficient L2 speakers were able to use small words to lengthen their speaking turns in pair speaking tasks, and L1 speakers used small words in this way better than L2 speakers.

In another study of the relationship between formulaic language and oral fluency, Ushigusa (2008) studied the association among multi-word units, holistic scores on theOral English Proficiency Test (OEPT), and speakers' temporal measures of fluency. He analyzed the monologic responses of L1 Chinese ITA's (n=38) and L1 English TA's (n=12) to a task in which examinees were to give advice. The ITA's were required to take the OEPT, which was rated on a

four-point scale (3-6) to certify their oral English proficiency for purposes of screening ITA's for teaching duties. The L1 English speakers were not rated but automatically assigned a holistic score of 7. In the responses, Ushigusa (2008) coded all instances of multi-word units (MWU's), which he broke down into idiomatic MWU's, phrasal verbs (a subset of idiomatic MWU's), verbal phrases, multi-word small words, multi-word discourse markers, and collocations. The study included the following temporal measures of fluency: total response time, articulation rate, phonation-time ratio, mean length of speech run, speech rate (measured in syllables and words), mean silent pause duration, silent pause time ratio, and number of pauses per minute. The statistical analysis that is of interest in the present literature review included a Spearman correlation analysis between multi-word units and temporal measures of oral fluency.

Results showed that frequency of all MWU's and idiomatic MWU's correlated moderately and positively with speech rate and mean length of speech run for both formulaic language variables. This finding suggests that there is some evidence in support of Levelt's model with regard to the importance of vocabulary knowledge for development of L2 oral fluency.

2.4.4 Lexical Features and Oral Proficiency

In a related study that analyzed monologic examinee responses to a computer-administered "Compare and Contrast" (p. 41) speaking task, You (2014) studied the relationship between various measures of lexical proficiency and holistic scores on an oral English proficiency test. You analyzed the oral responses of 303 test-takers at a large university, all of whom except the L1 English test-takers were required to take the test in order to be certified as oral English proficient for the purposes of serving as teaching assistants at the university. Responses were analyzed from examinees of four different L1's: Mandarin (N=111), Korean (N=100), Hindi (N=67), and English (N=25).The lexical proficiency measures included number of tokens, number of types, type-token ratio, the D measure, and words from the four different lexical frequency categories of Laufer and Nation's (1995) Lexical Frequency Profile: most frequent 1000 words (K1), second most frequent 1000 words (K2), Academic Word list (AWL), and off list words (OL). Her results showed that number of types of K1 words used increased at each successively higher holistic score point. In fact, variety of highly frequent (K1) words used distinguished better than any other variable between holistic score points within each L1 group.

This finding suggests that use of a variety of highly frequent words is associated with L2 oral English proficiency. While You (2014) did not include oral fluency measures in her analysis, Ginther et al. (2010), which is reviewed above, did analyze temporal measures of fluency in oral responses to the same test (the Oral English Proficiency Test) and found strong positive correlations between temporal measures (especially mean length of speech run and speech rate) and holistic scores on the test.

Studying the lexical diversity of texts in a somewhat more sophisticated manner, McCarthy and Jarvis (2010) investigated the validity of four measures of lexical diversity: the "measure of textual lexical diversity" (MTLD) (p. 381), vocd, Maas, and HD-D. To assess construct validity, they used each measure to gauge the lexical diversity of written texts drawn at random from a corpus comprised of 16 different registers. The authors used correlation analysis to test the extent to which each measure agreed with the other lexical diversity measures in the study, all of which are widely recognized in the field of applied linguistics as good measures of lexical diversity. Then, they tested the extent to which the measures disagree with a widely recognized poor measure of lexical diversity (the type-token ratio). Next, they measured the degree to which each measure could distinguish between written texts of high and low cohesion. Finally, they ran a correlation analysis to see which measure correlated the lowest with text length in terms of word count.

Results showed that MTLD was the only lexical diversity measure that showed satisfactory performance on all four tests of validity, including not correlating significantly with text length. The findings suggest that MTLD is a valid measure of lexical diversity. Of course, the question of whether MTLD is a valid measure of the lexical diversity of oral texts has not been answered. It would be interesting to see if large fluency gainers exhibit large increases in MTLD, possibly suggesting a link between this measure and oral fluency gains.

Section 2.4.1 introduced formulaic sequences. 2.4.2 reviewed two different academic vocabulary lists, both of which were used in the data analysis of the present dissertation. 2.4.3 summarized the research on the associations between formulaic language and L2 oral fluency. Finally, section 2.4.4 reviewed the research on the associations betwen lexical knowledge and L2 oral proficiency. Wray (2002) and Schmitt (2004) argued in 2.4.1 that multi-word units, by whatever name they are known, are important for the same processing reasons described by Pawley and Syder (1983).

Section 2.4.2 is relevant to the lexical analysis of the present dissertation. Simpson-Vlach & Ellis (2010) used corpus linguistics to create a list of spoken academic formulas that are more common in academic spoken discourse than written academic discourse and more common in academic spoken discourse than nonacademic spoken discourse. Laufer and Nation (1995) developed a list of the most frequently used single words in academic writing. The present study tested the extent to which large fluency gainers undergo changes in their frequency of use of words at different frequency levels.

The research summarized in 2.4.3 and 2.4.4 demonstrated that lexical knowledge is associated with L2 oral fluency and L2 oral proficiency, based on studies involving high-stakes oral test tasks. Two studies (one from 2.4.3 and one from 2.4.4) are particularly relevant to the current dissertation: Ushigusa (2008) and You (2014). Ushigusa's finding that use of multi-word units in L2 oral responses was associated with L2 oral fluency measures is important because it implies an association between the use of spoken formulaic sequences and temporal measures of oral fluency. You (2014) found that OEPT test-takers who used a wider variety of high-frequency vocabulary scored higher on the test. This is significant because it suggests a relationship between use of high frequency vocabulary and temporal measures of oral fluency like mean length of run, speech rate, and phonation time ratio because these temporal measures of fluency correlated strongly with OEPT scores according to Ginther et al. (2010). The present dissertation presents analyses pre-post change in the lexical frequency profile, lexical diversity, and formulaic language use of large fluency gainers.

2.5 Literature Review Summary

This chapter started by reviewing the literature related to oral fluency theory in section 2.2. Then, in section 2.3, some of the key concepts related to L2 oral fluency and some of the variables that have been studied in relation to L2 oral fluency over the past three decades. Section 2.3 went on to review the research on longitudinal development of L2 oral fluency and cross-sectional studies of listener perception of L2 oral fluency. Section 2.4 discussed vocabulary and how it relates to L2 oral English proficiency in general and L2 oral fluency in particular. Section 2.4 also discussed two specific aspects of vocabulary: formulaic language and lexical frequency. Section 2.4 reviewed some research findings that suggest that use of formulaic language is associated with L2 oral English proficiency and L2 oral fluency. Other research

findings reviewed suggested that use of individual words at different levels of lexical frequency distinguished L2 English writers at different levels of proficiency and L2 English speakers at different levels of proficiency.

The theory underlying oral fluency is primarily based on psycholinguistics and cognitive psychology. Goldman-Eisler (1958a, 1958b) presented freedom of choice as a stumbling block for the L1 speaker in regards to oral fluency. Pawley and Syder (1983), on the other hand, argued that the native speaker's large, well-organized storehouse of lexical knowledge allows the native speaker to circumvent this stumbling block. Anderson (1983) provided the theoretical framework that explained the development and proceduralization of such a large, well-organized storehouse of linguistic knowledge. Levelt (1989) applied Anderson's framework to the production of speech. Levelt also described a theoretical model consisting of different phases of speech production and argued that vocabulary and syntax played a key role in the efficient, automatic production of speech.

Oral fluency is a key aspect of L2 proficiency because it subsumes other core linguistic skills, especially syntax and vocabulary. SLA scholars (Ginther et al., 2010;; Lennon, 1990; Levelt, 1989; Norris & Ortega, 2009; Skehan, 2009; Towell et al., 1996) have long argued that fluency is a multi-faceted construct that should be studied using multiple objective measures; hence, the importance of temporal measures of fluency.

Longitudinal research findings related to L2 oral fluency development are consistent with Levelt's (1989) argument that vocabulary and syntax play an important part in oral fluency development. Towell et al.'s (1996) findings suggested that an increase in mean length of speech run, which they argued was associated with use of more complex syntax and collocations, was the driving force behind the oral fluency development of L2 learners of French. Segalowitz et al (2004) provided some indirect evidence that syntactic development over a study abroad period enables oral fluency development. Furthermore, Huensch and Tracy-Ventura's (2017) findings that mean length of speech run and speech rate increase gradually and are associated with lasting gains are consistent with the argument that these measures are associated with complex linguistic knowledge. However, these studies provided little direct evidence of a relationship between vocabulary, syntax, and L2 oral fluency development.

A review of the cross-sectional research on temporal measures of oral fluency and listener perception of L2 fluency yields some important findings. First, temporal measures of

oral fluency are moderately to strongly associated with listener perception of L2 oral fluency. Cross-sectional research findings also provide more evidence that temporal measures are associated with syntax and/or vocabulary. De Jong et al. (2013) found moderate correlations between a temporal measure of fluency (mean syllable duration) and both vocabulary knowledge and grammatical ability. Derwing et al. (2004) found that fluency ratings were strongly associated with temporal measures. They also provided evidence that fluency ratings are more strongly associated with comprehensibility than with accent. Since comprehensibility is primarily concerned with integrating vocabulary and syntax to form sentences that make sense, this finding provided evidence that vocabulary and syntax play an important role in L2 oral fluency. Iwashita et al. (2008) provided indirect evidence of an association between vocabulary and oral fluency, when they showed that two vocabulary measures and three temporal measures of fluency all increased in step-wise fashion at successively higher TOEFL iBT holistic speaking score levels. Finally, while Ginther et al. (2010) did not analyze vocabulary or syntax in oral responses, they did find very strong correlations between some temporal measures of fluency and holistic oral proficiency test scores, suggesting that these measures must gauge core linguistic knowledge and skills.

The theory related to formulaic sequences fits nicely within the cognitive models related to oral fluency. Wray (2002) and Schmitt (2004), in agreement with Pawley and Syder (1983) and echoing the psycholinguistic theory discussed earlier (Anderson, 1983; Levelt, 1989), argued that formulaic sequences play a large part in oral proficiency for processing reasons. Essentially, these scholars argue that speakers need to allocate scarce working memory to various tasks of daily living, one of which is speech production. Since formulaic language is retrieved as unanalyzed chunks, use of formulaic language frees up working memory for other cognitive tasks. This efficiency makes formulaic language an ideal mechanism for proceduralizing the functioning of L2 speech formulation in Levelt's Speech Production Model.

Some studies suggest that use of formulaic sequences is associated with L2 language proficiency and oral fluency. For example, Hasselgren (2002) found that L1-English speakers used more "small words" to lengthen speaking turns in dialogic speaking tasks. Moreover, Ushigusa (2008) found that use of multi-word units correlated moderately to strongly with temporal measures of fluency, especially mean length of speech run.

Lexical diversity and knowledge of words at different frequency levels may also facilitate L2 oral fluency development. Laufer and Nation's (1995) frequency-based framework for testing lexical knowledge may shed some light on how the L2 lexicon changes as oral fluency develops. Moreover, a similar argument might be made of lexical diversity. It stands to reason that as L2 learners increase their vocabulary, they can speak more fluently.

2.6 The Research Gaps

Although four studies (Huensch & Tracy-Ventura, 2017; Lennon, 1990; Segalowitz & Freed, 2004; Towell, et al., 1996) examined longitudinal development of L2 oral fluency, the existing studies have some limitations. First, they have relatively small sample sizes. In fact, the largest number of participants included in any one cohort was 26 (Huesch & Tracy-Ventura, 2017). Second, three of the four studies reviewed in this literature review included L1 English participants learning either Spanish (Huensch & Tracy-Ventura, 2017; Segalowitz & Freed, 2004) or French (Towell et el., 1996) in a study abroad experience. Only one of the four studies (Lennon, 1990) looked at L2-English learners, and this study had only four participants. No study has examined the longitudinal development of L2 English oral fluency in L1 Chinese students. This is a large demographic group, particularly at large STEM research universities, that faces particular challenges with regard to learning English. For this reason, a relatively large-scale (N=100) longitudinal study of oral fluency development in this demographic group is overdue. Hence, the present study will make an original contribution to the literature.

Another gap in the literature that the present study will fill relates to the pre-post development of formulaic language use and lexical diversity. No other longitudinal study has examined the pre-post change in the formulaic language use and lexical diversity of large fluency gainers. In fact, no study has focused on the pre-post lexico-syntactic change exhibited specifically by learners who make large fluency gains. Knowledge of how large L2 oral fluency gainers' lexico-syntax changes over time could have implications for L2 oral fluency theory as well as pedagogy and assessment.

CHAPTER 3. METHODOLOGY

3.1 Introduction to Methodology

In this chapter, the methodology of the present study is explained in terms of data collection, the test task, data analysis, and statistical analysis. I collected the data from 100 L1 Chinese students who took an exam called the Assessment of College English-International (ACE-In), a computer-administered English language proficiency test, at the beginning and end of a two-semester, two-course language and culture sequence. The test task from which I collected data was the "Express Your Opinion" speaking task. I analyzed all responses for eight different oral fluency measures (speech rate, mean length of speech run, phonation time ratio, articulation rate, mean silent pause length, mean filled pause length, silent pause frequency, and filled pause frequency). In phase 1, I conducted statistical analysis of examinees' pre-post change in oral fluency by computing pre-post descriptive and inferential statistics of pre-post change in oral fluency measures.

In the second phase of the study, I conducted lexical and syntactic analyses. In this phase, I first determined which oral fluency measure was associated with the largest pre-post change, in terms of Cohen's *d* effect size of pre-post change from phase 1. Then, I identified the ten participants who made the largest percentage-wise gains with regard to this oral fluency measure. Next, I conducted a linguistic analysis of these ten participants' oral responses, involving three lexical measures (frequency of use of words at different levels of lexical frequency, lexical diversity, formulaic sequence proportion (proportion of words in the response made up of formulaic sequences) and three syntactic measures (dependent clause ratio, coordinate clause ratio, and words per T-unit). In the linguistic analysis phase, I only included descriptive statistics of pre-test and post-test oral responses and percentage-wise pre-post change. I only used descriptive statistics for the linguistic analysis because the participants were not randomly chosen, and the sample size was small (n=10).

3.1.1 Data Collection

I collected the data for the present study by using the ACE-In Exam administration application. As a testing office assistant in the Purdue Language and Cultural Exchange

(PLaCE), I participated in test administrations and other administrative work related to the ACE-In during the large-scale test administration sessions from which I collected pre-test and post-test data for the study. Therefore, I had some involvement with standardization of test administration procedures and sole responsibility for data collection.

3.1.2 Participants

I randomly chose 100 L1-Chinese students enrolled in the Purdue ENGL 110 class in the Fall 2016 semester from a database of 245 L1-Chinese students who took the ACE-In twice: once at the beginning of their first semester of EAP instruction and once at the end of their second semester of EAP instruction. The gender breakdown of the participants was 55 males and 45 females, and all participants were between 17 and 21 years of age at the time of the pretest. Participants represented a broad cross-section of academic colleges from across the university, but most studied STEM majors. See figure 3.1 below for a detailed breakdown of the academic colleges in which the participants studied at the time that they took the pre-test.

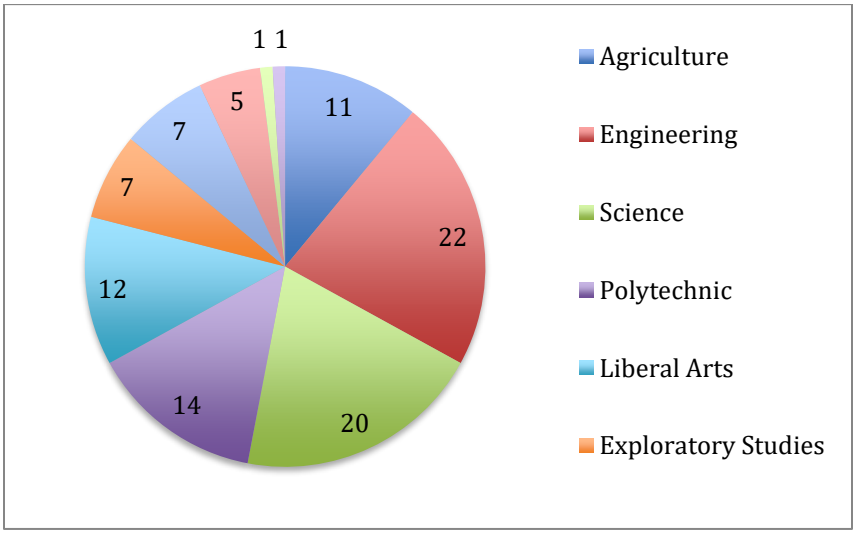

Figure 3.1. Academic Colleges of Participants

3.1.3 The Learning Context

ENGL 110 is an EAP course offered through the Purdue Language and Cultural Exchange (PLaCE). ENGL 110 is a language and cultural support class for L2 English students who have scored 100 and below on the Test of English as a Foreign Language internet-based

Test (TOEFL iBT). The range of students' TOEFL iBT scores was 80-100. The sampling was random from the L1-Chinese[8] ENGL 110 student population.

I analyzed these students' responses to the "Express Your Opinion" speaking task. PLaCE testing personnel administered the test task as part of the ACE-In, a post-admission, computer-administered English language proficiency test, which the students must take at the beginning and end of the Fall and Spring semester as a course requirement. I collected the data for the present study from the Fall 2016 pre-test and Spring 2017 post-test sessions. The speaking task was one of five speaking tasks administered to the students as a part of the second of three modules. PLaCE testing personnel administered modules 1 and 2 in the same session, and PLaCE instructors administered module 3 (the timed writing task) in class in a separate session. ENGL 110 students signed up to take the first two modules of the exam with other students in a computer lab on campus. Two proctors administered the first two modules of the ACE-In. One of the proctors was an ENGL 110 instructor and the other a trained PLaCE testing office assistant.

3.1.4 The Test Task

The speaking task prompts were designed to elicit a spoken response that is devoid of specialized knowledge or terminology. The task was meant to assess examinees' general speaking proficiency. During the task, a written prompt appears on the screen, which examinees hear read aloud. The prompt is a statement of opinion about a decision related to their studies or living situation. The prompt directs examinees to agree or disagree with a statement (e.g., Living with a roommate is a good idea), and examinees are expected to support their opinions with reasons and examples. Test-takers have two minutes to prepare their response and two minutes to respond. Each examinee can see how much time remains by looking at a countdown clock in the upper right-hand corner of the screen. During the preparation period they can write notes to use during their response.

[8] The L1-Chinese population is the largest non-English L1 sub-group at Purdue.

3.2 Data Analysis

I analyzed the data in two phases: oral fluency analysis and lexico-syntactic analysis. I analyzed examinees' oral fluency using a semi-automated transcription and annotation tool called *Fluencing* (Park, 2016). I used The Lextutor Vocab Profiler to analyze their lexical frequency profiles and the Text Inspector to measure MTLD (a measure of textual lexical diversity). I used a three-step algorithm to identify formulaic sequences. Each of the phases of data analysis is explained below in a separate sub-section.

3.2.1 Oral Fluency Analysis

To analyze oral fluency, I used a Python-based tool called *Fluencing*. This system was created by Park (2016), specifically for analyzing oral fluency. The steps of the oral fluency analysis process included pre-processing, segmentation, and transcription of examinee speech.

3.2.1.1 Pre-processing

Analyzing recorded oral performances by means of semi-direct computer administration involved some data cleaning. Some aspects of the oral response recorded in such a setting are the result of the computer-administration itself, as opposed to the examinee's oral proficiency. One such aspect is response latency. Response latency is simply the silent pause before the examinee begins responding and after the examinee finishes responding (Fazio, 1990). To explain, examinees rarely begin speaking immediately after pressing the "record" button, and they rarely speak right up until the time, when they press the "Stop" button at the end of their response. They usually exhibit a silent pause in both places.

Given the purpose of the present study, I removed these pauses from the speech samples before analyzing the responses. While some researchers have studied the psychological implications of this variable in L1 speakers (Fazio, 1990) and its psycholinguistic dimensions in L2 speakers (Cheng, 2016; MacIntyre and Gardner, 1994; Munro and Derwing, 1995), the present study was not concerned with this aspect, but rather with the features of L2 speech itself. Hence, in order to prepare the files for analysis, I removed the silence at the beginning and end of each speech sample by using *Audacity*, a popular audio editing tool available for download online free.

3.2.1.2 Oral fluency measures.

To answer research question 1, I calculated the following oral fluency measures. Table 3.1 below presents the formula for calculating these oral fluency measures.

Table 3.1 Oral fluency measures

Measures	Formula
Speech rate	Sixty times the number of syllables divided by response time.
Mean length of speech run	Number of syllables divided by number of speech runs.
Articulation rate	Number of syllables divided by speech time (response time minus filled pause time and silent pause time).
Phonation time ratio	Phonation time (response time excluding silent and filled pause time) divided by response time.
Mean silent pause length	Silent pause time divided by number of silent pauses.
Mean filled pause length	Filled pause time divided by number of filled pauses.
Silent pause frequency	Sixty times the number of silent pauses, divided by response time.
Filled pause frequency	Sixty times the number of filled pauses, divided by response time.

Calculation of the oral fluency variables in the left column above required precise measurement of six different variables: response time, number of syllables in the response, total silent pause time, total filled pause time, number of silent pauses, and number of filled pauses. The segmentation sub-section focuses on the procedures used to measure all these variables except number of syllables in the response and response time, which are automatically calculated by *Fluencing*.

3.2.1.3 Segmentation

After pre-processing the speech samples, the next step was to segment each speech sample into speech runs, silent pauses, and filled pauses. I did this by using Fluencing (Park, 2016). The Fluencing tool allows the researcher to examine each speech sample visually, audibly, and quantitatively. I simply opened each audio file, and the speech sample appeared in waveform at the top of the graphic user interface (See Figure 3.2 below). The horizontal line that runs from the far left extreme to the far right extreme of the interface, vertically half way between the top and bottom of the spectrogram, is a visual representation of absolute silence. The waveform is comprised of stretches of the speech sample with tightly bunched vertical lines (See Figure 3.2) and stretches with either no vertical lines or very short vertical lines. The former stretches denote sound, while the latter denote silence or near silence.

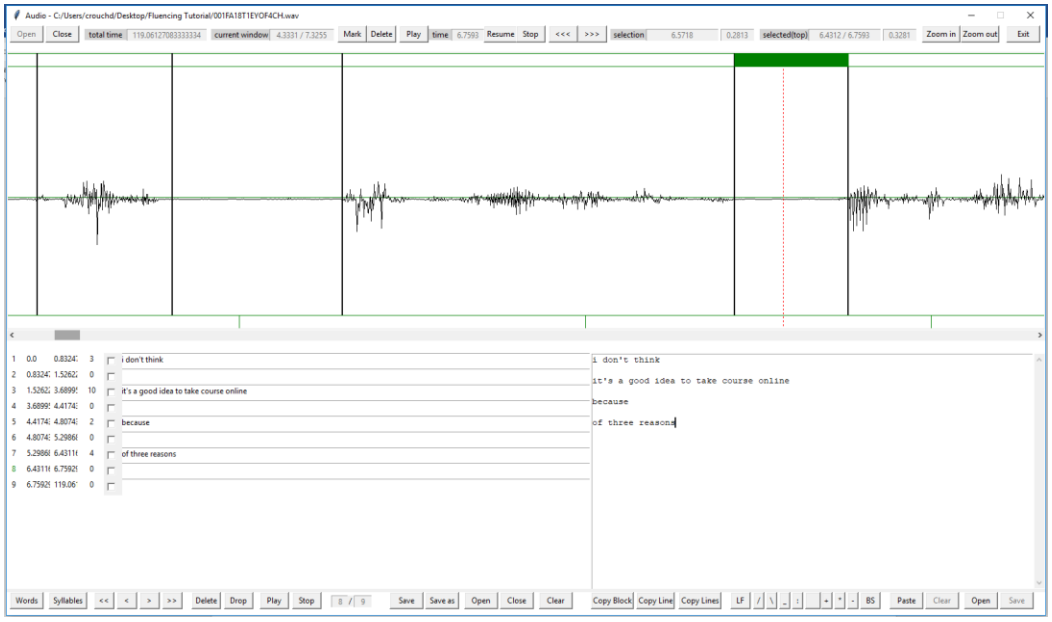

Figure 3.2. *Fluencing* User Interface

3.2.1.4 Definitions

A few key oral fluency terms are important for understanding the data analysis.

- silent pause: a portion of the response that is at least .25 seconds in duration with no speech sound (Kormos & Dénes, 2004).

- speech sound: any sound uttered by the speaker that contains lexical content or an attempt thereof.

- speech run: a portion of the response of any duration containing continuous speech sound between two silent pauses.

- filled pause: "all occurrences of the English hesitation devices [ɛ, æ, ə, r, a, m]" (Maclay & Osgood, 1959, p. 24).

It is important to make some distinctions and clarifications here. First, for practical purposes, a silent pause is almost never completely silent. There is almost always some kind of sound in a silent pause, whether it is other examinees speaking in the background, the speaker breathing into the microphone, aspiration from the final consonant of the previous speech run, the speaker rolling a pencil, etc. I filtered these sounds out mentally and focused attention only on the speaker's speech sounds and lack thereof. I marked absence of speech sound as silence, and in accordance with several previous researchers, (e.g., Ginther et al., 2010 Goldman-Eisler, 1958a, 1958b; Kormos & Dénes, 2004), I marked only silent pauses of .25 seconds or longer as silent pauses.

Silent pauses must meet a length threshold in order to be considered silent pauses, while a filled pause must meet no such threshold. The reason is that short pauses are usually not very noticeable to the listener because they do not break the flow of speech. In fact, they can actually help the listener to parse the message if placed appropriately (Clark & Tree, 2002). Although setting a minimum threshold level of duration for silent pauses at .25 is somewhat arbitrary, shorter pauses are reflected in the articulation rate variable. This is because the more short pauses a speaker makes within speech runs the longer it will take the speaker to articulate the syllables within each speech run. This, after all, is what articulation rate measures.

The second distinction relates to speech sounds and filled pauses. A speech sound must contain an attempt to convey linguistic content, as opposed to an attempt to delay. A non-lexical attempt to delay must be marked as a filled pause. Some discretion is necessary on the part of the annotator to make this distinction. Maclay and Osgood's (1959) guidelines for identification of English hesitation devices were used for lack of a foolproof method of identifying non-lexical fillers. Moreover, I considered utterance of a partial word or unintelligible sound that falls

outside of Maclay and Osgood's (1959) guidelines a speech sound because I considered it an attempt to convey linguistic content.

On the other hand, filled pauses represent neither an attempt to convey linguistic content nor a device used to help the listener parse speech, at least not in the monologic speaking mode of semi-direct testing. While Clark and Tree (2002) argued that filled pauses serve a rhetorical purpose in conversational speech, like "keeping the floor" and "ceding the floor" (p. 73), the same purpose does not apply in a situation in which the speaker (the examinee) is the only person who can possibly occupy the floor. Besides, a filled pause of any length is noticeable to the listener because in this context, the only purpose of a filled pause is to delay speech.

Simply by visually inspecting the waveform, I wa able to identify segments of the speech sample that I thought to be silent pauses. I marked each suspected silent pause with barriers on each side, which can be seen in Figure 3.2 as the solid vertical lines. Pause length would not affect speech rate or mean syllables per run; however, since some of the variables in the present study (mean silent pause length, mean filled pause length, articulation rate, and phonation time ratio) were derived from either silent pause length, filled pause length, or both, I measured and marked each pause length precisely.

I used the cursor to measure precisely. *Fluencing* allows the annotator to click and drag to highlight a stretch of audio, which allows the user to listen to just that portion of the response and see the length of just that portion at the top right of the user interface (See Figure 3.2). In this way, I was able to listen carefully and measure pause length precisely. This was especially important when measuring silent pauses, as these must be at least .25 seconds to qualify as such. I listened to each portion of audio that I suspected of being a silent pause equal to or greater than .25 to confirm that each one was in fact a silent pause, and I adjusted the boundaries of each silent pause accordingly. I listened to every potential silent pause before finalizing its length because certain sounds often looked very similar to silent pauses.

3.2.1.5 Transcription

After segmenting the speech samples, I manually transcribed the examinee responses. Using the Fluencing transcription tool (See Figure 3.3), I listened to each speech run individually and transcribed each intelligible word using Standard American English orthography. To mark disfluencies, I used the same transcription conventions (See Table 3.2 below) used by Park

(2016). For example, I marked each silent pause as an empty line in the transcription window and each filled pause as a '-'. Fluencing processes these symbols according to the transcription guidelines.

Table 3.2. *Fluencing* Transcription Conventions

Symbol	Meaning
Blank line	Silent Pause
-	Filled Pause
*	Partial word or unintelligible (one * per syllable)

After the annotator transcribes the response, *Fluencing* counts syllables automatically using a syllable dictionary. Every time the annotator transcribes and submits a word that is not in the syllable dictionary, *Fluencing* prompts the annotator to add the word and its syllable count to the syllable dictionary or delete the word if it is misspelled or a non-existent word. *Fluencing* only does this once per word. In other words, after I entered a word into the syllable dictionary, *Fluencing* counted the syllables in that word automatically and reliably every time I transcribed and submitted the same word. The syllable dictionary serves two very important purposes: automaticity and reliability.

I transcribed unintelligible speech sounds and partial words because such sounds are attempts at conveying linguistic content. *Fluencing* processes each asterisk (*) as one syllable of a partial word or one unintelligible syllable, and it includes all such syllables in the total syllable count for any given response. Use of the asterisk is illustrated below. In example 1 below, the speaker ends the speech run with two syllables, which are either unintelligible to the annotator or recognizable as two syllables of a partially spoken word. In this example, the two unintelligible or partial word syllables count as two syllables, making the total length of the speech run 13 syllables.

In an utterance in which the speaker utters a partial word or unintelligible speech sounds followed by a restart or self-correction[9], the partial word or unintelligible speech sounds and the

[9] Restarts and self-corrections were not marked because the present study did not include these variables.

words that make up the restart or self-correction are included in the total syllable count. In example 2 below, the speaker utters either two syllables of a partial word or two unintelligible syllables and then corrects himself by uttering a complete, intelligible word: "inefficient". In this example, both the unintelligible or incomplete word syllables and the self-correction are included in the syllable count.

Ex 1: Studying in a study group is very ** (13 syllables)

Ex 2: Studying in a study group is very ** inefficient (17 syllables)

Using the segmentation barriers in conjunction with the transcription conventions in Table 3.2 above and the syllable counts, *Fluencing* automatically calculated speech rate and mean length of speech run. To calculate speech rate, it simply divides the syllable count for a response by the response time. For mean syllables per speech run, it divides the syllable count by the number of speech runs in a response.

I performed some manual calculation to measure phonation time ratio, articulation rate, mean silent pause length, and mean filled pause length. This was necessary because *Fluencing* does not automatically calculate total silent and filled pause time. Mean silent pause length requires measurement of silent pause length; articulation rate requires measurement of total silent pause time; and phonation time ratio requires total silent pause time and total filled pause time. Therefore, I recorded the length of each silent and filled pause was manually in an Excel spreadsheet. After recording each pause length, the mean silent pause length, total silent pause time, and total filled pause time, I used the latter two variables to calculate articulation rate, phonation time ratio, silent pause frequency, and filled pause frequency, according to the formulas in Table 3.1 above.

Because the oral fluency analysis described above involved annotator judgement, I asked a trained testing office assistant to annotate 10% of the oral responses used in the present study, using the same process that I used to measure oral fluency. Then, I conducted a correlation analysis using the Pearson Product Moment Correlation.

3.2.2 Lexical Analysis

The ability to retrieve vocabulary from long-term memory involves multiple levels of lexical knowledge. Pawley and Syder (1983) emphasized the processing advantages of lexical chunks, describing the native speaker's large, well-organized, storehouse of multi-word units. It

stands to reason that as L2 vocabulary becomes more ample and better-organized, formulaic multi-word units become more frequent in L2 speech. Hence, formulaic multi-word units were included in the present study. Of course, it also makes sense that as L2 proficiency develops, the L2 speaker uses a wider variety of single words, and the L2 lexicon is increasingly made up of single words at different frequency levels (Laufer & Nation, 1995), some that are common to every situation and others that are more specialized and academic.

I performed lexical analyis on pre-test and post-test oral responses. Given the multiple levels of the L2 lexicon, I included three levels of lexical complexity analysis in the present study: lexical frequency profile (Cobb, 2002; Heatley, Nation, & Coxhead, 2002; Laufer & Nation, 1995), lexical diversity (McCarthy & Jarvis, 2010), and use of formulaic sequences (Hasselgren, 2002; Liu, 2003; Simpson-Vlach & Ellis, 2010 Ushigusa, 2008). Table 3.3 below summarizes the lexical complexity variables.

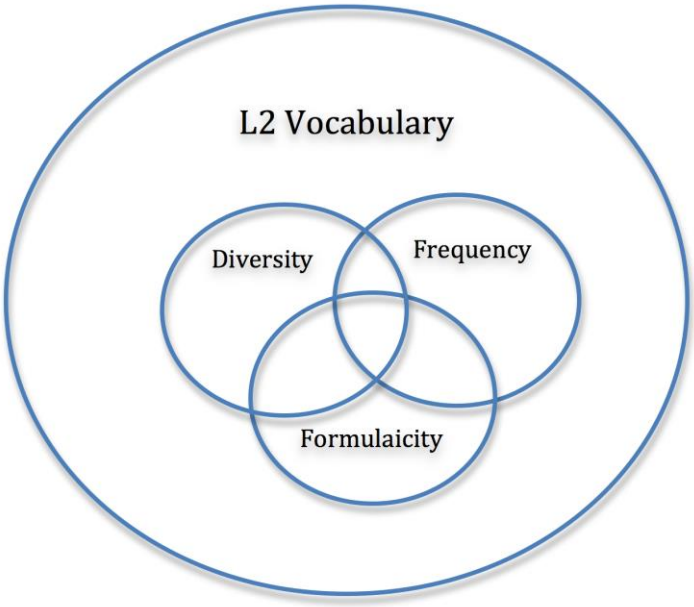

Figure 3.3 L2 Lexical Diagram

Table 3.3**.** Lexical Complexity Variables

Categories	Variables	Representation
Lexical frequency profile	K1 (%): first 1000 most common words divided by word count K2 (%): second 1000 most common words divided by word count AWL (%): Academic Word List words divided by word count OL (%): off list words, which are specialized words that are not on any of the other three lists, divided by word count	Ability to use single words at different levels of frequency.
Lexical diversity	MTLD: Measure of Textual Lexical Diversity	Ability to use a wide variety of single words.
Formulaic language	Formulaic Language (%): Number of words that make up formulaic sequences divided by word count	Ability to use multi-word lexical chunks.

I computed the lexical frequency profile for each of the ten largest fluency gainers automatically by means of the Compleat Lexical Tutor Vocab Profiler Classic (Cobb, 2002; Heatley et al., 2002), which is an online lexical analysis tool. I calculated the MTLD measure automatically using the Text Inspector online tool (McCarthy & Jarvis, 2010).

My process for calculating formulaic language proportion was by necessity more complicated. Wray (2002) noted that defining formulaic language involves automatic corpus analysis and annotator judgment. Moreover, Pawley and Syder (1983) asserted that dictionaries reflect the collective lexical knowledge of a language community. Hence, I conducted corpus analysis, English collocation dictionaries, and formulaic language lists from recent studies. I used word lists from recent studies (Hasselgren, 2002; Liu, 2003; Simpson-Vlach & Ellis, 2010; Ushigusa, 2008) on formulaic language to compile a master list from four lists that were the research product of those studies. Then, I used five popular online collocation dictionaries: *Oxford Learner's Dictionary* (2020), *Longman Dictionary of Contemporary English* (2020), *Cambridge Dictionary* (2020), *MacMillan Dict*ionary (2020), and *Collins Dictionary* (2020). I

also used the Corpus of Contemporary American English (Davies, 2008) for cross-referencing purposes.

Next, I went through each of the transcribed responses of the ten largest fluency gainers and identified all possible formulaic sequences of two words or more, being very inclusive. For each possible formulaic sequence, I searched the master list of formulaic sequences as well as the collocation dictionaries for each possible formulaic sequence. If the possible formulaic sequence appeared in *either* the master collocation list *or* any of the dictionaries, I then cross-referenced it using the COCA spoken corpus to verify that it was indeed a formulaic sequence.

I allowed for morphological modification, on the assumption that any morphological modification of a formulaic sequence could reasonably be considered evidence of knowledge thereof. For example, I counted the phrasal verb *figure out* if it appeared in a response as *figuring out* or *figured out*. Moreover, when calculating frequency and mutual information index in COCA, I included all morphological modifications of a given sequence by grouping results by lemmas.

A possible formulaic sequence had to meet the objective criteria used by Vlach-Simpson and Ellis (2010), which were frequency (> 10 occurrences per million words) and mutual information index (> 3). Mutual information index is a statistical measure of the extent to which the individual words in a sequence appear together more frequently than can be expected by random chance. I used this algorithm to ensure that formulaic sequences identified by dictionaries and/or research met the same objective criteria because different dictionaries and studies used different methodologies.

After identifying all formulaic sequences in a response, I calculated a formulaic sequence proportion by counting the number of words that made up all formulaic sequences in a response and dividing by the total number of words in the response, counting contractions as two words and excluding words that filled a slot in a separable phrasal verb. For example, I counted a contraction like *couldn't* as two words, and in a verbal phrase like *pick him up*, I excluded *him* from the formulaic sequence word count.

3.2.3 Syntactic Complexity Analysis

Similar to lexical complexity, reorganization of syntactic knowledge stored in long-term memory occurs on multiple levels. More specifically, L2 speech becomes more complex in a few

ways: 1) increasing subordination; 2) increasing coordination; and 3) lengthening of T-units. Therefore, the present study included three syntactic complexity variables (See Table 3.5 below): dependent clause ratio, coordinate clause ratio, and number of words per T-unit. Production units were defined according to Hunt's (1965) guidelines (See Table 3.4 below).

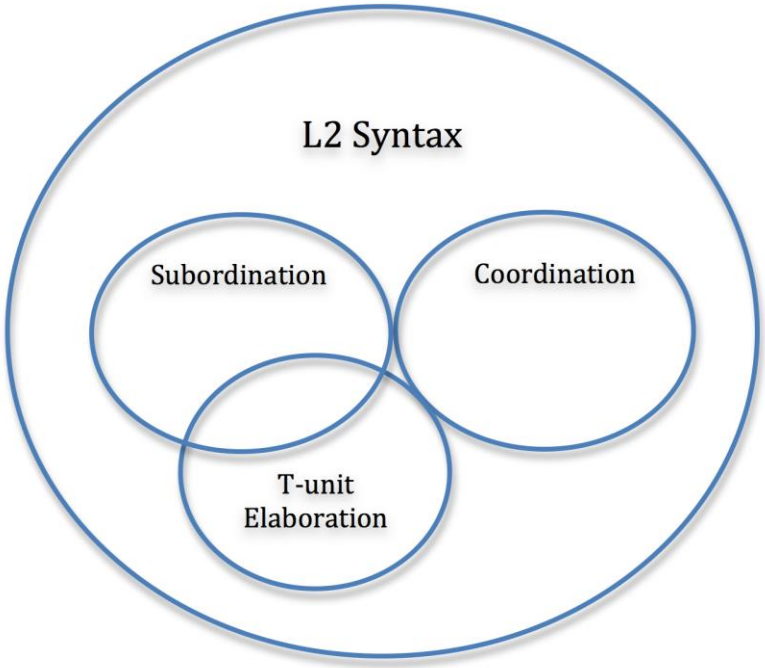

Figure 3.4. L2 Syntax Diagram

Table 3.4. Syntactic Complexity Production Unit Definitions

Unit	Definition
Clause	A group of words with a subject and a verb that shows tense.
Subordinate clause	A clause that begins with a subordinating conjunction and does not express a complete thought.
Coordinate clause	An independent clause that is connected to one or more other independent clause by a coordinating conjunction.
T-unit	A main clause and any subordinate clauses that are connected to it.

Table 3.5. Syntactic Complexity Ratio Formulas

Measure	Formula
Dependent clause ratio	Number of dependent clauses divided by total number of clauses.
Coordinate clause ratio	Number of coordinate clauses divided by total number of clauses.
Words per T-unit	Number of words divided by number of t-units.

Because the syntactic complexity analysis described above involved annotator judgement, I asked a trained testing office assistant to annotate 10% of the large fluency gainers' oral responses, using the same process that I used to measure syntactic complexity. Then, I conducted an inter-annotator agreement analysis using Brants' (2000) F-score.

3.3 Statistical Methods

As stated above, the statistical analysis for phase one (the longitudinal oral fluency development phase) involved descriptive statistics and inferential analysis. The descriptive analysis for both phase one and phase two involved computation of mean and standard deviation at T1 and T2. The inferential analysis of phase one involved paired sample t-tests of pre-post paired differences associated with each of the eight oral fluency variables. Moreover, the statistical analysis for phase two only included descriptive statistics at T1 and T2 and percentage-wise pre-post change. The research design of phase one is summarized in Table 3.6 below.

Table 3.6. Longitudinal Oral Fluency Design

Oral Fluency Variable	Pre-test	Post-test	Difference=Time 2-Time 1
Articulation rate	Syllables per minute	Syllables per minute	Syllables per minute
Speech rate	syl per min	Syllables per minute	Syllables per minute
Mean length of speech run	Syllables per run	Syllables run	Syllables per run
Phonation time ratio	Proportion[10]	Proportion	Proportion
Mean silent pause length	Seconds	Seconds	Seconds
Silent pause frequency	Silent pauses per minute	Silent pauses per minute	Silent pauses
Filled pause frequency	filled pauses per minute	filled pauses per minute	Filled Pauses

3.3.1 Statistical Tests

I used SPSS to calculate descriptive statistics (mean and standard deviation) for the sample for each oral fluency variable at T1 and T2. Then, I calculated the pre-post paired difference by subtracting the T1 mean from the T2 mean for each oral fluency variable. I went on to conduct an inferential statistical analysis of the pre-post difference for each oral fluency variable. In other words, I tested to see if the pre-post difference in examinees' oral fluency was significantly different from zero. For this purpose, I used a paired sample t-test with a family-wise alpha level of .05. This means that there was a 5% chance of concluding that there was a pre-post difference with regard to any of the eight oral fluency variables when in fact there was no such difference for that oral fluency variable. I used a Bonferroni[11] adjusted alpha level

[10] A proportion is the same as a percentage, but it is expressed in decimals between 0 and 1.
[11] A Bonferroni adjustment corrects for the fact that when researchers conduct k number of paired sample t-tests at an alpha level of α, the true probability of concluding that there is a pre-post difference with regard to one of the x tests is k*x. Hence, the alpha level should be set at α/k to adjust for the fact that multiple tests (one for each variable) were carried out.

of .00625 because it was necessary to control type I error, which is inflated when conducting multiple tests, each of which poses the same risk of finding a difference by random chance alone. Since I conducted eight paired sample t-tests, I divided the standard alpha level of .05 by eight to arrive at a Bonferroni-adjusted alpha level of .00625.

3.3.2 Statistical Assumptions

First, I analyzed the data to make sure that the measures of each paired dependent variable met the statistical assumptions of the paired sample t-test, which are summarized in Table 3.7 below.

Table 3.7. Paired Sample T-Test Statistical Assumptions

Assumption	Explanation	Test
Continuous Variable	Any value between minimum and maximum value is possible.	Researcher intuition
Independence of Observations	The measurement assigned to any observation does not depend on that assigned to any other observation.	Assumed because sample is a simple random sample drawn from the population.
Normality	Paired pre-post differences are normally distributed.	Shapiro-Wilk Test
Outliers	No extreme values.	Dunnett's Test

3.3.2.1 Outliers

The purpose of checking for outliers was three-fold: 1) to ensure that no participants' response time was significantly different from that of the overall distribution; 2) to ensure that all participants belonged to the same population; and 3) to ensure that the correct statistical analysis was conducted. In this section, each of these purposes and the procedures that I used to fulfill them are discussed individually.

First I checked each participant's response time at T1 and T2 for outliers. This was important because to adequately measure L2 oral fluency, or any other aspect of language, one

first has to have an oral response of sufficient length. To avoid setting an arbitrary cutoff length that could bias results, I used the Dunnett's outlier test to identify outliers that were statistically significant. Since multiple participants used nearly all of their allotted two-minute response time in the computer-administered task, the outliers with respect to response time were on the low side; I identified two participants with very short responses (one at T1 and one at T2) and replaced them with randomly-selected L1-Chinese examinees who took the same two-course EAP sequence and hence the same computer-administered exam at both T1 and T2. I again checked the distribution of response times for outliers, and I found none.

I then checked for participants who were not members of the population being studied, the population being L1-Chinese EAP learners in their first year of studying abroad as full-time college students. I did this by checking both the T1 and T2 distributions of oral fluency measures to identify participants with either very high or very low oral fluency measures. There were 100 participants, and I measured eight oral fluency variables at two time points, so the chances of finding one or more extreme value associated with one of the eight variables at either T1 or T2 were quite high. Therefore, to avoid identifying and removing too many participants who were in fact legitimate members of the population, I used a rigorous algorithm to identify outliers to be removed.

First, for each participant with an extreme value, I looked at both the T1 and T2 values for the associated oral fluency variable. If and only if both values were outliers, then I considered this consistent presence of extreme values sufficient evidence to conclude that the participant was not a member of the population. I found only one participant that fit this description; this participant exhibited extremely high (18 syllables<MSR<19 syllables) mean length of speech run at both T1 and T2.

Beyond looking at numerical values associated with oral fluency measures, I also examined the oral response associated with each extreme value to identify anomalies. I examined all such oral responses to make sure that all observations were recorded accurately. When I found recording errors, I corrected them. If an extreme value was not the result of a recording error, I closely examined the *Fluencing* analysis annotations to ensure that there were no errors. If there were no errors, I then checked the test administration notes for the testing session in which the participant took the test to make sure that the student did not experience any technical difficulties that may have resulted in the extreme value in question. When I found no technical

problems, I assumed that the extreme value was an accurate measure of a legitimate oral performance, and I retained the associated participant in the dataset.

Finally, I tested the paired differences associated with each oral fluency measure for outliers using the Dunnett's Test. I found that there was only one variable that had an outlying paired difference associated with it: mean length of silent pause. After checking the distribution of paired differences associated with mean length of silent pause for normality by means of the Shapiro-Wilk normality test, I found that the distribution also deviated significantly from the normal distribution. Hence, I used the Wilcoxon Signed Rank Test for this variable instead. The Wilcoxon Signed Rank Test is an acceptable non-parametric alternative to the paired sample t-test.

3.3.2.2 Normality assumption.

After testing normality of paired differences by means of the Shapiro-Wilk Test and the distributions for outliers by means of the Dunnett's Test, all assumptions of the paired sample t-test were met. Then, for the seven variables whose distributions met the paired sample t-test statistical assumptions, I ran paired sample t-tests, using the T1 and T2 data for each of the 100 participants. For mean silent pause length, I ran the Wilcoxon Signed Rank Test. I used SPSS to conduct the paired sample t-tests, and I reported the mean difference, the standard deviation of the difference, the standard error mean of the difference, a 95% confidence interval for the difference, the t statistic, the degrees of freedom (sample size minus one), and the p value. The p value, in this case, is the probability of finding a non-zero paired difference when in fact there is no such paired difference.

3.4 Summary of Methodology

This chapter described the methodology of the present dissertation in its two phases: the pre-post change in oral fluency phase and the linguistic analysis phase. The former phase involved detailed analysis to extract oral fluency measures from the oral responses of 100 L1 Chinese examinees at T1 (the beginning of a two semester EAP course sequence) and T 2 (the end of the EAP course sequence).

I conducted the detailed oral fluency analysis using a software tool called *Fluencing*, which was designed by Park (2016) specifically for oral fluency analysis. I first deleted the silence at the beginning and end of each oral response. Then, I segmented each response into silent pauses, filled pauses, and speech runs. Next, I transcribed each oral response using *Fluencing*'s transcription tool and syllable dictionary. I extracted the speech rate and mean length of speech run variables automatically from the *Fluencing* output. After that, I manually measured and extracted the filled and silent pause length information that was necessary for calculating the phonation time ratio, articulation rate, and mean silent pause length variables. I asked a colleague to conduct this same analysis process on ten examinee oral responses for the pre-test and post-test. Then, I calculated the inter-annotator agreement using Pearson Product Moment correlation for oral fluency measures and Brants' (2000) F-score for syntactic complexity ratios. Once ensuring acceptable inter-annotator agreement, I conducted a descriptive and inferential analysis of the longitudinal development of the participants' oral fluency development. The descriptive analysis involved descriptive statistics, and the inferential analysis involved paired sample t-tests.

In phase two, I conducted lexical and syntactic analyses on the ten participants who made the largest gains in terms of the oral fluency measure associated with the largest longitudinal effect size. The lexical analyses involved automatic analysis of large fluency gainer oral responses for two variables (lexical frequency profile and MTLD measure) and corpus-informed manual analysis of the same oral responses for one variable (formulaic language proportion). The syntactic analyses were entirely manual (hence requiring a second annotator), including three variables (dependent clause ratio, coordinate clause ratio, and words per T-unit). They included analysis of transcribed oral responses for identification of all main clauses, dependent clauses, coordinate clauses, and T-units, based on Hunt's (1965) definitions of all relevant production units, followed by computation of the syntactic ratios stated above. To measure pre-post change in lexical and syntactic variables I calculated descriptive statistics and subtracted pre-test results from post-test results, and I also calculated pre-post change in terms of percentage-wise change..

CHAPTER 4. RESULTS & DISCUSSION

4.1 Oral Fluency Results

For each research question, I will first present descriptive statistics, including mean and standard deviation for each variable at T1 and T2. Then, I will present inferential statistics, along with a hypothesis test for each pre-post difference in oral fluency measures. Before calculating inferential statistics and conducting hypothesis tests, I tested the statistical assumptions of the paired sample t-test, and those assumptions were met.

4.1.1 Preliminary Analysis

RQ1: How does the L2-English oral fluency of university-level L1-Chinese test-takers change over the course of two semesters of language and culture study?

I answered this research question by calculating the following oral fluency measures of the participants at T1 and T2: speech rate, mean length of speech run, phonation time ratio, articulation rate, filled pause frequency, silent pause frequency, mean length of silent pause, and mean length of filled pause.

4.1.1.1 Prompt comparability

As was mentioned earlier, I used four different prompts in this study, and participants responded to a different prompt at T2 than they did at T1. Therefore, it was necessary to verify that no prompt elicited responses exhibiting a level of oral fluency that was significantly different from any other prompt. This way it could be concluded that pre-post differences in the oral fluency of a participant's response reflected pre-post differences in oral fluency as opposed to a prompt effect.

To verify this fact, I conducted eight one-way ANOVA analyses, each with prompt as the effect and each of the eight oral fluency variables as the dependent variable. The fact that there was no statistically significant prompt effect ($p>.05$) in any of the ANOVA analyses represented sufficient evidence of prompt comparability with regard to L2 oral fluency elicited.

4.1.1.2 Inter-annotator agreement

Given the reliance on annotator judgement, a second trained annotator annotated both the T1 and T2 responses of ten randomly chosen participants. This second annotator also coded one randomly chosen large fluency gainer's T1 and T2 responses for each syntactic unit. For each fluency measure, I calculated a Pearson correlation (see table 4.1 below) to measure inter-annotator agreement. For each of the syntactic structures mentioned above, just as Lu (2010) did, I followed the procedures of Brants (2000) to calculate an F-score (see Table 4.2 below). Inter-annotator agreement was acceptable (see Tables 4.1 and 4.2 below).

Table 4.1**.** Inter-Annotator Oral Fluency Measure Pearson Correlations

Measure	Correlation
Mean length of speech run	.96**
Speech rate	.96**
Articulation rate	.91**
Phonation time ratio	.99**
Mean silent pause length	.98**
Mean filled pause length	.94**
Silent pause frequency	.94**
Filled pause frequency	.94**

Note: **p<.01

Table 4.2**.** Inter-Annotator Agreement on Syntactic Structure Identification

Structure	A1	A2	Identical	Precision	Recall	F-score
Dependent clauses	14	13	13	1.0	.93	.96
Coordinate clauses	15	13	13	.80	1.0	.89
T-units	21	17	17	.81	1.0	.89
Words	383	379	366	.96	.97	.96

As Table 4.3 shows, the means of six of the oral fluency measures changed in the expected directions: mean length of speech run, speech rate, articulation rate, and phonation time ratio increased; silent pause frequency and filled pause frequency decreased. The oral fluency measure that exhibited the largest pre-post percentage-wise increase was mean length of speech run, followed by (in descending order of percentage-wise difference) speech rate, articulation rate, and phonation time ratio; mean filled pause length and mean silent pause length increased marginally, when they could reasonably have been expected to decrease. On the other hand, filled pause frequency decreased the most percentage-wise, and silent pause frequency decreased the second most percentage-wise.

Table 4.3**.** Descriptive Statistics of Oral Fluency Measures

Measure	Mean (S.D.)		
	Pre-test	Post-test	% Difference
Mean length of speech run	7.72 (1.74)	8.46 (2.17)	+9.58%
Speech rate	172.20 (23.80)	179.45 (25.99)	+4.21%
Articulation rate	219.98 (25.81)	226.40 (26.02)	+2.92%
Phonation time ratio	.74 (.05)	.75 (.06)	+2.47%
Mean filled pause length	.34 (.08)	.34 (.08)	+.37%
Mean silent pause length	.55 (.11)	.56 (.13)	+0.29%
Silent pause frequency	23.57 (3.88)	22.39 (4.02)	-4.99%
Filled pause frequency	8.28 (4.61)	6.68 (4.62)	-19.28%

Notes: N=100 for all oral fluency measures, except mean silent pause length; N=96 for mean silent pause length because four participants used no filled pauses in either the pre-test or post-test.

4.1.1.3 Inferential Analysis

To test whether the pre-post differences with regard to oral fluency measures were statistically significant I chose the paired sample t-test, which was appropriate because the oral fluency measures of the same participants were taken at two different time points. Before conducting this test, I checked the statistical assumptions of the test: 1) continuous dependent variable: met because each variable could take any value between the minimum and maximum

value; 2) independence of observations: met because independence can be assumed if the sample is a simple random sample drawn from the population; 3) normality: met because the paired pre-post differences for each individual oral fluency variable, when submitted to a Shapiro-Wilk Test of Normality (see table 12 above), yielded a non-significant p-value ($p>.05$) for the paired differences associated with all variables, except mean length of silent pause ($p<.01$), which deviated significantly from normality; 4) no outliers: this assumption was met for all oral fluency variables, except mean silent pause duration. As stated/discussed in chapter 3, I conducted a Dunnett's outlier test on a) the paired differences associated with each dependent variable, in addition to response time and b) the values of each dependent variable, and response time, at both time points. After this analysis, I removed three outliers from the dataset and replaced them: one with extremely high mean length of speech run at both T1 and T2 and two participants with very short responses (one at T1 and the other at T2). After analyzing replacements and performing outlier tests again, the results indicated an absence of outliers, and I proceeded with the two-tailed paired sample t-test[12] for all oral fluency variables except mean silent pause length, for which the Wilcoxon Signed Rank test was used.

The paired sample t-test results appear in table 4.4 below. Results of the Wilcoxon Signed Rank Test for the paired pre-post differences in mean silent pause length indicated no statistically significant difference between T1 and T2, $Z=.76$, $p=.445$.

Table 4.4 shows the mean, standard deviation, a 95% confidence interval, t statistic, and degrees of freedom for each of the seven paired differences, one for each oral fluency measure. As Table 4.4 shows, the respective p values for six paired differences fall below the Bonferroni-adjusted alpha level of .00625, meaning these variables were associated with statistically significant pre-post changes. Four (mean length of speech run, speech rate, phonation time ratio, and articulation rate) were associated with statistically significant pre-post increases. In contrast, silent pause frequency and filled pause frequency were associated with statistically significant

[12] While the paired sample t-test only requires an absence of outliers with regard to paired differences, I wanted to make sure that no participants included in the study exhibited outlying oral fluency measures at either Time 1 or Time 2, simply because I wanted to ensure that every participant was truly a member of the population being studied. This approach was suggested to me by a language testing expert (Ginther, 2017, personal communication).

decreases. Mean length of silent pause and mean length of filled pause were not associated with a statistically significant change.

Table 4.4**.** Oral Fluency Paired Sample T-Tests

	Pre-test		Post-test			95% CI		Cohen's d
Measures	M	SD	M	SD	t(99)	LL	UL	
Mean length of speech run	7.72	1.74	8.46	2.17	4.25*	.31	.87	.42
Speech rate	172.14	23.75	179.45	25.99	3.95*	3.65	10.99	.40
Phonation time ratio	.74	.05	.75	.06	4.05*	.01	.03	.40
Articulation rate	219.90	25.85	226.41	25.89	3.69*	3.01	10.02	.37
Mean filled pause length	.34	.08	.34	.08	.22	-.02	.02	n.s.
Silent pause frequency	23.57	3.88	22.30	4.02	-2.99*	-1.96	-.40	.30
Filled pause frequency	8.26	4.59	6.70	4.65	-3.66*	-2.41	-.71	.37

Note: Bonferroni adjusted α=. 00625. *<. 00625. N=100 for all oral fluency measures, except mean filled pause length; N=96 for mean filled pause length because four participants used no filled pauses in either the pre-test or post-test.

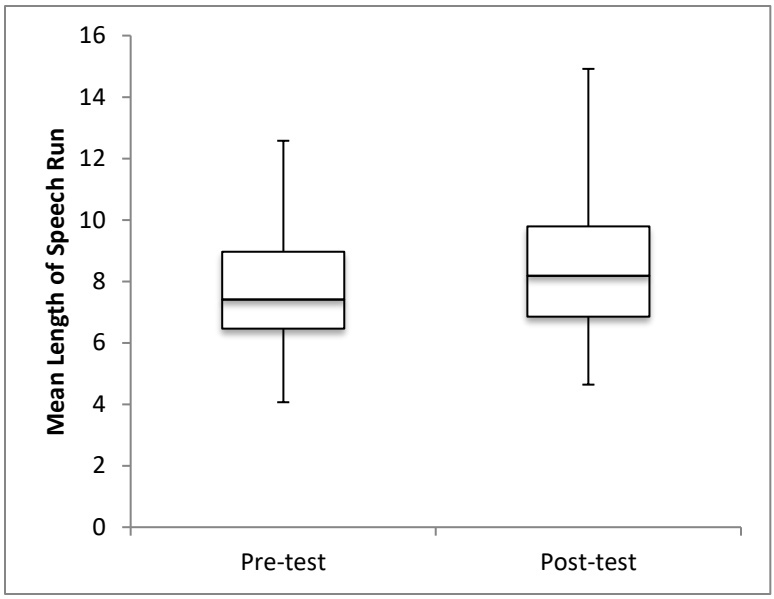

Figure 4.1. Mean Length of Speech Run Pre-Post Box Plots

I then calculated the Cohen's *d* effect size (see figure 4.2 below) associated with each statistically significant pre-post difference. The paired difference for mean length of speech run was associated with the largest effect size, followed by (in descending order of effect size) phonation time ratio, speech rate, filled pause frequency, articulation rate, and silent pause frequency.

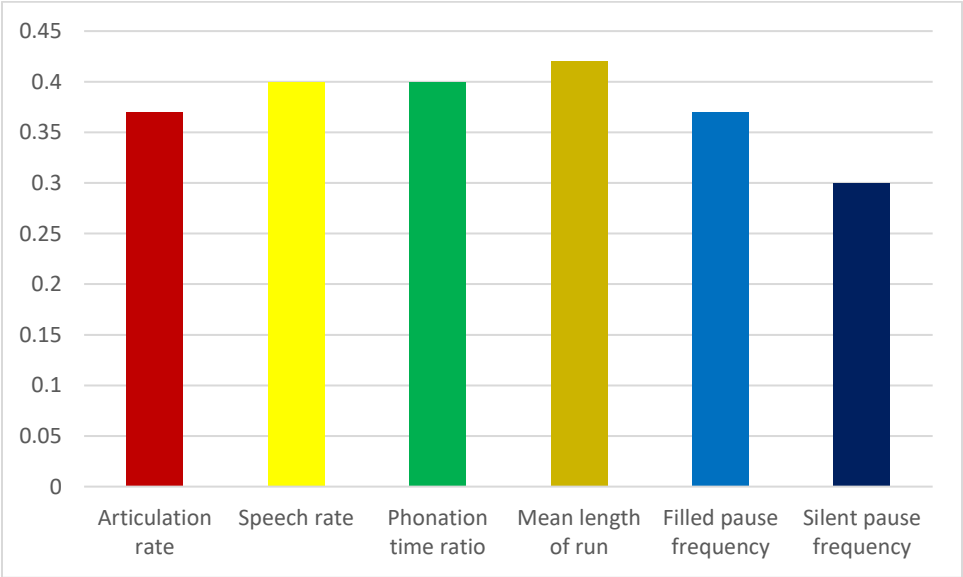

Figure 4.2 Oral Fluency Pre-Post Cohen's *d* Effect Sizes of Gains

The principal findings of the first phase of the present study have important implications, which will be explained later in the discussion. Moreover, the results of the oral fluency analysis were the point of departure for the second phase of the study. In the first phase of the study, I found that, of eight oral fluency variables, mean length of speech run increased the most in terms of Cohen's *d* effect size.

This finding is notable because mean length of speech run has emerged as one of the most important variables in L2 oral fluency research. Levelt (1989, 1999) hypothesized that L2 oral fluency is supported by vocabulary and syntax. Subsequent empirical work by Towell et al. (1996) provided some evidence in support of this hypothesis. The latter study argued that mean length of speech run was the L2 oral fluency measure most closely associated with L2 formulaic language and syntax, after comparing L2 responses involving the retelling of a short movie before and after a study abroad period. Upon analyzing the discourse of these responses, Towell et al. concluded that participants who increased their mean length of speech run by the largest magnitude did so by proceduralizing syntax and formulaic language.

4.2 Lexico-Syntactic Results

The purpose of the second phase of the study was exploratory, seeking a direction for future research, as opposed to generalizability of findings. Mean length of speech run has emerged as an important variable in L2 oral fluency research, and research suggests that 1) L2 speakers with higher proficiency exhibit higher mean length of speech run than L2 speakers of lower proficiency (Ginther et al., 2010; Kormos & Dénes, 2004) and 2) L2 speakers increase their mean length of speech run over time (Towell et al., 1996). Therefore, one question arises: by what means do L2 learners lengthen their speech runs as their L2 proficiency improves? The answer provided by previous research seems to be that L2 speakers increase their mean length of speech run by learning L2 syntax and vocabulary.

Given the results of the first phase of the present study and the findings of Towell et al. (1996), it makes sense to assume that mean length of speech run is the L2 oral fluency measure that best encapsulates the lexico-syntactic development that occurs concurrently with L2 oral fluency development. Moreover, making this assumption, it stands to reason that lexico-syntactic development over a certain time period is most observable in those L2 speakers who have increased their mean length of speech run a great deal over that time period.

Hence, for the second phase of the study, I chose a much smaller subset of the participants. I identified the ten participants who made the largest percentage-wise gains in mean length of speech run ("largest fluency gainers" henceforth), and I analyzed the oral responses of these participants linguistically. Since the above-mentioned scholars hypothesized that L2 oral fluency, in general, and mean length of speech run in particular are largely a function of vocabulary and syntax, this leads to an examination of the nature of the changes that take place in the lexico-syntactic systems of the largest fluency gainers. Lexis and syntax are the focus of research questions 2 and 3, respectively.

RQ2: How does the L2-English oral lexical ability of the largest fluency gainers change over the course of two semesters of language and culture study?

4.2.1 Pre-Post Change in Lexical Ability

Having found that the largest longitudinal change in L2 oral fluency was associated with mean length of speech run, it was then important to find out what changes took place in the vocabulary of those participants who increased their mean length of speech run the most. It was important to identify the examinees with the greatest gains to explore what is *possible* in terms of oral fluency, as opposed to what is *expected*. It is expected that the group of examinees as a whole would make some oral fluency gains, on average. Of course, some learners out of the large group (N=100) could be expected to make large oral fluency gains. By identifying and analyzing these large fluency gainers, it could be determined what lexico-syntactic and discourse changes took place over the course of two semesters of EAP language and culture instruction and university study. The findings could shed some light on the nature of individual lexico-syntactic and discourse features that may tend to accompany large fluency gains in L2 EAP learners. By extension, the findings could also inform curriculum development. In other words, curriculum developers could emphasize and enhance those aspects of language instruction that enable large fluency gains.

Therefore, the large fluency gainers' oral responses were analyzed for three different aspects of lexical use: 1) lexical frequency profile, 2) lexical diversity, and 3) formulaic sequence proportion. Since the sample size for the vocabulary and syntax phases of the study was small,

and the selection was not random, only descriptive statistics are presented henceforth. The respective pre-post results for lexical and syntactic variables appear in tables 4.6 and 4.7 below.

As can be seen in table 4.5 below, there are no large pre-post changes in the lexical analysis results. While the K2%, AWL%, and OL% exhibited double-digit percentage-wise changes, it should be noted that words in all three frequency categories were quite infrequent, each making up <5% of word count at both T1 and T2. Similarly, the 10.84% decrease in formulaic sequence proportion seems somewhat large, but formulaic language made up only a small percentage of the spoken words at both T1 and T2. The ten largest fluency gainers did not exhibit any large changes in their use of vocabulary. The lexical frequency profile changed little; and lexical diversity, as measured by the MTLD measure, changed little.

Table 4.5**.** Lexical Variable Descriptive Statistics

Measures	Mean (S.D.)		
	Pre-test	Post-test	% Difference
K1 percentage	89.12 (2.23)	92.41 (2.55)	+3.69%
K2 percentage	4.62 (2.33)	2.70 (2.12)	-41.67%
Academic Word List percentage	2.18 (1.12)	1.31 (1.15)	-39.86%
Off list percentage	4.08 (2.65)	3.58 (2.70)	-12.19%
Measure of textual lexical diversity	37.35 (4.84)	40.55 (6.17)	+8.55%
Formulaic sequence proportion	.10 (.04)	.09 (.03)	-10.84%

Note: n=10.

RQ3: How does the L2-English oral syntactic complexity of the largest fluency gainers change over the course of two semesters of language and culture study?

4.2.2 Pre-Post Change in Syntactic Complexity

Analyzing the same ten examinees as described above, I measured the extent to which they increased their oral syntactic complexity. Table 4.6 below shows the pre-post results for dependent clause ratio, coordinate clause ratio, and mean length of T-unit. Figure 4.3 displays the pre-post change in dependent clause ratio and coordinate clause ratio in a line graph.

Table 4.6**.** Syntactic Complexity Descriptive Statistics

Measures	Mean (S.D.)	Mean (S.D.)	% Difference
	Pre-test	Post-test	
Coordinate clause ratio	.40 (.13)	.49 (.10)	+23.54%
Dependent clause ratio	.43 (.11)	.45 (.10)	+2.97%
Words per T-unit	14.88 (3.11)	16.47 (3.18)	+10.70%

Note: N=10.

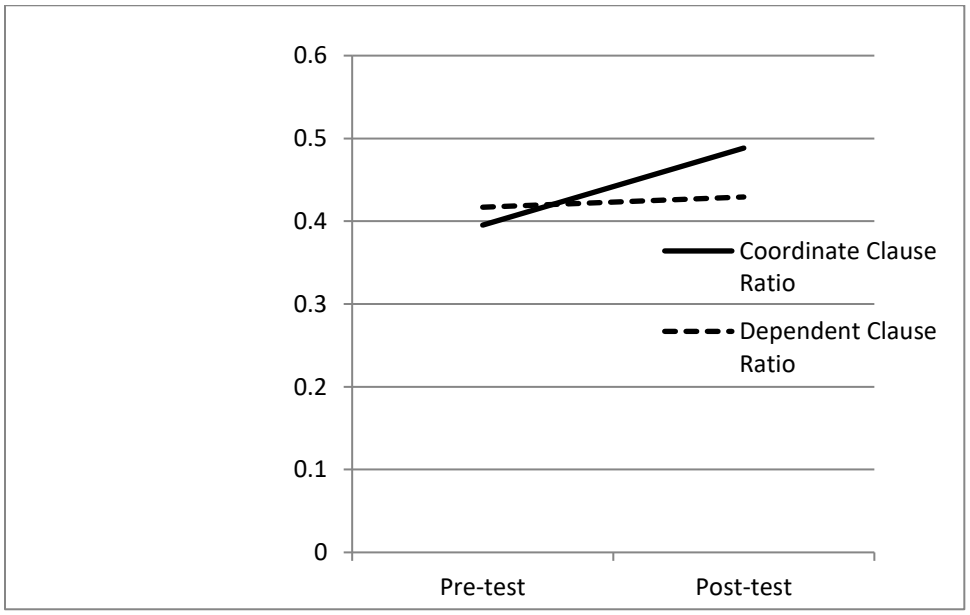

Figure 4.3. Line Graph of Pre-Post Change in Coordinate Clause Ratio and Dependent Clause Ratio

As table 4.6 above shows, there is a large pre-post change only in one syntactic complexity variable: coordinate clause ratio. Dependent clause ratio increased slightly, and mean length of T-unit increased moderately, but coordinate clause ratio exhibited the largest pre-post change (+23.54%). In fact, of the six lexico-syntactic variables, coordinate clause ratio was the only variable to exhibit large pre-post changes in the oral responses of the large fluency gainers. Of course, to gain deeper insight into how this increase in coordinate clause ratio manifested itself in oral responses, it was useful to take a closer look at the oral transcripts of large fluency

gainers. Tables 4.7-4.10 below summarize the pre-post changes in oral fluency and syntax of two exemplar large fluency gainers, who both increased their coordinate clause ratio considerably.

4.2.3 Exemplar Large Fluency Gainers

Table 4.7. Exemplar 1 Oral Fluency Measures

Measure	Pre-test	Post-test	% Change
Mean length of speech run	7.02	13.00	+85.12%
Phonation time ratio	.72	.85	+18.41%
Speech rate	162.15	213.84	+31.88%
Articulation rate	211.24	244.01	+15.51%
Mean silent pause length	.57	.47	-17.40%
Mean filled pause length	.27	.21	-21.06%
Silent pause frequency	24.63	15.86	-35.60%
Filled pause frequency	11.29	7.64	-32.35%

Table 4.8. Exemplar 1 Syntactic Complexity Measures

Syntactic Complexity Measures	Pre-test	Post-test	% Change
Coordinate clause ratio	.2083	.44	+113.33%
Dependent clause ratio	.5000	.48	-3.70%
Words per T-unit	20.27	19.46	-4.00%

What follows is a pre-post comparison of the syntax of two large fluency gainers in terms of the discourse that they produced. These exemplars' general patterns of development in terms of syntax and overall discourse were typical of those seen in the large fluency gainer group. The discourse of Exemplar 1's oral response at T1 and T2 is analyzed below as an example of how L2 syntax, fluency, and discourse develop together. To view the pre-post oral fluency and syntactic complexity means for exemplar 1, see tables 4.8 and 4.9 above, respectively. Exemplar 1, with an 85.12% increase in mean length of speech run, was the second largest oral fluency gainer of the 100 participants in the study. Moreover, of the 100 participants, Exemplar 1 made the second largest gains in coordinate clause ratio, with a 113.33% increase. (Notes on Annotation: Coordinate clauses in bold; SP=Silent Pause; FP=Filled Pause).

4.2.3.1 Exemplar 1 T1 transcript

1 basically I agree with the statement
(SP)
2 joining a (partial word)
(SP)
3 student club is a good idea
(SP)
4 there are
(SP)
5 three main reasons
(SP)
6 the first reason is that (FP)
(SP)
7 joining
(SP)
8 (Restart) joining a student club
(SP)
9 can improve our skills and broaden our horizons
(SP)

10 for instance
(SP)
11 if you are interested in statistics or mathematics
(SP) (FP) (SP)
12 you join a statistics club
(SP)
13 then **it is very likely** that you will have
(SP) (FP) (SP)
14 (Self-correction) you will get access to resources that you
(SP) (FP) (SP)
15 will never
(SP)
16 (FP) (Self-correction) you never touch before
(SP)
17 **and**
(SP)
18 (FP) through
(SP)
19 (FP) (Repetition) through conferences and (FP)
(SP)
20 competitions **you will also improve your skills of solving problems**
(SP)
21 and also if you choose to
(SP) (FP) (SP)
22 take part in
(SP)
23 activities and join a club (FP) like
(SP) (FP) (SP)
24 (FP) charities (FP) **and then you may get a chance of working out**
(FP) **your**
(SP)

25 **home**

(SP)

26 (FP) **(Self-correction) hometown**

(SP)

27 **(Repetition) working out your country**

(SP)

28 (FP) to see more things

(SP)

29 the second reason is that it is a

(SP)

30 good idea because

(SP) (FP) (SP)

31 (FP) it is very important for you to communicate with others on this

(SP)

32 projects you are working on within the club

(SP)

33 (FP) being a leader is

(SP)

34 (FP) very challenging

(SP)

35 (FP) if you

(SP)

36 are not (Unintelligible)

(SP)

37 (Repetition) are not good at communicating with other people

(SP)

38 (FP) then it is hard to lead a club

(SP)

39 (FP) thirdly

(SP)

40 **many students complain about college life** and think it is a bad idea to join

(SP)

41 that they don't know how to manage

(SP)

42 time properly

(SP)

43 **and it's important to manage time properly**

(SP)

44 to join a club

4.2.3.2 Exemplar 1 T1 analysis.

Similar to the overall large fluency gainer group, the most striking longitudinal change in exemplar 1's syntax involves the use of coordinate clauses. In the T1 response, the first coordinate clause does not appear until speech run 13. The speaker uses the coordinating conjunction *and* in speech runs 17, 21, and 24 to change topics in a somewhat incoherent attempt at developing the main idea that "joining a student club is a good idea". The speaker uses only the coordinating conjunction *and*, and the content of these (notably disfluent) coordinate clauses reveals no clear semantic connection, or even an obvious attempt to draw such a connection between the details that form the support of the argument: "access to resources", "conferences and competitions", "charities", and then "working out (of) your (home) country". Nor is there much semantic connection between the rest of the ideas that make up the argument: communication, leadership, and time management.

There are also examples in this response that may reflect lack of proceduralization of syntax at the phrasal level. For example, in speech runs 2, 3, 7, and 8, the speaker has trouble producing the gerund phrase in subject position: "joining a student club". The difficulty is particularly noteworthy because the participant is simply repeating prompt language.

4.2.3.3 Exemplar 1 T2 transcript.

1 in my opinion (FP) studying with a study group is a good idea (FP) because studying with a study group can help you to share your knowledges and ask questions from the people who have the same interests with you
SP

2 (FP) in the first semester I came to the university **I found myself interested in a particular area**
(SP)

3 **and I tried to learn the area by myself**
(SP)

4 by surfing the internet and reading the relevant books
(SP)

5 and taking the relevant online courses
(SP)

6 **and**
(SP)

7 yes indeed after a semester **I found myself**
(SP)

8 (FP) **the skills of that particular area** (FP) **improved greatly**
(SP)
9 **but I also have some problems**
(SP)
10 that (FP) when I (partial word) studies
(SP)
11 and learn something fun
(SP)
12 (FP) I have no one to share with
(SP)
13 because I don't know anyone who have the same interest with me
(SP)
14 also if I encounter some problems that (FP) I cannot solve (unintelligible) (FP) through the internet
(SP)
15 (FP) **I have (Self-correction) I also could have no one to turn to**
(SP)
16 (FP) **but studying within a study group is different**
(SP)
17 within a study group **you will meet the students** who have the same interests with you
(SP)
18 **and you can learn things together**
(SP)
19 **and you can discuss** what you learn together to
(SP)
20 share your interesting opinions
(SP)
21 **and** also if you have any problems **you can always turn to the student within that study group for help**
(SP)
22 helping students to
(SP)
23 (FP) make friends with each other and (FP)
(SP)
24 growing each other in that area
(SP)
25 **so**
(SP
26 in conclusion **I think study with a study group is a good idea**

4.2.3.4 The interface of exemplar 1's discourse, syntax, and fluency.

Notes on Annotation: X=a complete clause; X= coordinate clauses; >=an incomplete clause that is interrupted by a pause;<= the completion of a clause previously interrupted by a pause or an incomplete clause added onto a preceding completed clause; S=silent pause; F=filled pause)

Figure 4.4. Exemplar 1 T1 Multi-Level Diagram

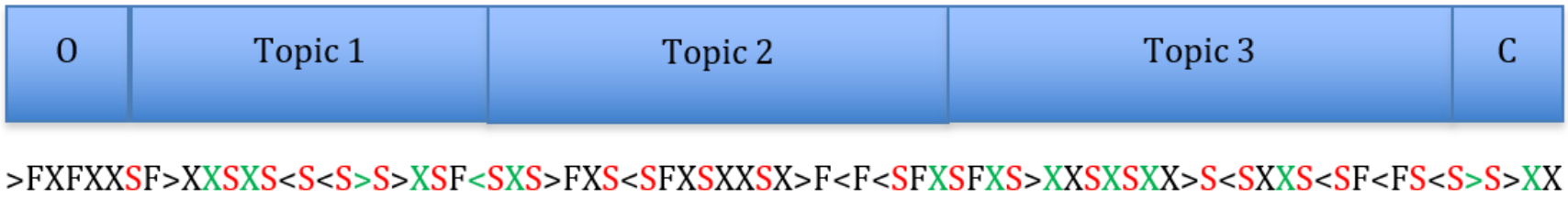

Figure 4.5. Exemplar 1 T2 Multi-Level Diagram

4.2.3.5 Exemplar 1 T2 analysis

At T2, the speaker's improved use of coordinate clauses is obvious. The speaker uses a wider variety of coordinating conjunctions, earlier, more frequently, and in more rhetorically sophisticated ways. The first pair of coordinate clauses appear in speech runs two and three. These two coordinate clauses, connected by the coordinating conjunction and, begin to convey a complex, personalized message that defines the problem to which "studying with a study group" is the solution.

The speaker goes on to use coordinate clauses to develop ideas and transition. Separate coordinate clauses begin in speech runs 2, 3, 6, 9, 15-19, 21, and 25, respectively. In speech runs nine and 16, the speaker uses the coordinating conjunction *but* to connect coordinate clause pairs that transition effectively. The first of these coordinate clause pairs (made up of coordinate clauses starting in speech runs seven and nine, respectively) transitions between two main ideas: the rewards of individual study and the problems involved in studying alone. The second pair of coordinate clauses connected by *but* (starting in speech runs 15 and 16, respectively) transition between two main ideas: problems with individual study and the solution: group study. As can be seen visually in the multi-level diagram for Exemplar 1's T2 response, coordinate clause pairs overlap the boundaries between topics one and two and two and three. The speaker follows up with four consecutive coordinate clauses, which develop topic three in long, syntactically well-formed speech runs. In speech run 25, the speaker uses a final coordinating conjunction, *so*, to transition to a long, syntactically complex, closing speech run.

At T2, there are clear signs of proceduralization of syntax at the clausal level and below. Looking at the T2 multi-level diagram, the most obvious indication of greater automaticity at T2 than at T1 is that there are so many more runs that include two or more complete clauses. In fact, there are eight examples of such speech runs in the T2 response, versus only three occurrences at T1. These more syntactically complex speech runs increase the overall mean length of speech run at T1 and T2. The mean length of speech run for exemplar 1's multi-clausal runs at T1 is 14.00 syllables, versus 7.02 syllables for theT1 response overall. Multi-clausal runs at T2 averaged 25.25 syllables, versus an overall mean length of speech run of 13.00. Moreover, two of the three multi-clausal speech runs at T1 contained a coordinate clause, and of the eight multi-clausal speech runs in the T2 response, five speech runs contained a coordinate clause.

Furthermore, in topic three of the T2 response alone, there were three multi-clausal speech runs, all of which contained a coordinate clause.

Some examples of greater proceduralization at T2 than at T1 involve production of the gerund phrase. In contrast to the opening speech runs of the speaker's response at T1, the speaker produces the gerund in subject position, "studying with a study group" in this case, twice fluently in speech run one to build a very long speech run that serves as the thesis of the response. Furthermore, in speech runs one, four, and five, exemplar 1 uses coordinate phrases to produce long (>10 syllable) speech runs. It is also noteworthy that at T2, almost all pauses take place at clause boundaries, whereas at T1, many pauses occurr mid-clause.

Table 4.9**.** Exemplar 2 Oral Fluency Measures

Oral Fluency Measure	Pre-test	Post-test	% Change
Mean length of speech run	7.46	10.42	+39.68%
Phonation time ratio	.75	.80	+7.16%
Speech rate	158.28	175.52	+10.89%
Articulation rate	196.03	206.41	+5.30%
Mean silent pause length	.49	.54	+8.89%
Mean filled pause length	.27	.40	+46.83%
Silent pause frequency	23.39	17.28	-26.82%
Filled pause frequency	13.05	6.48	-50.36%

Table 4.10**.** Exemplar 2 Syntactic Complexity Measures

Syntactic Complexity Measures	Pre-test	Post-test	% Change
Coordinate Clause Ratio	.28	.65	+132.48%
Dependent Clause Ratio	.53	.42	-20.36%
Words per T-unit	18.58	14.94	-19.62%

*4.2.*3.6 *Exemplar 2*

Tables 4.9 and 4.10 summarize the oral fluency and syntactic complexity means, respectively, of exemplar 2. Of the 100 participants, exemplar 2 was the eighth largest fluency gainer (+39.68%)

and made the largest gains in coordinate clause ratio (+132.48%) of the ten largest fluency gainers.

4.2.3.7 Exemplar 2 T1 transcript.

(FP) (SP)
1 actually I think living off campus is not a good idea
(SP)
2 the first reason is living off campus
(SP)
3 this (FP) (Self-correction) **it means** we have to live in an apartment
(SP)
4 which is far from campus
(SP)
5 **so** if we have to go to class
(SP)
6 **it will take a long time to get to the classes**
(SP) (FP) (SP)
7 **the second reason is**
(SP)
8 if we live
(SP)
9 out
(SP)
10 (Self-correction) off campus we have to cook by ourselves
(SP) (FP) (SP)
11 which means we have to buy vegetables and some food
(SP)
12 (FP) from supermarket
(SP)
13 **and it will**
(SP)
14 (FP) **both take time and take our energy**
(SP)
15 which (FP)
(SP)
16 (FP) (Restart) which will
(SP)
17 (FP) waste our time and (FP)
(SP) (FP) (SP)
18 **it's** (two partial words)
(SP)
19 (Self-correction) **inconvenient for us**
(SP)

20 (FP) the second reason is
(SP)
21 (FP) living off campus (FP) also means we have to
(SP)
22 (FP) deal with the renting things with the manage (FP) (self-correction) manager of apartment
(SP)

23 like **we have to talk about the electronic fee the water fee or something**
(SP)
24 and it will cost
(SP)
25 **and**
(SP)
26 **it will cost lots of energy** which
(SP)
27 (FP) (Repetition) which we cannot concentrate most of our time on study
(SP)
28 as for
(SP)
29 (FP) (Repetition) as for living campus
(SP)
30 living off campus is
(SP)
31 (FP) cheaper than living in campus
(SP) (FP) (SP)
32 **I don't think** it is (FP) important cause
(SP)
33 the time we saved
(SP)
34 (FP) when we live in campus
(SP)
35 will help us to learn more
(SP)
36 which can
(SP)
37 (FP) (Restart) which can help us to earn more
(SP)
38 more money in the future
(SP)
39 and
(SP)
40 **so living off campus is not a good idea**

4.2.3.8 Exemplar 1 T2 analysis

Upon analyzing the discourse of exemplar 2's T1 response, two interrelated observations come to mind: the syntax and the rhetoric are quite one-dimensional. Stated another way, the speaker fails to craft a complex message, which is reflected in lack of elaboration on each individual point, and the speaker relies heavily on clausal subordination without much clausal coordination. Short speech runs and frequent pauses accompany the rhetorical and syntactic simplicity at T1.

The lack of syntactic variety is apparent in the distribution of clause types. Coordinate clauses are few. Separate coordinate clauses begin in speech runs 3, 5, 7, 13, 18, 23, 25, 32, and 40, respectively. In contrast, over half of the clauses are subordinate clauses, with separate subordinate clauses beginning in speech runs 1-5, 8, 11, 16, 20, 27, 30, 32, and 37, some of which are broken into multiple short speech runs. Some of these speech runs containing subordinate clauses are quite long, for example the two non-restrictive clauses in speech runs 11 and 27, respectively; however, the speaker's use of subordination does not lead to sustained speech run elongation beyond one or two speech runs at a time.

The lack of syntactic variety goes hand-in-hand with rhetorical one dimensionality and lack of elaboration. First, the speaker seems to have only one main argument in support of the thesis that "living off campus is not a good idea", and that is that living off campus wastes time. Each topic (time spent driving, cooking, and paying bills, respectively) is made up of only 4-6 speech runs, some of which are very short. The speaker does not elaborate much on any point, weigh the pros and cons of the position chosen, address counterarguments, or convey a complex personalized message in defense of the chosen position.

Just as in the exemplar 1 T1 response, there is also evidence that exemplar 2 has not proceduralized syntax at the clausal level and below. For example, there are mid-clause pauses after speech runs 8-9, 15-6, 18, 20, 21, 24, 26, 30, and 32, and 36. The speaker also struggles with the gerund in subject position. After using gerund in subject position to repeat the prompt in speech run 1, the speaker has trouble using that same gerund in subject position to create a novel utterance, leading to breakdown in speech runs 2-3.

4.2.3.9 Exemplar 2 T2 transcript

1 (FP) **I think** taking courses online is not a good idea
(SP) (FP) (SP)
2 because I am not a person who can control themselves really well
(SP)
3 **and also I take the online course before**
(SP)
4 **and that is not work on me**
(SP)
5 **so first reason is** that I always miss the classes
(SP)
6 like (FP) **I was supposed to watch that video before** (FP) like Sunday
(SP)
7 **but I always just**
(SP)
8 **think there's no (Repetition) no other important things**
(SP)
9 (FP) to do with that class **there's no homework due on Sunday so I can just push off that**
(SP)
10 (Repetition) **that time so i always**
(SP)
11 (FP) **watch that video like on Tuesday or even Wednesday**
(SP)
12 **or sometimes just not watch that**
(SP)
13 **so**
(SP)
14 **it also makes me keep missing the homework progress**
(SP)
15 Like
(SP)
16 **i only** (Unintelligible) (FP) **watch that video before the homework due**
(SP)
17 **and it is a not good way to study**
(SP)
18 also **i think** online course is also
(SP) (FP) (SP)
19 make us lose the opportunity to
(SP)
20 meet with professor to talk with them to share our opinion with them
(SP)
21 cause online course is poor at community with (FP)
(SP)
22 (Self-correction) have some conversation with others
(SP)

23 **so someone may say that the online course is effective**
(SP)
24 cause it save time
(SP)
25 from (FP) like (Unintelligible) walking to class
(SP)
26 or taking bus something
(SP)
27 **but i think** (FP) **walking to the classroom or taking the bus is also good for us** to
(SP)
28 refresh our mind
(SP)
29 to like exercise ourselves
(SP)
30 **so**
(SP)
31 (FP) **this is a reason why i think taking course online is not a good idea**

4.2.3.10 The interface of exemplar 2's discourse, syntax, and fluency

Notes on Annotation: X=a complete clause; X= coordinate clauses***; >=an incomplete clause that is interrupted by a pause;<= the completion of a clause previously interrupted by a pause*** *or* an incomplete clause added onto a preceding completed clause; S=silent pause; F=filled pause

Figure 4.6. Exemplar 2 T1 Multi-Level Diagram

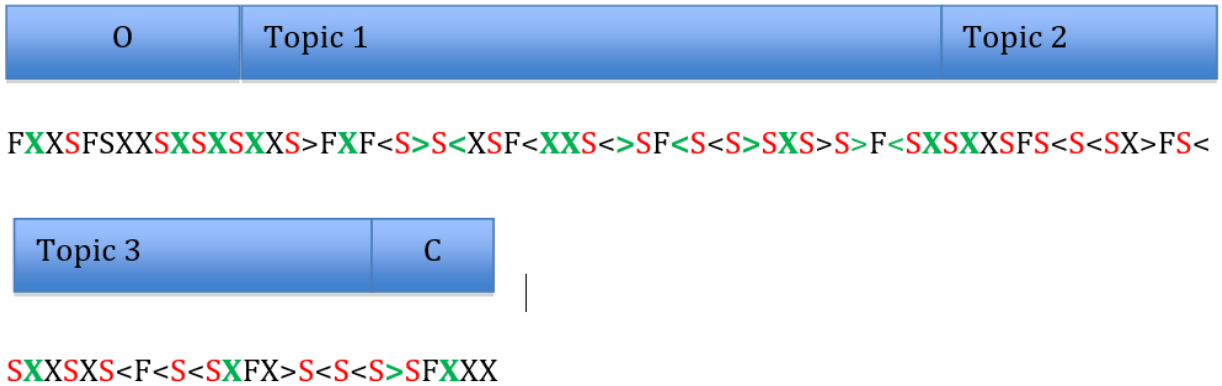

Figure 4.7. Exemplar 2 T2 Multi-Level Diagram

4.2.3.11 Exemplar 2 T2 analysis

In comparison to T1, exemplar 2's more frequent and qualitatively superior use of coordinate clauses at T2 is apparent from the beginning. Not only does the speaker use a wider variety of coordinating conjunctions to construct more coordinate clauses at T2 than at T1, but those coordinate clauses at T2 also suit the speaker's rhetorical purpose better than the (primarily) subordinate clauses in the T1 response. In the T2 response, separate coordinate clauses begin in speech runs 1, 3-7, 9-10, 13, 16-8, 23, 27, and 30. Six of the first seven speech runs (1, 3-7) contain coordinate clauses. The first speech run takes a clear position on the prompt: "taking courses online is not a good idea"; the second speech run briefly explains why the speaker takes that position: lack of self-control; then, speech runs 3-5 transition to a complex, personalized message illustrating the thesis: the speaker's lack of self-control when it comes to taking online courses.

Two coordinate clauses connected by *but* begin in speech runs 6 and 7, respectively. These define the source of the problem: online course expectations and the ease with which the speaker ignored these expectations, when taking an online course. The coordinate clauses beginning in speech runs 10 and 13 transition from the source of the problem to the consequence: missed homework. Finally, the coordinate clauses beginning in speech runs 23 and 27, respectively, state the counter-argument (the fact that online courses save time) and then refute that counter-argument. The speaker is clearly putting coordinate clauses to work in more rhetorically effective discourse at T2 than at T1.

Similar to exemplar 1, exemplar 2 exhibited greater syntactic complexity at the speech run level at T2 than at T1. Evidence of this pre-post change is the fact that, at T1, exemplar 2 produced five multi-clausal speech runs, while exemplar 2 produced eight at T2. There is also evidence at T2 that coordinate clauses facilitated the production of longer, more syntactically complex speech runs. In fact, at T1, two of the five multi-clausal speech runs contained at least one coordinate clause, and at T2, seven of eight multi-clausal speech runs contained at least one coordinate clause.

Just as in exemplar 1's responses, there is evidence in exemplar 2's T1 and T2 response that these multi-clausal speech runs increased the overall mean length of speech run of each response. In fact, in exemplar 2's T1 response, multi-clausal speech runs had a mean length of speech run of 12.20 syllables per speech run, versus an overall mean length of speech run of 7.46

syllables. Moreover, at T2, exemplar 2's multi-clausal speech runs had a mean length of speech run of 16.13 syllables, versus an overall mean length of speech run of 10.42 syllables.

One fact stands out in exemplar 2's T2 response. That is that coordinate clauses tend to serve an organizational purpose in the online production of complex, personalized discourse made up of multiple long speech runs. At T2, exemplar 2 uses coordinate clauses to elaborate on the speaker's complex thought processes, creating logical connections between past experience and the thesis. The result is a more sophisticated argument built on a foundation of personal experience, each idea building on the one before it.

4.2.3.12 Summary of exemplar analysis

From the syntactic and discourse analysis of the large fluency gainers' responses, a few key takeaways can be drawn. First, as was just mentioned, coordinate clauses seem to play more of a discourse organizing function at T2 than at T1. They occur at points where the speaker is constructing a complex, personalized message in response to the prompt. Moreover, they appear to allow the speaker to transition between main points and build sophisticated discourse structures, like pro's-cons, problem-solution, and counterargument-refutation.

Second, most of the multi-clausal speech runs contain a coordinate clause, even though most of the multi-clausal speech runs do not contain a pair of coordinate clauses, but rather a main clause and dependent clause. In other words, coordinate clauses do not appear to be directly contributing to syntactic complexity at the speech run level, but this does not preclude a contribution at the discourse level, which brings us to the third takeaway.

When analyzing the pre-post change in L2 syntax and speech run length, it is necessary to consider multiple levels of analysis: the clause, the utterance, and the discourse. An increase in the quantity and quality of coordinate clause use may result in speech run elongation over multiple speech runs (the discourse level). The possible psycholinguistic implications of this finding will be discussed further in the discussion section below.

4.3 Oral Fluency Discussion

Viewed holistically, the findings of the pre-post oral fluency analysis confirm and expand upon the claims made by Towell et al. (1996). To review, Towell et al. (1996) provided evidence

from 12 L2 learners of French studying abroad that L2 oral fluency development is characterized by an increase in automaticity, which takes the form of the following changes:

- Learners produced more syllables per minute (increased speech rate)
- *and* elongated their speech runs (higher mean length of speech run)
- *without* slowing their articulation rate (no decrease in articulation rate)
- *or* spending a lower proportion of the response time speaking (no decrease in phonation time ratio)
- *or* pausing longer (no increase in mean length of silent pause) to plan.

For comparison's sake, the present study found that mean length of speech run, speech rate, phonation time ratio, and articulation rate were each associated with a statistically significant increase, while filled pause frequency and silent pause frequency were each associated with a statistically significant decrease, and the pre-post differences associated with mean length of silent pause and mean length of filled pause, respectively, were not statistically significant. Therefore, the present study adds considerably to the evidence provided by Towell et al. (1996) in favor of the validity of the L2 Speech Production Model (De Bot, 1992; Levelt, 1989, 1999). At the level of individual pre-post oral fluency variables, there are also interesting findings.

4.3.1 Fluency as Flow

One important finding in the pre-post oral fluency analysis was that participants made the largest gains in mean length of speech run. This finding is consistent with the findings of Kormos and Dénes (2004), who found that mean length of speech run was the oral fluency measure that correlated most strongly with the L2 oral English fluency ratings of both trained L1-English raters and trained L1-Hungarian raters. Kormos and Dénes also found that mean length of speech run distinguished better than any other oral fluency measure between advanced and low intermediate L2 English learners.

The finding with regard to mean length of speech run in the present study was also consistent with Ginther et al. (2010). This was a language testing study showing that, of 15 L2 oral fluency variables, mean length of speech run was associated with the strongest overall

correlation to Oral English Proficiency Test scores, which were based on ratings performed by trained raters, and that correlation was positive. It stands to reason that if mean length of speech run is the L2 oral fluency variable that is most strongly and positively correlated with both oral fluency ratings and overall oral English proficiency test scores, then mean length of speech run should be associated with the largest pre-post gain, given sufficient time and adequate instruction and immersion in the target language. The findings of the present study show that the L1 Chinese participants analyzed did, in fact, exhibit their largest gains in mean length of speech run. The mean length of speech run finding is similar, but slightly different, from the finding of Towell et al. (1996) that mean length of speech run was associated with a smaller effect size than speech rate. The fact that in the present study, mean length of speech run was associated with a slightly larger effect size than speech rate supports Ginther et al.'s (2010) argument that mean length of speech run is a better representation of global L2 oral fluency than speech rate is.

The findings of the present study also bolster Segalowitz's (2010) argument that the interconnected processes involved in L2 oral fluency favor continuance of the flow of speech. In support of this argument, Segalowitz (2010) cited Filmore (1979), who wrote that "the ability to talk at length with a minimum of pauses" (p. 4) was one of the fluency-related abilities by which people judge fluent speech. Segalowitz (2010) also cited Freed's (2000) finding from a study involving the perceptions of six L1-French speakers who judged the fluency of L2 spoken French. In this study, most of the judges mentioned “smoother speech" and “fewer pauses/hesitancies” (p. 4) as criteria that they considered when evaluating fluency.

The present study provides more pieces of the L2 oral fluency puzzle, which fit with the research discussed above. Mean length of speech run may be considered a measure of ability to speak "at length" or the ability to produce speech "smoothly". Others (Ginther et al., 2010) have referred to it as a measure of "density". To the extent that any (or all) of these characterizations of fluent speech are psycholinguistically valid measures of fluently delivered L2 speech, one would expect these features to increase over time in the speech of L2 EAP learners. Further assuming that mean length of speech run is a valid measure of "length", "smoothness", and/or "density", the findings of the present study provide empirical support for characterizing these features as emblematic of fluent L2 speech because they did in fact increase over time in a group of 100 L1-Chinese EAP learners.

The finding of a gain in phonation time ratio lends more evidence to this argument that the essence of L2 oral fluency is continuance of the flow of speech. Phonation time ratio, it should be remembered, is the ratio of speech time (excluding pause time) to response time. Since response time is comprised of phonation (speech) time and pause time, an increase in phonation time ratio is synonymous with a decrease in pause time ratio because pause-time ratio equals one minus phonation time ratio. The fact that the effect size associated with the longitudinal increase in phonation time ratio is almost as high as that for mean length of speech run (.40 vs .42) may suggest that lengthening speech runs and spending more time speaking, as opposed to pausing, are complementary developments.

Before explaining this complementarity more fully, it is necessary to discuss pause time ratio, a variable which was only measured indirectly in the present study. It should be noted that pause time is the sum of two variables: filled pause time and silent pause time. To further break these two variables down, filled pause time and silent pause time are each a product of the length and frequency of the pauses in each respective pause category. Hence, pause-time ratio can decrease in only four ways: 1) a decrease in mean filled pause length, 2) a decrease in filled pause frequency, 3) a decrease in mean silent pause length, or 4) a decrease in silent pause frequency. It might be expected that a decrease in pause-time ratio would occur by means of some combination of the four trends just described. In point of fact, the decrease was entirely a function of two of these four trends: 2) and 4) above.

Evidence of the complementarity between mean length of speech run and absence of pausing are the pre-post trends in pause variables. It is noteworthy that while participants spent a smaller proportion of their response time pausing, mean length of both silent and filled pauses changed very little. The fact that these two measures changed very little pre-post, while pause-time ratio decreased, means that all of the decrease in time spent pausing was accounted for by decreases in pause frequency. Moreover, the findings of the present study support this conclusion, with both filled pauses per minute ($p<.001$) and silent pauses per minute ($p=.003$) associated with statistically significant decreases. Clearly, if a speaker pauses less frequently, speech runs, by necessity, become longer. This finding aligns with the literature discussed earlier in this section (Fillmore, 1979; Freed, 2000; Segalowitz, 2010), which argued that absence of pausing is an important feature of fluent speech. In other words, the present study provides pre-

post evidence that progress towards "absence of pausing" means *fewer* pauses, not *shorter* pauses.

4.3.2 Fluency as Speed

The final two oral fluency variables worth mentioning are the speed variables: articulation rate and speech rate. The Cohen's *d* effect size of the increase associated with speech rate is larger than that for articulation rate. This difference implies that the L2 speakers were able to produce more speech at T2 than at T1, not only by speaking faster within speech runs, but also by spending less time pausing. The findings associated with speech rate and articulation rate suggest that proceduralization of the speech production mechanisms in the articulation phase of the L2 Speech Production Model (De Bot, 1992; Levelt, 1989, 1999) occurs, but it is not the driving force behind L2 oral fluency development. This finding makes sense because there is an upper limit to the speed of articulation, even for L1 speakers. Moreover, advanced L2 speakers who have already met the TOEFL speaking cut score at Purdue presumably already meet a fairly high threshold of articulation rate at T1. This fact alone limited the range of possible articulation rate gains that they could make. The rest of the oral fluency gains would need to come from decreased pausing and lengthening of speech runs.

4.3.3 Lingering Questions About Mean Length of Speech Run

The findings with regard to mean length of speech run present some interesting questions, which can be considered by comparing mean length of speech run to speech rate. Participants made gains of a similar magnitude in mean length of speech run and speech rate. These findings align with multiple studies that have shown a great deal of *overlap* between these variables; however, conceptually, these two variables are quite different. Speech rate directly involves speed as well as absence of pausing. On the other hand, while rapid articulation may enable some degree of speech run elongation, mean length of speech run is not directly related to speed. In contrast to speech rate, mean length of speech run, to a greater extent, entails the ability to continue speaking at the individual utterance level without silent pausing.

Hence, the component measures underlying speech rate and mean length of speech run are different. Speech rate increases as articulation becomes more rapid, but speech rate also

increases as pausing, however measured, becomes less prevalent. In contrast, the mechanisms underlying mean length of speech run are less obvious. Since a speech run is a run of continuous speech between two silent pauses, only silent pause frequency directly affects mean length of speech run. This makes sense because a silent pause is the only fluency feature that can interrupt a speech run, thus preventing a speaker from continuing the flow of speech.

Intuitively, the ability to continue speaking at the utterance level aids listener comprehension. After all, oral communication largely takes place one or a few utterances at a time. Speakers convey meaning orally in phrasal and clausal units, and the listener can process these units more easily when they are delivered as complete units in the same utterance, with pauses placed between units instead of mid-unit. In fact, listeners expect them to be delivered in this way. Clarke and Tree (2002) discussed how fluency features of speech attend to listener expectations. Moreover, empirical findings suggest that L2 speakers with higher mean length of speech run are perceived as more fluent (Kormos & Dénes, 2004) and rated as more proficient (Ginther et al., 2010).

That mean length of speech run has proven such a good measure of L2 oral proficiency makes it even more important to find out how it develops over time. The fact that mean length of speech run has so few constituent parts makes it difficult to understand solely in terms of sound and silence. The rest of the present discussion will move beyond sound and silence. The question of what drives gains in mean length of speech run certainly involves linguistic change, but it may also involve cognitive developments in the conceptual preparation phase of the Levelt Speech Production Model (1989, 1999). It certainly makes sense that EAP learners would improve their conceptual preparation over the course of two semesters studying in a language and culture curriculum during their first year of college. Learning more about this abstract process of cognitive development requires a closer look at pre-post changes in lexico-syntax as well as discourse organization. Clues may be found at the nexus of vocabulary, syntax, and discourse.

4.4 Lexico-Syntactic Analysis

The most important finding from the pre-post linguistic analysis of the top ten fluency gainers was the large increase (+23.54%) in coordinate clause ratio from T1 (M=.40; SD=.13) to T2 (M=.49; SD=.10). In fact, surprisingly, this pre-post difference was the only large change observed in any of the six variables involving syntactic complexity or lexical ability. The finding

with regard to coordinate clause ratio aligns with Collentine (2004), who found that study abroad learners of L2-Spanish increased their use of coordinate clauses more than the control group who studied at home. Noteworthy as well is the fact that the study abroad group made oral fluency gains at the same time that they increased their use of coordinate clauses, while the at home group exhibited little change in either oral fluency or coordinate clause use.

Notable by comparison to the large increase in coordinate clause ratio in the present study is the small increase in dependent clause ratio. The dependent clause ratio for the largest fluency gainers actually increased by only 1.35% from T1 (M=0.43; SD=.11) to T2 (M=.45; SD=.10). While Biber, Gray and Poonpon (2011) argued that subordination is a characteristic of complexity seen in L1 conversational English; the findings of the present study suggest that subordination does not necessarily change much over the course of two semesters in L2 speakers who make large oral fluency gains. This finding is consistent with the cross-sectional findings of Iwashita et al. (2008), who found that dependent clause ratio did not distinguish the five holistic score levels of TOEFL iBT speaking tasks. Taken together, these two syntactic findings may suggest that L2 speakers who make large fluency gains tend to restructure their syntactic knowledge horizontally as opposed to hierarchically. To shed some light on why this might be so, it is necessary to take a closer look at changes in the discourse.

4.5 Discourse Analysis of Exemplar Large Fluency Gainers

Before discussing the discourse analysis of exemplar large fluency gainers, a caveat is necessary. This analysis is exploratory in nature. The purpose of the analysis is to begin to elucidate the nature of the syntactic proceduralization that takes place in the speech production processes of L2 EAP learners who increase their mean length of speech run by a large magnitude. The analysis is not meant to provide any generalizable findings of how psycholinguistic processes become more automatic in the speech production process. On the contrary, the analysis is merely intended to push inquiry in the strand of L2 oral fluency in new directions. More specifically, it is hoped that researchers will begin drawing connections among L2 discourse, syntax, and oral fluency development. More specifically, it is hoped that this discussion can do for L2 syntax and the conceptual preparation phase of the L2 Speech Production Model (De Bot, 1992; Levelt, 1989, 1999) what Towell et al. (1996) did for formulaic language and the lexico-syntactic formulation phase of the original Speech Production

Model (Levelt, 1989). One purpose is to generate some hypotheses and begin a search for answers about how L2 syntax supports and/or reflects a psycholinguistic reorganization of L2 knowledge that results in sustained lengthening of speech runs at the discourse level.

It is necessary to acknowledge the pioneering work of Towell et al. (1996), who first conducted this kind of analysis more than two decades ago on L2 learners of French. Their work inspired much of the present study, and the present study would not have been possible without their work. Moreover, the present study builds on their work by 1) identifying speech run elongation as the largest source of longitudinal L2 oral fluency development, based on analysis of a large pre-post sample; 2) leveraging that large sample size to identify a larger group of large fluency gainers than Towell et al. was able to do with their much smaller sample; 3) describing the lexico-syntactic changes that had taken place longitudinally in this group of large fluency gainers by means of descriptive statistics; and 4) moving the discussion beyond the speech run level to the discourse level.

4.5.1 L2 Syntax and Discourse Models

That said, a closer look at the longitudinal changes in syntax and discourse of the largest fluency gainers reveals some interesting possible connections among increased coordinate clause use, longer speech runs, and more sophisticated discourse organization. The participants who made the largest gains in mean length of speech run not only tended to exhibit large increases in the proportion of their clauses that were coordinate clauses, but they also seemed to use coordinate clauses to build more coherent, sophisticated arguments at T2 than at T1. More specifically, large fluency gainers used coordinate clauses to develop main ideas and transition between main ideas in the process of building more sophisticated discourse models. For instance, large gainers used coordinate clauses as integral parts of arguments based on analysis of pros-cons, problem-solution, counter-argument-refutation, and, most commonly, extended elaboration on the speaker's personal experiences related to the speaking prompt. Hence, discourse structure, coherence, idea development, and syntax seem to be closely related constructs in the oral responses of large fluency gainers.

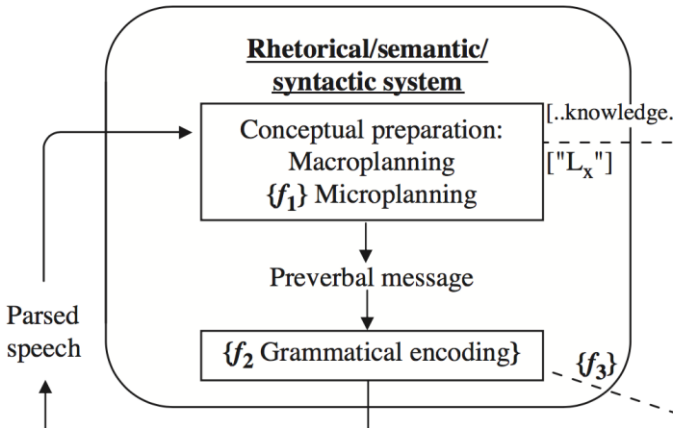

Figure 4.8. Conceptual Preparation Phase of the L2 Speech Production Model

This finding may have important implications for the theory related to cognitive fluency at the abstract level of conceptual preparation (*f1* in the L2 Speech Production Model, pictured above). The scholars most responsible for this influential model (De Bot, 1992; Levelt, 1989, 1999) theorized that syntax plays a rhetorical as well as a semantic role in conceptual planning (See Figure 4.7 above). The findings of the present study suggest that coordinate clauses indeed play an important role in L2 rhetorical discourse. It is possible that as L2 speakers learn L2 discourse models, their procedural knowledge related to conceptual preparation reorganizes to accommodate more sophisticated rhetoric; L2 syntax may develop in support of this process of L2 maturation.

This cognitive reorganization may be similar to that theorized by Pawley and Syder (1983) regarding formulaic language. Pawley and Syder argued that native speakers attain "native-like fluency" (p. 191) by memorizing chunks of language, which they called "lexicalized sentence stems" (p. 191). They argued that the native speaker draws on a large, well-organized mental repository of lexicalized sentence stems to use in a range of communicative contexts.

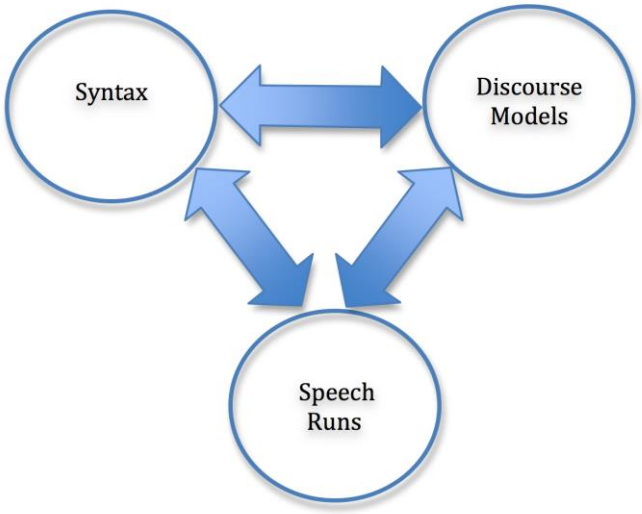

Figure 4.9. Syntax, Discourse Models, and Speech Runs

Applying a similar idea to conceptual preparation, advanced L2 EAP learners may draw on a growing collection of discourse models, choosing the appropriate ones to use for any given speaking situation. For example, similar to the way that a native speaker can quickly choose the lexicalized sentence stem "That's easier said than done, (p. 206)" when describing a difficult task, possibly a highly fluent L2 EAP learner knows that in an argumentative speaking situation, it is appropriate to define the problem, weigh pro's and cons, state the counterargument, and then refute it. To take this analogy a step forward, just as a native speaker knows the lexico-grammatical usage of formulaic language, for example, which phrasal verbs are separable, a highly fluent L2 EAP learner may know the different elements of each discourse model. Furthermore, she may know what kind of syntax to use when transitioning smoothly between elements within a discourse model and between discourse models. Moreover, this expanded discourse competence may apply not only to argumentative speaking, but also other speaking contexts.

It makes sense that L2 oral fluency gains accompany this expansion of the L2 discourse repertoire. Consistent with Segalowitz's (2010) description of parallel processing, as one phase of the L2 speech production process becomes more automatized, processing capacity is reallocated to other phases: lexico-grammatical encoding, articulation, etc. As the L2 speaker develops more sophisticated L2 rhetorical schema, L2 speech production at all phases may

become more efficient. Coordinate clauses may also convey a processing advantage that allows production of more syntactically complex speech runs. Evidence of this advantage is the fact that most of the bi-clausal speech runs in the T2 responses of the exemplar large fluency gainers contained at least one coordinate clause.

4.5.2 Syntax & Temporal Cycles

Also related to the efficiency of L2 speech production, the finding of increased coordinate clause ratio relates to Roberts and Kirsner's (2000) findings with regard to "temporal cycles in speech production" (p. 129). Roberts and Kirsner described a temporal cycle as periods during which "(L1) speakers alternate between phases of low fluency, during which they prepare macroplans, and high fluency, during which macroplans are executed in speech" (p. 129). Interestingly, they found that speakers spoke less fluently when topic shifts occurred, and then fluency increased after each topic shift. Roberts and Kirsner argued that these temporal cycles in speech production represent evidence of competition for processing resources between speech production sub-processes: macro-planning and micro-planning. In other words, they argued that while the speaker is planning what to say in the next stretch of discourse (macro-planning), s/he has less processing capacity available to attend to the planning and execution of specific decisions related to vocabulary, grammar, morphology, articulation, etc., the result being speech breakdown.

The findings of the present study may provide evidence that pre-post restructuring of L2 syntax smoothens out L2 temporal cycles. The present study found that the participants who increased their mean length of speech run the most (the large fluency gainers) also increased their coordinate clause ratio considerably. When comparing the pre-post discourse of the large fluency gainers, I found that, at T2, they were using these additional coordinate clauses to build more sophisticated discourse models and transition more effectively between main ideas. These sophisticated discourse models (pro's and cons, problem-solution, counterargument-refutation, complex personalized message) appeared to facilitate elaboration and idea development in the oral responses, thus increasing mean length of speech run.

Coordinate clauses played a particularly important role in transitioning from topic to topic. This phenomenon was described earlier in the exemplar discourse analysis section. There are at least two possible psycholinguistic explanations for this phenomenon. The first

explanation is that the syntax of the large fluency gainers became more proceduralized, restructuring over time in favor of more coordinate clauses as an organizational tool to support the construction of more sophisticated discourse models. According to explanation one, this restructuring facilitates smooth transition from main idea to main idea and from detail to detail within each main idea. In other words, the increased coordinate clause use is a syntactic development that supports an underlying cognitive restructuring of the macro-planning mechanisms associated with L2 speech production. If this explanation is valid, increased use of coordinate clauses may allow L2 speakers to change topics more efficiently, with less pausing for online micro-planning at points of topic shift.

Another possibility is that increased coordinate clause use is merely a reflection of an increase in the number of ideas that the L2 speaker can devise in the macroplanning phase and/or express successfully due to more automaticity in a wide range of microplanning processes. Stated another way, at T2, the L2 speaker may be able to devise more ideas and/or articulate more of the ideas that she comes up with; therefore, the result is that more ideas make it through the L2 speaker's lexico-grammatical filter as fully formed thoughts in the oral response. In this explanation, the increased prevalence of coordinate clauses may be just one noticeable change that is ancillary to the underlying cognitive restructuring that is taking place at one or multiple phases of speech production. According to explanation two, since more thoughts are available to be expressed and the L2 speaker is able to formulate and articulate more of these thoughts intelligibly and comprehensibly, the result is that there are more thoughts to connect and thus the simplest, most common syntactic means of connecting ideas, coordinate clauses, are more prevalent. In the event that explanation two is valid, large fluency gainers would be expected to exhibit increases in use of other English connectors (e.g., conjunctive adverbials) that are comparable to that seen in coordinate clauses.

4.5.3 A Possible Counter-Argument

A skeptic of the findings of the present study might offer task-related objections. In other words, a skeptic might argue that the 100 participants' oral fluency gains and the ten large fluency gainers' use of more complex syntax and more sophisticated discourse models simply reflects the fact that the "express your opinion" task prompts are confined to a narrow range of topics related to university life choices. The skeptic might say that any student would be able to

answer such a question using more fluent, sophisticated speech after nearly two semesters of college life than after just arriving on campus at the beginning of the first year of college. This skeptic might also suspect that this increase in fluency, complexity, and discourse sophistication only reflects greater knowledge of and experience with topics related to the prompts, not gains in language proficiency. Hence, according to this objection, the task suffers from construct irrelevant variance (Messick, 1996); therefore, the assumption that increases in oral fluency, syntactic complexity, and discourse sophistication reflect proficiency gains is not warranted.

Of course, this argument makes a flawed assumption. The assumption is that the goal of the first year university EAP program that the participants in the present study completed is general language proficiency. It is not, at least not exclusively. The stated mission of the program is as follows (PLaCE, 2019): "Our mission at PLaCE is to provide a strong instructional and assessment program. We'll help you develop the academic, linguistic and cultural competencies needed to participate in university life and to compete for graduate school and employment opportunities" (p. 1). The program assistant director (Allen, personal communication) also emphasized the importance of "meaning-focused input and output" as a cornerstone of the PLaCE curriculum. The findings of the present study align with this goal. The findings suggest that examinees who made the largest fluency gains over the course of the two semester PLaCE Program also improved their ability to discuss university-related issues in a complex, coherent, well-organized, personalized way. This is another way of saying that they demonstrated sophisticated discourse competence and possibly improved critical thinking skills. Hence, the ability to discuss university-related living choices at length is quite relevant to the construct being tested. This brings us to the implications of the study for teaching and learning.

CHAPTER 5. CONCLUSION

5.1 Implications for EAP Teaching and Learning

The findings of the present study have important implications for EAP pedagogy. The debate between proponents of meaning-focused instruction and form-focused instruction has raged for decades, and it will certainly continue; however, the findings of the present study provide some evidence that meaning-focused approaches may confer oral fluency benefits that have not been previously discussed in the literature. More specifically, the findings suggest that students who make large oral fluency gains tend to improve their use of discourse models to frame their oral responses and organize ideas.

In terms of curriculum development, this finding strengthens the argument for meaning-focused instruction. In the event that future research, both longitudinal and cross-sectional, can provide more evidence of a connection between the use of discourse models and L2 oral fluency, then this evidence might further justify instruction that gives students the tools to think through ideas and organize their thoughts. For example, the "describe, analyze, evaluate" (DAE) method of intercultural learning is used in the curriculum of the PLaCE Program EAP course sequence from which the data for the present study were collected. Hence, it is not so surprising that students in this program who make large fluency gains over two semesters are organizing their oral responses in a way that is consistent with the DAE method. Furthermore, in the face of such evidence, it may be more difficult for critics of meaning-focused instruction to argue that meaning-focused input and output represent an inefficient use of EAP class time that could be better spent on form-focused exercises or vocabulary building.

5.2 Theoretical Implications

If a connection can be drawn between meaning focused intercultural pedagogy and oral fluency gains, along the lines just discussed, cognitivists and socioculturalists might find common ground. Proponents of the cognitivist school of applied linguistics believe that L2 development occurs when learners are engaged in the process of forming connections between new knowledge and old knowledge and strengthening those connections through focused language practice. One of the primary aims of cognitivists is to reorganize knowledge in the

learner's system in ways that enable ever more efficient L2 processing. An example of a cognitivist language learning activity is studying formulaic language, which cognitivists believe confers processing efficiency because language users can process and produce these multi-word units as un-analyzed chunks of language, rather than operating on a word-by-word basis.

Socioculturalists, on the other hand, do not value processing efficiency as much as they value the social process of collaborative meaning making. In terms of SLA, sociolculturalists believe that learners develop L2 knowledge by collaborating with other learners to achieve shared goals. An example of a sociocultural activity is a group decision-making activity in which group members must cooperate to reach an optimal outcome. The sociocultural learning process, particularly in the ESL context, requires bridging the divides between individuals with diverse L1's and cultural perspectives. Although linguistic diversity can lead to misunderstanding, it also offers ample pedagogical opportunities. The multicultural reality of the North American University both incentivizes and provides opportunities for intercultural collaboration and learning in the EAP classroom. Discourse models may facilitate reorganization of students' L2 knowledge in service of greater efficiency at the same time that they provide a common frame of reference for intercultural learning. To the extent that this is true, both cognitivist and sociocultural aims can be met.

Discourse models may be effective tools for EAP learning because they have emerged from the social marketplace of ideas as efficient ways to organize ideas. Discourse models like problem-solution, pros and cons, counterargument-refutation, and the personal anecdote have emerged in our collective consciousness precisely because they facilitate communication of complex ideas in ways that all educated adults can understand. The same might be said of the DAE model, Bloom's Taxonomy, and other sequential step learning methods. Hence, EAP educators should include discourse models in the curriculum. Moreover, if these models are too Westernized, then EAP educators should include models from the global East, South, and Middle East as well.

5.3 Implications for Language Testing

The most important implication that the present study has for language testing relates to testing for pre-post gains. Understandably, some universities now require language programs operating under their purview to provide credible evidence of L2 learner gains in language

proficiency. Therefore, it is beneficial to identify objective measures of language proficiency that research has shown to a) distinguish proficiency levels and b) change significantly over the duration of instruction. The pre-post measurement of such variables can assist language program directors in providing credible evidence of language proficiency gains to stakeholders, especially for the purpose of program evaluation.

Of course, it is important to choose objective measures that satisfy both criteria just mentioned. After all, some measures may distinguish proficiency levels without having the potential to measurably change in the language performances of a group of students of a particular proficiency level over the duration of a course or course sequence. Using such a measure to gauge pre-post gains may lead to unrealistic expectations of growth. This being the case, the findings suggest that pause length measures like mean silent pause length and mean filled pause length should not be used to measure pre-post gains for program evaluation purposes, while global measures like mean length of speech run and speech rate would better serve this purpose. The fact that the latter two variables have been shown to distinguish proficiency levels (Ginther et al., 2010; Kormos & Dénes, 2004) and increase over time (Towell et al., 1996) makes them suitable variables for measurement of pre-post gains.

Moreover, the findings of the present study imply that when assessing L2 academic speaking proficiency, it is necessary to integrate oral fluency, syntactic complexity, discourse competence, coherence, and organization. In fact, it may be very difficult to separate these various criteria, to the extent that they tend to co-occur. After all, critical thinking skills are seemingly indistinguishable from the macro-planning process described by Levelt (1989, 1999).

More specifically, the findings of the present study have implications for how speaking should be rated. First and foremost, the findings provide further empirical support for the validity of certain oral fluency characteristics as predictors of oral English proficiency, namely continuation of the flow of speech and frequency of pauses, rather than length of pauses. More specifically, the finding regarding mean length of speech run supports the validity of mean length of speech run as a measure of oral English proficiency. The finding that mean length of speech run was associated with the largest L2 pre-post oral fluency gains over a two semester EAP course sequence provides longitudinal evidence in favor of the argument that this measure could be used in automated assessment of oral English proficiency (Ginther et al., 2010). As was discussed earlier, not only does this variable correlate strongly with the fluency ratings of both

trained L1 and L2 raters (Kormos & Dénes, 2004), but it also correlates strongly with oral English proficiency test scores (Ginther et al., 2010). Moreover, free response speaking scales should include descriptors related to speech run length.

The findings may also suggest a rethinking of the notion of sophistication in regards to rating of L2 academic speech. The findings provide some preliminary evidence that sustained speech run elongation may be related to examinee use of discourse models in response to the speaking prompt. Oral proficiency rating scales have often included descriptors involving sophistication, and more specifically lexico-syntactic sophistication. This notion of sophistication has often been associated with use of complex sentence structure, modality, academic vocabulary, and formulaic sequences. In light of the findings of the present study regarding use of discourse models, coordinate clauses, and speech run length, maybe scale descriptors at the high end of the scale should include language related to sophistication at the discourse level. One example of such a descriptor comes from the Oral English Proficiency Test (Ginther et al., 2010): "the speaker is able to elaborate a complex personalized message using a variety of tenses/aspects and moods" (p. #).

The findings of the present study also may point in a new direction for oral language construct definition. Many speaking test scales mention syntactic complexity or lexico-syntactic sophistication, but the literature has not clearly identified what aspects of oral syntax constitute such complexity or sophistication. Some attempts to identify oral syntactic measures that distinguish high stakes language test holistic scores have failed (Biber, Gray, & Staples, 2014; Iwashita et al., 2008). Hence, such approaches may be either looking at the wrong syntactic measures (neither included coordinate clause ratio), or L2 oral syntactic complexity alone may be too narrow a construct to distinguish high stakes holistic scores. To the extent that L2 oral syntactic sophistication is related to discourse sophistication, coherence, and organization, a more integrated approach to construct definition may be necessary.

One final point about implications for language testing is worth making, and that is related to so-called free response "templates". Whether the reorganization of discourse knowledge discussed earlier can be characterized as improved use of response templates or improved critical thinking skills is a matter of semantics. Of course, the term "response template"

would imply the use of testing strategies in a high stakes testing setting. This term seems less appropriate in the testing setting from which the data were collected for the present study, the participants of which were only given a completion grade that amounted to two percent of their final grade for each semester, and most of the responses were never rated. The pejorative term "response template" may be a misnomer in the case of the large fluency gainers just discussed. After all, if a response technique allows an examinee to speak in a more coherent, sophisticated, fluent manner on a language test, then it is hard to argue that this technique should be discouraged.

Admittedly, in the present study, only ten large fluency gainers were analyzed in terms of lexico-syntactic complexity, so the linguistic findings of the present study should be examined further in future research.

5.4 Future Research and Limitations

Future research should examine further the relationships among L2 syntax, vocabulary, oral fluency, and discourse. The present study only scratched the surface of this potentially fruitful sub-strand. A more systematic analysis of how coordination relates to the development of advanced rhetoric is beyond the scope of the present study, but such research is worthwhile.

The present study had some limitations. First, it did not include a control group, so it is possible that an equivalent group of L2 English learners studying English in their home country or L2 learners at the same university, but without the benefit of language support, would have made the same L2 oral fluency and complexity gains. Furthermore, it may also be true that an equivalent group of participants would have made greater L2 gains if their primary purpose in studying abroad had been to learn the L2, instead of to learn non-language-related (primarily STEM) academic content. Future research examining differences in L2 development between STEM students and students of other disciplines would be worthwhile. Second, the second phase of the lexico-syntactic development phase of the study only included analysis of the large fluency gainers. It is possible that these participants were not representative of the group as a whole. Finally, the present study only collected data at two time points and only over a seven-month time period. Future research should be designed in such a way as to avoid these limitations.

5.5 Conclusions

Returning to the theoretical discussion of L2 oral fluency development, the findings of the present study provide further evidence in support of the L2 Speech Production Model (De Bot, 1992; Levelt, 1989, 1999) . The present study analyzed a relatively large sample size (N=100) of participants in a large, understudied student population (L1-Chinese undergraduates at a large STEM university in the US) and examined the same oral fluency variables as Towell et al (1996) did, in addition to mean filled pause length, silent pauses per minute, and filled pauses per minute over the course of two semesters of EAP language instruction and mainstream university enrollment. Results showed that the participants increased their oral fluency by displaying gains in the following measures (in descending order of effect size): mean length of speech run, phonation time ratio, speech rate, filled pauses per minute, articulation rate, and silent pauses per minute. Consistent with Towell et al. (1996), mean silent pause length and mean filled pause length (which was not included in Towell et al., 1996) changed very little pre-post.

The overall trend in oral fluency development aligns with Segalowitz's (2010) argument that L2 speakers favor continuance of the flow of speech. A few findings from the oral fluency analysis support this conclusion: 1) participants made their largest gains in mean length of speech run; 2) participants made their second largest gains in phonation time ratio; and 3) all of the participants' gains in phonation time ratio were accounted for by pausing less frequently, as opposed to shortening their pauses; in fact, participants decreased both filled pause frequency and silent pause frequency, while their mean silent pause length and mean filled pause length changed little longitudinally.

The findings of the linguistic analysis provided evidence that L2 syntax becomes more complex as L2 speakers lengthen their speech runs. More specifically, the major finding with regard to syntax was that the ten participants who made the largest gains in mean length of speech run made large gains in coordinate clause ratio. After examining how large fluency gainers used coordinate clauses in their oral responses, it became apparent that, at T2, they used coordinate clauses to build more sophisticated arguments, involving transitions and logical development of main ideas. Of course, the relationships among coordinate clause use, rhetorical sophistication, and L2 oral fluency are not entirely clear from the results of the present study. Future research should attempt to replicate with a larger sample size the finding of increased

coordinate clause use over time. More generally, future research should take a discourse analysis approach to the development of L2 oral fluency, syntax, and vocabulary because it is not enough to know that lexical/syntactic variables increase/decrease at the same time that L2 oral fluency develops. Researchers must also scrutinize the function of lexical/syntactic structures in oral discourse at multiple levels: the clausal, utterance, and discourse levels.

Finally, it is surprising that the largest oral fluency gainers did not exhibit large changes in any of the lexical variables studied: lexical frequency profile, formulaic language proportion, or lexical diversity. While one cross-sectional study (Ushigusa, 2008) has shown a relationship between the use of formulaic language and L2 oral fluency, and other studies have shown that L2 lexical variables involving use of single words (Collentine, 2004; Freed, 2004; Kim et al., 2015; Mora & Valls-Ferrer, 2012 ; Segalowitz & Freed, 2004) change over time, the findings of the present study suggest that large changes in L2 vocabulary may not be necessary to make large gains in L2 oral fluency. However, this finding should not be interpreted as evidence against a relationship between L2 vocabulary and L2 oral fluency; it may only mean that one academic year is not a long enough period of time for large changes in L2 vocabulary to occur.

LIST OF REFERENCES

Anderson, J. R. (1983). Cognitive science series. The architecture of cognition. Hillsdale, NJ, US.

Biber, D., Gray, B., & Poonpon, K. (2011). Should we use characteristics of conversation to measure grammatical complexity in L2 writing development?. *TESOL Quarterly*, *45*(1), 5-35.

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.

Brants, T. (2000, May). Inter-annotator Agreement for a German Newspaper Corpus. In *LREC*.

Bybee, J. (2008). Usage-based grammar and second language acquisition. In *Handbook of cognitive linguistics and second language acquisition* (pp. 226-246). Routledge.

Cambridge Dictionary. (2020). Retrieved February 15, 2019, from https://dictionary.cambridge.org/us/.

Cheng, L. (2014). *Effects of pragmatic task features, English proficiency, and learning setting on Chinese ESL/EFL spoken performance of requests* (Doctoral dissertation, Purdue University).

Clark, H. H., & Tree, J. E. F. (2002). Using *uh* and *um* in spontaneous speaking. *Cognition*, *84*(1), 73-111.

Chomsky, N. (1957). Syntactic Structures Mouton. *The Hague*, *19573*.

Cobb, T. Web Vocabprofile [accessed 15 January 2019 from http://www.lextutor.ca/vp/], an adaptation of Heatley, Nation & Coxhead's (2002) Range.

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in second language acquisition*, *26*(2), 227-248.

Collins Online English Dictionary. (2020). Retrieved February 15, 2019, from https://www.collinsdictionary.com/us/dictionary/english.

Davies, M. (2008-) *The Corpus of Contemporary American English (COCA): 600 million words, 1990-present*. Available online at https://www.english-corpora.org/coca/.

De Bot, K. (1992). A bilingual production model: Levelt's' speaking'model adapted. applied linguistics 13 [1], 1-24.

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.

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.

Fazio, R. H. (1990). A practical guide to the use of response latency in social psychological research.

Fillmore, C. J. (1979). 'On fluency'. In D. Kempler & W. S. Y. Wang (Eds.), Individual differences in language ability and language behavior. New York: Academic Press, 85–102.

Firth, J. R. (1957). Ethnographic analysis and language with reference to Malinowski's views. *Man and Culture: an evaluation of the work of Bronislaw Malinowski*, 93-118.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language: A unit for all reasons. *Applied linguistics*, *21*(3), 354-375.

Freed, B. 2000. Is fluency in the eyes (and ears) of the beholder? In H. Riggenbach (ed.) Perspectives on Fluency. University of Michigan Press: Ann Arbor: 243–265

Ginther, A., Dimova, S., & Yang, R. (2010). Conceptual and empirical relationships between temporal measures of fluency and oral English proficiency with implications for automated scoring. *Language Testing*, *27*(3), 379-399.

Goldman-Eisler, F. (1958a). Speech production and the predictability of words in context. *Quarterly Journal of Experimental Psychology*, *10*(2), 96-106.

Goldman-Eisler, F. (1958b). The predictability of words in context and the length of pauses in speech. *Language and Speech*, *1*(3), 226-231.

Goldman-Eisler, F. (1968). Psycholinguistics: Experiments in spontaneous speech.

Halliday, M. A. K. (1975). Learning how to mean. In *Foundations of language development* (pp. 239-265). Academic Press.

Hasselgren, A. (2002). Learner corpora and language testing: Small words as markers of learner fluency. *Computer learner corpora, second language acquisition and foreign language teaching*, 143-174.

Huensch, A., & Tracy–Ventura, N. (2017). L2 utterance fluency development before, during, and after residence abroad: A multidimensional investigation. *The Modern Language Journal*, *101*(2), 275-293.

Heatley, A., Nation, I.S.P. & Coxhead, A. (2002). Range and frequency programs. Available at http://www.victoria.ac.nz/lals/staff/paul-nation.aspx .

Hunt, K. W. (1965). *Grammatical structures written at three grade levels* (No. 3). Champaign, IL: National Council of Teachers of English.

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.

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.

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

Laufer, B., & Nation, P. (1995). Vocabulary size and use: Lexical richness in L2 written production. *Applied linguistics*, *16*(3), 307-322.

Leaper, D. A., & Riazi, M. (2014). The influence of prompt on group oral tests. *Language Testing*, *31*(2), 177-204.

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

Levelt, W. (1989). Speaking : From intention to articulation (ACL-MIT Press series in natural-language processing). Cambridge, Mass.: MIT Press.

Levelt, W. (1999). Producing spoken language. *The neurocognition of language*, 83-122.

Liu, D. (2003). The most frequently used spoken American English idioms: A corpus analysis and its implications. *TESOL Quarterly*, *37*(4), 671-700.

Longman Dictionary of Contemporary American English Online. (2020). Retrieved February 15, 2019, from https://www.ldoceonline.com/.

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International journal of corpus linguistics*, *15*(4), 474-496.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal*, *96*(2), 190-208.

MacIntyre, P. D., & Gardner, R. C. (1994). The subtle effects of language anxiety on cognitive processing in the second language. *Language learning*, *44*(2), 283-305.

Maclay, H., & Osgood, C. E. (1959). Hesitation phenomena in spontaneous English speech. *Word*, *15*(1), 19-44.

Macmillan Dictionary. (2020). Retrieved February 15, 2019, from https://www.macmillandictionary.com/us.

McCarthy, P. M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, *42*(2), 381-392.

Messick, S. (1996). Validity and washback in language testing. *Language testing*, *13*(3), 241-256.

Möhle, D., & Raupach, M. (1987). The representation problem in interlanguage theory. *Perpectives on Language in Performance. Tübingen: Gunter Narr*, 1158-1173.

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.

Munro, M. J., & Derwing, T. M. (1995). Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language learning*, *45*(1), 73-97.

Nation, P. (2007). The four strands. *International Journal of Innovation in Language Learning and Teaching*, *1*(1), 2-13.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press.

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.

Oller Jr, J. W. (1974). Expectancy for successive elements: Key ingredient to language use. *Foreign Language Annals*, *7*(4), 443-452.

Oxford Collocations Dictionary. (2020). Retrieved February 15, 2019, from https://www.oxfordlearnersdictionaries.com/us/.

Park, S. (2016). Measuring fluency: Temporal variables and pausing patterns in L2 English speech.

Pawley, A., Syder, F. H., Richards, J. C., & Schmidt, R. W. (1983). Language and communication. *London: Longman*, 191-195.

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

Roberts, B., & Kirsner, K. (2000). Temporal cycles in speech production. *Language and Cognitive Processes*, *15*(2), 129-157.

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

Rossiter, M. J., Derwing, T. M., Manimtim, L. G., & Thomson, R. I. (2010). Oral fluency: The neglected component in the communicative language classroom. *Canadian Modern Language Review*, *66*(4), 583-606.

Schmitt, N. (Ed.). (2004). *Formulaic sequences: Acquisition, processing and use* (Vol. 9). John Benjamins Publishing.

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*, *26*(2), 173-199.

Segalowitz, N. (2010). *Cognitive bases of second language fluency*. Routledge.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied linguistics*, *31*(4), 487-512.

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

Towell, R., Hawkins, R., & Bazergui, N. (1996). The development of fluency in advanced learners of French. *Applied linguistics*, *17*(1), 84-119.

Ushigusa, S. (2008). *The relationships between oral fluency, multiword units, and proficiency scores* (Doctoral dissertation, Purdue University).

van Gelderen, A. (1994). Prediction of global ratings of fluency and delivery in narrative discourse by linguistic and phonetic measures-oral performances of students aged 11-12 years. *Language Testing*, *11*(3), 291-319.

Wray, A. (2002). Formulaic language and the lexicon. Cambridge: Cambridge University Press.

Xue, G., & Nation, I. S. P. (1984). 1984: A university word list. Language Learning and Communication 3, 215-229.

You, Y. (2014). Relationships between lexical proficiency and L2 oral proficiency.

VITA

David Crouch received his Bachelor of Arts, majoring in Spanish and Business Administration from Rhodes College in Memphis, Tennessee and a Master's of Arts in TESOL from Murray State University in Murray, Kentucky. He pursued a doctoral degree in English, with a concentration in Second Language Studies at Purdue University in West Lafayette, Indiana.

David conducted research on a few topics during his time studying at Purdue, concentrating mostly on L2 oral fluency. With regards to L2 oral fluency and language testing, he presented his studies, as the sole author at the American Association of Applied Linguistics (AAAL) Conference, the Midwest Association of Language Testing (MwALT) Conference, and the Purdue Linguistics, Literature, and Second Language Studies Conference. He also co-presented in a technology demonstration of the *Fluencing* oral fluency analysis tool at the Language Assessment Research Conference.

David was born and raised in Murray, Kentucky, where his love of ESL teaching began at the Murray State University ESL Program. Next, he taught English as a foreign language as a Foreign Expert at Qingdao Agricultural University in Qingdao, China He worked as a Teaching Assistant in the English Department at Purdue, teaching Introduction to Composition and Introduction to Composition for International Students. Throughout most of his doctoral program at Purdue, he also tutored international teaching assistants in the Oral English Proficiency Program (OEPP), earning three Excellence in Teaching Awards. Towards the end of his doctoral studies, he became interested in teaching K-12 English Language Learners. In his spare time, he tutored local (West Lafayette) K-12 ELL's in English writing. During the COVID-19 pandemic of 2020, he supported the English language development of local ELL's with interactive online instruction.

He conducted language test administration and language test development at the Purdue Language and Cultural Exchange (PLaCE). At PLaCE, he wrote test items for different tasks on the Assessment of College English International (ACE-In) Exam, which is the PLaCE Program's in-house post-admission English language proficiency exam for first year undergraduate international students enrolled in English 110 and 111 classes. His most extensive test

development work involved upgrading the test specifications of the ACE-In cloze-elide and elicited imitation tasks and writing test items in accordance with the new test specifications.