The Longitudinal Development of Oral Linguistic Complexity and Accuracy of Spanish Learner Language in Second-Year University Students

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Abby Louise Bajuniemi

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Carol A. Klee, Adviser

September 2015

Acknowledgments

First and foremost, I have to thank my family, especially my mom and Eric, for their unwavering support. Writing a dissertation is not only difficult and trying for the dissertator, but also for her support system, and I will forever be grateful to you for helping me through some truly difficult times.

I'd also like to thank my friends and colleagues, both within and outside of my department—especially those of you involved in the 50,000+ Facebook message thread. I don't know how I could have gotten through everything without you. I'll miss the camaraderie in class and in the office, laughing ridiculously long and loudly at the wonders of the Internet. Sitting with you at conferences. Studying and throwing out ideas together. Commiserating together. Supporting one another. I feel so lucky to have been a part of such a great group of linguists and nonlinguists.

To my participants: It was a pleasure to get to know each of you. Without you, this project would not exist. I am grateful for your willingness to be a part of my project. I also am grateful to the instructors who allowed me to intrude into their classrooms for an entire semester. I learned so much, and my own teaching has benefitted from your expertise!

I would be remiss if I did not (profusely) thank Dr. Andrea Révész for her guidance and support in reimagining this project. Without her, I would not have gotten to the finish line. Thank you, Andrea, for your kindness and generosity. Your contributions were critical to getting this project to the end. You are truly a gem in academia, and I wish you a long, successful career.

Finally, I would like to thank my interim home at Macalester College. My colleagues, friends, and students were an especially bright light during the most trying of times. Wherever the future takes me, my time at Macalester will be a time I will truly treasure.

I can’t forget the four-leggeds in my life, either. Your insistence upon being in my lap, your need for exercise, and your silly antics were a major comfort and reason to smile.

Dedication

To grandma Betty and Lee. I wish both of you could have seen the culmination of years of hard work. I miss you both.

Abstract

To date, few studies have investigated the production of linguistic complexity and accuracy with naturalistic data. Very often, the data are obtained in a laboratory or laboratory-style settings, with tasks defined by the researcher rather than the instructor (Hatch, 1978; Seedhouse, 2004). Additionally, replicability of studies that investigate the production of complexity, accuracy, and/or fluency (CAF) has been made difficult by the myriad ways that researchers have operationalized the constructs in their research (c.f. Ellis & Barkhuizen, 2005; Housen & Kuiken, 2009; Housen, Kuiken, & Vedder, 2012; Norris & Ortega, 2009; Pallotti, 2009). Further, few studies use a Dynamic Systems Theory lens when researching CAF and the development of CAF. The current longitudinal case study investigates how task affects the production of linguistic complexity and accuracy by three intermediate students of Spanish using data collected in the participants' regular classrooms over one academic year. A DST framework is used to reflect upon each student's developmental trajectory over the course of the study.

The data were transcribed and separated into AS-units, and then further coded using two global syntactic complexity measures and one specific syntactic complexity measure, and one global accuracy measure as well as one specific accuracy measure. In order to determine task effects, ANOVAs were performed on each student's data. A multi-level mixed effects model was used to determine whether there were any interactions between time and task type. Results of the ANOVAs showed that task affects each student's production in a slightly different way, while the multi-level mixed effects

modeling showed that verbal accuracy alone showed an interaction between time and task type.

Results of the longitudinal analysis of the oral production of linguistic complexity and accuracy using a DST lens showed that the students' production did vary over time and that each student followed her or his own trajectory over the course of an academic year. These results also showed that there were some trade-off effects with the measures of linguistic complexity and accuracy, in that when complexity measures increased, there was a tendency for accuracy measures to decrease.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Statement of the Problem

Many studies of learner language in the field of Second Language Acquisition (SLA) attempt to analyze learner language in a way that is predominantly etic and prescribed (Hatch, 1978, 1992; Markee, 2000; Seedhouse, 2004). That is, the researcher goes into the study knowing what she is going to be testing or looking for by prescribing specific tasks and contexts for the participants and then imposing categories of her own making onto the data and the analysis of the language subsequently produced during these tasks. This is true of much of the research on Spanish within the field of SLA as well. While these studies certainly have merit and a place in the literature to describe and theorize how languages are learned, what is currently needed is analysis of data that are collected in its naturally occurring setting, the classroom. This project accomplishes this by examining task-based teaching and learning within the field of SLA via a Dynamic Systems Theory theoretical lens.

Task-based Learning and Teaching

Task-based learning and teaching (TBLT) has been widely studied in the field of SLA (Ellis, 2003). According to Ellis, the definition of what constitutes a "task" in the classroom has been inconsistent (p. 2). The current study will use Ellis's definition, described in detail in Chapter 2 of this dissertation, which basically states that a task is some sort of meaningful activity undertaken in a classroom that results in some specific linguistic and nonlinguistic outcome (Ellis, 2003). That said, there is a very large body of

research within TBLT, with a subset of this research focusing on how learners produce linguistic complexity, accuracy, and fluency (CAF). Much of the literature that deals with CAF has focused on EFL (English as a Foreign Language) or ESL (English as a Second Language) learners, and, as previously mentioned, takes place in experimental conditions, either in that there is a "treatment" condition in a classroom environment or the participants go to an actual laboratory setting to perform the tasks created by the investigators. The foci of these studies have been varied, from lexical use, use of grammatical structures, or general performance on some kind of assessment, be it written or oral (cf. Housen et al., 2012, for an extensive, but not complete, examination of the ways in which linguistic accuracy and complexity have been operationalized in the literature). The important thing to note is that in many of these studies, the researcher designs and manipulates these tasks to attempt to determine how cognitive complexity will affect output (Robinson, 2001a, 2001b, 2003, 2005, 2007; Robinson & Gilabert, 2007; Skehan, 1996, 1998, 2003; Skehan & Foster, 1997, 2012). However, these treatment conditions do not exactly replicate how students interact in regular, day-to-day classroom interactions. Because the researcher is often not the instructor of the class, any task that is researcher designed may not be a task that students are accustomed to or would do in their daily routines. Additionally, many studies often utilize a one-time, cross-sectional data collection approach; occasionally data are collected over the course of several weeks, possibly a whole semester, but often the same tasks are performed at each data collection. As Seedhouse (2004) notes, many teaching approaches are based on "task-as-workplan" pedagogical theory, but what happens in the classroom may diverge

greatly from what the instructor had intended, and look very different as "task-in-progress" (p. 264). Further, these treatments and carefully constructed conditions run the risk of reductivism in trying to explicate the results obtained with the resultant data (Larsen-Freeman & Cameron, 2008a, 2008b).

L2 Development

There has not been, at the time of writing, much said about the longitudinal development of linguistic accuracy and complexity in adult Spanish L2 learners based on analysis of naturally occurring classroom data. Moreover, an important tenet of studying language acquisition with a Dynamic Systems Theory (DST) lens is that these laboratory-style investigations separate the system from the context (Larsen-Freeman & Cameron, 2008a, p. 39). The fundamental thought behind DST is that the system, in this case the language repertoire of the learner, is connected to the context of learning. DST demonstrates how factors in a system will vary over time. DST is interested in describing the variation within the individual as the system (the language) organizes and reorganizes itself as it develops.

Experimentally-based investigations in Second Language Acquisition (SLA) present an "idealized" version of the representation of the learner's productive capacities (Larsen-Freeman & Cameron, 2008a, 2008b). As Larsen-Freeman and Cameron state, "Idealization of complex systems has often involved the removing of 'noise' from data: for example, removing individual variation by averaging across samples" (p. 40). They go on to argue that this individual variation is precisely the data needed to observe how learning happens in real-world activities, and that idealizing away the "noise" of context

may be causing the research to lose essential pieces that may produce more accurate and more useful descriptions and explanations of language use and acquisition (p. 40). The present study fills that gap with the first step of observing how such students' linguistic accuracy and complexity develop over the course of an academic year, and how (or if) accuracy and complexity vary based on task type and over time. This line of inquiry also adds to the growing body of knowledge that extends into non-English L2 language learning to better understand the general learning processes that occur with L2 learners, by describing the development of linguistic complexity and accuracy in Spanish as an L2 over time using a DST lens. This data set, rather than artificially "testing" what the students can do or know how to do, shows what they actually do, using what has been presented to them in language instruction. This data set also shows the individual variation of the language produced by each participant and the developmental trajectory, using a Dynamic Systems Theory framework, of each as they navigated the tasks they were asked to do in their intermediate Spanish classrooms.

Significance of Study

The findings of this study contribute to our understanding of how task affects linguistic output of learners of intermediate Spanish, as well as how linguistic complexity and accuracy interact over time in the developmental trajectory of the 3 participants. Dynamic Systems Theory (DST; Polat & Kim, 2014; Larsen-Freeman, 2006, 2009, 2011, 2014; Larsen-Freeman & Cameron, 2008a, 2008b; Spoelman & Verspoor, 2010; Verspoor, Lowie, & van Dijk, 2008) as applied to L2 output is still understudied with respect to both Spanish language acquisition and use as well as the study of the

development of linguistic complexity and accuracy. Even though this is a case study, and thus not generalizable to the general population of learners of intermediate Spanish, it is a first step in investigating how students perform in teacher-designed tasks as well as how factors of linguistic complexity and accuracy may or may not interact in the development of the L2. Finally, Norris and Ortega (2009) have made calls for a more "organic" approach to the coding and analysis of the production of the L2 in order to better study linguistic complexity and accuracy. This study responds to this call for more detailed analyses by including both granular and global measures of linguistic complexity and accuracy. In this way, the present study fills the gaps in the literature on how students perform in teacher-designed tasks with respect to both granular and global measures of linguistic complexity and accuracy, and is a first step in the analysis of L2 Spanish data with a DST lens.

Research Questions

Because the results in the previous research on how task affects linguistic output are varied, this study will, instead of imposing task categories and operationalizing task complexity before the students perform them, investigate how students are producing language in tasks that were not manipulated by the researcher. This study will also investigate if and how the dependent variables outlined in Chapter 3 interact and vary within the individual as a function of time. Thus, the research questions that guided this study are the following:

1. Does task type affect oral linguistic complexity and accuracy of learner language?

a. Does time interact with task type to affect oral linguistic complexity and accuracy of learner language?

2. Does oral linguistic complexity and accuracy exhibit change in trajectory over time?

Overview

The present study is a longitudinal case study of classroom data from two second-year Spanish classes, one in a large, urban research university and the other at a smaller, liberal arts college, both in the Midwest, to determine what, if any, differences occur in the linguistic complexity and accuracy produced by learners in the task types used in those classes and over time. The rest of the dissertation is organized as follows: Chapter 2 reviews and discusses the relevant literature on the acquisition and production of linguistic complexity and accuracy by task as well as within a Dynamic Systems Theory framework for the longitudinal analysis of variation. A discussion of the relevant literature that describes adults' acquisition of Spanish L2 verbal structures also appears in Chapter 2. Chapter 3 outlines the participant profiles, the institutional contexts, and describes the operationalization of the tasks found in the data as well as provides a justification for the data analysis measures used in the determination of linguistic complexity and accuracy. Chapter 3 also outlines the method used for coding and subsequent analysis of the data. Results are presented in Chapter 4, and the discussion and conclusion are presented in Chapter 5.

Chapter 2

Literature Review

Introduction

This dissertation focuses on how adult L2 learners of Spanish develop oral linguistic accuracy and complexity over the course of an academic year in two contexts: one at a university that employs a hybrid model of language instruction based on a communicative approach to language teaching and one that uses an approach that falls within a content-based approach to language teaching. While most previous studies of oral linguistic accuracy and complexity have examined the relationship of these variables in researcher-designed tasks, there has been little research on the development of these two variables during normal classroom activities over the course of 1 academic year.

In order to get a clear picture of the previous ways in which learner language has been theorized, it is important to outline the various theoretical underpinnings that have guided previous research on oral L2 production and how it changes over time and by context. It is also necessary to outline what previous research has found on the development of L2 Spanish in order to be able to situate the results of this study. Therefore, the first section will first briefly discuss variation in the L2 and the rationale for the particular theoretical framework, Dynamic Systems Theory that will be used in this study. The subsequent sections will discuss task-based instruction and how tasks have been used in the SLA literature to document how complexity and accuracy in learner language may be accounted for and described, where complexity is defined as "the stage and elaboration of the underlying interlanguage system" (Skehan, 1996, p. 46)

and accuracy as the "learner's capacity to handle whatever level of interlanguage complexity s/he has currently attained" (Skehan, 1996, p. 46). Then, verbal and syntactic development in L2 Spanish will be discussed in the context of the intermediate language learner and what forms are normally expected and used by learners at this stage in order to understand the kind of data collected in the current study, concluding with the research questions that frame the analysis of the data presented in this dissertation.

Variation in SLA

According to Polat and Kim (2014), researchers are beginning to accept the idea that variation has an internal, cognitively-based component in which the various constructs within L2 performance interact and cause variation over time (p. 186). This variation shows that language acquisition is not a linear endeavor, and that there is a dynamic system at work as learners develop their L2 systems (Larsen-Freeman, 2009; van Geert, 2008). Additionally, Dynamic Systems Theory (DST) rejects that there is any kind of "end state" to acquisition (Larsen-Freeman & Cameron, 2008a, 2008b). Larsen-Freeman and Cameron argue that language systems are complex systems that are constantly in flux, with no determinable end point to the development. In this way, DST expects variation, as does variationist SLA, but instead of comparing the output with a target norm, DST scholars seek to describe the variation and the individual trajectory of language use and development. DST has thus been shown to be an appropriate framework for use in the analysis of the variation exhibited by L2 learners.

One of the first SLA scholars to introduce the idea of using DST with learner language was van Geert, who said, "in essence, cognition, thinking, and action are

explained as dynamic patterns unfolding from the continuous, 'here-&-now' interaction between the person and the immediate environment" (p. 184). Larsen-Freeman (2006, 2009) was the first to argue further that a DST approach to the analysis of learner language is especially needed in order to tease apart the interactions that happen among the variables measured in the analyses as well as with the environment. DST allows the researcher to investigate how the variables are interacting over time, as the way in which they interact will change over time, again reinforcing the fact that language learning is not a linear (or terminal) process. DST illuminates the complexity and dynamicity of the learning process (van Geert, 2008).

DST and Complex Systems

Larsen-Freeman and Cameron (2008a, 2008b) make a case for how and why SLA scholars can and should apply a complex systems approach to the analysis of L2 data, first by discussing what it means to have or participate in a complex, dynamic system, and then explaining how these rather abstract, scientific concepts may be applied to L2 contexts, as the concept of a complex, dynamic system originated mostly in biological and natural sciences (p. 1).

According to the authors, language systems are dynamic systems because of one specific feature: *change* (p. 25). This is not to say that there is no stability in a dynamic system, but the system as a whole is in a constant state of flux, and any part of the system may change from stable to variable at any time, given the correct conditions for change. This view is a reflective view of production: this approach is not meant to predict behavior, but rather to explain it. According to DST and a theory of complex systems, we

cannot predict, we can only reflect and, as Larsen-Freeman and Cameron state, "postdict" what occurred in any given context.

These changes, however, are motivated by pressures that cause a phase shift, or a sudden change to a radically different mode. Larsen-Freeman and Cameron describe the phase shift with the example of a horse that shifts from a trot to a gallop. In going from a trot to a gallop, the horse changes his whole manner of moving his legs as well as his speed (p. 45). However, in order to have a phase shift, one must have a state space. This state space is kept by three different kinds of attractors, i.e. fixed point, cyclic, and chaotic. Attractors are behaviors or states that the system prefers (p. 49). Phase shifts can occur because of some force exerted on an attractor state, or they can come about spontaneously because the system itself, being dynamic, has shifted into a new attractor state on its own (p. 58).

This self-organization can sometimes lead to what is called *emergence*, which is a state that is at a higher level of organization than the previous. These emergent states "are new stabilities of behavior (sometimes emerging from previous disorder), which are open to further change and which have different degrees of variability or flexibility around them" (p. 59). Again, relating to SLA, we can imagine the learner who has been introduced to the Spanish subjunctive, but up until a certain point, has not been able to use the form reliably in her spontaneous production. *Emergence* would occur when her verb system has reorganized to allow her to use this mood reliably in spontaneous production. This new state now has the subjunctive as a possibility for mood expression in verbal structures.

These concepts are the basis for the analysis of L2 data within a DST lens. Larsen-Freeman and Cameron state that any one person's language production is the way that it is in any given moment because of what has happened before that moment (p. 80). It is important to note that the initial state of a system, any complex system, is impossible to know with complete certainty, as is the exact cause of any given effect (pp. 230–231). Care needs to be taken when interpreting results via a DST lens, as there could be many reasons for the data at hand. What is of interest is the interaction between the items being measured (p. 231).

One of the questions among SLA scholars centers on repetition effects and whether practice effects affect performance results. According to DST, as explained by Larsen-Freeman (2009) and Larsen-Freeman and Cameron (2008a, 2008b), every time a learner begins a task, she starts from a different starting point than the last time she attempted the same or a similar activity (Larsen-Freeman, 2009, p. 584). Larsen-Freeman questions whether practice effects are even of concern because of this change in starting point, being careful to note that that does not mean that repetition is not desirable or needed. She states:

> In fact, from a complexity theory standpoint, using a task more than once is what drives learning. When it comes to language learning, revisiting the same, or similar, territory again and again is essential. All I'm saying is that each time the task is used, the learners' experience of it will be different, in part because learners will orient to it differently. Besides, learning is not the taking in of different linguistic forms in an aggregative manner; it is changing the system. This happens best when learners are engaged in enacting the meaning potential of the language, as they do with each iteration of a task. (Larsen-Freeman, 2009, p. 584)

Following this logic, then, DST will show how the system is reorganizing over time, showing the advances and regressions, and thus the trend lines of increase or

decrease of the different variables being examined. Each time a student undertakes an activity, she is starting from a point of changed experience/changed system from the previous attempt at engaging in the same or similar activities. Further, using various measures of complexity, accuracy, and fluency (CAF)[1] within a DST framework can show interactions within the system that could not necessarily be captured if there were only one measure and can reveal the kinds of interactions present in the data with global and granular measures of CAF. DST allows visualization of the complex nature of acquisition that shows the "coupled system" (Larsen-Freeman, 2009, p. 585) that is the task and the learner.

Larsen-Freeman (2006), in her examination of learner language through the DST lens, found that, in general, her learners exhibited steady improvement in CAF, but DST was able to show what group averages did not: that there were advances and regressions in the individual that the group averages were not able to capture. She was also able to describe the interactions between the measures that group averages were not able to capture.[2] Each learner traversed her own path of development at different rates. Similarly, Verspoor, Lowie, and van Dijk (2008), Spoelman and Verspoor (2010), and Polat and Kim (2014) found that the trajectories of learners' development varied over time, and the CAF measures also varied in the ways in which they interacted.

---

[1] Fluency is often grouped together with complexity and accuracy, forming the well-known acronym CAF, but fluency is not one of the constructs under analysis in this dissertation. See, for example, Skehan (1998) and Ellis and Barkhuizen (2005) for more on the construct of CAF.

[2] Please see Hakuta 1974 and 1976 for previous research on trajectories not using a DST framework.

DST, therefore, seems an especially apt theoretical framework to apply to the L2 production of the participants in this study because it is a case study that will allow for the type of fine-grained longitudinal study called for by researchers who have advocated for and/or used DST with L2 data (cf. de Bot, 2008; Larsen-Freeman, 2006, 2009, 2012; Larson-Freeman & Cameron, 2008a, 2008b; Polat & Kim, 2014; van Geert, 2008) in order to discover the individual development trajectories of learners as well as the relationship(s) and interaction(s) of the linguistic complexity and accuracy constructs in this development. According to Ortega and Byrnes (2008), this framework is especially suited to case studies in the elucidation of L2 development of CAF, enabling the researcher to provide evidence for Skehan's (1996) assertion that

> language learning is not any sort of simple, linear, cumulative process. Instead, learners must be able to develop their interlanguage systems in more complex ways, through cycles of analysis and synthesis revisiting some areas as they are seen to require complexification, learning others in a simple, straightforward manner, developing others by simply relexicalizing that which is available syntactically, but which need not be used on such a basis. (p. 58)

Task-based Language Instruction

Tasks in the foreign language classroom began to gain attention in the 1970s and have been the subject of continued research since then (e.g., Ellis, 2003; Seyyedi, 2012; Skehan, 1996, 2003; Skehan & Foster, 1997). The fact that tasks and task-based instruction have garnered so much attention in the literature has resulted in varying interpretations and definitions of what a task is. For the purposes of this dissertation, the definition given by Ellis (2003) is the one that will be considered here. According to Ellis, a task must have six properties:

1. a work plan/planned activity,

2. attention to meaning,
3. real-world processes of language use,
4. requires use of any of the four modalities: reading, writing, speaking, or listening,
5. requires students to engage in "selecting, classifying, ordering, reasoning, and evaluating information," and
6. has a required outcome or way of assessing task completion. (pp. 9–10)

This is in agreement with Skehan's (1996) definition of what constitutes a task. In this study, Ellis's (2003) definition will be used to determine which of the classroom activities count as a task and can be included in the analysis. See Chapter 3 for a detailed description of further categorization of task type.

Tasks in the L2 classroom seem to be most associated with the communicative approach to language teaching, which appears to be complementary to the definition of a task as an "endeavor that requires learners to . . . manipulate and/or produce the target language." The point of communicative language teaching is to encourage learners to develop L2 skills via communication, or interaction, with other learners or native speakers (NSs) of the L2. This idea that interaction is necessary in the acquisition of an L2 has led to a very large body of research on interaction between NSs and nonnative speakers (NSSs) as well as in NNS–NNS dyads. The seminal work on interaction, which led to the Interaction Hypothesis, was done by Long (1983, 1985, 1996), using NS–NNS dyads as his prototype. The Interaction Hypothesis states that for acquisition to occur, learners need to engage in interaction and in the subsequent negotiation of meaning that

occurs when there is some kind of misunderstanding between the conversation partners. This negotiation makes the input comprehensible, which will then facilitate the acquisition of that input and allow the learners to use it. Varonis and Gass (1985) take Long's hypothesis and describe a model for the negotiation that occurs between NNS–NNS dyads, stating that there is more negotiation and subsequent comprehensible input in this type of dyad due to the participants having the same "shared incompetence" (p. 85), making learners more willing to take the risk of being misunderstood by their interlocutors. This research, in turn, led to Swain (1995) arguing for the role of output in L2 acquisition, another factor that has received much attention. Her Output Hypothesis states that interaction, and the desire to be understood, pushes learners to attend to their output and take risks in order to accurately express what they want to say. These theories tend to be utilized as underpinnings for much of the research on task-based learning and teaching, and while the areas related to these two hypotheses are important, they are also very broad, and not within the scope of the current work. However, it is important that they be briefly mentioned and explained in order to understand the background for the current research on CAF. What is important to note about this line of research is that context and interlocutor (other than NS or NNS status) are often not taken into consideration as possible variables to explain the language data obtained. In other words, according to Long, language learning is decontextualized, and the same negotiation of meaning and subsequent learning should happen no matter the context of the interaction. I disagree with this conceptualization of learner interaction; there is more that affects learner acquisition and output than just whether the interlocutor is a NS or NNS or how

much negotiation occurs in an interaction. This view of decontextualized learning seems to hold true in some of the task-based research on CAF; when individual differences are considered as variables, they tend to be categorized in terms of L2 proficiency or psychological characteristics such as motivation or affect. Context is often considered only in the difficulty or cognitive complexity of the task being performed by the participants.

Theoretical Models that Account for Linguistic Complexity and Accuracy and the Relation to Cognitive Complexity

Within the body of research on the effects of task-based teaching and learning, there has been a focus on the effects of task on the linguistic output of learners. Specifically, a cognitive approach to the analysis of this linguistic output has been frequently used to determine how the cognitive load of the task being performed has affected the linguistic complexity, accuracy, and fluency of learners' performance. Even though the current work is not investigating the effects of cognitive load, the theories underlying its effect on oral L2 performance will be described in order to provide a more complete background on the previous CAF research. This cognitive load, also called *task complexity*, has been theorized in two ways, as the Limited Attentional Capacity Model or the Trade Off Hypothesis (Skehan, 1996, 1998, 2003; Skehan & Foster, 1997) and as the Cognition Hypothesis (Robinson, 2001a, 2001b, 2003, 2005, 2007). Both of these theoretical models are based on the model for information processing put forth by Levelt (1989), which states that there are three conceptually hierarchical processing components. The first component is what he called the Conceptualizer, where the speaker begins the

processing sequence by establishing a communicative goal. The goal then goes through a macro- and microplanning process whereby the goal is broken down into subgoals with the related information retrieved and then these subgoals and information chunks are processed and assigned linguistic representations of the information, respectively. The next step involves the information being put through the Formulator, which maps the phonologic, lexical, and grammatical features onto information chunks. From there, this information goes into the Articulator, where actual speech is realized. According to Levelt, this sequence does not occur in a linear, stepwise fashion, but is a series of processes that occur in a parallel fashion.

However, it is important to note that Levelt's model is meant to describe the processes involved in L1 speech, not L2 speech. This caveat does not mean that the same sorts of processes and/or constraints would not hold true for L2 speech. Kormos (2006) and Skehan (2009) attempted to design an approach to describing how Levelt's model may apply to L2 speech. In both L1 and L2 speech, the Conceptualizer is taxed in the formation of utterances while the Formulator is mostly automatic in L1 speech but appears to be susceptible to influence in an L2 because the L2 linguistic system is not as well developed and automatized as an L1 system (de Jong, Steinel, Florijn, Schoonen, & Hulstijn, 2012, p. 136; Skehan, 2009). According to Skehan's approach, during L2 speech, the Conceptualizer is pressured by planning, more complex cognitive operations, abstract and/or dynamic information, or simply a greater quantity of information (Skehan, 2009, p. 525). This pressure placed on the Conceptualizer will lead to more complex speech production because the Conceptualizer has to work to manage more complex

ideas and formulate a message appropriate to the context. On the other hand, the Formulator is pressured by the need for less frequently occurring lexical items, the nonnegotiability of a task, time pressure, heavy input presence, and monologic tasks (p. 525). Because the Formulator takes the message from Conceptualizer and translates it into the linguistic output, a more complex message from the Conceptualizer would push the Formulator to access more complex syntax and lexical items. So, it can be said that the pressure on the Conceptualizer indirectly creates pressure on the Formulator, as the L2 learner would not have as well developed a language system to draw upon in order to create the linguistic form of the message created in the Conceptualizer. This is what is referred to by the nonnegotiability of the task: the more complex the message from the Conceptualizer, the less room the Formulator has to choose a less difficult way to encode the message. Time pressure and a monologic task pressure the Formulator similarly by limiting the time to complete the task and restricting the ability to rely on an interlocutor; the speaker has less time and assistance to create complex language out of the message received from the Conceptualizer. These processes take up more attentional resources since it is not as automatized in an L2 learner as it would be for a NS, thus resulting in an inability for the L2 learner to attend to all the features of complexity, accuracy, and fluency at once.

Skehan also outlines the factors that may ease pressure on the Conceptualizer and Formulator (Skehan, 2009). On the level of the Conceptualizer, tasks that contain concrete and static information, that contain less information, and that require less complex cognitive operations will all ease the pressure on the Conceptualizer in the

formation of the message. The Formulator is eased by including planning time to organize ideas and rehearse, structured tasks, and a dialogic condition. These conditions allow the speaker support in the production of the L2, in that the messages the Conceptualizer creates are simple and do not require excessive attentional resources, and once sent to the Formulator, support the speaker once again with an interlocutor who could offer scaffolding and/or priming and time to attend to the linguistic output. This reduction on pressure at these two stages releases attentional resources to be able to attend to more of the features of linguistic CAF. This is the basis for Skehan's Trade-Off Hypothesis (Skehan, 1996, 1998, 2003; Skehan & Foster, 1997), which will be described next.

This trade-off in attentional resources is the basic assumption of the Limited Attentional Capacity Model (LACM) put forth by Skehan and Foster (1997), i.e., that resources are limited, and when learners are confronted with a difficult task, they will not be able to attend to all the features of CAF. There will be a type of trade-off where learners will only be able to attend to one or two of those elements instead of all three. In fact, the model states that when cognitive complexity is increased, then learners will prioritize meaning over CAF, because they will be placing all their attentional resources on the meaning of the task instead of the language they are producing (the form) and they will not be able to produce language in as fluid a manner. However, more recently Skehan and Foster (2012) have modified this model somewhat to state that increased task complexity will not always cause one aspect to receive more attentional resources at the expense of others; it is just a tendency. This new, modified model, which they call the

Extended Trade-off Hypothesis, differs from Robinson's (2001a, 2001b, 2003, 2005, 2007) model, to be described next, in that it attempts to account for the fact that some studies have shown that accuracy and complexity seem to be tied together and affected similarly by task effects.

The other model, the Cognition Hypothesis, is a more fine-grained model that uses what Robinson calls the Triadic Componential Framework to explicate the effects of task conditions ("interactional factors"), task complexity ("cognitive factors"), and task difficulty. In Robinson and Gilabert (2007), the difference between task difficulty and task complexity is described, with task difficulty defined as "the learners' perceptions of task demands" (p. 163); this "will contribute to between learner variation in performing any one (simple or more complex) task, in the same way differences in aptitude for Math will distinguish the speed and success of those solving calculus or geometry problem" (p. 163). Task complexity "contributes to intralearner variation in performing any two tasks, such as doing simple addition versus calculus" (p. 163). So, here we can see that there are three dimensions to this framework: cognitive factors (task complexity), interactional factors (task conditions), and learner factors (task difficulty), which all affect how CAF are realized in the oral discourse of L2 learners.

Task complexity, then, is aimed at describing the processing factors that use resources in the production of the L2, which Robinson (2005, p. 5) divides into two types: resource-directing (e.g., +/- few elements, +/- here-and-now, +/- no reasoning demands) or resource-dispersing (e.g., +/- planning, +/- single task, +/- prior knowledge). An increase in resource-directing variables will cause an increase in task complexity,

causing L2 oral production to increase in both accuracy and complexity, but will cause fluency to decrease, due to the increased conceptual demands, forcing the learners to reach into their repertoire of more complex language to complete the interaction within the task. This demand for more complex language will cause the attentional resources available to learners to focus more on the linguistic form that they need, causing a decrease in speed with which the subsequent language is produced. However, when there is an increase in resource-dispersing variables, attention is not directed at one specific aspect of the linguistic system, but rather is dispersed across all aspects. This means that the Cognition Hypothesis would predict a decrease in complexity and accuracy, but an increase in fluency, when resource-dispersing variables are increased.

It is important to note that while Robinson's model claims to account for interactional factors, dividing these into participation variables (e.g., open/closed interaction, two-way/one-way interaction) and participant variables (e.g., gender, familiarity, power/solidarity) and learner factors, divided into affective variables (e.g., motivation, anxiety, confidence) and ability variables (e.g., aptitude, working memory, intelligence; Robinson, 2005, p. 5), a good portion of his work on CAF does not satisfactorily address these variables. In Robinson and Gilabert (2007), there is only a short paragraph on these variables, where they state, "Finally, the Cognition Hypothesis acknowledges that learner factors (contributing to perceived difficulty) interact with task factors (contributing to their complexity) in determining the extent of the above predicted effects," of the resource-directing and resource dispersing variables as well as interactional variables on CAF (p. 168). They go on to state that interactional and learner

factors will affect the interaction in that if there are many present, the results will show less clearly how the resource-directing and resource-dispersing variables are affecting CAF. In other words, when these variables are not present, it is easier to account for the effect of task complexity on CAF than when they are. However, the authors do make a call for more empirical SLA studies that include a more rigorous look at these interactions and learner factors to determine how they affect CAF (p. 168).

These two approaches or frameworks to the study of CAF, while both working under the assumption that attentional resources are limited, differ in fundamental ways. The most important difference is how the two models describe the allocation of attentional resources for L2 performance. The LACM is much more conservative with its description about how limited a learner's attentional resources are. According to this model, "humans have a limited information processing capacity and L2 learners must therefore prioritize where they allocate their attention during task performance, so that attention allocated to one dimension of language production will be lost on others" (Housen et al., 2012, p. 6). Therefore, when the complexity of a task is increased, gains in one aspect of CAF would seem to promote losses in the other two, since all of the attentional resources would be focused on that one specific aspect, with few to none left over to attend to the remaining two parts of the construct. The demands on the learner to make meaning and be understood will become prioritized over linguistic form in an interaction, thus resulting in the "trade off" of meaning vs. form (Kuiken & Vedder, 2012).

In contrast, Robinson's Cognition Hypothesis works under the assumption that attentional resources are not as limited as the LACM asserts, and that "learners draw on multiple attention pools simultaneously," (Housen et al., 2012, p. 6). When a task becomes more complex along the "resource-directing" variables, the learner draws upon those resources that allow the learner to express herself with more varied (complex) and more accurate language in order to communicate the resultant complex ideas. This means that, when a task is made more complex with "resource-directing" variables, learners will be able to produce L2 output that is both more complex and more accurate. In this model, the two are linked and increase together in the context of a complex task, while fluency may suffer due to the attentional resources allocated to the other dimensions of CAF (Housen et al., 2012; Kuiken & Vedder, 2012).

The two approaches also differ in how they define task complexity. Skehan (1996) cites Brown, Anderson, Shillcock, and Yule (1984), stating that certain features of tasks have different levels of difficulty for learners. For example, static tasks such as description are easier than dynamic tasks such as narration, and those tasks that are more abstract, such as those that require critical thinking or opinion giving, are more difficult (p. 40). For Skehan, tasks that are more difficult and require different kinds of processing are also more complex. For Robinson, the two are not conflated (Robinson, 2001a, 2001b, 2007; Robinson, Cadierno, & Shirai, 2009). Within the Triadic Componential Framework, task difficulty and task complexity are measured in different ways, because it is assumed that individual learners will perform differently on a task that is difficult, and the perception of difficulty will vary among learners, while the complexity of a task

may remain somewhat stable. However, as Kuiken and Vedder (2012) note, manipulating tasks in the realms of difficulty and complexity together may create even more confounding factors which will cause difficulty in determining factor effects. More elements may create more chance for interaction between the different variables (pp. 150–151).

Now that we have examined task complexity and theories that attempt to explain how CAF are realized in learner language, the variables CAF themselves and studies that look at the effect of task on oral CAF will be explained and described.

Complexity, Accuracy, and Fluency

Housen, Kuiken, and Vedder (2012) describe accuracy as "the ability to produce target-like and error-free language" (p. 2). However, complexity is a bit more problematic to define, since this term is used to refer to both cognitive complexity and linguistic complexity. Cognitive complexity, according to the authors, is "the relative difficulty with which language elements are processed during L2 performance and L2 learning" (p. 4). They define linguistic complexity, on the other hand, as

> an important component of cognitive complexity (or difficulty), but it does not coincide with it. Linguistic complexity is an objective given, independent from the learner, which refers to intrinsic formal or semantic–functional properties of L2 elements (e.g., forms, meanings, and form–meaning mappings) or to properties of (sub-)systems of L2 elements. (p. 4)

This definition of linguistic complexity seems to match Bulté and Housen's (2012) definition of linguistic complexity, which has to do with the degree of elaboration of the linguistic system (p. 25) and is called *global* or *system complexity* (p. 25). Bulté and Housen provide a quite elaborate discussion on the operationalization of the notion of

cognitive complexity, but, for the purposes of this dissertation, linguistic complexity will be the only type of complexity considered in both the interpretation and analysis of the data.

Ellis and Barkhuizen (2005), Housen and Kuiken (2009), Norris and Ortega (2009), and Housen et al. (2012) discuss the problems that can arise with studies that do not adequately or consistently operationalize CAF. The main critique is the variety of ways in which CAF have been operationalized causes difficulty in comparison across studies. Bulté and Housen (2012) argue that there are even studies that do not outline how any of the CAF constructions have been operationalized, further causing difficulty for comparison across studies. Ellis and Barkhuizen outline the most frequent ways, at the time of writing, that these terms have been operationalized in SLA research.[3] The main ways of measuring accuracy in CAF research, according to Ellis and Barkhuizen, are number of self-corrections, percentage of error-free clauses, errors per 100 words, percentage of target-like verbal morphology, percentage of target-like use of plurals, and target-like use of vocabulary (p. 150). None of these measures is without issue. Since grammatical accuracy, including percentage of target-like verbal morphology, is one of the measures that will be used here, its validity will be briefly discussed. There is a large body of literature on the acquisition of verbal morphology in Spanish, as will be discussed in the following section, allowing the researcher to be able to make comparisons to what is already known about the developmental stages of, for example,

[3] Because the current work will not look at fluency, that variable will be left out of this discussion, but complexity and accuracy and how they have been measured in previous work will be discussed.

tense and aspect, in the oral production of L2 Spanish. Ellis and Barkhuizen caution that the use of this measure of grammatical accuracy "rests on the extent to which learners' ability to use the verb tenses/plurals correctly correlates with their overall grammatical competence" (p. 151). They go on to state that learners do not acquire grammatical features concurrently, but I argue that since we do know some information about when and how verbal morphology is acquired,[4] this is a valid measure for the current study.

The operationalization of linguistic complexity is also outlined in Ellis and Barkhuizen (2005) who list five commonly used types of linguistic complexity: interactional, which is operationalized as number of turns and mean turn length; propositional complexity, evaluated by number of idea units encoded; functional complexity, described in terms of frequency of some specific language function (e.g., hypothesizing); grammatical complexity, conceptualized as amount of subordination, use of particular linguistic features (e.g., different verb forms), and mean number of verb arguments; and finally, lexical complexity, operationalized as type-token ratio (pp. 153–154). It should be noted that type-token ratio is not frequently employed anymore as a measure of lexical complexity. More common measures are MTLD (measure of textual lexical diversity; Schmid & Jarvis, 2014, p. 731) or D-value, a calculation based on a mathematical probabilistic model (Malvern & Richards, 2000, 2009, 2012; Malvern, Richards, Chipere, & Durán, 2004). The authors state that grammatical complexity is the most common way of operationalizing linguistic complexity, and further state that a measure such as the use of some specific linguistic feature is a good measure of

[4] See the discussion on the acquisition of verbal morphology in this chapter.

complexity in that it occurs in all levels of L2 learners, irrespective of their developmental level (p. 155). Bulté and Housen (2012) add to the task of summarizing the many ways that linguistic complexity has been operationalized in the research. Their survey showed more than 40 different ways of measuring linguistic complexity under the umbrella categories of grammatical (syntactic or morphological) or lexical (diversity or density) complexity (pp. 30–31). Again, they cite this veritable smorgasbord of options as one of the causes of the lack of comparability across research studies on CAF.

Norris and Ortega (2009), noting this lack of consistency in research methods, describe the ways in which linguistic complexity and accuracy have been defined and used in the research as they make a call for more consistency as well as a more "organic" approach to the measurement of linguistic complexity and accuracy. Further, they outline which global and which specific or fine-grained measures are most appropriate for each learner level, as certain features will be more easily measured, either because of frequency of appearance or amount of variation able to be measured. For example, they state that coordination is the most useful measure of linguistic complexity in beginning learners, but that subordination becomes more useful for intermediate learners, while phrasal-level complexity is the best measure for advanced learners (p. 564). They also call for multiple measures of linguistic complexity, as "depending on the proficiency or developmental levels of learners, if we focus only on anticipated changes in one area, we may be missing the really important changes (or lack thereof) going on in another" (p. 574). They end with a call for longitudinal studies that "employ measurement practices that engage with the construct reality of multidimensionality" (p. 574).

Now that the theoretical frameworks surrounding CAF have been outlined, the following section will discuss how task type has been conceptualized in the literature, followed by a description of the SLA studies done specifically on Spanish as a foreign or second language.

Conceptualizations of Task Type in the Literature

Just as there is much variation in the conceptualization of the constructs of CAF in the literature, so is there variation both in the number of definitions of "task" (Ellis, 2003) as well as descriptions of what sorts of tasks should, can, and have been employed in research and in the classroom (Mackey, 2012; Skehan, 2003; Skehan & Foster, 1997). According to Mackey (2012), the most commonly used task types in previous research are picture description, spot-the-difference, story completion, jigsaw tasks, and consensus tasks (p. 22). She further describes that tasks in interaction research can be differentiated further by, for example, whether a task is open or closed, one-way or two-way, and whether the information exchange is optional or required (p. 22). Pica, Kanagy, and Falodun (1993) include these items in their taxonomy of task type, but also include the dimension of whether the task is convergent (requires collaboration) or divergent (can be done independently; pp. 13–15), and instead of classifying a task as open or closed, they list the number of possible outcome options. One possible outcome would be a closed task, and 1+ outcomes corresponds to an open task.

The authors also include another task type, the information gap task. At first glance, this appears to be a very similar task to the jigsaw task, but the most important distinction between the two is that a jigsaw task is a two-way task, in that both

participants are contributing equally to reach the predetermined goal, while in an information-gap task, it is a one-way flow of information. That is, only one participant has information, and the other participant must elicit it from her partner (pp. 20–21).

Ellis (2003) also mentions information-gap activities (p. 88), but in his definition of what is an information-gap task, he states that there are both one-way and two-way information-gap tasks. The two-way information-gap tasks, according to this definition, appear to be very similar to Pica et al.'s (1993) definition of a jigsaw task. However, later in the manuscript, Ellis (2003) does define a jigsaw task similarly to Pica et al. (1993), stating that it is a two-way task that requires collaboration among the participants and pooling of information in order to comply with the goal of the task (p. 215). It is at this point that he also adds problem solving tasks, decision-making tasks, and opinion exchange tasks, stating that all of these remaining tasks can be either one- or two-way, all are optional for participation, and all but the last are convergent tasks, while all but the first is open-ended.

It is worth noting that a good portion of the research done with these operationalizations and categorizations of tasks have measured task effects by analyzing amount of negotiation in an interaction, number of recasts and language related episodes (LREs; Ellis, 2003; Mackey, 2012). That is, the tasks themselves have been manipulated along the lines previously mentioned (broad category such as information-gap, one vs two-way, open vs. closed, etc.), but the main goal was to determine what kind of interaction resulted with respect to negotiation of meaning, recasts, or LREs, for example. The reason for this type of analysis is that this research is done in large part by

interactionist scholars, using the Interaction Hypothesis, described previously, which works under the assumption that in order for language to be acquired, it must be used in interaction. Therefore, the idea is that the research should show what sorts of tasks promote the things that the Interaction Hypothesis claims will facilitate learning. The results, it is hoped, can inform language teaching on best practices for promoting language learning in the classroom by describing what kinds of tasks will result in, for example, the most negotiation of meaning, LREs and/or recasts. This type of approach rarely investigates the effects of task type on CAF, unlike the cognitive approach to task-based research.

Those scholars who work under a more cognitive approach to task based teaching and learning have taken a slightly different approach to the measurement of task effects (Robinson, 2001a, 2001b, 2003, 2005, 2007; Skehan, 1996, 1998, 2003; Skehan & Foster, 1997) on CAF. They, too, wish to inform language pedagogy, but they also wish to uncover the mental processes that govern language production. As previously outlined, there has been effort to determine whether certain factors related to task, such as complexity or difficulty, would affect linguistic performance in the tasks. There has also been a great effort to determine how planning time affects linguistic production by task as well. The results of these studies, as has been outlined in the previous section, are inconsistent at best, with less comparability across studies because of the lack of consistency in the operationalization of the variables (Bulté & Housen, 2012; Norris & Ortega, 2009).

Unfortunately, as much as the research has been interested in informing pedagogy, there seems to be some discrepancy between how this information is presented to aspiring language teachers vs. how it is presented in the research-based literature. For example, one of the pedagogy books used in the present study in the categorization of teacher-designed tasks, Willis and Willis (2007), makes no mention of whether studies were done using their categorization methodology. In general, books such as Willis and Willis (2007) are meant to instruct teachers on how to employ task-based teaching and learning (TBLT) into their pedagogical repertoire. Thus, their focus is not on confirming or refuting research findings on task-based instruction.

However, this is not to say that there is no overlap between how tasks are operationalized and described in the instructional materials. Shrum and Glisan (2009), is a popular instructional book for Foreign Language Education programs, and as will be seen in the descriptions of task types for the current study, they do make use of some of the tasks outlined in the literature, such as information gap and jigsaw tasks. This facilitates somewhat the ability to compare the results of the present study with research-based findings.

Since the development of linguistic accuracy and complexity over time is the goal of the current work, it is appropriate to follow this description of previous work done on CAF with a discussion on what is known about how verbal structures in L2 Spanish develop.

The Development of Verbal Morphology in Spanish as L2

Verbal complexity and accuracy and how this develops over time were chosen as a part of the linguistic analysis for this study because verb use is often employed as a barometer by which proficiency is measured. For example, the ACTFL proficiency guidelines (http://actflproficiencyguidelines2012.org/speaking) cite ability to manipulate different tenses in each of the ratings, such as the description of an advanced speaker, which states, "The topics are handled concretely by means of narration and description in the major time frames of past, present, and future." Verbal accuracy and use of target structures from previous lessons are also often used as a gauge for grading oral examinations. Because the use of verbal structures appears to be weighted so heavily in assessments, it was deemed an important variable for a longitudinal study of language acquisition and development.

There have been several attempts to determine how students utilize verbal morphology to mark tense and aspect and how their acquisition of tense and aspect develops over time. Andersen (1989, 1991), Andersen and Shirai (1996), and Bardovi-Harlig (1992, 1995), though in the latter case the studies focused on ESL, investigated the effects of lexical aspect, the relationship of semantic aspect built into the meaning of the verb, on the choice of verbal morphology. Andersen (1989, 1991), whose work focused on L2 Spanish learners, proposed that imperfect aspect begins with stative verbs and then extends to nonstative verbs while perfective aspect (preterit) begins with punctual verbs and then spreads to nonpunctual verbs and eventually to stative verbs. Specifically, in his 1991 work, he outlines a developmental sequence for tense and aspect morphological

marking based on data obtained from two children learning Spanish in a naturalistic environment in Puerto Rico in 1978 and then again in 1980. He admits that the intermediate stages were recreated and were not evidenced in his actual data, but according to his developmental sequence, learners will use the simple present for all types of verbs in stage 1. In stage 2, learners will begin to mark perfective aspect with the preterit with punctual events, but still will use the simple present for all other types of verbs. In stage 3, the imperfect is introduced with stative verbs, but activities and telic events (accomplishments in other literature) will still be expressed with the simple present. It is not until stage 4 that the imperfect extends to activities and the preterit extends to telic events. At this point, learners tend not to vary in their use of verbal morphology: stative and activity verbs are always marked with the imperfect, while telic events and punctual events are always marked with the preterit. However, in stage 5, the learners begin to alternate between preterit and imperfect with telic events only. In stage 6, the two aspects alternate with telic events and activities. Stative verbs and punctual events are still only expressed with imperfect and preterit, respectively. In stage 7, punctual events acquire both the imperfect and preterit morphology, leaving stative verbs the only category that is not marked by learners for both types of verbal morphology. In the final stage, stage 8, learners alternate between preterit and imperfect, depending on the meaning that they wish to convey, in all four categories of verbs. Andersen believes that this sequence occurs because learners are attending to lexical aspect first, and it is that factor which conditions their morphological marking of verbs, otherwise known as the Lexical Aspect Hypothesis. It is only later that they acquire grammatical aspect,

morphological marking, and are able to attend to the aspect of the whole proposition/predicate and mark verbs in a more target-like way.

Some years later, Andersen and Shirai (1996) put forth a more detailed staging of the development of verbal morphology in learners, though much of their data come from children, and the L2 acquisition in many of the cited studies was naturalistic rather than classroom based. They posited that learners first use perfective marking; in the case of Spanish, this would be the morphological preterit, with achievement and accomplishment verbs, later extending to activity and stative verbs. In contrast, in Spanish, the imperfect appears much later than the preterit, and first appearing with stative and activity verbs and then extending to accomplishment and achievement verbs. In addition to perfective and imperfective, they also proposed that progressive aspect will be marked with activity verbs first, and then later extend to accomplishment and achievement verbs, which then, not incorrectly, overextend to stative verbs (p. 533).

Keeping in mind that a good portion of the studies done by Andersen (and later, with Shirai) were based on data collected by prepubertal children, Salaberry (1999, 2002, 2003, 2011, 2013) expanded upon these studies with adult intermediate L2 Spanish learners taking part in classroom-based learning in order to determine whether the Lexical Aspect Hypothesis exerted any effect on the production of verbal structures in L2 Spanish, utilizing both a cross-sectional selection of Spanish L2 students at beginning, intermediate, and advanced level of study (Salaberry, 1999, 2002, 2003, 2013) and a group of advanced bilinguals in comparison to a group of Spanish monolinguals (Salaberry, 2011). He found that, in the beginning, there seemed to be an

overgeneralization of preterit use among learners, and hypothesized that this may be due to transfer from English, where perfective aspect is marked with –ed, and imperfective aspect is minimally marked in English verbal morphology. However, he also found that as proficiency increases, the Lexical Aspect Hypothesis tends to exert more of an effect. Liskin-Gasparro (2000), in a study meant to expand upon Salaberry's previous work, investigated whether students' production of verbal structures would differ in two narrative tasks: one a retelling of a personal experience, and one retelling of a silent film clip. Her participants were 8 advanced learners of Spanish who participated in the retell tasks, and then immediately afterward were interviewed to determine why they used the verbal structures they used in each retell task. She found results similar to Salaberry's; students' intuitions of why they use imperfect versus preterit had more to do with the lexical aspect of the verb than the context in which they were using it.

Bardovi-Harlig (1995) found that, in addition to lexical aspect, grounding may be important in the expression of verbal morphology, though as mentioned previously, her work was done on ESL learners. Her study found that foregrounded information in narrative speech tended to be expressed in perfective aspect (in the case of Spanish, the equivalent would be preterit), and background information tended to be expressed in imperfective aspect (which would be the imperfect in Spanish). Liskin-Gasparro (2000) found similar results, but also found that that narrative type had an effect on the production of verbal aspect. When students were recounting a personal narrative, they tended toward more imperfect verbal structures, stating that the stories were situated more proximally or personally to them, while a movie narrative produced many more

preterit constructions. The movie narrative, naturally, was more objective and distal to the students as they retold the plot. She concluded that her participants were using aspect to situate themselves as participants, or not, in each narrative. None of these studies, however, were performed using truly naturalistic classroom data. They used carefully designed tasks meant to elicit certain structures in order to be comparable across participants and settings.

Gudmestad and Geeslin (2012) used a cross-sectional design to determine the variable use of future time with L2 Spanish learners. They coded verbal structures for morphological (synthetic) future, periphrastic future, and present indicative, as it may be used to indicate future time. They compared the learner data to NS data in order to determine if the learners were using the structures in ways similar to the NSs and in similar contexts. For the purposes of the current study, only the results of Level 1 and Level 2 learners will be described, as they correspond to second semester, first year, and second semester, second year students. Their data were obtained via a word completion task (WCT) that elicited future time. The linguistic variables manipulated in the task were language time indicators, temporal distance, and certainty markers. The results showed that Level 1 learners used the present indicative 48.3% of the time, the periphrastic future 39.5%, and the morphological future 12.2% of the time. The Level 2 students seemed to flip the present indicative use with the morphological future use, with the former used 14.1% of the time and the latter 40.7% of the time, and the periphrastic future 45.3% of the time. Gudmestad and Geeslin also found that starting at Level 2, when there was a certainty marker, the learners performed similar to NSs with respect to choosing the

present indicative or morphological future. However, the other two conditions, the language time indicator and temporal distance, did not show such consistent similarity to NS norms, causing the authors to posit that the variation with these markers will take longer to develop. Because of this, it can be expected that the participants in the current study may show NS-like production of future time, but more accurately in some contexts than others, namely in contexts of certainty.

Silva-Corvalán (1996), in her study of the contact between Spanish and English in Los Angeles, discussed the attrition of the tense–aspect–mood system in Spanish/English bilinguals. Within this context, she highlighted how English-dominant bilinguals used Spanish verb structures to express tense, aspect, and mood, especially when their verbal repertoire was limited. Keeping in mind that this is in the context of language contact and attrition, it is still worth discussing her results as it is possible that these trends may also be found in this study's learners. First, she found that there was substitution of the closest verb in tense/mood/aspect to the one that would have been judged "correct" in a context, such as the use of the imperfect in place of the imperfect subjunctive (p. 42). She also found a trend that her participants often used periphrastic and auxiliary constructions with nonstative verbs and that the perfective/imperfective opposition disappeared with stative verbs (p. 47). This is most certainly a trend that could be seen in intermediate L2 learners of Spanish, since they tend to use what they have to make meaning as their tense/aspect/mood system is in development, as seen in the discussion above. Interestingly, she also found that syntactic complexity, as measured by subordination, decreased with increasing length of time living in the United States (p. 70). In general,

the participants in Groups 2 and 3 were those who were English dominant, so it follows that those that have a more limited proficiency in their nondominant language may not be able to express the same kind of syntactic complexity as more advanced proficiency speakers.

To summarize, because the learners in the current study may range in proficiency from novice–high to intermediate–high, and travel through different levels of development over the course of the year, the verbal structures that can be expected to appear in their learner language are present indicative, preterit, imperfect, morphological, and periphrastic future. There may be instances of present perfect due to transfer from the L1 as there is an equivalent structure in English. They may also produce more types of periphrastic or auxiliary verbal structures due to their inability to express themselves in a more native-like way because of an underdeveloped tense/mood/aspect system.

Chapter 3

Methodology

This study is a longitudinal case study of the development and production of linguistic complexity and accuracy in the oral Spanish of 3 intermediate-level university students. This chapter will describe the 3 participants as well as the contexts in which they were learning Spanish, the data collection techniques, the operationalization of the constructs in the research questions, as well as the method and rationale for the coding, and the types of analyses used to illustrate task effects and development.

Participants

This study's participants were 3 second-year Spanish language students in the third and fourth semesters (i.e., the second year) of university-level language study from two Midwestern institutions of higher education. University A, as previously mentioned, is a large, urban, public research institution while College B is a smaller, more rural liberal arts college. All participants were 18–25 years of age. Placement into the second-year courses was made based on entrance exams taken by all students at their respective institutions of higher learning. University A's placement exam includes three modes: reading, writing, and listening. College B's placement exam tests only reading and writing. Students in the first semester of the second-year sequence typically are at novice–high or intermediate–low proficiency in speaking according to ACTFL Proficiency Guidelines (2012), and should achieve intermediate–mid proficiency in speaking by the end of the second semester of the series. All participants were self-selected for participation based on instructor willingness to allow classroom-based

research to be conducted in their section of Spanish. Because this is a case study, each participant will be introduced and described in detail, with a discussion of each institution and description of the classroom environment and course activities to follow.

Mike

Mike was a 19-year-old Hmong–American male student at University A. He was in his first Spanish course at the university after having taken a 2-year break from language study. However, he had studied Spanish for 4 years in high school, and was thus in his fifth year of Spanish language study. He reported having spoken Hmong up to and during elementary school, but had never received any formal instruction in the language, and did not speak it any longer. He considered English his dominant language. During the initial interview that was performed with each of the participants, Mike described himself as shy, with speaking being the most difficult part of language learning for him. He reported speaking Spanish most with his classmates, his apartment mates, who were nonnative speakers of Spanish, and his Spanish professor, in that order. Mike did not express a preference for speaking with any particular kind of Spanish speaker; he found value in all kinds of interlocutors. For example, in the questionnaire item that asks if nonnative or native-speaking partners are preferred, his response was, "both because having a native speaker allows me to learn spanish [sic] more easily and having a nonnative speaker partner allows us to learn together knowing that we're not fluent at spanish [sic]." He also responded that he liked more, same, and less proficient speaking partners equally because he could learn different things from all of them, and a same or less proficient partner allowed him to teach them what he knows. However, despite his

enthusiasm for working with all kinds of Spanish-speaking interlocutors, he also expressed that his shyness prevented him from seeking out interlocutors outside of his roommates and classmates in order to practice Spanish. In fact, he described interacting with native Spanish speakers as being scary.

This shyness also manifested itself in his classroom interactions in that he would be hesitant to try something that he perceived to be too challenging. That said, he did make small goals for himself in his Spanish class, such as asking at least one question per day, and if he was able to attain that goal, his next goal would be to elaborate on something else. These small goals allowed him to take on small personal challenges and feel good about the progress he was making. Mike also enjoyed the challenge of working in small groups with people whose life experiences or worldviews differed from his. Speaking with someone who had different opinions, who showed him some aspect of a problem that he had never thought of before and allowed him to see some problem or issue in a new light was very exciting to him. This kind of situation inspired him to interact more in a group, to express his opinions and feelings more openly. According to Mike, if he is in a group with interlocutors who hold a somewhat homogeneous view of the world, he is much less likely to interact with other group members.

Mike's desire to do well carried over into the spring semester, and in his final interview, he expressed many of the same ideas about his own language learning. He continued to try to compete with himself to do better each class, especially because he did not earn the grade he would have liked to have earned during his fall Spanish course. According to Mike,

> I think I had a greater desire this semester to, to do better because last semester my grades weren't so good . . . Sort of an incentive for me to kind of put myself out there and usually I don't talk that much in class. I think the first part of the second semester, I barely talked at all. It was towards this last part, where, you know, I was getting worried about my grades and then I was like, "OK, you gotta put some more effort," so, so yeah.

Additionally, during the spring semester, he befriended a native Spanish-speaking coworker from Mexico with whom he enjoyed speaking Spanish. This interlocutor was very patient with Mike and would slow his rate of speech when they would interact in Spanish. Because of this increased effort and the additional practice outside of his Spanish class, he felt more comfortable speaking Spanish during the spring semester than the fall semester.

Mike's openness with regard to what type of interlocutor he preferred to work with remained stable from the fall semester to the spring semester. He continued to feel as though he could benefit from an advanced or native Spanish speaker, a same-proficiency speaker, or a less proficient speaker, and that his interlocutor did not have an effect on his language production. He continued to feel that the interlocutors' attitudes toward the task or interaction were the most important factor in whether he wanted to interact with them or not. He was much more excited about working with a partner who was enthusiastic about learning Spanish or speaking Spanish no matter the proficiency of the person. For example, when asked if there was anyone in his spring semester class with whom he would prefer not to work, he responded,

> There's a student, [redacted], and I can tell he's smart, but he's like one of those people who doesn't really care about the material, he wants to get it done, he wants to get it over with. But he doesn't really have that, you know, kind of attitude where it's like, "You know, let's make this fun," or "Let's enjoy it a little bit."

He also continued to think that one of his biggest weaknesses with respect to speaking Spanish was the ability to express complex ideas. He said he felt comfortable expressing his ideas when they were simple ideas. However, he experienced more difficulty when he tried to express more complex ideas that would require the use of different structures.

> I know there are a lot of different [verb] tenses but sometimes I prefer to use only one and it's, it's kind of complicated or complex for me to kind of integrate, to say, for example, using like a future tense with, like, imperfect or stuff like that.

When he was asked how he thinks he could improve that aspect of his Spanish, he responded with,

> I think the way to make it better is just to, you know, practice it more, get involved with like the Spanish culture, getting to situations where you're forced to or you have to use those kinds of ideas and then also, just—what's the word I'm looking for? Being able to, you know, be open to just different kinds of arguments I guess, because in arguments you find really complex ideas. And the way to be exposed to those are in this case, you know, like arguments in Spanish or so, or you can see all these different kinds of conflicts and try to piece them together.

However, when asked if he thought it would be a good idea to have more of these types of interactions in the classroom, he responded that he thought they weren't appropriate for second-year language classes. Mike also thought that he got more out of the class meeting 4 days a week in the spring semester versus the fall class that met 3 days a week because he had more chance to practice and use what he was learning. He did not enjoy the online activities, preferring to interact face to face.

Finally, when asked where he would rate himself in comparison to his classmates in the fall semester, he was hesitant to compare himself to other students. His reasoning was that he couldn't know how proficient his classmates really were, for example, if they

didn't speak much in class or if he did not work with them. That said, he did say that he would rate himself about in the middle with regards to his Spanish speaking skills in his fall semester class, and about in the top quarter in his spring semester class.

Mike, at the time of the final interview, was majoring in a social science, and thought Spanish would be helpful, but was not sure if he would continue his language learning.

Teresa

Teresa was an 18-year-old female freshman at College B. She was a heritage speaker of Spanish who reported English as the language spoken in the home after age 5, and that since age 5, she spoke no other languages in the home and had not had any other formal instruction in languages other than Spanish and English. Teresa responded that she had had 4 years of formal instruction in Spanish in high school at the time she filled out the questionnaire, with the year of data collection, her freshman year in college, being her fifth year of formal study. Teresa is a special case, because her mother was a native Spanish speaker from Argentina, but due to some difficult experiences that both she and her mother endured in the United States, her mother stopped speaking Spanish to Teresa when she was around 5 years old. Teresa said, when asked if she spoke Spanish with her mother, "It was frowned upon because I, Spanish was my first language originally but people were really weird towards my mom because I spoke English with an accent, so she stopped and I gradually just forgot." However, she wished she had spoken more with her mother, "because then I would be feeling obviously better now if we were to be speaking since I was little." She said that she has tried to initiate Spanish conversations

with her mother, but felt that perhaps her mother wasn't very receptive to it because it does take effort, saying, "I've tried, especially when I studied for the AP exam I tried to obviously speak Spanish but English was just easier because she was busy and it takes a lot of time and dedication."

Teresa expressed that she preferred to speak with native or highly proficient Spanish speakers because she felt a more proficient interlocutor would help her learn more, become better, and improve her accent. She mentioned that while in a Spanish-speaking country (in the past, she traveled once to Spain with her high school class and to Argentina to visit family an unknown number of times), she tried to sound like the native speakers of the area. "And I love learning Spanish without an American accent and I loved when people thought I was actually from Spain. And I would, you know, be up all night imitating their lisp; it always sounded like a lisp to me but…" When asked why she wanted to sound Spanish, she responded with, "Just because I was there. If I was in Argentina I'd try to sound the same. Pronounce how they pronounce and use their slang which is very difficult." She did seek out native speaking interlocutors at her college, because she felt they would help her in all areas of her oral Spanish—fluency, accuracy, and accent.

Speaking "correctly" seemed to be a concern for Teresa, as she used this as a descriptor of native speakers' Spanish as well as the reason why she preferred to speak with high proficiency and native Spanish speakers. She said, regarding speaking with same proficiency interlocutors, "I think it's fun at times and I'll talk but I don't necessarily learn a lot because if I need to ask questions or just want to listen to someone

talk, their grammar is just as full of mistakes as mine." She also said she felt that speaking with high-proficiency nonnative Spanish speakers gave her hope that she would improve her own Spanish. This is an interesting statement because it seems to imply that she did not consider herself a native speaker of Spanish, even though she did say Spanish was her first language. However, it may point toward some linguistic insecurity that she expressed in both the initial and final interviews as well as during some of the class periods. For example, in the final interview, she expressed that occasionally she felt "completely inadequate" speaking Spanish. She also said, in the same interview, "If I'm not paying attention, if I'm really tired I can speak Spanish very well because I don't worry, but when I worry I stress and then I just shut down because I'm worried that I'm using the wrong reflexive verb or tense or something like that," and "I don't think I do all that much better [than my classmates]. I think a lot of them usually think I speak, my Spanish is really good because I just sound like I know what I'm doing, so the whole 'fake it 'til you make it' but…"

She also recounted an example of an interaction with a highly proficient Spanish speaker that made her feel badly about her Spanish. "I over the summer met someone who was from Italy but also spoke Spanish and I get really subconscious [sic] of my mistakes and I said something wrong and he pointed it out. And I thought, 'Well, that's intimidating. You can let it slide!'" However, even though she preferred native and near-native (or high proficiency) interlocutors, she did not feel that working with lower proficiency partners was a negative experience. She stated that the mistakes that her interlocutors made could help her learn as she tried to correct their mistakes in her head.

So it seems that she might have suffered from some linguistic insecurity, but she was also aware of some advantages she had over the other students who did not start out speaking Spanish in the home at a young age. For example, she rated herself in the "middle towards the top" half of her class as far as Spanish speaking ability in her fall semester class. When asked why she thought that, she responded, "The fact that I could just, at least if I wasn't grammatically correct I could just sound like I was grammatically correct. And that people would ask me a lot of questions all the time. And I didn't mind at all. It was helpful to explain things." Those she rated as better than her were described as being better "grammatically" in their Spanish. However, she rated herself as being in "the middle" of her spring semester class as far as Spanish skills. She felt as though that particular group of students was much stronger than her with respect to control over the grammar. She also did not feel as though she had improved all that much from the fall to the spring semester, citing the difficulty of learning the verbs and grammar. However, she did report that her mother commented on her improvement from fall to spring semester. Teresa and her mother took a trip to Argentina during spring break for approximately 1 week in mid-March to visit family. While in Argentina, Teresa reported that she spent a considerable amount of time speaking Spanish with family members, though she did not reveal whether or not she stayed with family. She felt that communicating in Spanish, both with her family on the trip and after she returned to the United States, was much easier than it had been. However, when asked how she thought she could improve her weaknesses, she replied that practice was an important factor in improving. It seems that this was a great motivator for Teresa, and why she sought out native and advanced

interlocutors. She was very invested in improving her Spanish and seemed to like to be perceived as not having an American accent; she had a desire to speak "correctly" as a native speaker would. She expressed interest in continuing her studies in Spanish to a major in Latin American Studies at her college.

Rebecca

Rebecca is also a female student, an 18-year-old freshman at College B with Teresa. She had taken 3 years and 1 semester of Spanish in junior high and high school until 10th grade, resulting in a 2-year break from her language studies before taking Spanish at College B. She also reported no languages other than English being spoken in the home and no other formal instruction in a language that was not Spanish or English. Rebecca, much like Teresa, was concerned with how well she used Spanish grammar when she spoke, talking about her own grammatical accuracy and "fixing" mistakes. "I try to correct myself, like if I can, if I can catch myself saying something wrong, if I catch it I try to correct myself or I think, or just talking or asking questions definitely like, 'Is this subjunctive, not subjunctive?'" She was also aware of the difference between oral and written Spanish, with respect to accuracy, saying,

> I do wish that sometimes we can do a little more on the grammar front just because, you know, when you start talking to people in class you're just trying to summarize stuff, a lot of the grammar goes out the window and you're lucky if you get it all what's in the past tense in the past tense when you talk.

When asked if she preferred native or nonnative interlocutors, she stated she preferred native speakers because she felt she could learn more from them. She also stated that, with a native speaker, she would know she cannot fall back on her English, and would thus have to try harder to express herself in Spanish. Within the nonnative

speaker category, she preferred same proficiency interlocutors because, as she said, "then I don't feel like I'm doing all the work with less proficient partners or like an idiot with speaking experts. We feel comfortable talking to someone at the same level and as a result, are more likely to talk and improve." Rebecca spoke Spanish mostly with her classmates and professors, and she made an effort to attend the Spanish Conversation Table at her college. This was a conversation hour where Spanish learners and speakers of all levels could go and speak Spanish outside of the classroom. It was attended by students, native speakers and professors, providing participants with a wide variety of interlocutors. Rebecca was very motivated to learn Spanish and do well, as she thought it would be really "cool" to be bilingual, saying, "I have nothing but high respect for people who can speak two languages because it's really difficult and I really want to be able to do that." Though Rebecca asserted that she prefers native speakers to nonnatives, but same-proficiency speakers in a nonnative speaker, she did say that she was open to working with any proficiency speaker in her classes. Like Mike, what was most important to her was that the interlocutor be willing to interact with her and try. She did not care to work with people who did not want to be in class or did not want to participate. Like Teresa, she felt that even if her interlocutor was struggling, it was helpful to her to have to explain things. She was glad to offer support to a classmate who was having difficulty. Like the other 2 participants, she also believed that opportunity to practice was extremely important. "Outside of class there aren't really any opportunities [to speak Spanish] and then everything you've learned almost, you lose a lot of it. Because you're not applying it constantly." She tried to speak Spanish in class as much as

possible, even when she and her classmates had finished an activity. “I want to practice my Spanish and people who just drop into English whenever there’s down time, I’m like, ‘Oh, you could try. You could try to keep that in Spanish.’ Just need practice.” She also said that she felt pressure to speak more and contribute more when her partner was weaker or not willing to participate, but when she had a same proficiency or higher proficiency partner or a partner who was more active in the conversation, she did not feel this same pressure to contribute. She also felt as though she got more out of an interaction with a same or higher proficiency partner.

Like Teresa, Rebecca, too, felt she improved from fall semester to spring semester, saying that finding the “right verb” and expressing herself was coming easier to her, though she felt that grammar was still a weakness and a challenge for her. She felt more comfortable speaking Spanish, and felt as if her ideas were becoming more complex. She described being able to “go deeper” when discussing the texts that were used in her spring class.

> I think this semester as far as ideas went I think I was going more below the surface than just like the simple repetitive stuff. Like last semester it was just, “This is what the article says, blah blah blah …” and this semester is more like, “Oh, this was interesting. I wonder why…” This is what’s coming up.

According to Rebecca, her ability to think critically about Spanish texts and express her own ideas about them was much better in spring semester than fall semester.

With respect to her self-rating compared to her classmates in fall and spring semesters, she, like Teresa, rated herself higher in comparison to her other classmates in fall semester (middle to top of the class) than in spring semester (about middle of the class). She, too, attributed this change to the fact that she felt her classmates during the

spring semester were, in general, much stronger than those in the fall semester class. Rebecca also provided support for Teresa's assertion that their classmates thought she was good at Spanish, saying, "Teresa usually sounds really good because she has a very distinctive accent and she sounds like she knows what she's doing." Rebecca worked with Teresa more in fall semester than in spring semester, but Rebecca did not say whether she felt that working with someone she perceived to be really good affected how well she spoke Spanish. As previously mentioned, she said that what mattered to her was whether the interlocutor was engaging with her or not, and what was affected was the amount of language she produced. What affected the quality of what she produced, according to her, was how much sleep she had gotten and how many things were on her mind when she was speaking Spanish. She said that the more worried she was or the more she had on her mind, the less accurate or able to produce language she was.

Though she was very motivated to learn Spanish and was very interested in being bilingual, Rebecca was not planning to major in Spanish or Latin American studies. She was planning to take one more Spanish class at her college, but was unsure if she would continue after that time, though she was open to studying abroad in a Spanish-speaking country.

Classroom Experiences

The contexts in which the students were learning Spanish differed slightly in pedagogical approach. The description of each context that follows is not exhaustive and is meant to give a general overview of each classroom's approach, not a detailed account of every single task used in each. The goal of the present research was not to compare the

two contexts, but rather to describe how different types of tasks elicit (or not) linguistic complexity and accuracy as well as to describe the participants' development over time.

University A used a communicative approach to language teaching, in which students were presented with comprehensible input, then asked to use the form being introduced to communicate with classmates in a personally relevant way by describing, for example, personal habits, desires, activities, plans or something else that is pertinent to their surroundings. The ultimate goal of this university's approach is communicative competence and the achievement of intermediate–mid level proficiency on the ACTFL scale in speaking and writing skills, and an intermediate–high level of proficiency in listening and reading skills. The students were also presented with culturally relevant materials such as music and fictional prose written by native Spanish speakers, but this was not the bulk of the instruction that they received. The particular section of Spanish in which Mike was enrolled in fall semester, 2012, was a hybrid course. This hybrid course met face to face 3 days a week for 50 minutes and made use of a Moodle site (an online course management website) that included activities such as forum posts and short writing activities. This class also utilized an online grammar and writing practice workbook. The Moodle activities were both individual and interactive with classmates, and involved a variety of media, from visual to audio. Students also made recordings of themselves speaking for some of the assignments on Moodle. Many of the online activities were based on cultural items such as songs, images or other cultural productions. For example, one of the first online activities was related to the Spanish

Movida.[5] The students were walked through various visual images and texts related to the Movida, and ended with a forum post based on the images and texts they have viewed. These topics did not get discussed in detail in the face-to-face meetings in the section in which Mike was enrolled. In the face-to-face meetings, students performed communicative-based grammar activities, viewed/listened to the musical selection from the textbook and completed its corresponding comprehension activities, and read a text selection from the textbook and, again, completed the associated comprehension activities. The music and reading activities took up 1 class period each, and started with either watching the music video or reading the text, then would focus on completing the comprehension questions, and then follow up with a teacher-fronted question and answer session. They may or may not have completed additional activities in pairs, but every activity was followed by a teacher-fronted comprehension check. Many of the activities that were observed during the fall semester tended to be largely teacher fronted with little opportunity for the students to interact face to face with one another in the target language. For example, of the "regular" class recordings during fall semester (that is, those classes that did not have some other goal such as an oral assessment or presentation), only one class offered more than 15 minutes of small group work. However, it is important to note that this was a small sample of the total number of

[5] The Movida in Madrid was a countercultural movement that erupted after the death of the dictator Francisco Franco in 1975. It was a rebellion or freeing of the culture from Franco's oppression and symbolized the new identity of Spain. It was somewhat similar to the movement in the 1960's in the United States that celebrated freedom of expression, drug use increased, and the celebration of that which was taboo under Franco's regime. For a complete description, see Lechado, J. M. (2005). *La Movida: Una Crónica de los 80*. Madrid, Spain: Algaba Ediciones.

classes this instructor gave, and no data were obtained on the activities during classes in which Mike was not being recorded.

During the spring semester, Mike enrolled in a different type of language class, called *4+1*, which was similar to the hybrid course except that students met face to face 4 days a week for 50 minutes instead of 3 days. This was also taught using a communicative approach. In contrast to the fall semester, the instructor of the spring semester was a more experienced language instructor with 10+ years of experience teaching at the university level, while the fall semester course was taught by a graduate student with much less experience in the classroom. The students in the spring course had much more opportunity to interact with one another and many fewer or shorter teacher-fronted activities during the days I was present in the classroom. Most of the "regular" class days included more than 15 minutes' worth of small group interaction during the spring semester.

One of the main differences between the hybrid and the 4+1 contexts was the *tertulia* activities that were a part of the hybrid course. These activities included all four modalities: listening, speaking, reading and writing, presented and completed in an online module on Moodle. The materials with which the students interacted included readings from the textbook, but also included texts, songs, or other materials that were not a part of the course textbook, but still related to the themes presented in each chapter. The tertulias exposed students to several authentic texts that they would not have normally encountered using only the materials provided with the textbook and online grammar workbook. In the place of this extra material, the 4+1 had what was called *lectura guiada*

(guided reading), that only included reading activities: prereading, reading, and postreading activities focused on the text selection for the chapter. The listening and speaking activities that these online modules lacked in the 4+1 class were to be completed during face-to-face class time. Since the hybrid class met only 3 days per week, and 4+1 was a 4-day-a-week class, the idea was that the hybrid students "lost" one class period in which to interact with other students, making the inclusion of the four modes in the online tertulias a replacement for that fourth day of face-to-face instruction and speaking/listening practice.

Another difference between the two types of classes at University A had to do with the types of writing assignments that were given to students. Students in a hybrid class had an online writing tool called *cuaderno* (notebook), where they would write short essays in addition to the formal compositions required. The cuaderno activities were completed wholly online, with the instructor providing feedback within the application. These essays focused on the material from the tertulias, which were materials that were not procured from the course textbook, and students were expected to integrate this new information into a short, coherent essay in a formal space (the cuaderno). The 4+1 students did not have the *cuaderno* application. Instead, they wrote short essays on the weeks they were not doing *lectura guiada* on one of the short cultural readings found in the textbook. These short essays were also completed online, but not in the cuaderno application. The students were not expected to integrate new, outside information in these essays but rather to answer questions based on that reading from the textbook. They were expected to process that one short text in relation to their own life experience. For

example, the short essay on one of the *Comparaciones* (comparisons) selections, "Mark López hace que Google se dirija al público hispano" [Mark López makes it so that Google is directed toward the Hispanic public], includes the following directions:

> —Basado en la información del artículo, ¿cuáles son las características del público hispano en los Estados Unidos que se deben considerar en una campaña de marketing? (*Note: you must explain these characteristics in your own words; do not copy word for word from the text.)* Personalmente, ¿cuál de esas características te parece a ti más importante? ¿Por qué?"
>
> *–Based on the information in the article, what are the characteristics of the Hispanic public in the US that should be considered in a marketing campaign? Personally, which of these characteristics seems to you the most important? Why?*

The directions explicitly ask students to base their essays on the text and then ask them to describe their personal opinions about the information presented and then justify those opinions.

The last difference between the two types of classes at University A was the frequency with which the student completed the *foros*, or online discussion forums. The hybrid students had 4–6 of these online discussions per week, while the *4+1* students had to complete a *foro* once per week. These were sentence to paragraph level (depending on the student) amounts of discourse, and were meant to be interactive. In other words, these *foro* activities were not monologic tasks, and the topics were always related to the textbook chapter themes such as discrimination or crime. They were not based on outside, supplemental information like the *tertulias*, which regularly incorporated materials from outside the course textbook and accompanying online workbook.

The class section of Spanish language at College B utilized a content-based approach to language teaching, as outlined in Barnes-Karol and Broner (2010) and

Barnes-Karol (2010), in which the instructors employ visual images and authentic texts with the goal that students will learn to critically view them and produce complex academic language at the same time. This method intends to go beyond making communication in the L2 pertinent to the students' everyday lives; rather, in this class the goal is to help them develop a critical eye, think about the producer(s) of the media, the consumer(s) of the media, who is included and excluded and possible reasons why and how to analyze the contents through and outside of the lens of the native culture. It is hoped that this will allow students to develop academically rich and complex language as they are developing critical thinking skills in order to make them true global citizens who are culturally and linguistically sensitive to the target language and culture. This instructional approach also hopes to stimulate the ability to and desire to use this critical lens in the students' everyday life and learning.

During the course of the academic year, students completed a variety of activities: discussing images, authentic news stories both watched online and printed from the Internet, watching movies, reading excerpts of novels, discussing demographic information, among others. All the examination and composition materials were contextualized within the topics discussed in class, whether that was a movie, a country and its demographics, or current events. The exams were not grammar based, but any grammar that appeared was contextualized within the content that was being studied at the time. They did occasionally work with grammar in the form of contextualized worksheets (completed in and/or outside of class) and contextualized focus on grammatical form and performed metalinguistic analysis of their peers' compositions in

addition to the analysis of the content of the compositions. There were a variety of listening activities, most of which were viewed/completed at home or otherwise outside of the regular classroom. Students in this context watched movies, news clips, and were encouraged (and often required, at the discretion of the instructor) to participate in the college's Spanish Conversation Table. Spanish classes at this institution met 3 days a week for 55 minutes per period. They did not have a website similar to Moodle where students do culture-based activities. They completed individual writing assignments and reflections on the videos they were to watch that they handed in as homework assignments. The bulk of the classes at College B were devoted to allowing students to actively use the target language in a variety of ways. For example, students might be presented with demographic information about a Spanish-speaking country, and then asked to discuss this information. One task that was used in this context was an information gap-type task, where each member of a group of students had different sets of information, and was asked to share their information with the members of their group. They would all share until they had a complete picture of the set of demographic data. Once the data set was complete, they would then discuss the data, making comparisons to their own lives and drawing on their knowledge, asked to guess why the information may be different from what they previously thought, whether this information was presented in a positive or negative light and why that might be, etc. Another example of an activity that was used was the picture description. In pairs or groups, one member was given a postcard from a Spanish-speaking country. The holder of the postcard then had to describe the scene to the members of her group, who would then try to recreate the scene

on a separate piece of paper. At the end of the class, the cardholder would show the members of her group the card to see how close they came to the actual photo.

The two contexts from which the participants were obtained are very different, with one being a private liberal arts college and one being a public research institution. Therefore, the differences in teaching approaches are not outlined here to compare teaching approaches or institutions, as they are not comparable. The goal of this study is simply to determine if certain types of activities promote or inhibit linguistic complexity and/or accuracy in oral Spanish.

Data Collection

Prior to starting the classroom-based data collection and recording, a structured interview with the student participants was performed, and each participant filled out a questionnaire in English via the internet in order to determine attitudes toward Spanish, language learning, and interaction with native Spanish speakers as well as low, intermediate, and high proficiency nonnative Spanish speakers. The questionnaire and the interview protocol instruments can be found in Appendices A and B. Even though the participants were interviewed and completed questionnaires regarding their language use, stimulated recalls were not performed, because the researcher had limited access to the students during the academic year and no access at all to the students beyond the academic year. Additionally, no proficiency examination was given to the students due to time constraints.

The data were collected by giving each participant a digital recorder with a lapel microphone to wear on approximately a biweekly basis at each institution during the

entirety of a Spanish language class period, and consisted of transcribed recordings of the interactions between the participants and the other members of the classroom, including the instructor. The nonparticipant students were asked to refrain from working with or sitting near the participants if they did not wish to be recorded. Oral examinations and presentations were recorded and transcribed when available. The transcriptions were performed by the researcher in the style described by Hatch (1978) during spring semester, 2013 and over the summer, 2013.

During the data collection, the researcher played the role of observer in University A and as a classroom assistant at College B, taking notes during classroom interactions to note body position, visible signs of affect (nervousness, annoyance, etc.) and whether the participants actively avoided certain students or tended to work with certain students only. The researcher assisted with lessons when requested by the instructor and interacted with students when requested to do so. In the context of College B, the researcher was requested to function as an assistant in order to make the students more comfortable with her presence and to hopefully remove the appearance of being scrutinized in a laboratory setting.

Operationalizing the Constructs and Coding the Data

Linguistic complexity and accuracy have been widely studied in the field of second language acquisition, and one of the major challenges of this type of work, as described in Chapter 2, is the operationalization of these constructs and providing consistency and replication across studies (Ellis & Barkhuizen, 2005; Housen & Kuiken, 2009; Norris & Ortega, 2009; Pallotti, 2009; *inter alia*). Housen et al. (2012) provide an

overview of the ways in which linguistic complexity and accuracy have been operationalized in the literature. This variation in methods and constructs can cause problems when one tries to compare studies across the discipline and has been criticized by Housen and Kuiken (2009), but this also demonstrates a richness in the methods available to study linguistic complexity and accuracy. Further, Ellis and Barkhuisen (2005), in a discussion of how linguistic complexity and accuracy may be measured in SLA research, note that even though researchers may define the terms slightly differently, there seems to be enough correlation among the findings (pp. 163–164), that as long as the researcher is clear and consistent in how she is coding and analyzing the data, these differences should not cause problems in the analysis. As Norris and Ortega (2009) stated, " our measurements must provide multivariate, longitudinal, and descriptive accounts of constructs in L2 performance in order to capture the complex, dynamic, and development of CAF phenomena" (p. 574). With this in mind, the units of analysis and operationalization of the constructs will be presented and explicated.

The main unit of analysis for the present study will be the AS-unit, as outlined in Foster, Tonkyn, and Wigglesworth (2000). There are many different ways to segment spoken data in order to perform analysis (Foster, Tonkyn, & Wigglesworth, 2000), each with its own set of limitations and benefits. Because the variables under analysis are morphosyntactic structures, it was deemed appropriate to choose among the syntactic units already in use in the analysis of oral data. Among these, the AS-unit was determined the most appropriate for the type of oral data obtained from the participants. According to Foster et al. (2000), an AS-unit is defined as, "a single speaker's utterance

consisting of an *independent clause, or subclausal unit*, together with any *subordinate clause*(s) associated with either" (p. 365). They further define the terms independent clause, subclausal unit and subordinate clauses. "*An independent clause* will be minimally a clause including a finite verb" (p. 365). "*An independent subclausal unit* will consist of: *either* one or more phrases which can be elaborated to a full clause by means of recovery of ellipted elements from the context of the discourse or situation" (p. 366). "*A subordinate clause* will consist minimally of finite or nonfinite Verb element plus at least one other clause element (Subject, Object, Complement, or Adverbial)" (p. 366). These definitions were adopted in this study in the initial segmentation of the data into analyzable units.

With respect to the application of this framework to the current data, Foster et al. also created a principled way to exclude unhelpful data or data that would distort any analysis.[6] This method has three levels: Level 1, Level 2, and Level 3. The analysis of these data used their Level 2, which they state, is "to be used for highly interactional data. This is for researchers who are working with interactional data which can yield a high proportion of minimal units (e.g., one-word minor utterances and echoic responses) whose inclusion in an analysis could distort the perception of the nature of the performance" (p. 370). In other words, Level 1 is to include "everything except untranscribable data, although single inaudible words of identifiable word class should be included" (p. 370), and Level 2 incorporates all of Level 1, excluding "one-word minor utterances" such as "sí," "uhhuh," "claro," or other backchanneling utterances, echoic

[6] Please note that all words or phrases in English were left out of the analysis.

responses, such as a repetition of a word given to the learner by the instructor or repetition verbatim of a partner's utterance or self-repetition (p. 370). This approach to the segmentation of the data obtained was able to account for the highly interactive nature of the linguistic production by the participants as well as the monologic tasks. Next, the different measures of linguistic complexity and accuracy will be described.

Linguistic Complexity

Given the previously discussed variation in ways of operationalizing linguistic complexity and accuracy, Norris and Ortega (2009) make a call for a more "organic and sustainable" approach to the study of CAF, stating that,

> measurement practices in relation to CAF must become considerably more organic, in the sense that they need to capture the fully integrated ecology of CAF development in specific learning contexts over time, so as to help us understand how and why language develops or not within them. (p. 556)

They further call for "sustainable" measurement practices that allow researchers to understand how other research fits into a global understanding of the study of CAF (p. 556). Norris and Ortega go on to use syntactic complexity as an example of how research can move forward in such a way with recommendations for measuring syntactic complexity. The most important takeaway message from Norris and Ortega is the importance of using more than one measure of syntactic complexity in order to capture the nuances of interlanguage development. Specifically, they recommend using measures of coordination in beginning level language learners, subordination in intermediate level learners, and phrasal-level complexification for advanced-level learners (p. 563). They also recommend using a global measure with a more fine-grained, specific measure of syntactic complexity in order to capture both large-scale or long-term changes that a fine-

grained measure may not be able to capture (p. 568). With their recommendations in mind, the data were coded for number of clauses per AS-unit as the global syntactic complexity measure, number of words per clause as a global phrasal complexity measure, and then, as a fine-grained complexity measure, all verbal structures were coded for type, that is, the number of verb forms, including forms distinguished by tense or mood, out of the total number of verbs used. For example, if a student used the simple present, the present perfect, and the preterit in one utterance, that utterance would have a verbal complexity score of 3. The reason for the inclusion of this final measure of linguistic complexity is to observe how the learners develop their repertoire of verbal structures; an increase in the number of different verb forms indicates an increase in verbal complexity. As Norris and Ortega stated,

> not all traits of CAF will have an equally predictive value for all proficiency levels. Development is a long-term and multifaceted process, and data must be interpreted within awareness of where along that process the evidence is being collected and analyzed. CAF, it seems, consists of a variety of dynamically related indices which do not all advance hand-in-hand towards an ideally complex, accurate, and fluent performance. Indeed, depending on the proficiency or developmental levels of learners, if we focus only on anticipated changes in one area, we may be missing the really important changes (or lack thereof) going on in another. (p. 573)

Students are also rated for proficiency based on ability to express themselves in real vs. hypothetical situations and their ability to express time frames other than the present (http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/english/speaking), which are often expressed with the tense/mood aspect system of verbal structures. According to Ågren, Granfeldt, and Schlyter (2012), who also used number or variety of verb forms in their study, a learner who uses a wide variety of verbal morphology has a much more complex and developed

tense/aspect/mood system than a learner who, for example, uses only the simple present and present perfect to express all events in the present, past, and future (p. 100). These three measures were deemed appropriate to determine the learners' linguistic complexity and to measure its development. Additionally, these measures are not novel in the literature of CAF. Bulté and Housen (2012) created an outline of many of the ways in which linguistic complexity and accuracy have been measured in the literature, with all of these measures appearing in a minimum of three previous studies. Their list is not exhaustive by any means, meaning that there are likely many more studies that use the same measurements.

Linguistic Accuracy

With regard to linguistic accuracy, the global measure of accuracy is the number of errors per 100 words, and the specific measure is the "correctness" of the verbal structures. An error at the global level would be any error that violates the guides set forth in instructional grammar books for Spanish. For example, a noun-adjective agreement error would be one error. A pronunciation error would count as one error. The use of the wrong gendered article with a noun would also be considered an error. Previous research has used number of error-free clauses as a measure of global accuracy, but according to Kuiken and Vedder (2012), it would be very difficult to find error-free clauses in the speech of an intermediate language learner. Therefore, number of errors per 100 words was determined to be an acceptable measurement of global errors in the intermediate learners in the present study. With respect to verb phrases, accuracy was measured as a percentage of correct verbal structures used of the total number of verbal

structures used in a particular task or recording. A verb was considered correct if it was conjugated correctly in person, tense, aspect and mood. Any verb that was lacking in one of those items was coded as incorrect, and any verb that was not an appropriate word choice (i.e., *ser* vs. *estar*) was also coded as incorrect.

Task Categorization

As previously described in Chapter 2, much of the previous research on task-based performance and learning involved the use of researcher-designed tasks that manipulate certain variables in order to determine their effects on L2 production.[7] This study, in contrast, examined the effects of teacher-designed tasks on the production of the participants' oral Spanish. As such, a post-hoc categorization of the types of tasks that were present in the data was needed and will be explicated in this section. Any activity that fell under Ellis's (2003) generalized definition of task, explained in Chapter 2, was categorized for data analysis.

This categorization was made more difficult by the two different learning contexts in which the students participated, but there were commonalities across the two contexts allowing for categories that accounted for the tasks that appeared in both contexts. In general, there were two main types of interaction in each context: teacher-fronted question and answer-type activities and student–student small group work. The number of people in the small groups varied from dyads to groups of 4 or 5 students, occasionally reaching as many as 6 or 7, but groups of 2 to 4 students were the most common.

[7] See previous discussion on Robinson's Cognition Hypothesis and Triadic Componential Framework for task categorization.

Therefore, data were coded as being teacher-fronted or student groups. Because the teacher-fronted activities included the entire class, it follows that there was less opportunity for student production in these contexts, and that was borne out in this data set. The most production by the 3 participants happened when they were in small groups.

Another commonality between the two contexts was a task-based sequence of activities that combined teacher-fronted activities with small group, student-centered activities that often progressed in difficulty and expectations of student involvement. The teacher-fronted activities usually sandwiched the small group activities in order to first "prime" the students and then to share the results of their small group work. However, this was not characteristic of all teacher-fronted activities. In cases where a sequence is clearly related, e.g., an instructor was "priming" students with a pre-activity warm-up and closed with a large-group share of the results of the small-group discussion/task, those tasks were considered together for the purposes of evaluation of the production of the participants. For example, Mike participated in an activity that started with a teacher-fronted activity wherein the students were presented with questions on a PowerPoint. The instructor asked these questions to the class, and the class responded with short answers from a reading in the textbook. The students then were asked to complete a multiple choice comprehension activity based on the same reading, and then the final step was a sharing of the answers as a class, which was also teacher fronted. These tasks did not fall within the prototypical *think-pair-share* model, described below, because during the priming portion of an activity, the students were asked either simple yes/no questions or one-word response questions with no requirement to elaborate in order to "warm up" to

the, often open-ended, questions, different from the question presented in the "warm up," that they would discuss in their pairs or small groups. The large group share at the end was based on the small group work, and not on the "warm-up" activities that preceded them. Additionally, these priming, "yes/no" questions rarely allowed much language production and could not be analyzed on their own. If the students did not produce any language in a teacher-fronted priming activity, the teacher-fronted portion was left out of the analysis as the focus of this study is on learner production. The ultimate categorization of these tasks followed the description in Table 1, which gives a snapshot of the different categories used in this analysis.

Table 1

Task Categorizations Used in Coding Task Type

Task name	Source	Definition
Think-pair-share	Shrum and Glisan (2009)	"(1) listen while the teacher poses a question; (2) are given time to think of a response; (3) are told to pair with a classmate and discuss their responses; and (4) share their responses with the whole group" (p. 268).
Jigsaw	Shrum and Glisan (2009)	A two-way activity in which each member of the group contributes equally to an end-goal.
Information-gap	Shrum and Glisan (2009)	A one or two-way activity in which there is only one student who has the information needed to complete a task. The other members of the group need to obtain this information in order to complete the task.
Role-plays	Shrum and Glisan (2009)	Students must act out a scene, such as a visit to the doctor, in the TL.
Sharing personal experiences: storytelling, anecdotes, and reminiscences	Willis and Willis (2007)	Students relate personal experiences.
Discussion tasks	Willis and Willis (2007)	Students share opinions, debate, narrate, describe and explain.
Ordering and sorting	Willis and Willis (2007)	Sequencing, ranking, and classifying items
Matching	Willis and Willis (2007)	For example, matching words and phrases to pictures or directions to a map
Comparing	Willis and Willis (2007)	Finding differences or similarities between two items.
Listing	Willis and Willis (2007)	Brainstorming or fact-finding, for example.
Metalinguistic/LRE	Ellis (2003), Mackey (2012)	Any language that deals specifically with language itself.

In order to categorize the tasks in the data, a combination of frameworks for task categorization was used with the present work's data. Shrum and Glisan (2009) outline the different types of effective foreign language pedagogy tasks that are recommended for use in the FL classroom (pp. 268–275), while Willis and Willis (2007) outline a taxonomy of task types (p. 108). These frameworks were deemed adequate, in part, because they are taught to and used by foreign language instructors, as well as being complimentary in nature. More importantly, they accounted for the data in this study. The instructors in the study did not follow one or the other, but rather a combination of different types of activities that are best served by this combined framework. This categorization is also useful in the analysis of the data in that it can be replicated in classrooms in other contexts since these types of tasks are commonly used by FL instructors. Laboratory-based conditions are often artificial and highly manipulated and do not reflect the way that many FL classrooms actually operate.

Shrum and Glisan's (2009) task description will be outlined first. *Think-pair-share* is described as a task wherein students "(1) listen while the teacher poses a question; (2) are given time to think of a response; (3) are told to pair with a classmate and discuss their responses; and (4) share their responses with the whole group" (p. 268). Rebecca and Teresa engaged in a classic *think-pair-share* activity wherein their instructor asked them to think about what they would do if Canada invaded the United States and they had to either defect from the United States and claim Canadian citizenship or move to one of the remaining United States territories to keep their U.S. citizenship. The

students then discussed this hypothetical in small groups and then shared the results of their discussion with the entire class.

Jigsaw activities (Shrum & Glisan, 2009) are those in which each member of the group assumes responsibility for a given portion of the lesson, becoming the "expert" in the group on that particular portion of the lesson. The "experts" must share their information with other members of the group, with the goal being that each student learns from the other "experts" on each section, creating a whole lesson from each expert's parts (pp. 268–269). Again, this was more common in College B, with an example being that the students were given 1 of 4 different articles to read. The next class period, they joined with people who read the articles they did not read, and presented their article's information to their group. The group then created a summary of all four articles and shared their results with the class.

An *information-gap* activity (Shrum & Glisan, 2009) is one in which one student has more information than her classmates, and the classmates who do not have that information need it in order to complete some task. The students who do not have the information ask questions of the student who does until they all have enough information to complete the task (pp. 269–270). The information-gap activity differs in one important way from the jigsaw task: in the jigsaw task, all of the participants have an equal amount of the information needed to complete the task, and it is a two-way task. Even though jigsaw tasks are a subcategory of information-gap tasks, the two are differentiated here by whether they are a one-way or two-way task. As mentioned, in an information-gap task, only one student has enough information to finish a task, and it is often one-way in nature

(Ellis, 2003; Mackey, 2012; Pica, Kanagy, & Falodun, 1993; Shrum & Glisan, 2009). An example of this type of activity from the current data was an activity from College B in which students, in groups of four, were given one postcard. The student with the postcard was directed to describe the postcard to her classmates while they attempted to draw it. The classmates that were drawing could ask questions of the postcard holder in order to fine-tune their knowledge of the scene being described.

Finally, Shrum and Glisan also outlined *role-plays*, where students act out situations, such as a restaurant scene or a visit to a doctor. They note that it is important to display the role-play situation in the native language of the learners (p. 274). University A uses this type of activity when doing paired oral assessments, in which the students were given a role, either question-asker or question-answerer, and were directed to play that role while their partners played the other role. The students then switched roles and completed the role-play again in the opposite role.

Willis and Willis (2007) outline a taxonomy of tasks, many of which were used in the classrooms under study (p. 108). The first is *problem solving*, which requires students to offer advice and recommendations on problems ranging from general issues such as global warming to more specific problems such as what to do if your neighbor's dog is digging up your garden. According to Willis and Willis, these activities serve as the basis for writing activities such as note taking, drafting, and finalizing proposals for solutions (p. 93). This type of task was not captured in the data set.

Second, they offer *sharing personal experiences: storytelling, anecdotes, and reminiscences*. These activities include things such as recreating a familiar story or

sharing their own experiences, often making them entertaining and dramatic. This happened more often at University A, with an example being the activity in which Mike described his composition, which was a narration of a notable event in his past, to a group of fellow students. He was not allowed to read off his composition, so this was a basic retelling of a past event.

Next, they offer *discussion tasks*, in which students share opinions, debate, narrate, describe and explain. For example, students may share an opinion of a photograph or a text. University A used this type of activity, for example, with the formal debate at the end of the spring semester. College B used this type of activity, for example, in their oral assessments when students were to give their opinions on the novel they had read. Discussion tasks were very frequent in both contexts.

Willis and Willis also outline *ordering and sorting* tasks (sequencing, ranking, and classifying items), *matching* (e.g., words and phrases to pictures or directions to a map), *comparing* (finding differences or similarities), and *listing* (brainstorming or fact-finding, for example; p. 108). An example of a listing activity occurred in College B's class, where Rebecca and Teresa were asked to find information in the reading selection that would allow them to fill out a worksheet with the names of characters, places or events that had influence on the main character.

Lastly, any task that was form based and/or dealt with language explicitly was termed a metalinguistic task,[8] and was categorized as such. For example, Mike completed

[8] Frequently referred to in the literature as LREs, or Language Related Episodes (Ellis, 2003; Mackey, 2012)

an activity in which his class was given subjects, and he and his classmates were to create sentences with the passive voice. Another example of a metalinguistic task occurred in Teresa and Rebecca's classroom, in which they were to conjugate the given verbs in contextualized sentences to the correct form of the imperfect indicative or imperfect subjunctive, according to the context. Any discussion that occurred was directly related to which verb form they were supposed to use to make the sentence complete. Metalinguistic tasks do fall within the definition of a task as outlined by Ellis (2003), since the metalinguistic tasks in this data set did result in reasoning about what forms to be used as well as a discussion among the students as to the meaning of the sentence and how it constrains what form(s) could be used. To use a previous example, in College B, Teresa and Rebecca were engaged in a task that had the goal of completing a worksheet on the use of present subjunctive vs. imperfect (past) subjunctive. The students, while they did not form opinions or discuss content, were discussing which form should be used in each blank, engaging in a negotiation of meaning of the context and whether they thought the form should be in the present or the imperfect. These are certainly "real world processes of language use" (Ellis, 2003, p. 10), and thus fall under the definition of "task" given by Ellis, allowing them to be included in the analysis.

There were instances in which more than one category could be applied to a task. In those cases, the more elaborate or overarching category was the one selected for that task. For example, in one class session, Mike participated in an activity in which he was to discuss the United Farm Workers' logo, but within that activity was an embedded communicative form-based activity using the construction, "lo que" (that which, what) to

make sentences that expressed what the students liked best or least about the image (e.g., "lo que me gusta más de la imagen es…" "*What I like the best about this image is…"*). However, the overarching goal of the activity was to discuss the logo, to describe it, and to give opinions about it, so this activity was classified as a "discussion" activity according to the discussion task outlined by Willis and Willis (2007).

In order to determine reliability in task categorization, 21% of the tasks were randomly selected in order to calculate interrater reliability. These tasks were coded independently by the researcher as well as a professor of Foreign Language Education. Both the researcher and the professor of Foreign Language Education used the task categorization description presented in this chapter to in order to guide the classification of the tasks that were extracted to make the interrater reliability calculation. Simple percentage agreement was used to calculate agreement, following Polat and Kim (2014). The resultant agreement between the coders was 95% for task categorization, and all disagreements in categorization were resolved through discussion.

Now that the constructs under analyses and the coding structure have been outlined, the statistical analyses used to investigate the data for each of the research questions will be outlined.

Statistical Analyses

Because this was a case study, there were some limitations as to how the data collected could be analyzed and interpreted. These limitations will be taken into account in the following section that includes a description of the variables used, and the rationale for the statistical analyses used.

Variables

For each of the 3 participants, Mike, Rebecca, and Teresa, there are two independent variables: task type and points in time. These data do not have longitudinal date intervals as would be found in a traditional longitudinal study. That is, time 1 is not the same for all participants, time 2 is not the same for all participants, etc., so points in time were used here for the analysis of the longitudinal data, but these are not regularized, common time points across the academic year. This type of measurement is a monotonic time trend, which is a linear time estimate that tracks, chronologically, the nature of the longitudinal data collection (Ritchey, 2008). Task type will include the categories outlined in the previous section.

The following dependent variables were also analyzed for each participant: variety of verbal structures, average number of clauses per AS-unit, average number of words per clause, number of errors per 100 words, and percent correct verbal structures per total number of verbal structures. These variables and constructs were also described in the previous section. These variables were used in the analyses to answer the following research questions.

RQ 1: Does task type affect oral linguistic complexity and accuracy of learner language?

An ANOVA was run for each participant with task type as the independent variable and verb complexity, phrasal complexity, syntactic complexity and verbal accuracy serving as the dependent variables. The individual ANOVAs were conducted as a way to be comprehensive. That is, the individual ANOVAs were computed in order to

ascertain granular effects at the individual level with respect to the variables in question. Given that the data did not have sufficient variability to support multilevel modeling at the individual level, one-way ANOVAs are the method of choice for investigating granular effects within the data.

Post-hoc tests were then run in order to reveal details on any statistically significant relationships. The Tukey post-hoc test was run for homoscedastic data (as evidenced by a statistically nonsignificant Levene's test of equality of variance), while the Games-Howell post-hoc test was run for heteroscedastic data (as evidenced by a statistically significant Levene's test of equality of variance). The ANOVA will determine whether there is any statistically significant difference between the different task types and show trends that are present in the data. However, ANOVAs do not reveal the granular details of these differences, and therefore, post-hoc statistical measures are needed to provide the more detailed information about the size and direction of the differences between the task types as well as to help explain the graphs and tables that may show differences.

RQ 1a: Does time interact with task type to affect oral linguistic complexity and accuracy of learner language?

Because these data are correlated, however, a more robust method of analysis needed to be employed to test RQ 1a. To that end, multilevel mixed-effects modeling was performed on the aggregated data in order to test whether there was any effect of time on the dependent variables within task type. These data could not be broken down into

individual analyses by participant because the variation by participant was very small, and the number of participants was also small. Chapter four will outline these results.

RQ 2: Does oral linguistic complexity and accuracy exhibit change in trajectory over time?

A Pearson's correlation with points in time as the independent variable and each of the dependent variables was performed. A Pearson's correlation was determined to be the appropriate test because all of the variables in question are continuous variables, and this test will show the relationship between continuous variables over time (Ritchey, 2008). Pearson correlations were computed as a way to ascertain the bivariate relationships between time and the five key variables for all three respondents. The correlational analyses were conducted at the granular level for the three participants to see if the bivariate relationships between time and the five key variables were present on an individual participant basis.

Descriptive statistics, results from the ANOVAs, post-hoc tests, multilevel modeling, Pearson's correlation, and visual representations of the data in the form of graphs and tables are produced and explained in Chapter 4.

Chapter 4

Results

This chapter will outline the results of the statistical analyses that were carried out in answering the research questions that guided this research. The data will be presented in the forms of tables and figures, with a brief explanation of the results. Full discussion of the results is found in Chapter 5.

Participants

This data set constitutes a case study of 3 participants: P1 = Mike, P2 = Rebecca, and P3 = Teresa. Detailed descriptions of each participant can be found in Chapter 3. Mike was a student at a large, urban, public university, while Rebecca and Teresa were students at a smaller, rural liberal arts college. Both institutions were situated in the Midwest area of the United States.

Independent Variables

There were two independent variables under consideration in this research: task type and time. The tasks were categorized into eleven categories, described in detail in Chapter 3, which were *think-pair-share*, *information gap*, *jigsaw*, *discussion*, *metalinguistic*, *listing*, *ordering and sorting*, *personal experiences*, *role play*, *comparing*, and *matching*. These categories were developed based on a hybrid of two frameworks presented by both Shrum and Glisan (2009) and Willis and Willis (2007). The hybrid of the two categorizations was used because of the need for a post-hoc categorization of task types due to the researcher being uninvolved in the development of the lessons in which

she was a part and also because these frameworks are commonly taught to and used by instructors of language.

Upon initial investigation of the data, it was determined that some of the categories had too few data points to analyze separately in the statistical analyses. Table 2 presents the descriptive statistics that show the categories with too few data points to analyze. As can be seen in Table 2, the following categories occurred six times or less in the data from all participants: Thinkpairshare; Infogap; Jigsaw; Orderinsorting; Personalexperiences; Roleplay; Comparing; Matching.

Table 2

Percentages and Frequencies for All Categorical Variables

Variable	Frequency	%
Task type		
Thinkpairshare	2	2.1
Infogap	1	1.1
Jigsaw	5	5.3
Discussion	46	48.9
Metalinguistic	11	11.7
Listing	12	12.8
Orderingsorting	5	5.3
Personalexperiences	3	3.2
Roleplay	2	2.1
Comparing	6	6.4
Matching	1	1.1

Based on these distributions, it was therefore determined that the categories with the fewest number of data points would need to be collapsed to make the statistical analysis viable in order to be able to run the tests that will be presented in subsequent sections. The categories *think-pair-share*, *jigsaw*, *personal experience*, and *role play* were grouped together in the category renamed *thinkplay*, because these tasks all seemed to be, based on review of the data, more open-ended in nature. For example, *personal experience* does not necessarily entail arriving at a "correct" answer, but would include things such as narrating a past event, as Mike did when recounting his first experience riding a roller coaster at a local theme park. The *think-pair-share* example that was obtained from this data set was a prototypical example where the instructor asked the students if Canada and the U.S. were in a war and Canada won, if they would stay in their state and learn French or if they would move to another state that was still under control of the U.S. The students discussed this question in small groups and then shared their answers with the class as a whole.[9] For examples of *role play* and *jigsaw* tasks, see excerpts (1) and (2) below, which are excerpts from Mike and Rebecca, respectively.

(1) Students are discussing the use of 5-Hour Energy drinks

Male student: Uh, los corazo-corazones son muy más rápido? con 5-hour ENERGY? (makes sound of rapid heartbeat)
Mike: //Sí. Sí.
Male student: //Sí.
Male student: Yeah. No-no es bueno.
Male student: No. Um:

[9] These data are not provided due to the overwhelming amount of identifying information contained within this task.

Mike: Y-Yo pienso que tengo, um, tienes, tienes que necesitan, uh, tomar muchas, um, 5-hour, ENERGY, uh, bebidos para, para, um, um, mor-mortar?[10] muerte? Sí?
Male student: //Sí.
Male student: //Sí. Um? no sé. Um, probablemente. (laugh) espero que: necesitan //muchos 5-hour para mortar,

Male student: Uh, the hear-hearts are very much faster? With 5-hour ENERGY? (makes sound of rapid heartbeat)
Mike: //Yes. Yes.
Male student: //Yes.
Male student: Yeah. It's not–it's not good.
Male student: No. Um:
Mike: I–I think that I have, um, you have, you have to they need, uh, to drink many, um, 5-hour, ENERGY, uh, drinks in order to, in order to, um, um, [to di-to die]?[11] Death? Yes?
Male student: //yes.
Male student: //Yes. Um? I don't know. Um, probably. (laugh) I hope that they need //many 5-hour to die,

(2) Each student in Rebecca's class looked up a Spanish language news story, and was to share this story in groups of 3–4 students.

Rebecca: mhm. Uh:, escribo:, uh, inventores de nuevos tipos de instrumentos zapatos y ropa explica sobre la tecnología de 3D systems y:, tres-D system, y, que ellos usan para:, hes-hacer sus, e-invenciones. Sí.
Female student: Um:, y quién or quiénes las personas más importantes? Um, yo creo que es que a un compositor?
Rebecca: Um-hmm.
Female student: Es importante, um, que: su banda es: el primer, primera banda a usar los instrumentos de://
Rebecca: //Um-hmm.

Rebecca: mhm. Uh:, I write, uh, inventors of new types of shoe instruments and clothes explains about the technology of 3D sysems and, three-D system, and, that they use for, hes-make their, e-inventions. Yes.

[10] The correct verb for "to die" is *morir*.

Female student: Um, and who or who(plural) the people most important? Um, I think that it is that to a composer?
Rebecca: Um-hmm.
Female student: It is important, um, that their band is the first, first band to use the instruments of//
Rebecca: //Um-hmm.

Contrast these examples with the more closed and structured activities that were grouped in the *logic* category, which included *information gap, ordering and sorting, comparing,* and *matching* tasks. These tasks tended not to elicit longer discourse in Spanish from the students, especially in tasks in which the goal was to match items together, such as choosing the correct phrase that is associated with some numbered item, such as a name or date. The data in this category tended to be very brief. See, for example, *ordering and sorting* in (3) from Mike, *comparing* in (4) from Mike, *matching* in (5) from Mike, and *information gap* (6) from Rebecca.[12]

(3) Mike is participating in an activity that asks students to rate listed items in the textbook.

Mike: Okay. Uh, los derechos y libertades fundamentes? Segundo. Segunda. El uso de los avances, Tercero. Tercera. La libertad de expresión, la libertad. Primera. La se-guridad social. Igualidad? Sí. Igual, uh, segunda.
Male student: Segunda?
Mike: Sí. Uh, po-porque es social. Uh, la (correzación)[13] del medio ambiente?
Male student: Uh

Mike: Okay. Uh, the fundamental rights and freedoms? Second. Second. The use of advances, third. Third. The freedom of expression, the

[12] It is important to note that the person who is relaying information in an information gap activity will create longer utterances, but in this data set, there was only one instance of an information gap activity, and my participants did not take the information-relaying role.

[13] I was unable to determine what word he intended, so this was left as-is in the translation below.

freedom. First. The social se-curity. Equality? Yes. Equal, uh, second.
Male student: Second?
Mike: Yes. Uh, be-because it is social. Uh, the (correzación) of the environment?
Male student: Uh

(4) Mike and his classmates were given a prompt on a PowerPoint slide, and were to use comparisons to answer the given questions.

Mike: All right. Chris Christians es más alto que mí.
Male student: Alto que yo.
Mike: Then, oh, it could be.
Male student: Right. Um, Quién es el más alto que todos?
Mike: Ooh. Kareem Abdul Jabbar? Posiblemente? Más alto de todos.

Mike: All right. Chris Christians is taller than me.
Male student: Taller than I.
Mike: Then, oh, it could be.
Male student: Right. Um, who is the tallest of everyone?
Mike: Ooh. Kareem Abdul Jabbar? Possibly? Taller than everyone.

(5) Mike and classmates are doing a matching activity from their textbook.

Mike: Okay. Uh, Cuándo se fundó?
Male student: Um? Cuándo se fundó (unintelligible) [mumbling] B.
Mike: Sí.
Male student: Sí.
Female student: No sé.
Male student: No sé. Which one is C. No sé.

Mike: Okay. Uh, when was it founded?
Male student: Um? When was it founded (unintelligible) [mumbling] B.
Mike: Yes.
Male student: Yes.
Female student: I don't know.
Male student: I don't know. Which one is C. I don't know.

(6) Rebecca and her classmates are attempting to draw a scene that appears on a postcard. One student in each group of four has a postcard, and the other members of the

group have to draw it according to the oral description given to them by the card holder.

The drawers may ask questions.

> Female student: El calle tiene una, uh, hotel que existen en la, la izquierda. Aquí está en la izquierda. En el frente de la foto, la calle está a la izquierda.
> Rebecca: Enfrente, en el primer plano?
> Female student: En el fondo.
> Rebecca: Primer fondo?
> Female student: Primer..
> Rebecca: Plano.
>
> *Female student: The street has a, uh, hotel that they exist en the, the left. Here it is on the left. In the front of the photo, the street is on the left.*
> *Rebecca: In front, in the first plane?*
> *Female student: In the background.*
> *Rebecca: First background?*
> *Female student: First..*
> *Rebecca: Plane.*

Once the combination of categories was completed, percentages and frequencies were recalculated for all participants. These data are presented below in Table 3.

Table 3

Percentages and Frequencies for All Categorical Variables

Variable	Frequency	%
Task type		
Discussion	46	48.9
Metalinguistic	11	11.7
Listing	12	12.8
Thinkplay	12	12.8
Logic	13	13.8

The second independent variable was points in time. This variable is a continuous, monotonic variable, but it is important to note that each participant had a different number of points in time in which data were collected. This was partly because the three students were in residence at two different institutions, but also because the two students, Rebecca and Teresa, at the liberal arts college did not always attend class on the same days. In particular, Teresa has fewer data points for time because she had many more absences than Rebecca, resulting in both fewer points in time and overall less data than either Rebecca or Mike.

Dependent Variables

This study included five dependent variables: verb complexity, or variety of verbal structures used; a measure of syntactic/global complexity measured by the average number of clauses per AS unit; phrasal complexity, measured as the average number of words per clause; verbal accuracy as expressed by the percent of correctly used verbal structures per total number of verbs used; and finally, number of errors per 100 words uttered. All but the last were analyzed to answer research question 1. This was because each task may not have included 100 words, making it impossible to use this measure in the determination of task effects. Therefore, number of errors per 100 words was only used for research question 2.

Descriptive Statistics

This section will present the descriptive statistics for all the variables under analysis in this research.

Table 4

Means and Standard Deviations for all Continuous Variables, All Participants

	Variety of verbal structures	Average number of clauses per AS unit	Average number of words per clause	Number of errors per 100 words	Ratio correct verbal structures per total number of verbal structures
Mike					
Mean	4.32	1.30	4.22	13.52	0.87
Standard Deviation	3.16	0.40	1.73	7.40	0.12
Rebecca					
Mean	3.76	1.25	3.60	13.76	0.80
Standard Deviation	2.26	0.22	1.47	5.80	0.16
Teresa					
Mean	5.36	1.26	3.77	12.27	0.78
Standard Deviation	3.38	0.29	1.33	7.12	0.16
All three participants					
Mean	4.38	1.27	3.84	13.27	0.82
Standard Deviation	2.94	0.31	1.53	6.73	0.15

Table 4 presents the means and standard deviations for all continuous variables within the dataset, both for each participant as well as showing the totals for all three. Each variable represents the average number of times a particular event occurred for all 3 participants. For example, the total average variety of verbal structures was 4.38, whereas the average number of clauses per AS unit was 1.27. The average number of words per

clause for all 3 participants was 3.84, while the average number of errors per 100 words was 13.27. Finally, the mean of the variable that measures percent correct verbal structures per total number of verbal structures can be interpreted as a percentage. In other words, the average percent correct was 82 percent.

It should also be noted that, with respect to time points, there was a total number of 28 unique time points in which data were collected during one academic year. This variable was measured as a continuous monotonic time trend that represents the 28 unique time points.

Table 5 presents the frequencies and percentages for the categorical variables used in the current investigation, both as totals for all three participants as well as individually. As can be seen, there is a roughly even distribution among 4 of the 5 categories of task type. The category of discussion is overrepresented within the data.

Table 5

Percentages and Frequencies for all Categorical Variables, All Participants

	Discussion	Metalinguistic	Listing	Thinkplay	Logic
Mike					
Frequency	15	2	2	5	8
%	46.9	6.3	6.3	15.6	25
Rebecca					
Frequency	17	6	6	4	4
%	45.9	16.2	16.2	10.8	10.8
Teresa					
Frequency	14	3	4	3	1
%	56.0	12.0	16.0	12.0	4.0
All three participants					
Frequency	46	11	12	12	13
%	48.9	11.7	12.8	12.8	13.8

The following section will discuss the results of the ANOVA and post-hoc tests run in the investigation of research question 1.

Research Question 1: ANOVA Results

The results of the ANOVAs run on the data each participant individually with their respective post-hoc tests will be presented in this section.

Table 6

One-way Analysis of Variance (ANOVA), Mike

	Discussion		Metalinguistic		Listing		Thinkplay		Logic		
	M	SD	M	SD	M	SD	M	SD	M	SD	*F*
Variety of verbal structures	5.14	3.18	4.00	1.41	1.50	0.71	6.80	3.19	1.43	0.54	4.054**
Average number of clauses per AS unit	1.40	0.53	1.31	0.15	1.29	0.30	1.35	0.35	1.07	0.09	0.877
Average number of words per clause	4.67	1.64	5.75	3.14	2.63	0.88	4.74	1.42	2.90	1.15	2.891*
Percent correct verbal structures per total number of verbal structures	0.87	0.12	0.74	0.17	0.90	0.14	0.78	0.12	0.98	0.05	3.268*

Note: *=p<.05, **=p<.01, ***=p<.001, two-tailed tests.

Table 6 presents the results of the one-way ANOVA for Mike. It should be noted that Levene's test for homogeneity of variance was statistically significant for variety of verbal structures ($F = 3.863$; $df = 4, 86$; $p = .014$). As such, the Games-Howell approach was used for decomposition of effects in the case of variety of verbal structures.

Table 6 shows three statistically significant differences within the data for Mike. Mean scores for variety of verbal structures ($F = 4.054$; $df = 4, 86$; $p = .011$), average number of words per clause ($F = 2.891$; $df = 4, 86$; $p = .041$) and percent correct verbal structures per total number of verbal structures ($F = 3.286$; $df = 4, 86$; $p = .027$) vary as a function of the independent variable.

Decomposition of effects via the Games-Howell post-hoc test shows that with respect to variety of verbal structures, the score for discussion (M = 5.14) is significantly higher than the score for listing (M = 1.50) and logic (M = 1.43).

Decomposition of effects via the Tukey's HSD post-hoc test shows that with respect to average number of words per clause, the significant omnibus *F*-value is actually a statistical fluke. That is, even though the ANOVA showed statistical significance, when the Tukey's HSD post-hoc test was run, no significant differences were detected within the data. As previously mentioned, an ANOVA test, when there are more than three categories, will occasionally show significant differences if there is any significance between the categories. Occasionally, this significant relationship is not borne out in the results of the post-hoc tests because the decomposition of effects reveals that the relationship is not significant. As such, the significant *F*-value for average number of words per clause should be disregarded.

Decomposition of effects via the Tukey's HSD post-hoc test shows that with respect to percent correct verbal structures per total number of verbal structures, the score for logic (M = 0.98) was significantly higher than the score for thinkplay (M = 0.78).

Table 7

One-way Analysis of Variance (ANOVA), Rebecca

	Discussion		Metalinguistic		Listing		Thinkplay		Logic		
	M	SD	M	SD	M	SD	M	SD	M	SD	*F*
Variety of verbal structures	3.75	3.04	4.83	1.47	3.50	1.38	4.25	1.89	2.00	0.82	0.931
Average number of clauses per AS unit	1.35	0.22	1.09	0.11	1.08	0.03	1.43	0.35	1.20	0.17	3.847*
Average number of words per clause	4.36	1.73	2.23	0.64	3.09	0.74	4.51	1.12	3.16	0.70	3.642*
Percent correct verbal structures per total number of verbal structures	0.82	0.08	0.76	0.18	0.82	0.23	0.75	0.27	0.85	0.21	0.323

Note: *=p<.05, **=p<.01, ***=p<.001, two-tailed tests.

Table 7 presents the results of the one-way ANOVA for Rebecca. It should be noted that Levene's test for homogeneity of variance was statistically significant for average number of clauses per unit ($F = 4.694$; $df = 4, 86$; $p = .004$) and for percent correct verbal structures per total number of verbal structures ($F = 3.288$; $df = 4, 86$; $p = .023$). As such, the Games-Howell approach was used for decomposition of effects in the case of average number of clauses per unit and for percent correct verbal structures per total number of verbal structures.

Table 7 shows two statistically significant differences within the data. Mean scores for average number of clauses per AS unit ($F = 3.847$; $df = 4, 86$; $p = .012$) and average number of words per clause ($F = 3.642$; $df = 4, 86$; $p = .015$) vary as a function of the independent variable.

Decomposition of effects via the Games-Howell post-hoc test shows that with respect to average number of clauses per AS unit, the score for discussion (M = 1.35) is significantly higher than the score for metalinguistic (M = 1.09) and listing (M = 1.08).

Decomposition of effects via the Tukey's HSD post-hoc test shows that with respect to average number of words per clause, the score for discussion (M = 4.36) was significantly higher than the score for metalinguistic (M = 2.23).

Table 8

One-way Analysis of Variance (ANOVA), Teresa

	Discussion		Metalinguistic		Listing		Thinkplay		Logic		
	M	SD	M	SD	M	SD	M	SD	M	SD	*F*
Variety of verbal structures	4.79	3.53	8.67	3.06	3.00	2.31	5.67	0.58	3.00	.	1.576
Average number of clauses per AS unit	1.34	0.35	1.05	0.02	1.12	0.12	1.40	0.38	1.09	.	1.013
Average number of words per clause	3.84	0.99	2.56	0.62	2.95	1.00	3.93	1.49	4.18	.	0.238
Percent correct verbal structures per total number of verbal structures	0.80	0.14	0.74	0.13	0.85	0.18	0.80	0.07	0.29	.	3.465*

Note: *=p<.05, **=p<.01, ***=p<.001, two-tailed tests.

Table 8 presents the results of the one-way ANOVA for Teresa. It should be noted that Levene's test for homogeneity of variance was statistically significant for none of the four variables. As previously noted, Levene's test is only used to measure whether the variance of the dependent variables under examination can be considered heterogeneous or homogeneous in nature. According to Ritchey (2008), when heterogeneity of variance is present within the data (as indicated by a statistically significant Levene's test), the appropriate post-hoc test is the Games-Howell approach. In the case of homogeneity of data (as indicated by the statistically nonsignificant Levene's test results for the four ANOVA equations in Table 15), the Tukey's HSD post-hoc test is appropriate (Ritchey, 2008). As such, the Tukey's HSD approach would have been appropriate for decomposition of effects in Table 15 under normal circumstances with respect to the one statistically significant result associated with percent correct verbal structures per total number of verbal structures ($F = 3.465$; $df = 4, 29$; $p = .026$). However, because there was only one data point for the *logic* category, there is no variance and thus no standard deviation, and thus a decomposition of effects via post-hoc tests was not possible.

In summary once the data for individual participants were analyzed, the results differed. Mike's data showed that verbal complexity and verbal accuracy were significantly different by task category. Specifically, for Mike's variety of verbal structures, the score for discussion was significantly higher than the score for listing and logic. With respect to percent correct verbal structures per total number of verbal structures, Mike's score for logic was significantly higher than the score for thinkplay.

Rebecca's data only showed significant differences by task for average number of clauses per AS unit and average number of words per clause. As previously noted, with respect to average number of clauses per AS unit, Rebecca's score for discussion is significantly higher than the score for metalinguistic and listing. With respect to average number of words per clause, Rebecca's score for discussion was significantly higher than the score for metalinguistic. Finally, because Teresa's data could not be decomposed via post-hoc tests, no firm claims can be made about the effects of task on the variables under study for Teresa's data. In other words, task does appear to matter in the production of linguistic complexity and accuracy, but it matters in different ways for different participants.

Research Question 1a: Multilevel Modeling Results

In order to test whether there was any significant change within task type over time, a linear mixed effects model was used to analyze the data from the three participants. It was not possible to analyze the participants separately due to the small amount of data for each individual, but the variation due to participant was very small across all 4 analyses. The following tables and figures outline the results from this model for the variables verbal complexity, syntactic global complexity (number of clauses per AS-Unit), phrasal complexity (number of words per clause), and verbal accuracy both by task type and by time.

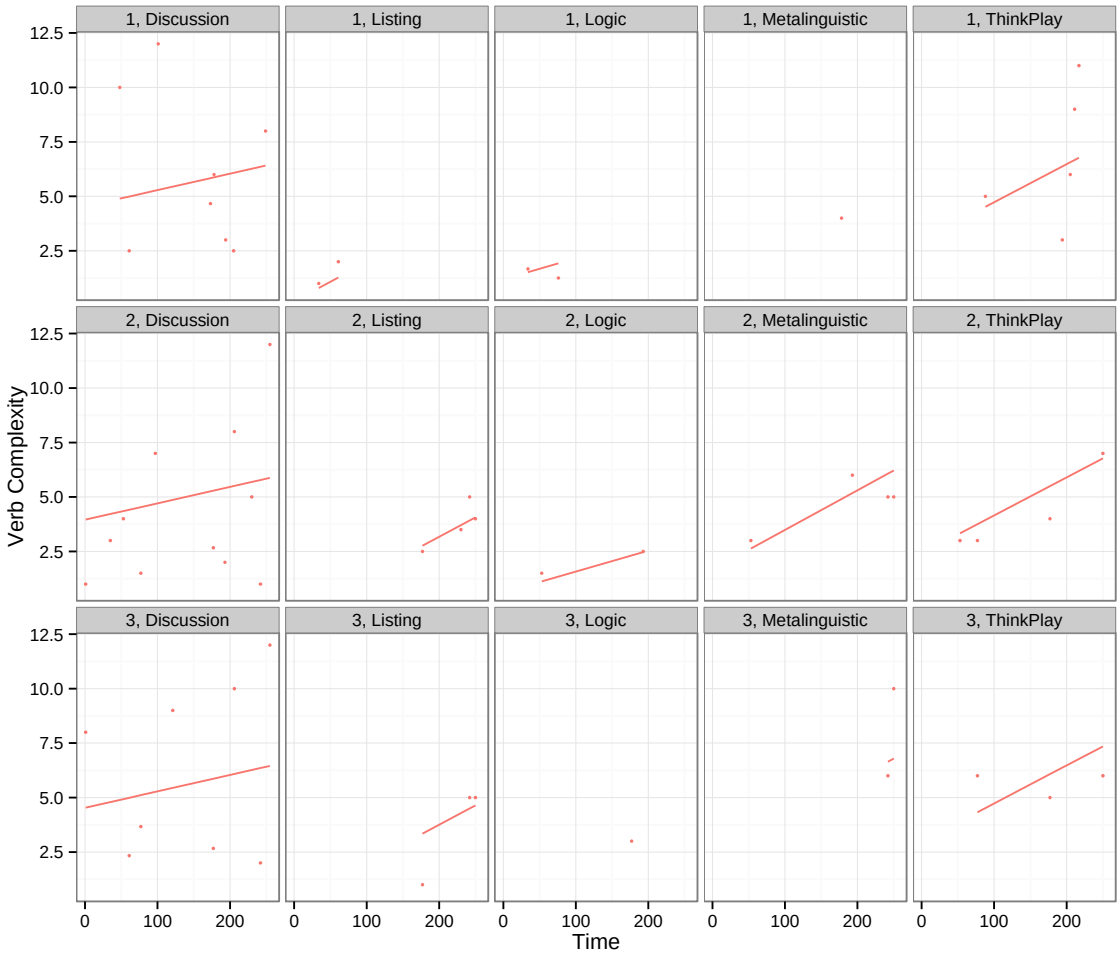

Figure 1. Fitted values from restricted maximum likelihood estimated maximum likelihood models: Verbal complexity.

Table 9

Maximum Likelihood Parameter Estimates for a Linear Mixed-Effects Model Describing Changes in Verb Complexity, All Three Participants

Parameter	Estimate (s.e.)	*t*-value	*p*-value
Intercept	4.31 (1.13)	3.82	<.001
Time interval	0.01 (0.01)	1.14	0.262
Listing	-4.19 (2.68)	-1.56	0.126
Logic	-3.28 (2.68)	-1.23	0.226
Metalinguistic	-2.46 (3.62)	-0.68	0.500
ThinkPlay	-1.61 (2.41)	-0.67	0.509
Time Interval * Listing	0.01 (0.01)	0.67	0.505
Time Interval * Logic	0.00 (0.02)	0.08	0.940
Time Interval * Metalinguistic	0.01 (0.02)	0.64	0.525
Time Interval * ThinkPlay	0.01 (0.01)	0.76	0.453
Variance Components	Chi Square	*df*	*p*-value
Time Interval	7.19	1	0.007
Task Type	11.75	4	0.019
Time Interval * Task Type	1.26	4	0.868

Table 9 indicates that both time interval (p=0.007) and task type (p=0.019) are significantly associated with verb complexity. There is an increasing trend toward greater verb complexity over time. In regards to task type, intercept differences indicate that Discussion and ThinkPlay both tended to start with higher levels of verb complexity on average, but there is an increase in verb complexity across all task types. The interaction

of time and task is not significant (p=0.868). The estimated variability contributed by participant is 0.538 and the estimated error is 2.82.which indicates that participant differences account for 3.5% of the variance. The likelihood ratio test versus a model with no random effect is not significant (p=0.775). This indicates a fixed effect model is statistically equivalent to the model with the random effect by participant.

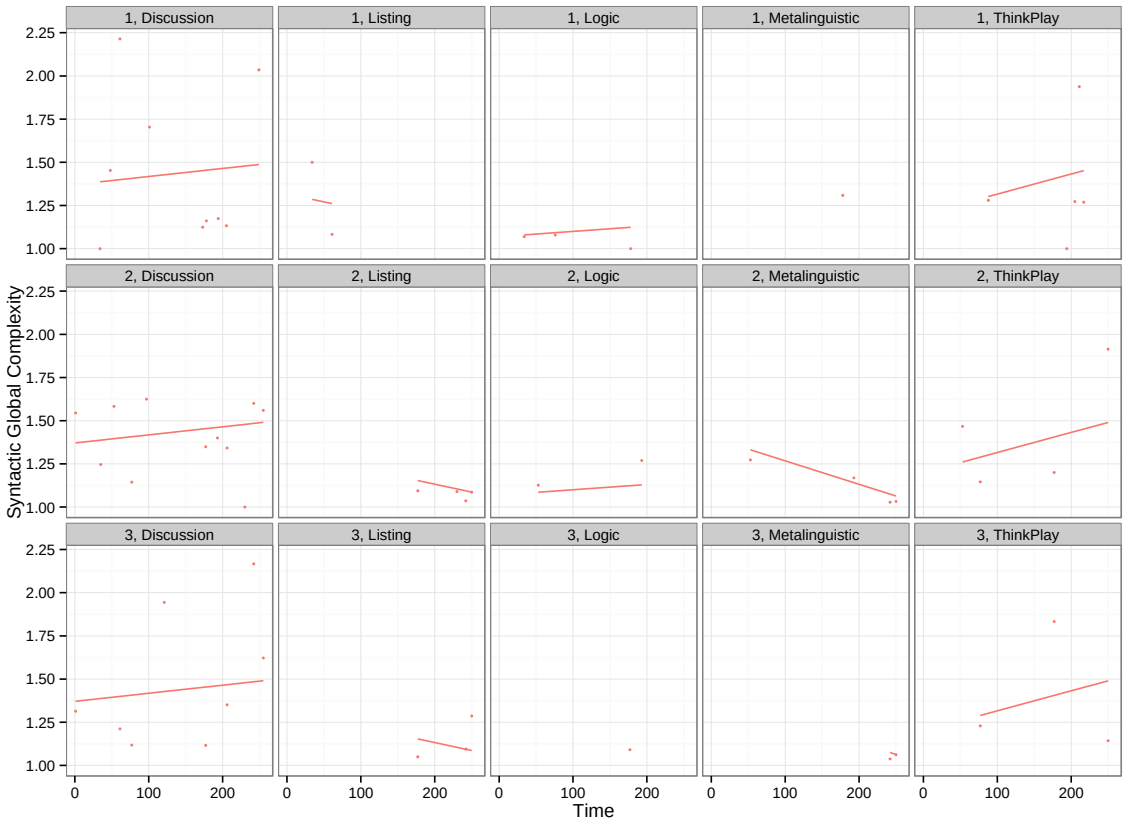

Figure 2. Fitted values from restricted maximum likelihood estimated maximum likelihood models: Global Syntactic Complexity.

Table 10

Maximum Likelihood Parameter Estimates for a Linear Mixed-Effects Model Describing Changes in Global Syntactic Complexity, All Three Participants

Parameter	Estimate (s.e.)	*t*-value	*p*-value
Intercept	1.37 (0.11)	12.72	<.001
Time interval	0.00 (0.00)	0.71	0.483
Listing	-0.05 (0.26)	-0.20	0.841
Logic	-0.31 (0.27)	-1.13	0.263
Metalinguistic	0.03 (0.37)	0.09	0.930
ThinkPlay	-0.17 (0.24)	-0.71	0.483
Time Interval * Listing	0.00 (0.00)	-1.00	0.324
Time Interval * Logic	0.00 (0.00)	-0.08	0.934
Time Interval * Metalinguistic	0.00 (0.00)	-1.03	0.310
Time Interval * ThinkPlay	0.00 (0.00)	0.50	0.619
Variance Components	Chi Square	*df*	*p*-value
Time Interval	0.19	1	0.659
Task Type	17.90	4	0.001
Time Interval * Task Type	2.99	4	0.559

Table 10 shows that task type (p=0.001) is significantly associated with syntactic global complexity. This is indicated by the higher estimated intercepts within Discussion and Thinkplay. Neither time (p=0.659) nor interaction of time and task type (p=0.559) are significant. Participant differences account for <1% of the variation. The likelihood ratio test versus a model with no random effect is not significant (p=0.999). This

indicates a fixed effect model is statistically equivalent to the model with the random effect by participant.

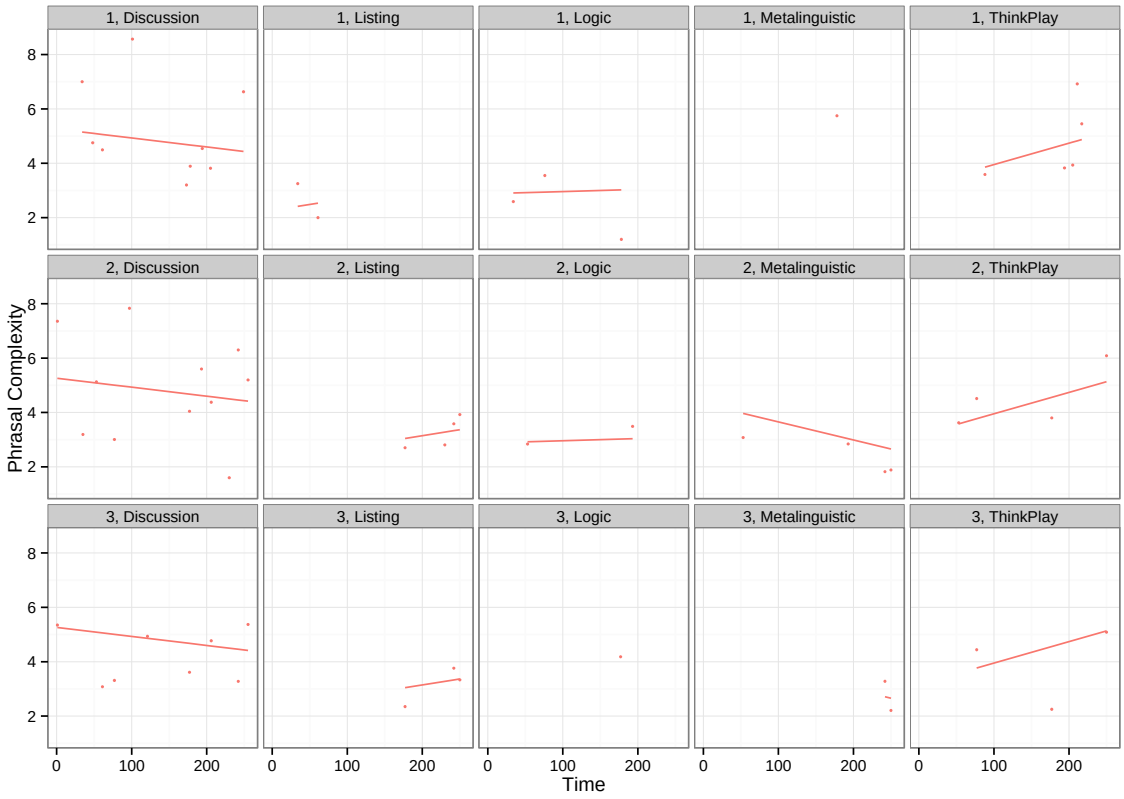

Figure 3. Fitted values from restricted maximum likelihood estimated maximum likelihood models: Phrasal Complexity.

Table 11

Maximum Likelihood Parameter Estimates for a Linear Mixed-Effects Model Describing Changes in Phrasal Complexity, All Three Participants

Parameter	Estimate (s.e.)	*t*-value	*p*-value
Intercept	5.26 (0.52)	10.10	<.001
Time interval	0.00 (0.00)	-1.04	0.30
Listing	-3.00 (1.30)	-2.30	0.03
Logic	-2.38 (1.29)	-1.85	0.07
Metalinguistic	-0.94 (1.77)	-0.53	0.60
ThinkPlay	-2.10 (1.17)	-1.79	0.08
Time Interval * Listing	0.01 (0.01)	1.15	0.26
Time Interval * Logic	0.00 (0.01)	0.44	0.66
Time Interval * Metalinguistic	0.00 (0.01)	-0.39	0.700
Time Interval * ThinkPlay	0.01 (0.01)	1.67	0.101
Variance Components	Chi Square	*df*	*p*-value
Time Interval	0.03	1	0.859
Task Type	24.10	4	<.001
Time Interval * Task Type	4.88	4	0.300

Table 11 indicates that task type ($p < 0.001$) is significantly associated with phrasal complexity. This is indicated by the higher estimated intercepts within Discussion and Thinkplay. The interaction of the time and task type ($p=0.300$) and time ($p=0.859$) are not significant. Participant differences account for <1% of the variance. The likelihood ratio test versus a model with no random effect is not significant ($p=0.999$).

This indicates a fixed effect model is statistically equivalent to the model with the random effect by participant.

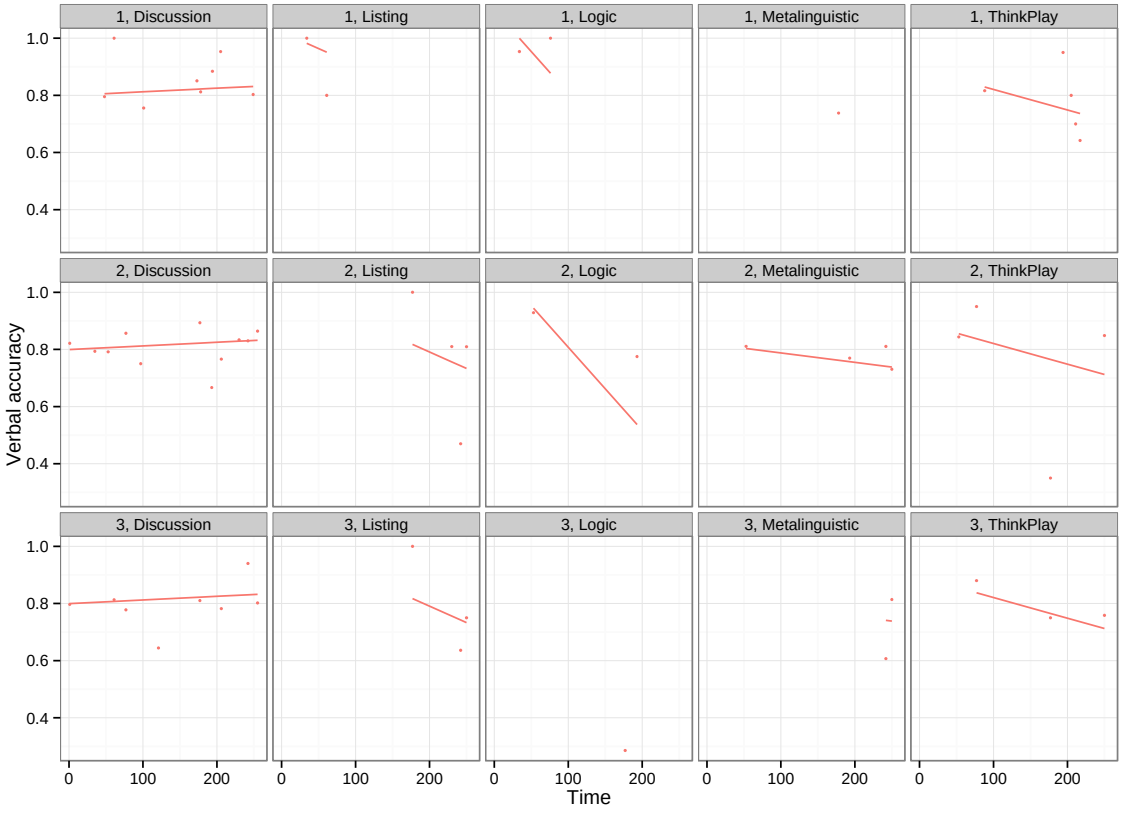

Figure 4. Fitted values from restricted maximum likelihood estimated maximum likelihood models: Verbal Accuracy.

Table 12

Maximum Likelihood Parameter Estimates for a Linear Mixed-Effects Model Describing Changes in Verbal Accuracy, All Three Participants

Parameter	Estimate (s.e.)	*t*-value	*p*-value
Intercept	0.80 (0.05)	16.28	<.001
Time interval	0.00 (0.00)	0.43	0.67
Listing	0.22 (0.12)	1.88	0.669
Logic	0.30 (0.12)	2.54	0.01
Metalinguistic	0.02 (0.16)	0.13	0.90
ThinkPlay	0.09 (0.10)	0.88	0.386
Time Interval * Listing	0.00 (0.00)	-2.10	0.041
Time Interval * Logic	0.00 (0.00)	-3.36	0.002
Time Interval * Metalinguistic	0.00 (0.00)	-0.59	0.559
Time Interval * ThinkPlay	0.00 (0.00)	-1.40	0.168
Variance Components	Chi Square	*df*	*p*-value
Time Interval	5.25	1	0.022
Task Type	1.88	4	0.757
Time Interval * Task Type	16.94	4	0.002

Table 12 indicates that the interaction of the time and task type (p=0.002) and time (p=0.022) are significantly associated with verbal accuracy. As indicated within the fitted values, the slopes of the lines over time are clearly different. However, some of this could be anomalous due to the Logic category for Teresa, who only had one task that fell within the Logic category. Essentially, the fixed coefficients indicate a small increase in

verbal accuracy over time for Discussion, but a decrease over time for all other task types. Task type (p=0.757) alone is not significant. Participant differences account for <1% of the variance. The likelihood ratio test versus a model with no random effect is not significant (p=1.000). This indicates a fixed effect model is statistically equivalent to the model with the random effect by participant.[14]

To summarize, verbal complexity was significantly associated with both task type and time, but there was no significant interaction between task type, time, and verbal complexity. In other words, task type affected verbal complexity, and students did become more verbally complex over time, but they did not become more complex over time within task types.

The global measure of complexity, number of clauses per AS-unit, was also significantly associated with task type. However, there was no significant relationship between global syntactic complexity and time, nor was the interaction between time and task type significant.

Similarly, phrasal complexity, or average number of words per clause, showed a significant relationship to task type, but as with the global measure of syntactic complexity, was not significantly related to time or the interaction between time and task type.

---

[14] The assumption of a normal random error showed some issue. A transformation of the response model was computed to investigate whether there was normal random error. Results of the arcsine transformation model indicate that a fixed effect model is statistically equivalent to the model with the random effect by participant. The assumption of a normal random error in the arcsine model reaches the same conclusions as the untransformed version reported above Results of the arcsine transformations are included in Appendix C.

Finally, the measure of verbal accuracy did show significant relationships for time and for the interaction between time and task type. Task type alone was not significant, although the fixed coefficients indicate a small increase in verbal accuracy over time for Discussion, but a decrease over time for all other task types.

The next section will report the results of the analysis of research question 2.

Research Question 2: Pearson Correlation

Table 13

Pearson Correlations, All Participants

Variables	Variety of verbal structures (N=100)	Average number of clauses per AS unit (N=103)	Average number of words per clause (N=103)	Number of errors per 100 words (N=184)	Percent correct verbal structures per total number of verbal structures (N=100)
Time	0.304**	-.011	-.033	-.226**	-.212*

Note: *=p<.05, **=p<.01, two-tailed tests.

Table 13 contains the Pearson correlation coefficients between the independent variable of time and the five dependent variables for all 3 participants. As can be seen in Table 13, the correlation between time and variety of verbal structures is positive and statistically significant ($r = 0.304$, $p < .01$), meaning that as time increases, the variety of verbal structures among all 3 participants also increases. Time is negatively correlated with number of errors per 100 words ($r = -0.226$, $p < .01$) and percent correct verbal structures per total number of verbal structures ($r = -0.212$, $p < .05$), which means that as

time increases, both number of errors per 100 words and percent correct verbal structures will both decrease. While this may seem a contradiction, it is not. As was described previously in Chapter 3, errors per 100 words encompassed any sort of error that a student may commit: subject–verb agreement, adjective–noun agreement, word choice, etc. Therefore, it is entirely possible for a student to become globally more accurate while decreasing in verbal accuracy.

Table 14 presents the same correlations for only Mike. As can be seen in Table 14, only 2 of the 3 previously statistically significant relationships in Table 13 remain significant. Specifically, the correlation between time and variety of verbal structures is positive and statistically significant ($r = 0.380$, $p < .01$), and the correlation between time and percent correct verbal structures per total number of verbal structures is negative and statistically significant ($r = -0.433$, $p < .01$).

Table 14

Pearson Correlations, Mike

Variables	Variety of verbal structures (N =31)	Average number of clauses per AS unit (N =33)	Average number of words per clause (N =33)	Number of errors per 100 words (N =62)	Percent correct verbal structures per total number of verbal structures (N =31)
Time	0.380**	.002	.193	.076	-.433**

Note: *=p<.05, **=p<.01, two-tailed tests.

Table 15 presents the correlations for only Rebecca. As can be seen in Table 15, only 2 of the 3 previously statistically significant relationships in Table 13 remain significant. Specifically, the correlation between time and variety of verbal structures is positive and statistically significant ($r = 0.405$, $p < .01$), and the correlation between time and number of errors per 100 words is negative and statistically significant ($r = -0.435$, $p < .01$).

Table 15

Pearson Correlations, Rebecca

Variables	Variety of verbal structures (N =41)	Average number of clauses per AS unit (N =41)	Average number of words per clause (N =42)	Number of errors per 100 words (N =71)	Percent correct verbal structures per total number of verbal structures (N =41)
Time	0.405**	-.092	-.080	-.435**	-.144

Note: *=p<.05, **=p<.01, two-tailed tests.

Table 16 presents the correlations for only Teresa. As can be seen in Table 16, only one of the three previously statistically significant relationships in Table 13 remain significant. Specifically, the correlation between time and number of errors per 100 words is negative and statistically significant ($r = -0.315$, $p < .05$).

Table 16

Pearson Correlations, Teresa

Variables	Variety of verbal structures (N =28)	Average number of clauses per AS unit (N =28)	Average number of words per clause (N =28)	Number of errors per 100 words (N =51)	Percent correct verbal structures per total number of verbal structures (N =28)
Time	.156	.097	-.190	-.315*	-.080

Note: *=p<.05, **=p<.01, two-tailed tests.

Individual Developmental Trajectory

Figures 5, 6, and 7 below show the individual developmental trajectories of each participant for each of the dependent variables over the course of the academic year. Additionally, in order to compare the data collection dates, Tables 17, 18, and 19 have been included to show the dates that corresponded with the time points. This will be important when looking at Rebecca and Teresa's data, as they were attending the same Spanish classes, but were not always recorded at the same time due to Teresa's absences and her participation in an extra individual monologic task.

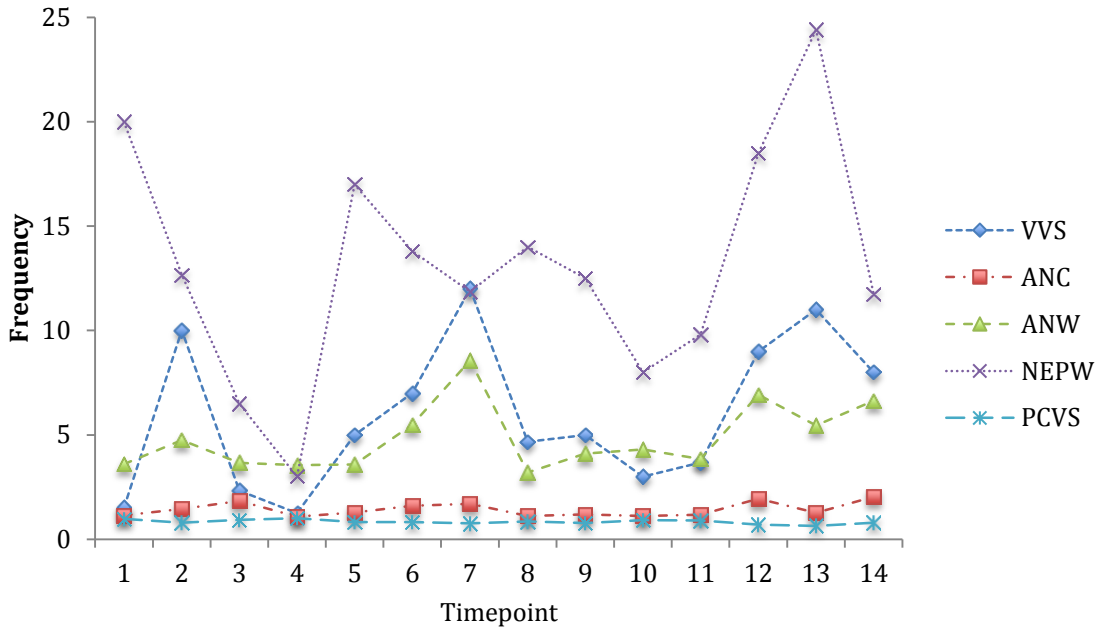

Figure 5. Developmental trajectory, Mike.

Table 17

Number of Task Types by Date of Data Collection, Mike

	Task				
Time Point	Discussion	Metalinguistic	Listing	Thinkplay	Logic
1	1	1	0	0	3
2	1	0	0	0	0
3	2	0	1	0	0
4	0	0	0	0	4
5	0	0	0	1	0
6	1	0	0	0	0
7	1	0	0	0	0
8	2	0	0	0	0
9	2	2	0	0	1
10	2	0	0	1	0
11	2	0	0	1	0
12	0	0	0	1	0
13	0	0	0	1	0
14	1	0	0	0	0

Table 18

Time Points by Date of Data Collection, Mike

Time	Date	Time	Date
1	10/4/12	8	2/21/13
2	10/18/12	9	2/26/13
3	10/29/12	10	3/14/13
4	11/15/12	11	4/4/13
5	11/27/12	12	4/10/13
6	11/29/12	13	4/16/13
7	12/11/12	14	5/10/13

Figure 5 above presents the frequencies for variety of verbal structures (VVS), average number of clauses per AS unit (ANC), average number of words per clause (ANW), number of errors per 100 words (NEPW), and percent correct verbal structures per total number of verbal structures (PCVS) for Mike. The frequency of occurrence for each variable is detailed along the vertical axis of the figure, and the points in time where values were collected are detailed along the horizontal axis of the figure. It should be noted that whenever multiple values were observed for a particular time point, the average of the values was taken. Table 17 shows the number of each task type completed by Mike during each one of the days of data collection. Table 18 shows the dates that corresponded with the time points, for reference.

Mike showed several spikes in the variety of verbal structures. Notably, times 2 and 13, which corresponded to dates 10/18/12 and 4/16/13 were the days in which he

performed the “mesa redonda” (round table) discussions, in which he presented his composition topics in small groups. These tended to include more narration in the past and more subordination, as evidenced by (7), taken from 4/16/13, seen below.

> (7) La, la carta par, parar. Sí, por, por, por, por alguna, algún razón, la la carta, um para. Y, y todos, todas las personas, um, gritan, “Oh, no! What’s happening?” Y, y l-la carta fue, um, parado por, por dos or tres minutos? Y, y no, no sé por qué. Um. Yo yo pensaba que, um, que los conductores, um, supieron que yo:, yo, mi cam, yeah sí, mi mi cámara y:. Pero-pe-pero, um, por, por suert-suerte, um, (laughing) la carta, com, después del tres minutos, um, comienza a, um, cont-continue on.
>
> *The, the car sto, to stop. Yes, for, for, for, for some, some reason, the the car, um, stops. And, and everyone, all the people, um, scream, “Oh, no! What’s happening?” And, and t-the car was, um, stopped for, for two or three minutes? And, and I don’t, I don’t know why. Um. I I was thinking that, um, that the conductors, um, found our that I, I, my cam, yeah yes, my my camera and. But-bu-but, um, by, by luc-luck, um, (laughing), the car, com, after the three minutes, um, starts to, um, cont-continue on.*

Time 7 corresponded to the date 12/11/12, which was Mike’s oral presentation. His topic was the history, geography, and climate of Lima, Peru, with some discussion of the Nazca Lines, a series of geoglyphs in Southern Peru. Because the presentation concerned the history of Peru, there was quite a bit of narration in the past and subordination, as can be seen in the following excerpt. However, it should be noted that Mike was reading off of his Powerpoint presentation the majority of the time, so this was not a real reflection of his spontaneous oral production. An excerpt from this presentation can be seen in (8) below.

> (8)
>
> Y ahora hablaré, uh, sobre la breve historia. El país de Perú comenzó como, uh, las comiunidades rurales pequeñas? Pero:, por un mil cuatrocientos sesenta, um, fue incorporado al imperio Inca. Uh, sin embargo, en un mil, uh, quinientos veinte, Francisco Pizarro y su pareja, Diego de Amagro uh obtuvieron un permiso del, uh, gobernador de

Panemá? y, uh, hicieron una expedición hacía Perú, um, a donde: descubrieron, los Incas. Uh, los españoles, uh, fueron vistos a los dioses, como los dioses, uh, y tratados con respeto. Pero. Cuando Pizarro vio la ricueza, um, él, él, uh quería toda para sí mismo. Uh, como un resultado, uh, ellos se volvieron, uh, a España, uh, para, uh, recaudar fondos y conseguir apoyo para sus planes a, controlar el imperio Inca.

And now I will talk, uh, about the brief history. The country of Peru began, like, uh, the small rural communities? But, through one thousand four hundred sixty[15]*, um, it was incorporated to the Inca empire. Uh, nevertheless, in one thousand, uh, five hundred twenty, Francisco Pizarro and his partner, Diego de Amagro uh obtained a permission from the, uh, governor of Panama? And, uh, they made an expedition [toward]*[16] *Peru, um, to where they discovered, the Incas. Uh, the Spaniards, uh, were seen to the gods, like the gods, uh, and treated with respect. But. When Pizarro saw the richness, um, he, he, uh, wanted it all for himself. Uh, as a result, uh, they returned themselves, uh, to Spain, uh, in order to, uh, collect money and secure support for their plans to, control the Incan empire.*

Time 12 was Mike's paired oral interview, as seen in excerpt (9). In this interview, he was explicitly instructed to use the past tense as well as to "show what he knows" in terms of verb structures and vocabulary, and this is likely the reason for the result in the spike in verbal complexity at this time point.

(9)

Male student: Muy bien. (3) U:m, en: tu:, ciudad, um (5) uh, viste, un:, diferencia de, um, dos, um, personas de, etnias diferencias, uh, en (6.5) en el pa:sado? Um. Qué, viste? Um, discriminación.

Mike: Um, no, no he, um, no he via-jado, um, a mi ciudad mucho, um, así que no, no veí [sic], um, no he veído [sic], um, mucho discriminación, um, porque: yo sé:, um, yo sé que:, hay: algún: discriminación, um, en la ciudad. Um, por ejemplo, um, la discriminación en contra del, um, las razas y las etnias, um, son, um, más la forma de:, um, escribir? or, uh, como-como el grafiti: en los, baños, um, en las puertas, en las ventanas. Um, y, um, no-

[15] Mike was attempting to say 1460, but the "un" in front is not required in Spanish.

[16] The pronunciation used by Mike here was the imperfect indicative form of the verb *hacer*, "to make." The word *hacia* means "toward." These pronunciations frequently get confused by students in oral production.

no pienso, que, la discriminación es un, um, un gran problema? En, nuestro:, uh, ciudad? Uh pero: no sé si es la verdad o no.

Male student: Very good. (3) U:m, in your city, um (5) uh, did you see, a, difference of, um, two, um, people of, ethnicities differences, uh, en (6.5) in the past? Um. what, did you see? Um, discrimination.

Mike: Um, no, I haven't, um, I haven't tra-veled, um, to my city much, um, so I didn't, I didn't see, um, I haven't seen, um, much discrimination, um, because I know:, um, I know that, there is some discrimination, um, in the city. Um, for example, um, the discrimination against the, um, the races and the ethnicities, um, are, um, more the form of, um, writing? or, uh, like-like the graffiti in the, bathrooms, um, in the doors, in the windows. Um, and, um, I don't-I don't think, that, the discrimination is a, um, a big problem? In our, uh, city? Uh but I don't know if it's the truth or not.

The individual developmental trajectories for Rebecca are shown below.

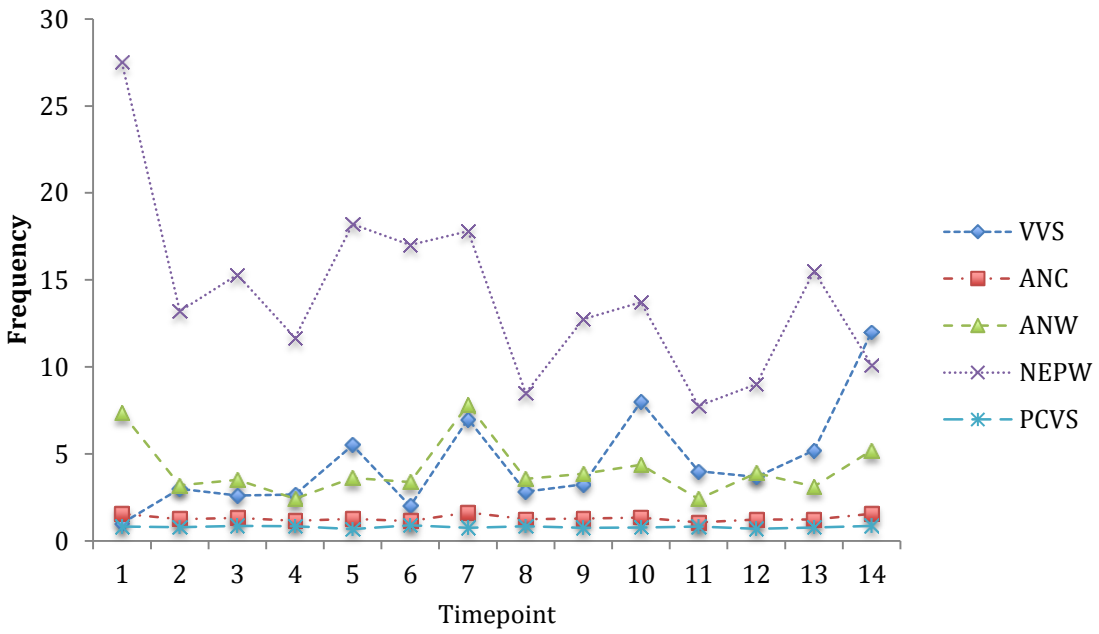

Figure 6. Developmental trajectory, Rebecca.

Table 19

Number of Task Types by Date of Data Collection, Rebecca

	Task				
Time Point	Discussion	Metalinguistic	Listing	Thinkplay	Logic
1	1	0	0	0	0
2	2	0	0	0	0
3	1	1	0	0	2
4	2	0	0	1	0
5	1	0	0	1	0
6	2	0	0	1	0
7	1	0	0	0	0
8	3	0	2	0	1
9	1	1	0	0	1
10	1	0	0	0	0
11	1	0	2	0	0
12	1	1	1	0	0
13	0	3	1	1	0
14	1	0	0	0	0

Table 20

Time Points by Date of Data Collection, Rebecca

Time	Date	Time	Date
1	9/1/12	8	2/25/13
2	10/5/12	9	3/13/13
3	10/23/12	10	4/5/13
4	10/31/12	11	4/19/13
5	11/9/12	12	5/1/13
6	11/16/12	13	5/8/13
7	12/7/12	14	5/13/13

Figure 6 above presents the frequencies for variety of verbal structures (VVS), average number of clauses per AS unit (ANC), average number of words per clause (ANW), number of errors per 100 words (NEPW), and percent correct verbal structures per total number of verbal structures (PCVS) for Rebecca. The frequency of occurrence for each variable is detailed along the vertical axis of the figure, and the points in time where values were collected are detailed along the horizontal axis of the figure. It should be noted that whenever multiple values were observed for a particular time point, the average of the values was taken. Table 19 shows the number of each task type completed by Rebecca during each one of the days of data collection. Table 20 shows the dates that corresponded with the time points at which Rebecca was measured. The trends in Rebecca's data will be discussed below together with those in Teresa's data as they

engage in some of the same activities on the same dates. The individual developmental trajectories for Teresa are shown below.

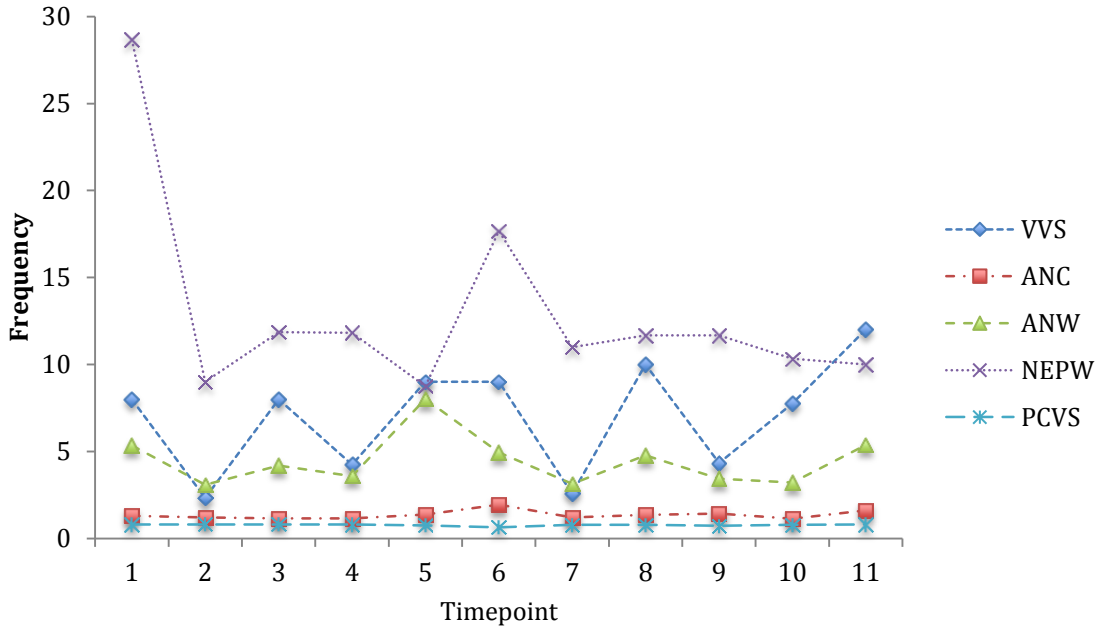

Figure 7. Developmental trajectory, Teresa.

Table 21

Number of Task Types by Date of Data Collection, Teresa

	Task				
Time Point	Discussion	Metalinguistic	Listing	Thinkplay	Logic
1	1	0	0	0	0
2	2	0	0	1	0
3	1	0	0	1	0
4	2	0	0	1	0
5	1	0	0	0	0
6	1	0	0	0	0
7	3	0	2	0	1
8	1	0	0	0	0
9	1	1	1	0	0
10	0	2	1	1	0
11	1	0	0	0	0

Table 22

Time Points by Date of Data Collection, Teresa

Time	Date	Time	Date
1	9/1/12	7	2/25/13
2	10/31/12	8	4/5/13
3	11/9/12	9	5/1/13
4	11/16/12	10	5/8/13
5	12/7/12	11	5/13/12
6	12/31/12		

Figure 7 above presents the frequencies for variety of verbal structures (VVS), average number of clauses per AS unit (ANC), average number of words per clause (ANW), number of errors per 100 words (NEPW), and percent correct verbal structures per total number of verbal structures (PCVS) for Teresa. The frequency of occurrence for each variable is detailed along the vertical axis of the figure, and the points in time where values were collected are detailed along the horizontal axis of the figure. It should be noted that whenever multiple values were observed for a particular time point, the average of the values was taken. Table 21 shows the number of each task type completed by Teresa during each one of the days of data collection. Table 22 details the dates that correspond to the time points at which Teresa's data were collected.

Each figure above shows how each variable behaved over the course of the academic year and how they interacted for each student. The time measures for all 3 participants are not the same, as can be seen in Tables 18, 20, and 22. Mike's data were

collected on different dates than Teresa and Rebecca's data, and Teresa was absent some of the days Rebecca's data were collected, as previously mentioned. Teresa had one additional monologic oral task recording in December (corresponding to 12/31/12) that Rebecca did not have. The instructor did not make it clear to the researcher if Rebecca did not participate or was unavailable for that recording. This additional monologic task required Teresa to discuss a movie she had seen and to relate this movie to the actual situation of immigration in Spain. This produced more hypothesizing, opinions, and subordination, as evidenced by the following excerpt.

(10)

Hola, me llamo Teresa. Voy a hablar sobre la situación inmigrante ilegal, en España. El estrecho, es la vía más pupa-popular para inmigrante ilegales. Durante un año, las gotas españoles tienen muchos inmigrantes ilegales. Y hay varios, y hay situaciones, muy triste. Por ejemplo, un barco, con más de, cuarenta personas, fueron, fue en, la agua, pero, treinta y sieto personas mueren. Esto es un ejemplo, de la situación, de imi-inmigración ilegal, en España. La situación ha llevado la implatación de tecnología nueva. Como cámaras nocturnas. En otro situaciones. Más de, veinte:, mujeres. Que fueron embarazada? y tiene hijos, fueron deternados, en España.

Hello, my name is Teresa. I am going to talk about the situation illegal immigrant, in Spain. The strait is the most pupa-popular route for illegal immigrant. During one year, the Spanish [drops] have many illegal immigrants. And there are various, and there are very sad situations. For example, a boat, with more than, forty people, were, was, in the water, but, thirty seven people die. This is an example, of the situation, of illegal im-imigration, in Spain. The situation has brought the implantation of new technology. Like night cameras. In other situations. More than, twenty:, women. That were pregnant? And have children, were [detained][17], in Spain.

[17] The word Teresa used, "deternados," does not exist. This is my approximation of what I believe she was trying to say.

The time points in the graphs are not meant to be compared among participants, but rather to show development and interaction as time, in terms of when each participant's data were collected, progresses. Therefore, Time 1 for Mike is not the same as Time 1 for Rebecca or Teresa. The data are displayed in this way for ease of understanding of how each variable behaved for each participant as time progressed.

However, even though the data were not meant to be compared across participants and time points, Teresa and Rebecca did show similar trends in similar tasks. For example, the time points that correspond to 12/7/12 were oral presentations for Rebecca and Teresa's class. Teresa did not perform as well as Rebecca, but they both showed increases in verbal complexity (variety of verbal structures) on that date. The topics of their presentations were, respectively, "Japanese influence in Perú" and "Hispanic literature in the world today." Both of these topics required them to narrate in the past, to use more advanced structures, and speak alone for a long period of time. It can be seen that there is a corresponding decrease in verbal accuracy for both women. The difference in performance is attributed to the fact that Rebecca attended class and participated very regularly, while Teresa had many absences and occasionally did not participate as much as her classmates.

Again, for Rebecca and Teresa, spikes in variety of verbal structures can again be seen on 4/5/13 and 5/13/13, which correspond with the group oral assessments. The nature of the oral assessments were such that the instructor required the groups of students to discuss a series of questions about the novel they read in class or the movies they had seen. They were to compare the movies and novel, to draw conclusions about

the relevance to society and make hypotheses about the connections between them. This resulted in hypothesizing language, opinions, and some narration in the past. The figures show that there is slight increase in average number of words per clause and average number of clauses per AS unit, as well, which supports the hypothesis that such activities would promote more complex language production. This can be seen, for example, in the segment from Rebecca's oral assessment from 5/13, below.

(11)

Rebecca: El inmigrante es quien no ha dejado del todo el lugar, del que se fue y ha terminando por el dat, adaptarse completamente el sitio donde llegó. No es de aquí, ni es de allá.
Female student: Me gusta.
Rebecca: (laughs) Te gusta, mm-pero (1.0) creo que no es el (1.2) verdad completame-mente porque hay, mm, mm
Female student: Primero? Es pregunta?
Rebecca: Sí, sí. Uh, la, un inmigrante no es, solamente una persona que no, tiene una, place para, llamar (0.9) a, su hogar? (0.8) o, un, su casa?

Rebecca: The immigrant is whom has not left of everything the place, from which he left and has finished by the dat, adapting himself completely [to] the site where he arrived. He is not from here, nor from there.
Female student: I like that.
Rebecca: (laughs) You like it, mm-but (1.0), I believe that it's not the (1.2) truth complete-ly because there is, mm, mm
Female student: First one? Is that a question?
Rebecca: Yes, yes. Uh, the, an immigrant is not, only a person who doesn't, have a, place to, call (0.9) to, his home? (0.8) or, a, his house?

As can be seen from the tables and figures above, there is variation over time in how the students produced linguistic complexity and accuracy. Whether these changes were significant varied by participant as well. Mike showed a significant decrease in verbal accuracy and a significant increase in verbal complexity. Rebecca showed a

significant increase in variety of verbal structures and a significant decrease in number of errors per 100 words over time. Teresa showed only a significant decrease in number of errors per 100 words over time. In other words, she became significantly more globally accurate over the course of one semester. The rest of the variables did not show a significant change over time. Each student followed her or his own individual trajectory in the development of their oral linguistic complexity and accuracy over time.

The next chapter, Chapter 5, will discuss these results in detail and how these data can be interpreted in the context of each participant.

Chapter 5

Discussion and Conclusions

This chapter will discuss the research questions that guided this investigation. Specifically, how did task type affect the linguistic complexity and accuracy of the oral production of Spanish by 3 learners of intermediate Spanish, and how did their oral linguistic complexity and accuracy change over time? The results of the first research question, "Does task type affect oral complexity and accuracy of learner language?" will be discussed first, followed by a discussion of the sub question, "Does time interact with task type to affect oral linguistic complexity and accuracy of learner language?" Finally, the second research question, "Does oral linguistic complexity and accuracy change over time?" will be discussed both in the context of the statistical analyses presented in Chapter 4 as well as within a Dynamic Systems Theory framework in order to describe each student's development of the five dependent variables over time. This chapter will conclude with a discussion of the limitations of this study and future directions based both on the limitations and the findings of this research.

RQ1: Does task type affect oral linguistic complexity and accuracy of learner language?

As was described in Chapter 2, there are two main theoretical frameworks that describe the effects of task on linguistic complexity, accuracy, and fluency (CAF). Even though cognitive or task complexity was not a variable in this study, it is important to describe how task effects have been theorized in the literature before discussing the results obtained here. The first hypotheses that are relevant are Skehan and Foster's

Trade-off Hypothesis (Skehan, 1996, 1998, 2003; Skehan & Foster, 1997), and later the Extended Trade-off Hypothesis (Skehan & Foster, 2012), which states that attentional resources are limited, and when a learner is confronted with a difficult task, that is, when too many attentional resources are directed toward a cognitively complex task, she will not be able to attend to all the features of CAF at once. The tendency, under this model, is that when one of the features is taxed, the other features will show decreases, though this is not always borne out in the research. It is because of the inconsistency of results that Skehan and Foster adjusted the Trade-off hypothesis to the Extended Trade-off Hypothesis (Skehan & Foster, 2012) to reflect that their previous model was predictive of tendencies only, and that actual language production may not always follow it.

Another hypothesis, the Cognition Hypothesis, put forth by Robinson (2001a, 2001b, 2003, 2005, 2007) and Robinson and Gilabert (2007), also attempts to describe how task effects are seen in the production of a second language. However, this model is more fine grained and describes how task factors such as cognitive factors, interactional factors, and learner factors condition learner output (Robinson & Gilabert, 2007). In this model, linguistic complexity and accuracy are similarly affected by task effects, in that when linguistic complexity increases, so does linguistic accuracy, while fluency decreases. In other words, linguistic complexity and accuracy are tied together and have an inverse relationship with fluency.

Just as with the Extended Trade-off Hypothesis, results using this framework have been mixed. That is, linguistic complexity and accuracy have not been shown to always behave in the same ways. These mixed results may be because the models aren't

accurately representing how language is conceptualized or produced, or this could be a byproduct of the inconsistency in the operationalization of the CAF variables (Ellis & Barkhuizen, 2005; Housen & Kuiken, 2009; Housen et al., 2012; Norris & Ortega, 2009). In the current study, a more "organic" and nuanced operationalization of the variables was attempted in order to determine how these variables interacted in the oral production of intermediate Spanish. The results of this data analysis show that the linguistic output of the three participants does, indeed, differ in different tasks. This is in accord with the previous literature that was not able to show definitively how task affects the linguistic output of L2 learners.

Because cognitive or task complexity was not a variable in the present study, no solid claims can be made about how the cognitive complexity of the tasks affected the production of Spanish. Therefore, task effects will be described in terms of the type of language that was produced by these learners and trade-off effects between linguistic complexity and accuracy. The next sections will discuss how tasks affected the dependent variables under examination, taking each variable separately, and then discussing the data as a whole.

Global Syntactic Complexity

As a reminder, global syntactic complexity is defined as the number of clauses per AS unit.

Mike. Mike's data showed the highest average number of clauses per AS unit in the categories *discussion* and *thinkplay*, with *discussion* being slightly higher than *thinkplay*, although the difference was not statistically significant. There are a few

possible confounding factors that could be contributing to the lack of significant findings in Mike's data. First, he did not produce as much language during the fall semester as he did during the spring semester, something to which he admitted in his interview with the researcher, stating he did not participate as much as he should have during fall semester and was making a concerted effort during spring semester to participate more. There was also a difference in instructors and instructional model between fall and spring semester for Mike. During the fall semester, he was in a hybrid course that met only 3 days a week, while he was in a "4+1" section during the spring, which met 4 days a week. In the hybrid class, he was taught by a less experienced instructor, and in the spring, he was taught by a more experienced instructor, who engaged the class in more paired and group activities. It follows that the quantity of language data that he produced in some tasks, especially during fall semester, may be too small to show significant differences in the number of clauses per AS unit.

Rebecca. Rebecca's data showed the same trend as Mike's, but her data showed that average number of clauses per AS unit was significantly higher for *discussion* than *metalinguistic* or *listing*. Like Mike, *thinkplay* was also higher, and logic also lower in average number of clauses per AS unit, but these two did not show significant differences.

Teresa. Teresa showed similar tendencies as Mike and Rebecca with respect to average number of clauses per AS unit. She produced the highest average in the categories *discussion* and *thinkplay*, but these differences were not significant. These two categories were also higher than *listing, metalinguistic,* and *logic*, but *logic* only had one

data point for Teresa, making this category difficult to analyze statistically. Again, none of these differences was statistically significant. The reason for this lack of statistical significance could possibly be attributed to her absences on data collection days, which resulted in fewer data points.

These results are not surprising given the nature of the data. The categories *discussion* and *thinkplay* included more open-ended activities that promoted longer strings of discourse. The other categories were much more closed in nature, producing language that was very brief. For example, *listing* included activities that asked students to fact-find, such as when Teresa and Rebecca were asked to look for words and/or phrases that would describe the protagonist of the novel they were reading. The language use in that task was very brief, often consisting of one- or two-word answers. The same sort of result was found with the *metalinguistic* category, in which students primarily offered conjugated verbs or short phrases like, "¿Es subjuntivo?" (*"Is it subjunctive?"*)[18].

The next section will report the results with respect to phrasal complexity, or average number of words per clause.

Phrasal Complexity

Mike. Mike's data showed statistical significance for average number of words per clause, but this finding turned out to be a statistical fluke once the post-hoc test was run, as no significant relationship was found upon decomposition of the results. This may be due to the fact that there was only one task in Mike's data that corresponded to the

[18] It is important to note that metalinguistic tasks were not the only tasks completed on the days in which they occurred. In fact, they were usually only a very small part of a day's interaction. Thus, there was usually no visible effect in the data on the global syntactic complexity for the days' data and there would be no meaningful effect on the graphic representations of the development of global complexity.

metalinguistic category. Interestingly, Mike has the highest number of average words per clause within the *metalinguistic* category at close to 6 words per clause, while Rebecca and Teresa's scores were closer to 3 words per clause. This may be because Mike only participated in one metalinguistic task, which asked students to conjugate correctly the verbs presented on a Powerpoint. In that task, he was asking questions of his classmates and his instructor, such as, "es una expresión, ¿no?" (*"it's an expression, isn't it?"*), rather than just giving an answer in the form of a conjugated verb. This longer response contrasts with the metalinguistic tasks that Rebecca and Teresa completed, where they were to conjugate verbs in sentences, instead of asking longer questions of one another. Their interaction was much briefer, consisting of comments such as, "Es subjuntivo." (*"It's subjunctive."*) or one-word responses.

Rebecca. Rebecca's data showed that *discussion* had a statistically significant higher average number of words per clause than *metalinguistic.* In addition, *thinkplay* was higher than *discussion*, and *logic* and *listing* were lower than *discussion* and *thinkplay*, but these relationships were not statistically significant. Again, given the nature of *discussion* compared to metalinguistic tasks, this result is not surprising.

Teresa. Teresa had no significant differences with respect to the average number of words per clause. However, similar to Mike and Rebecca, *discussion* and *thinkplay* had a higher number of words per clause than *metalinguistic* and *listing*. However, she had the highest average number of words per clause in the *logic* category, which is comprised of the subcategories *information gap*, *ordering and sorting*, *comparing*, and *matching*. Because of her absences, she was only present for *comparing*, in which the students were

to compare their first weeks at the college away from their families. This skewed the results toward a higher average number of words per clause for two possible reasons. One, this task occurred during the spring semester, when she was feeling a bit more confident, and two, she was providing more detail about her experience at the college than other participants did, making her utterances longer. Her partners seemed fascinated by her experience, possibly because they did not move as far away from home as she did, and asked her many follow-up questions about how her mother reacted to Teresa going to school that fall. See excerpt (12) from this task.

(12)

Teresa: Uh, mi mamá? fue un: poco triste: pero ella fue, bien. (4.5) Porque yo, um, experensa, casi todo los (emociones) el año, anterior?
Female student: Oh sí.
Teresa: So. Fue bien.

Teresa: Uh, my mom? Was a little sad but she was, good. (4.5) Because I, um, experience [sic], almost all the emotions the year, before?
Female student: Oh yes.
Teresa: So. It was good.

In sum, similarly to the global complexity measure, when more and longer utterances are produced, it follows that there will be a higher number of words per clause. Again, the nature of the data elicited by *metalinguistic*, *logic*, and *listing* would predict that the average number of words per clause would be lower, since the utterances in these categories tended to be short, such as offering a conjugated verb, or a very simple SVO sentence with no subordination.

Verbal Complexity

Verbal complexity, or variety of verbal structures employed by the participants, did show differences by task generally. *Metalinguistic* tasks in the data elicited a wider

variety of verbal structures because the majority of the metalinguistic tasks in which the students participated focused on verbal structures, such as when to use the present subjunctive or imperfect subjunctive. This resulted in a higher number of verbal structures with a lower percentage of accuracy because they were specifically instructed to work with verb structures they could not control spontaneously at that time, such as the imperfect subjunctive. With discussion, often narratives would fall under that task,[19] meaning that students were not only using the present indicative, but also the preterit and the imperfect, and often the present perfect. In *listing* tasks, there were often *no* verbal structures produced, and when they did occur, they were usually limited to the present indicative, present perfect, or preterit. In the *logic* category, the verbal structures were almost exclusively limited to present indicative, especially the third person singular conjugation of the verb *ser*, "to be." See excerpt (13) below from an *ordering and sorting* task where students were asked to rate in importance a number of items in a list.

(13)

Mike: Oh, es, uh, es, tercera. El obtener, asilo, asy-political asylum, uh, //primera.
Male student: //Primera.
Mike: Sí.

Mike: Oh, it's, uh, it's, third. The obtaining, asylum, asy-political asylum, uh //first.
Male student: //First.
Mike: Yes.

[19] In the case of the current data set, narration, especially in the case of Rebecca and Teresa, was not couched in personal experience necessarily. A narration could be a retelling of an event from one of the novels one of the movies that was assigned.

Mike. In Mike's data, *discussion* had a significantly higher verbal complexity score than *listing* or *logic*. *Thinkplay* scored higher on verbal complexity than *discussion,* but this relationship was not statistically significant. Both *thinkplay* and *discussion* were higher than *metalinguistic*, but again, these were not statistically significant findings. These results were likely due to the nature of the language elicited in these types of tasks. Mike tended to rely on present indicative in tasks that would fall under the categories of *listing* and *logic,* while he used a much wider variety of verbal structures in *discussion, thinkplay* and *metalinguistic* tasks.

Rebecca. Rebecca did not show any statistically significant relationships for number of types of verbal structures. However, *discussion, thinkplay,* and *metalinguistic* show the greatest average number of verbal structures. *Logic* and *listing* have the lowest values, again most likely for the same reasons as Mike: a reliance on the present indicative for utterances within these two latter categories, specifically third person singular of the verb *ser*.

Teresa. In Teresa's data, there weren't statistically significant differences between *discussion*, *logic*, *thinkplay, listing,* and *metalinguistic* for verbal complexity. However, post-hoc tests were not able to be run on these data due to insufficient data in the *logic* category. Even though she did not have statistically significant results, she showed the most verbal complexity in the metalinguistic tasks, with the next highest in *thinkplay* and *discussion*, respectively. These data also follow the trend shown by the group data and Rebecca and Mike's individual data analyses. The fact that *discussion* did

not follow the same trends as with Mike's or Rebecca's results could be a function of her insecurity (i.e., she may have felt shy about her abilities), her absences, or both.

Finally, the results of the analysis of verbal accuracy, or rate of correctly used verbal structures, will be discussed.

Verbal Accuracy

Mike. Mike showed a statistically significant difference between *logic* and *thinkplay*, in that he was more accurate in the former than the latter. Mike tended to over rely on the present indicative, and especially the third person singular of the verb *ser* in this category, whereas *thinkplay* had more elaborate utterances and he drew upon a wider variety of verbal structures in order to express himself. For example, the role-play oral assessments fell under the *thinkplay* category, and within that subcategory, he used preterit, imperfect, present perfect, and present indicative.[20] He also had higher scores for verbal accuracy in *listing* and *discussion*, though these were not statistically significant. The finding for *discussion* is especially interesting given that these tasks tended to be more elaborate, and given that he had a higher score for verbal complexity. Mike also tended to code-switch into English, which reduced the amount of analyzable language, since only Spanish language production was under analysis in the current study. Additionally, the verbal complexity measure did not count the number of times each verbal structure is used, just that it was used. Therefore, there could be overrepresentation of certain verb structures in this category that was not captured, though this is speculation on the part of the researcher and would need to be examined in future studies.

[20] Please see excerpt (9) in chapter 4 for an example.

Rebecca. Rebecca did not show any statistically significant results for any categories under investigation. Similar to Mike, her highest accuracy scores were found for *discussion, logic,* and *listing.* However, Rebecca rarely code-switched. Overall, she is similarly accurate in all tasks.

Teresa. Teresa showed no statistically significant differences in her scores for verbal accuracy, as with Rebecca. She was similarly accurate with verbal structures in all categories, with metalinguistic being the lowest with 74% accuracy. Logic is an anomaly, as previously explained, since there was only one task that fell under this category, and it was not typical of the types of tasks that usually occurred in the *comparing* task. She used many verbs in the preterit in that task, which lowered her accuracy score considerably.

RQ1: Summary

The data analyzed to answer RQ1 have shown that task does, indeed, affect linguistic output. However, the results varied by participant as well as by task. Those tasks, such as *discussion* and those categories falling under the umbrella category *thinkplay* tended to produce the most language, resulting in higher complexity scores for the various measures of complexity. Those tasks, such as *listing* and those categories subsumed by the *logic* category, which tended toward shorter strings of speech, tended to have higher rates of accuracy.

There were also interesting results that show some trade-off effects within tasks and across tasks. In Mike's data, the average number of clauses per AS unit was relatively stable, except in the *logic* category, which saw a dip to 1.05 from a range of 1.3–1.4 for the other categories. However, the other complexity measures, variety of

verbal structures and average number of words per clause did show some trade-off effects with the verbal accuracy scores. Specifically, the general trend was that when the two complexity measures decreased, the accuracy scores increased. The *logic* category, for Mike, was the only anomaly to this trend. In this category, all of his values decreased, and as previously explained, this could be due to the nature of the language produced in the tasks that fell under this category.

Rebecca's data showed similar trends to Mike's with respect to the complexity and accuracy measures. She, too, showed some trade-off effects where variety of verbal structures and average number of words per clause were higher when verbal accuracy scores were lower. However, in the *logic* category, unlike Mike's data, there appeared to be trade-off effects that followed the results for the other categories: as the complexity measures decreased, her verbal accuracy increased. In fact, she showed the highest verbal accuracy score in the *logic* category. This could be because of a possible overrepresentation of the third person singular form of *ser* ("to be"). Again, however, frequency of forms was not under investigation in the current study, so this is conjecture and would need follow-up in future studies.

Teresa's data were a bit more varied than Mike and Rebecca's. As previously discussed, Teresa missed several class periods, and had fewer recording sessions than Mike or Rebecca. This could be affecting her results, since there are fewer data points for her than for the other two participants. She also was not as participatory, especially during fall semester, than the other two participants. That said, she does exhibit some trade-off effects with linguistic complexity and accuracy, too, but this appears

consistently only with variety of verbal structures and verbal accuracy. When variety of verbal structures is higher, her scores for verbal accuracy are lower. There is no clear trend of an inverse relationship between average number of words per clause or average number of clauses per AS unit and verbal accuracy. It appears that, for Teresa, these two measures are not exhibiting trade-off effects with verbal accuracy.

These mixed results coincide with the mixed results in the previous literature. On one hand, Mike and Rebecca seem to provide support for Skehan (1996, 1998, 2003) and Skehan and Foster's (1997, 2012) LACM, which states that when students' attentional resources are directed at meaning rather than form, they will tend to see an inverse relationship between linguistic complexity and linguistic accuracy. That is, when complexity increases, accuracy will decrease. However, Teresa's data, the only measure that interacts in this way with linguistic accuracy is the variety of verbal structures. However, this still provides support for this model, as when more types of verbal structures were used, Teresa was more focused on the meaning she was trying to convey.

In addition to general trends, it is also important that the higher complexity scores (and resultant lower accuracy scores), were attributed to those tasks that were more open in nature, allowing for more expression of ideas and opinions. While each student had different scores for each category and measure, *thinkplay* and *discussion* tended to have the highest linguistic complexity scores and lower verbal accuracy scores.[21] This, too, follows the LACM, as these tasks required the students to put more of their attentional resources on expressing meaning and opinions, which, according to the LACM, would

[21] Though please note that *discussion* did show an increase in accuracy over time.

take resources away from the students' ability to express themselves accurately. As described previously in chapter 2, Skehan (2009) outlined how easing the pressure on the Conceptualizer and the Formulator frees up resources in the speaker's linguistic system to focus on aspects of CAF. And, as seen in the review of the literature, this can result in a focus on one aspect at the expense of others, or a focus on more than one aspect.

Because previous research has been so varied in both operationalization of the complexity and accuracy constructs (Ellis & Barkhuizen, 2005; Housen & Kuiken, 2009; Housen et al., 2012) as well as the high amount of variation in results found within the study of CAF, it is difficult to say definitively that these results are in line with those of previous studies. With more standardization of how the constructs of CAF are operationalized, more firm conclusions can be drawn on how task affects CAF (Foster et al., 2000; Norris & Ortega, 2009). This research is a first step in the direction toward a more standardized measurement of CAF production data across tasks in order to more effectively compare results across studies.

RQ1a: Does time interact with task type to affect oral linguistic complexity and accuracy of learner language?

Because the ANOVA analyses showed that there were some task effects on the oral production of linguistic complexity and accuracy by the three learners in this study, it was determined that a mixed model analysis would elucidate whether there was an interaction between task type and time on the different variables under study. Due to the small number of participants, this analysis was conducted on the data set as a whole, as previously mentioned.

Verbal Complexity

For the three participants, the analysis showed that time (p= 0.00734) and task type (p= 0.01937) affect verbal complexity for the three participants. There is an increasing trend toward greater verbal complexity over time across task types, even in Discussion and ThinkPlay which tended to start with higher levels of verbal complexity on average. The interaction between time and task was not significant (p= 0.86832).

Global Syntactic Complexity

Global syntactic complexity showed a significant difference by task type (p=0.001292), but neither time (p=0.659721) nor the interaction of time and task type (p=0.558692) were significant for this variable. This means that task type did affect the amount of clauses per AS unit in the oral production of the participants; there were higher estimated intercepts within Discussion and ThinkPlay. However, there was no significant change over time or within task type over time.

Phrasal Complexity

As with global syntactic complexity, there is a significant effect of task type ($p<0.0001$), but time (p=0.8591) and the interaction of time and task type (p=0.3001) are not significant. In regards to task, there are higher estimated intercepts within both Discussion and Thinkplay in relation to the other task types. Thus while there are task effects on the number of words produced per clause, change over time is not significant.

Verbal Accuracy

The results of the multilevel model on verbal accuracy are interesting, as they appear to be opposite to the results for global syntactic complexity and phrasal

complexity. That is, there was significance for time (p=0.021902) and the interaction of time and task type (p=0.001987). Task type (p=0.757147) alone is not significant. The results show a small increase in verbal accuracy over time within the *discussion* category, and decreases over time for all other task types. This means that, as a group, the participants became more verbally accurate in *discussion* tasks over time but decreased in verbal accuracy over time in all other task types, but that there were no differences in verbal accuracy between task types when time was not under consideration.

To summarize, the data showed that task type and time interact only for verbal accuracy. Verbal complexity, global syntactic complexity, and phrasal complexity showed a significant effect of task type, meaning that there were differences across tasks in the performance of the three participants as a group on these three measures. Overall measures of verbal complexity, global syntactic complexity and phrasal complexity were higher for *discussion* and *thinkplay* and lower for *listing* and *logic*.

Regarding time, only verbal complexity and verbal accuracy showed a significant increase over time. Global syntactic complexity and phrasal complexity were not significantly affected by time. In addition, there was an interaction of task type and time for verbal accuracy, with an increase in verbal accuracy across time only found in *discussion*, and decreases in verbal accuracy across time in all other task types. As previously discussed, this could be because students tended to over rely on the simple present in *discussion*, especially the third person singular *es*, "he/she/it is" and *tiene* "he/she/it has", and the third person plural of *son*, "they are" and *tienen*, "they have." This change could also be partially attributed to the fact that *discussion* tasks were the

most frequent tasks in the data set—the students participated in at least one *discussion* task per class period, while the other task types were less common. This result could also be indicative of students attempting to use other verb forms, such as the subjunctive, in other task categories. These more advanced structures would not be consistently used in an accurate way by an intermediate speaker of Spanish. In other words, perhaps the students were becoming more "adventurous" in their choice of verb structures in the other categories, while staying with a smaller range of structures for *discussion*, causing a decrease in accuracy in all other categories. However, this is conjecture and warrants further investigation in the future.

Now that effects of task and the interaction of task and time have been discussed, the next section will describe the results of the investigation of the participants' development over time.

RQ2: Does oral linguistic complexity and accuracy change over time?

When studying change over time, one is looking at how language varies over time. Specifically, within the field of SLA, one is looking at variation and acquisition processes. As Larsen-Freeman and Cameron (2008a, 2008b) and Larsen-Freeman (2014) point out, language learning is a complex, dynamic system in which the system (the language) is inextricably tied to the context. That is, the context is not a discrete factor that should be investigated as such, but rather something that interacts constantly with the system as it develops. This interaction over time is what both drives development and causes the variation that is seen in the production of the L2 over time (Larsen-Freeman, 2014; Larsen-Freeman & Cameron, 2008a, 2008b; Polat & Kim, 2014). According to

Larsen-Freeman and Cameron, a dynamic system is one that, "changes with time, and whose future state depends in some way on its present state" (p. 29). This change over time and interdependence of the factors in language production causes data to take a variable trajectory in the development of the L2 rather than a linear trajectory that is simply additive. This varied trajectory illustrates how the system is reorganizing itself over time, since, according to Larsen-Freeman (2009), each time a student undertakes an activity, she is starting from a point of changed experience/changed system from the previous engagement in similar activities. This difference in starting point means that language development is uneven and proceeds at multiple rates simultaneously (Larsen-Freeman & Cameron, 2008a, 2008b, p. 138). A student in the process of learning an L2 (or L3, etc.) will experience stages of stability, stages of regression, and stages of development, and this is the sort of variation that is of interest in the current study.

Further, Larsen-Freeman and Cameron (2008a, 2008b), Larsen-Freeman (2009), and Larsen-Freeman (2014) argue that when averages alone are used to describe how languages are learned, the individual is missing from the picture, saying, "It is well known that group averages can conceal a great deal of variability" (Larsen-Freeman & Cameron, 2008a, p. 145). They argue for the detailed and nuanced description of the individual trajectory to help describe the process of the development of the L2. That is not to say that these generalizations are not needed or of value, but that the individual is also important because this type of fine-grained analysis is necessary to help describe how the different components of the system are interacting in the development of the L2, writing, "Group data may often describe a process, or a functional relation, that has no

validity to any individual" (Larsen-Freeman & Cameron, 2008a, p. 145). With that in mind, statistical analyses were conducted, but descriptive figures were also created to show the trajectories of the students as they progressed through the academic year, and how the different variables changed over time. The results show that the students' linguistic output did, indeed, vary over time. The following sections will discuss the results of the Pearson's Correlations and the developmental trajectories of each student.

First, the results of the Pearson's Correlations will be discussed, and then the developmental trajectories will be discussed.

Global Syntactic Complexity

For all participants, as time increases, the number of clauses per AS Unit decreases, but this is not statistically significant or a meaningful change (-0.011).

Mike. Mike did not show a statistically significant correlation between time and number of clauses per AS unit, but he did show a very slight, increase in average number of clauses per AS unit. Although Mike participated much more in his Spanish class during the spring semester, this was not a meaningful increase, so it could be considered that Mike remained relatively stable across time in the average number of clauses per AS unit uttered.

Rebecca. Rebecca also did not show a statistically significant change in number of clauses per AS unit over time, but the trend was downward. That is, she showed a slight decrease of number of clauses AS units as time went on. Again, though, as with Mike, this was not a meaningful change, so she stayed basically stable over the course of the academic year with respect to the number of clauses per AS unit.

Teresa. Teresa, too, did not have a statistically significant finding for average number of clauses per AS unit over time, but she, like Mike, showed a slight increase in number of clauses per AS unit and, like Mike, she became more talkative during the spring semester in comparison with the fall semester. However, again, this did not result in a meaningful increase in average number of clauses per AS unit.

Phrasal Complexity

The Pearson's Correlation for all participants showed a very slight decrease in average number of words per clause as a function of time. This result is not statistically significant and means that they remained relatively stable over time with respect to number of words per clause.

Mike. Mike is the only one who showed a slight increase in phrasal complexity over time, but this was not statistically significant, meaning that he remained mostly stable over the course of the academic year.

Rebecca. Rebecca's phrasal complexity followed the group trend and decreased slightly over time, but this was not a significant finding.

Teresa. Like Rebecca, Teresa exhibited a slight decrease in average number of words per clause over time, but, as with the other two participants, this was not a significant or meaningful decrease.

Verbal Complexity

In the aggregated group data, there was a positive and statistically significant finding for the number of different types of verbal structures. As time increased, so did the number of different verbal structures the 3 students used as a group. Again, it is

important to note that this is not an accuracy measure, just that different verbal structures were attempted. However, when the data were decomposed into the individual students, the results varied slightly.

Mike. Mike showed a positive and statistically significant increase in the number of different verbal structures he used as time elapsed. This follows for two reasons: first, he did not create as much language during fall semester as he did in spring semester, and second, he was presented with more types of verbal structures as the year progressed and was asked to use those structures.

Rebecca. Rebecca also showed a positive and statistically significant increase in verbal complexity as time elapsed. She talked about the same amount during fall and spring semester but, like Mike, she was presented with more structures as the year elapsed and asked to use them.

Teresa. Teresa also showed a positive relationship between time and number of verbal structures produced. That is, she produced more types of verbal structures as the year progressed, but this was not a statistically significant result. This could be because she had the fewest days of data collection out of the 3 students, resulting in insufficient data for a significant difference in the analysis.

Verbal Accuracy

In the group data, the Pearson's Correlation showed that, for all participants, as time increases, verbal accuracy decreases, and this was a statistically significant finding. However, when the data were decomposed into the individual students, the significance disappeared in two cases.

Mike. Mike showed an inverse correlation between time and verbal accuracy that was statistically significant. That is, as time increased, verbal accuracy decreased. A possible reason for the increase in verbal complexity is the introduction of new verbal structures, such as the imperfect subjunctive, which can be late to be acquired (Montrul, 2008; Silva-Corvalán, 1996). These structures would not be used correctly when they were first introduced, and possibly not until much later.

Rebecca. Rebecca, too, showed a decrease in verbal accuracy over time; however, her result was not significant. This is an interesting finding because Rebecca did have a statistically significant finding for verbal complexity, so she showed a significant increase in the number of verb types used but not a significant decrease in verbal accuracy over time. This may be due to Rebecca's desire to sound "good" when speaking Spanish,[22] and this motivation could have translated into her performance in that she may have waited to use new verbal structures until she was sure they were "correct."

Teresa. Teresa also did not show a statistically significant relationship between verbal accuracy and time, though she, like Rebecca, did show a slight decrease in accuracy. However, unlike Rebecca, Teresa did not have a statistically significant increase in verbal complexity over time, which, again, could be attributable to the low number of data collection points as well as the low amount of data from the fall semester.

[22] Please see chapter 3, where each participant was described thoroughly with regard to her or his attitudes toward speaking Spanish and toward different types of interaction.

Accuracy per 100 Words

The group averages of the data on number of errors per 100 words showed that, as time increases, number of errors per 100 words decreases. That is, as a group, the 3 participants became more accurate as time passed. These errors were more than just verbal accuracy. An error could be a gender error, a number error, a pronunciation error, or a word choice error, for example. This was meant to be a more global measure of accuracy than the verbal accuracy measure. Once again, the decomposition of these data into individuals showed differences in effects.

Mike. Mike's data showed that as time increases, average number of errors per 100 words increases, but this increase is not significant. This increase in errors could be attributed to the fact that he was making more of an effort to be talkative in class. It follows that if one is attempting to produce more language, then there is a chance for more errors per word as a result of the attempt.

Rebecca. Unlike Mike, Rebecca's results were statistically significant. As time increases, number of errors per 100 words decreases. Rebecca was the most consistent student in terms of amount of language produced across the academic year and the participation in daily class activities. The fact that she was consistently interacting in the classroom, coupled with her desire to do well, earn a good grade, and her interest in speaking Spanish well, may have contributed to the decrease in number of errors per 100 words, but it cannot be asserted with certainty that this is why she improved in global accuracy over time.

Teresa. Teresa, similarly to Rebecca, showed a decrease in number of errors per 100 words as time increases, and this is a statistically significant finding. Teresa is a special case in that she is a heritage speaker of Spanish and spent a week in Argentina with family during spring break. In addition, she was feeling more positive about her relationship with her mother, a native Argentine, and her own Spanish after receiving compliments from Argentine family members about how good her Spanish was getting. She was highly motivated to sound "good" when speaking Spanish, and exhibited some dialectal features of Argentine Spanish such as /s/ aspiration and voseo. She also expressed, in her interviews with the researcher, that she really enjoyed sounding like the native speakers around her. For example, she attempted to use the north-central dialect of Spanish when her class took a short trip to Spain in high school. Her newfound confidence could have been a factor in her increase in global accuracy. According to de Bot, Larsen-Freeman, Verspoor and Lowie (2011), everything within a system is constantly reorganizing, and so even a change in attitude, toward one's self, toward the language, feelings of confidence or lack thereof will affect the other elements in a system and cause changes in the output. This means that her more positive self-assessment of her language skills may have thus prompted this improvement in performance.

Discussion of Developmental Trajectories

Now that the statistical analyses have been presented and explained, the development of the variables and interaction over time will be discussed. As previously mentioned, any change in the system, be that an introduction of a new grammatical form,

a change in attitude or motivation, or even state of wellness can affect change in the performance in the L2 (de Bot, Larsen-Freeman et al., 2011).

Mike. In Figure 5, there appear to be a few interactions that are taking place in Mike's development of linguistic complexity and accuracy. When looking at the plot for verbal complexity (variety of verbal structures), it appears to follow a similar path as the number of errors per 100 words. That is, when one peaks, generally the other appears to also peak. However, average number of words per clause, or phrasal complexity, seems to follow the verbal complexity plot even more closely. This is an interesting finding, but not entirely surprising; many verbal structures, such as present perfect, are compound structures that include a helping or auxiliary verb, which would increase the number of words in that clause. Average number of clauses per AS unit, the measure of global complexity, also seems to follow the same path as verbal complexity and phrasal complexity.

Verbal accuracy appears to be a much more stable value, with less variation, but at time point 7, there appears to be a dip in verbal accuracy down to 74%, with a concurrent increase in verbal complexity. This date corresponds to the day he gave an oral presentation at the end of his fall semester class. These data may be an anomaly in Mike's performance, since he was largely reading off his computer and the PowerPoint presentation he was using. He had written a script before the presentation, and this was artificially inflating his verbal complexity scores, but since he was trying to use structures he had not yet mastered, his verbal accuracy scores were lower than would be expected if he were using structures over which he had better control.

Another interesting data point occurred on time point 13, which corresponded with his last *mesa redonda* ("round table") activity, in which he described the contents of his composition with several groups of classmates. This composition dealt with his last trip to a local theme park, and his experience riding a roller coaster after having been afraid of heights for many years. Mike was really trying to inject a lot of drama into his narration, with shorter, exclamative sentences and gestures that would account for the dip in global complexity and phrasal complexity, but the fact that he was narrating things in the past, accounts for the increase in verbal complexity and the decrease in verbal accuracy (64%) and increase in number of errors per 100 words. The students were not supposed to read their compositions verbatim for this activity, which took up the whole class, and from the class observations as well as the data for this date, it is apparent that he was not reading his composition to his classmates. He was engaged, animated, and very excited to be telling this dramatic story about his first roller coaster ride. See example (14), an excerpt from this activity.

(14)

Mike: Um. And then, y, y entonces, entonces, like, it stops in midair, it was like, Oh my God! And then like, (clears throat) Sorry. It's it's emotional. (laughs)
Female student: No, it's okay.
Mike: So, so, um, la car, la carta, pausa en, en el medio. Y, y baja! rápidamente. Y, y es, estaba muy, emocionado porque, el, el aire just (makes whoosh sound),
Female student: Yeah!
Mike: y, y um, tu cuerpo, se sienten como like it's just. (sharp inhale) You know?
Female student: Yeah!
Mike: Like no hay, no hay, um, like, gravity? Or.

Mike: Um. And then, and, and then, then, like, it stops in midair. It was like, oh my god! And then like, (clears throat) Sorry. It's it's emotional. (laughs)
Female student: No, it's okay.
Mike: So, so, um, the ca[sic], the car, pauses in, in the middle. And, and drops! Rapidly. And, and it's, I was very, excited because, the, the air just (makes whoosh sound),
Female student: Yeah!
Mike: and, and um, your body, they feel like like it's just. (sharp inhale) You know?
Female student: Yeah!
Mike: Like there isn't, there isn't, um, like, gravity? Or.

Rebecca. Rebecca's data, graphed in Figure 6, are a bit different from Mike's data, and are suggestive of an inverse relationship between global accuracy and verbal complexity and phrasal complexity. The trend for global accuracy, or number of errors per 100 words, appears to decrease over time, while phrasal and verbal complexity seem to be trending upward. Verbal accuracy appears mostly stable, as does global complexity.

There are a few interesting data points in Figure 6. The first is the difference between time point 7 and 8. Time point 7 corresponds with Rebecca's oral presentation. The biggest difference between Mike's oral presentation and Rebecca's is that Rebecca did not have access to her computer: another student was controlling the laptop that was displaying the PowerPoint. Additionally, her PowerPoint presentation was extremely limited in the amount of text that could be included. Most of the slides were images from Perú, since her topic was the Japanese influence in Perú, and she had only a small note card to help her. She, like Mike, also had a jump in global accuracy, verbal complexity, and phrasal complexity the day of her presentation. Her accuracy was similar to Mike's at 75%. However, Rebecca had more errors per 100 words than Mike, at about 17 for his almost 12.

Contrasting time point 7 with time point 8, Rebecca's average number of errors per 100 words drops down to 8.5, her phrasal complexity and verbal complexity also drop, and her verbal accuracy increases slightly to 84%. This is notable because this was the first class that was recorded after the winter break, 2 weeks into the new semester. The tasks in which Rebecca participated that day were *listing*, *discussion*, and *thinkpairshare*. Most of the language she produced was limited to the present indicative and shorter utterances, which would account for the drops in verbal and phrasal complexity and the relative increase in verbal accuracy.

Teresa. Figure 7 displays the graphical representation of Teresa's developmental trajectory of linguistic complexity and accuracy. Teresa shows trends that are similar to Rebecca's, in that she seems to be decreasing in number of errors per 100 words as time increases, meaning her global accuracy increases over time. Verbal and phrasal complexity also appear to be increasing over time as seen in the trend in the plots for both of those variables. Like the 3 other participants, her global accuracy, or average number of clauses per AS unit appears to be somewhat stable over time. For Teresa, too, it appears that verbal and phrasal complexity seem to follow similar trajectories over time. The global accuracy plot is less clear than Rebecca's or Mike's, making it difficult to state whether there is a trend to decrease number of errors per 100 words over time or if this is somewhat stable for her at this time.

Much like Mike and Rebecca, Teresa also has some interesting data points that stand out. Time point 5 corresponds to the oral presentation that Teresa gave at the end of fall semester. She and Rebecca presented on the same day, but Teresa had had several

absences before that day, and was visibly less comfortable and less practiced than Rebecca was. Teresa was the last in her group to present, and also only had the use of a small note card to help her. Her verbal accuracy is the same as Rebecca's at 75%, and even though her number of errors per 100 words also jumped up to almost 9 errors per 100 words, it was much lower than Rebecca's score of almost 17 errors.

The next data point in Teresa's graph shows the values for an individual recording in December of 2012, where Teresa was directed to discuss the situation of immigration in Spain, and compare it to a movie, "Return to Hansala" that her class watched. Her number of errors per 100 words jumped up to almost 18, her variety of verbal structures jumped slightly from time point 5, while the number of words per clause, or phrasal complexity, dipped slightly. Her accuracy dipped to the lowest value of her entire data set: 64%. Teresa seemed exasperated during this task, frequently sighing, starting over, and repeating herself. She appeared to be frustrated greatly by this task, perhaps because she was trying to say very complicated things about a social problem in Spain, comparing it to representations of a similar narrative in the movie. See the excerpt from this recording in (15).

(15)

Mientras El Estrecho, es la vía más pupa-popular, para inmigrante ilegales, hay otros, métodos (big sigh) para las personas pueden ach, accesar, España. En una situación? Personas utilizar escalararas? Rudimentarios. Para, exit- para accesar a es, la costa de España. En muchos, en muchos cosas, es muy peligroso. Pero, las personas, quieren las oportunidades, de España. Y? La:s, ofre- uh las oportunidades de trabajo?

While the strait, is the most pup-popular route, for illegal immigrant, there are others, methods (big sigh) for the people they can ach, access, Spain. In one situation? People to use staircases? Rudimentary. In order to, exit-

in order to access a es, the coast of Spain. In many, in many things, it is very dangerous. But, the people, they want the opportunities, of Spain. And? The, ofre- uh the opportunities of work?

The next data point, time point 7, shows a subsequent decrease of phrasal and verbal complexity as well as a dip in the number of errors per 100 words. That day was a normal class period during the spring semester, where she was able to work in small groups. The monologic nature of the time point 6 recording task may have been exerting stress on Teresa, since she knew her instructor would be listening to it and possibly using it to evaluate her. Even though this task was not framed as an assignment, Teresa knew it might be used for research. Additionally, according to Skehan's (2009) interpretation of Levelt's model of information processing, the monologic nature of this task would place undue pressure on the Formulator, causing her to be unable to attend to all the features of CAF at once. Either type of pressure could have been the cause for the difficulty Teresa experienced while performing this task.

She was also very concerned, like Rebecca, about sounding "good" in Spanish, perhaps even more so because she was half Argentine and felt pressure to speak Spanish better than she did. The same reasons could be applied to Teresa's time point 1, which was a very similar task: a monologic, recorded task that would be evaluated by the instructor.

The last interesting time point for Teresa is time point 11, the last time point. This was the last oral assessment of the term, and Teresa had the highest score of verbal accuracy all spring semester at 80%, the lowest number of errors per 100 words during spring semester, the highest score for phrasal complexity during spring semester, and the highest verbal complexity score all year. This day, she was very bubbly, excited,

outgoing, and talkative. She was really feeling good and enjoying herself during this oral assessment, offering encouragement to her classmates, "*Sí se puede!*" "yes we can!", offering a lot of backchanneling and positive feedback during the assessment. See the excerpt from this day in (16).

(16)

Teresa: y:, um, al aprender inglés es como, um, adaptar a un nuevo cult-
cult ooh shoot!
All: (laughs)
Teresa: Una cultura! (laughing) Um (laughing) que es mucho más
diferente, de sus.
Male student: mmhm
Teresa: Y el sentido de la, las dos identidades también, um, por ejemplo
con Negi or Ana también, //Ana es un buen, buen ejemplo
Male student: //Sí, es un, sí.
Teresa: porque ella, está en hibridismo también, y su familia, no: le gusta
que ella quiere,
Male student: mmhm
Teresa: um, ir a la escue, a la escuela,

Teresa: And, um, upon learning English it's like, um, adapt to a new cult-
cult ooh shoot!
All: (laughs)
Teresa: A culture! (laughing) um, (laughing) that is very much different,
than their.
Male student: mmhm
Teresa: And the feeling that the, the two identities also, um, for example
with Negi or Ana too, //Ana is a good, good example
Male student: //Yes, she is one, yes.
Teresa: because she, is in hybridism also, and her family, does not like
that she wants,
Male student: mmhm
Teresa: um, to go to the scho, to the school.

Because of all the difficulties she experiences over the course of the academic year, the data on this date perhaps are indicative of her true abilities in Spanish that were partially hidden because of her health issues and the linguistic insecurity she felt at the beginning of the academic year. It is also possible that her trip to Argentina caused a

reorganization (Larsen-Freeman, 2009; Larsen-Freeman & Cameron, 2008a, 2008b) of her Spanish system, which allowed her to perform better. It is impossible to determine from the data, but it is a consideration when looking at her language development over the course of the year.

As can be seen from this discussion, the data show that the answer to RQ2, do linguistic complexity and accuracy change over time?, can be answered in the affirmative. The linguistic output of the three participants does, indeed, change over time. Additionally, the averages do indeed obscure the individual trajectories of the participants of this study, as predicted by Larsen-Freeman (2009) and Larsen-Freeman and Cameron (2008a, 2008b). The varied nature of the results for each student exemplify exactly what previous research has shown: that each student passes along his or her own individual trajectory, and that differences in cognition, emotional or physical state, and context will condition the oral production of linguistic complexity and accuracy (de Bot, 2008; Larsen-Freeman, 2006, 2009, 2012, 2014; Larsen-Freeman & Cameron, 2008a, 2008b; Polat & Kim, 2014; van Geert, 2008). It has also been shown that each student has advances and regressions, indicating that language learning does not travel along a linear path. These data provide further evidence that language development is uneven and proceeds at multiple rates simultaneously (Larsen-Freeman & Cameron, 2008a, p. 138), but that there is a general trajectory that can be seen despite this variation in production.

As with RQ1 and RQ1a, the variation in operationalization in the previous research on the complexity and accuracy constructs (Ellis & Barkhuizen, 2005; Housen & Kuiken, 2009; Housen et al., 2012) as well as the small number of previous studies that

look at CAF in a longitudinal way, create difficulty at the moment of comparison of research outcomes. As previously stated, as research becomes more standardized in how it operationalizes the constructs of CAF, the more comparable the results will be (Foster et al., 2000; Norris & Ortega, 2009), perhaps providing the field of SLA with more information on how students develop linguistic complexity and accuracy. Again, this research is a first step in that standardization. However, as Larsen-Freeman (2009) cautions, care must be taken not to attempt to generalize individual trajectories at the expense of the individual. It must always be kept in mind that, while generalizations are useful, so are the individual trajectories and variation.

Conclusions and Significance of this Study

Task Effects

This study has shown that learners' language will vary in linguistic complexity and accuracy based on the type of task that is being undertaken in the classroom. Those tasks that were more open ended seemed to elicit more complex, longer utterances and are those that require explanation or giving an opinion. Even though the tasks were not coded for complexity for this study, in general, tasks that require elaboration and discussion seem to promote more complex language. Discussion activities seem to correlate with an increase in aspects of their linguistic complexity over time as well as increase their verbal accuracy over time. Metalinguistic tasks also seemed to promote an increase in verbal complexity and a decrease in verbal accuracy. Those tasks that did not elicit longer discourse showed lower scores for linguistic complexity and higher scores for linguistic accuracy. Because the instructors were not consulted about the nature of the

tasks they designed for the classroom, it is impossible to determine what the "task-as-workplan" was, but these results show that tasks that require learners to give an opinion, discuss, or debate some item promoted more engagement with the language, while those that appeared more closed ended promoted less complex, shorter utterances. This is in alignment with Long (1983, 1985, 1996), Varonis and Gass (1985), and Swain (1995), who have all described the importance of interaction in language acquisition. Because Teresa and Mike's production improved when they began to participate more actively in their Spanish classrooms, it is shown that there is some level of importance to the act of interacting—both producing language as well as listening to and negotiating with interlocutor—in the language.

That said, task seems to be exerting an effect similar to what is described by Skehan (1996, 1998, 2003) and Skehan and Foster (1997) in the Trade-Off Hypothesis, which states that attentional resources are limited, resulting in an inability to attend to all the features of CAF when attentional resources are taxed. All three participants showed a tendency toward less accuracy when at least one of the measures of linguistic complexity increased, and this held true across the majority of the task types. Those tasks that seemed to encourage more complex linguistic output thus also seemed to cause a resultant decrease in linguistic accuracy. It should be noted that *discussion* did show an increase over time, but this task category did still result in lower accuracy scores than the other task types. So, while the participants did become more accurate over time, they were still less accurate and more linguistically complex than other tasks.

Even though these tasks were not coded for cognitive complexity, it is interesting to consider that open-ended tasks, especially *discussion*, resulted in more complex language. Because the tasks that were used in this study were not designed using measures of cognitive complexity, no conclusions can be drawn based on how cognitive complexity affected linguistic output. However, a post-hoc view of the data raises the question of whether the use of tasks that were designed specifically to be more cognitively complex would have produced similar results for these three participants. However, the present study wished to capture the language produced in a naturalistic environment, making the manipulation of the tasks before they were administered to the participants inadvisable since that would interfere with the naturalistic environment.

L2 Development

De Bot, Larsen-Freeman, Verspoor and Lowie (2011) and Larsen-Freeman (2014) argue for the use of the term "development" instead of "acquisition" in describing the trajectories of L2 use. This partially stems from the problematic nature of using the monolingual native speaker ideal as the end goal of language learning (Larsen-Freeman, 2014; Ortega, 2010). The monolingual "target" of language learning is erroneous for two reasons: it positions bilinguals as somehow deficient compared to monolinguals (Larsen-Freeman, 2014; Ortega, 2010), and monolinguals themselves experience destabilization of their systems and thus variation (de Bot, Larsen-Freeman et al., 2011; Larsen-Freeman, 2014; Montrul, 2008).

In addition, the question arises as to whether there is an "end-goal" at all when it comes to language learning (Larsen-Freeman, 2014). For this reason, L2 learning should

be approached as the development of the L2 system, since there is no real end point when one can say that the L2 has been learned. Larsen-Freeman (2014)[23], de Bot Larsen-Freeman, Verspoor and Lowie (2011) and Larsen-Freeman and Cameron (2008a, 2008b) assert that the term "acquisition" has a connotation that once a learner "has" some form, she will never lose it again, but, according to their research, this is not true. There are forms that are inherently unstable, as shown for Spanish by Silva-Corvalán in her 1996 book on Spanish/English contact in Los Angeles, California,[24] and the basic tenet of DST is that the system is constantly reorganizing, changing, and affecting and being affected by the components within the system. This means that all parts of the system cannot remain static: by definition, there is variability and change, making it impossible to say that a form has been "acquired." A DST framework describes this variability and change, or development, as a form of iteration in that every time a learner engages with the language, she is engaging with it in a new way, based on all her past experiences using the language. This does not mean that language development is additive; on the contrary, it simply means that every starting point is informed by the past, but is also affected by the state of all the interconnected parts within the system as well as the context in which the language use is situated (Larsen-Freeman & Cameron, 2008a, 2008b).

The results of the figures plotting the data over the course of the study (i.e., Figures 5, 6, and 7) show that the variables studied in this dissertation do indeed vary

---

[23] However, the counter to this argument is that when one is using the monolingual standard, the end point is implied. In addition, the student herself may have some sort of end point in mind as well.

[24] But note that in her most recent work in 2014, she has changed her opinion from that certain forms are unstable to the opinion that certain forms are incompletely acquired. See: http://www.linguisticsociety.org/system/files/abstracts/PlenaryCorvalan.pdf

over time. They also show general tendencies of certain variables to travel together, namely, phrasal complexity and verbal complexity. The data also show that there is an inverse relationship with linguistic complexity, at least on the granular level, and linguistic accuracy, at both the granular and global levels. That is, as phrasal and verbal complexity increase, verbal and global accuracy decrease. Once again, this appears to support Skehan (1996, 1998, 2003) and Skehan and Foster, (1997)'s Trade Off Hypothesis.

Global complexity remained somewhat stable over the course of the academic year, but this could be a related to how global accuracy was coded in the data. Perhaps more variation would have been evident if the type of clauses were also coded, such as nominal, adverbial, or relative, to see if there was variation in the types of clauses used and the number of each in each AS unit.

In terms of the developmental trajectory of linguistic accuracy, it can also be inferred from these figures (Figures 5, 6, and 7) and the Pearson's correlations that the trajectories are moving in a general direction, however slowly or nuanced that may be. For all 3 participants, global accuracy, or average number of errors per 100 words, seems to trend downward, while verbal accuracy, or percent of correctly used verbal structures, increases overall. In other words, the 3 participants generally seem to be becoming more accurate as time goes on.

The developmental trajectory of verbal complexity also seems to follow a general trend for all 3 participants in that verbal complexity appears to be trending upward as time goes on. Global complexity appears to be somewhat stable, but again, this could be a

byproduct of how the clauses were coded, or not coded in the case of types of subordinate clauses that appeared in the data. It could be that this measure was not sensitive enough to show a result for the participants of this study.

Significance of this Study

This study contributes to the field of SLA in a number of ways. It is one of few studies to date to have investigated the production of linguistic complexity and accuracy with naturalistic data, rather than in a laboratory or laboratory-style setting. The analysis used both global and specific measures of linguistic complexity and accuracy, operationalizing these constructs in a way that should facilitate the replicability of the study. Furthermore, the statistical analysis of the data included a newer statistical approach, a, multilevel mixed effects model, to determine whether there were any interactions between time and task type, Finally, the study is one of few to use a Dynamic Systems Framework to reflect upon each student's developmental trajectory over the course of an academic year. This nuanced, multi-faceted analysis of naturalistic Spanish data contributes to our understanding of how task affects the linguistic output of learners of intermediate Spanish, as well as how linguistic complexity and accuracy interact over time in the developmental trajectory of the 3 participants.

Limitations and Future Directions

As with any study, there are limitations. With only 3 participants, and only 1 from the University A context, it is impossible to generalize these findings to the broader population of L2 learners of intermediate Spanish, although as previously mentioned, this was not the goal of the description of individual development. More studies would need

to be conducted with more students to determine whether these results would hold generally across students in different types of learning contexts and with different experiences and backgrounds.

Firstly, as mentioned in the description of the data coding procedures, there were times in which the task included various types of activities, but were coded according to the overarching goal of the task. This action may have missed microlevel task effects that could show some sort of difference in the production of linguistic complexity and/or accuracy. Future studies could consider staging tasks as they are coded so that these differences, if there are any, could be captured in the data.

Furthermore, even though an attempt was made to investigate linguistic complexity and accuracy "organically" (Norris & Ortega, 2009), there are still ways that these data could have measured linguistic complexity and/or accuracy more granularly. For example, in order to investigate the development of different *types* of subordinate clauses, the clauses could have been coded as nominal, adverbial, or relative clauses to further differentiate what was being used and how the use of different types of clauses was developing over time (Verspoor & van Dijk, 2012). As was seen in the discussion of the results of this study, it is apparent that the measure for global complexity used in this study may not have been sensitive enough to account for any changes in global linguistic complexity for these participants. Perhaps the addition of *types* of subordinate clauses would have been more illustrative of the development of global linguistic complexity.

With respect to the results of linguistic accuracy in this study, it has been discussed that each student showed slightly different tendencies across time with respect

to linguistic accuracy. None of the participants was given a proficiency test by the researcher; they were selected as intermediate based on their placement into the first class in the sequence of Intermediate Spanish in their respective places of learning. The differences in accuracy rates and trends across students could be due to proficiency differences, and thus a proficiency test given at the beginning and end of each semester may have provided additional information to help in the interpretation of the results.

The study of L2 performance, as mentioned previously, tends to rely on the monolingual native speaker and/or the written standard of a language in order to determine linguistic complexity and accuracy. Indeed, this is the standard used in the current work. However, there is a problem with this standard or ideal, and that is that the bilingual's language system is not the same as the monolingual's (Larsen-Freeman, 2014; Ortega, 2010). Comparing an L2 learner of a language to a monolingual native speaker of that target language is comparing that learner to something she can never attain. The most she can attain is balanced bilingualism, the reality of which is even under question (Montrul, 2008). This ideal also implies that there is some sort of end-state to language learning, which is not the case (Larsen-Freeman, 2014). Though this topic is not within the scope of the current work, future studies may wish to consider something other than the monolingual native speaker ideal as the model for linguistic accuracy.

Additionally, this study took place over the course of 1 academic year. It is obvious in the results over time that it is quite possible that these students should have been observed for more than 1 academic year. An ideal, but lofty goal would be to study L2 learners' oral production throughout the first 2 years of language instruction to

determine developmental trajectories. A longer study may also elucidate whether the inverse relationship between linguistic complexity and accuracy with increasing cognitive complexity would change as the learners' interlanguage system matures and develops. However, because some of the structures the learners tried to use are much later-learned items, this may take quite a bit of time to achieve.

Since this research has dealt only with oral production, future studies should consider written forms of language as well, to determine whether oral production correlates with the participants' abilities in written expression. The written production of the learners was not available for research in the current study. However, whenever possible, the researcher consulted with the instructors of Mike, Rebecca and Teresa's classes to determine whether trends seen in the oral production matched the students' written language. According to the instructors, in general, all 3 performed similarly in oral and written forms, with slightly better performance in the written form than orally, which is a common finding (Verspoor & van Dijk, 2012). Even so, written data should be included in future studies to triangulate the oral data and to determine whether this medium follows the same developmental patterns as oral language.

The data could also be coded for fluency in order to determine whether the addition of this variable would further support or not either of the cognitive models of the production of CAF by learners.

Lastly, Mike, in particular, tended to code-switch frequently between Spanish and English. Along these lines, Mike used English and Spanish in different ways and for different purposes, some of which was for play or joking. A detailed look at code-

switching in the oral Spanish of these learners could provide insight into how intermediate learners of Spanish choose to utilize code-switching to maintain fluency or express meaning in their oral discourse.

Future studies that attempt to replicate the current study could take a variety of forms. First, an entire class could be studied, allowing the researcher more access to the class dynamic and interactions among the students in order to document the development of tutored Spanish. A longitudinal, fine-grained analysis of the development of Spanish would also provide a better picture of what the trajectories of this development look like for intermediate Spanish learners. Despite objections to generalizations (Larsen-Freeman & Cameron, 2008a, 2008b; Verspoor & van Dijk, 2012), they are useful and can give a better picture of how various factors interact during L2 development for a larger number of students. While the granularity would be lost, these averages may show general trends in trajectories that would help researchers determine what paths students take generally in L2 development.

Lastly, it is important to include heritage language learners in any future studies, because they may develop differently than L2 learners (Carreira & Potowski, 2011; Lynch, 2008). A more cognitively complex classroom may be beneficial for heritage language learners as it's been shown by Potowski (2003, 2004) that these types of learners can benefit from language instruction that is more akin to an English Language Arts class that L2 learners of Spanish have taken in their L1. For heritage language learners, the communicative classroom can feel repetitive, simplistic, and boring. Because there are often not enough heritage language learners in many geographical

areas to warrant a specific section of Spanish (or another language) for heritage learners, a change to a format that is more similar to a content-based class may provide a better environment for heritage language learners to gain literacy skills and/or improve their heritage language production.

Final Summary

This dissertation investigated the oral Spanish production of 3 intermediate learners of Spanish over the course of 1 academic year. The language of the students was analyzed for effects of task; that is, whether certain tasks seem to elicit more or less complex or accurate language. Their language was also analyzed as a function of time in order to look at their developmental trajectories and to determine how their linguistic complexity and accuracy vary over time.

The results showed that these participants did perform differently in different kinds of tasks. When engaged in more open-ended tasks, such as *discussion* and *thinkplay*, students tended to produce language with greater linguistic complexity generally. In regards to verbal accuracy, in the production of two students, Rebecca and Teresa, there were no differences according to task. However, Mike's production did reflect greater verbal accuracy in *discussion* and *thinkplay*. With respect to measures of complexity, metalinguistic tasks tended to show low scores for global linguistic complexity, while showing higher scores for verbal complexity, which follows since a majority of the metalinguistic tasks engaged in by the participants were focused on verbal forms such as the difference between the indicative and the subjunctive. More closed tasks that asked students to search for words or phrases, to match items, or create lists

tended to produce lower scores for linguistic complexity, both global and verbal, because there was an overreliance on the present indicative in those contexts, while verbal accuracy was not significantly affected. So, while task did seem to matter in the production of oral linguistic complexity and accuracy, the factors that were significant varied by student. Task did not affect production in the same way for all the students.

Longitudinally, the graphic presentation of the data showed that, even though these results were not always statistically significant, there was a general trend of overall increase in linguistic complexity, in verbal complexity and phrasal complexity, while global complexity remained relatively stable over time. There was an inverse relationship with verbal complexity and verbal accuracy, at the granular level of percent of correctly used verbal structures. That is, as linguistic complexity increases, linguistic accuracy decreases.

Future analysis of this type of data with different coding, such as differentiation of types of clauses, or addition of more factors, such as affective factors or interlocutor, may show that other factors are also conditioning the results. Even though the results of this analysis are not generalizable to L2 learners of Spanish due to the small sample size, it is a beginning step in the analysis of what sorts of activities may promote more linguistic complexity and accuracy and what developmental trajectory linguistic complexity and accuracy take over time. More studies are needed to determine whether these results are replicable with a larger population of students and in different locations.

References

ACTFL Proficiency Guidelines (2012). Retrieved from: http://actflproficiencyguidelines2012.org/speaking

Ågren, M., Granfeldt, J., & Schlyter, S. (2012). The growth of complexity and accuracy in L2 French: Past observations and recent applications of developmental stages. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 95–120). Amsterdam, Netherlands: John Benjamins.

Andersen, R. (1989). La adquisición de la morfología verbal. *Lingüística, 1*, 90–142.

Andersen, R. (1991). Developmental sequences: The emergence of aspect marking in second language acquisition. In T. Huebner & C.A. Ferguson (Eds.), *Crosscurrents in second language acquisition and linguistic theories* (pp. 305–324). Amsterdam, Netherlands: John Benjamins.

Andersen, R., & Shirai, Y. (1996). The primacy of aspect in first and second language acquisition: The pidgin-creole connection. In B. Laufer & W. Ritchie (Eds.), *Handbook of second language acquisition* (pp. 527–570). San Diego, CA: Academic Press.

Bardovi-Harlig, K. (1992). The relationship of form and meaning: A cross-sectional study of tense and aspect in the interlanguage of learners of English as a second language. *Applied Psycholinguistics, 13*, 253–278.

Bardovi-Harlig, K. (1995). A narrative perspective on the developing of the tense/aspect system in second language acquisition. *Studies in Second Language Acquisition, 17,* 263–289.

Barnes-Karol, G. (2010). Reading (literature) in, across, and beyond the undergraduate Spanish curriculum. *Hispania*, *93*(1), 90–95.

Barnes-Karol, G., & Broner, M. A. (2010). Using images as springboards to teach cultural perspectives in light of the ideals of the MLA report. *Foreign Language Annals*, *43*(3), 422–445.

Brown, G., Anderson, A., Shillcock, R., & Yule, G. (1984). *Teaching talk*. Cambridge, England: CUP.

Bulté, B., & Housen, A. (2012). Defining and operationalising L2 complexity. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 21–46). Amsterdam, Netherlands: John Benjamins.

Carreira, M., & Potowski, K. (2011). Commentary: Pedagogical implications of experimental SNS research. *Heritage Language Journal*, *8*(1), 134–151.

de Bot, K. (2008). Introduction: Second language development as a dynamic process. *The Modern Language Journal*, *92*(2), 166–178.

de Bot, K., Larsen-Freeman, D., Verspoor, M., & Lowie, W. (2011). Researching second language development from a dynamic systems theory perspective. In M. Verspoor, K. de Bot, & W. Lowie (Eds.), *A dynamic approach to second*

language development. Methods and techniques (pp. 5–23). Amsterdam, Netherlands: John Benjamins.

de Jong, N. H., Steinel, M. P., Florijn, A., Schoonen, R., & Hulstijn, J. H. (2012). The effect of task complexity on functional adequacy, fluency, and lexical diversity in speaking performances of native and non-native speakers. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 121–142). Amsterdam, Netherlands: John Benjamins.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford, England: Oxford University Press.

Ellis, R., & Barkhuizen, G. (2005). *Analysing learner language*. Oxford, England: Oxford University Press.

Foster, P., Tonkyn, A., & Wigglesworth, G. (2000). Measuring spoken language : A unit for all reasons. *Applied Linguistics*, *21*(3), 354–376.

Gudmestad, A., & Geeslin, K. (2012). Second language development of variable future-time expression in Spanish. *Selected proceedings of the 6th international workshop on Spanish sociolinguistics*. Somerville, MA: Cascadilla Press.

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning*, *24*, 287–297. doi:10.1111/j.1467-1770.1974.tb00509.x

Hakuta, K. (1976). A case study of a Japanese child learning English as a second language. *Language Learning*, *26*(2), 321–351.

Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 401–435). Rowley, MA: Newbury House.

Hatch, E. (1992). *Discourse and language education*. Cambridge, England: Cambridge University Press.

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

Housen, A., Kuiken, F., & Vedder, I. (2012). Complexity, accuracy and fluency: Defininitions, measurement and research. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 1–20). Amsterdam, Netherlands: John Benjamins.

Kormos, J. (2006). *Speech production and second language acquisition.* Mahwah, NJ: Lawrence Erlbaum.

Kuiken, F., & Vedder, I. (2012). Syntactic complexity, lexical variation and accuracy as a function of task complexity and proficiency level in L2 writing and speaking. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 143–170). Amsterdam, Netherlands: John Benjamins.

Larsen-Freeman, D. (2006). The emergence of complexity, fluency, and accuracy in the oral and written production of five Chinese learners of English. *Applied Linguistics*, *27*(4), 590–619.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 579–589.

Larsen-Freeman, D. (2011). Complex, dynamic systems: A new transdisciplinary theme for applied linguistics? *Language Teaching, 45*(2), 202–214.

Larsen-Freeman, D. (2012). Complexity theory. In S. Gass & A. Mackey (Eds.), *Routledge handbooks in applied linguistics: Routledge handbook of second language acquisition* (pp. 73–87). Florence, KY: Routledge.

Larsen-Freeman, D. (2014). Saying what we mean: Making a case for 'language acquisition' to become 'language development.' *Language Teaching.* doi:10.1017/S0261444814000019

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

Larsen-Freeman, D., & Cameron, L. (2008b). Research methodology on language development from a complex systems perspective. *Modern Language Journal, 92*(2), 200–213.

Levelt, W. J. M. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Liskin-Gasparro, J. (2000). The use of tense-aspect morphology in Spanish oral narratives: Exploring the perceptions of advanced learners. *Hispania, 83*(4), 830–844.

Long, M. (1983). Linguistics and conversational adjustments to nonnative speakers. *Studies in Second Language Acquisition, 5*, 177–193.

Long, M. (1985). A role for instruction in second language acquisition: Task-based language training. In K. Hyltenstam & M. Pienemann (Eds.), *Modelling and assessing second language acquisition* (pp. 77–99). Clevedon, England: Multilingual Matters.

Long, M. H. (1985). Input and second language acquisition theory. In S. M. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 377–393). Rowley, MA: Newbury House.

Long, M. H. (1996). The role of the linguistic environment in second language acquisition. In W. Ritchie & T. Bhatia (Eds.), *Handbook of second language acquisition* (pp. 413–468). New York, NY: Academic Press.

Lynch, A. (2008). The linguistic similarities of Spanish heritage and second language learners. *Foreign Language Annals*, *41*, 252–281.

Mackey, A. (2012). *Input, interaction and corrective feedback in L2 learning*. Oxford, England: Oxford University Press.

Malvern, D., & Richards, B. (2000). Validation of a new measure of lexical diversity. In M. Beers, B. van den Bogaerde, G. Bol, J. de Jong, & C. Rooijmans (Eds.), *From sound to sentence: Studies on first language acquisition* (pp. 81–96). Groningen, Netherlands: Centre for Language and Cognition.

Malvern, D., & Richards, B. (2009). A new method of measuring rare word diversity: The example of L2 learners of French. In B. Richards, M. Daller, D. Malvern, P. Meara, J. Milton, & J. Treffers-Daller (Eds.), *Vocabulary studies in first and*

second language acquisition: The interface between theory and application (pp. 164–78). Basingstoke, England: Palgrave Macmillan.

Malvern, D., & Richards, B. (2012). Measures of lexical richness. In C. A. Chapelle (Ed.), *The encyclopedia of applied linguistics* (pp. 1–5). Chichester, England: Blackwell. doi:10.1002/9781405198431.wbeal0755

Malvern, D., Richards, B., Chipere, N., & Durán, P. (2004). *Lexical diversity and language development: Quantification and assessment*. Basingstoke, England: Palgrave Macmillan.

Markee, N. (2000). *Conversation analysis.* New York, NY: Routledge.

Montrul, S. (2008). *Incomplete acquisition in bilingualism: Re-examining the age factor*. Amsterdam, Netherlands: John Benjamins.

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. doi:10.1093/applin/amp044

Ortega, L. (2010, March). The bilingual turn in SLA. In *Plenary delivered at the annual conference of the American Association for Applied Linguistics,* Atlanta, GA (pp. 6–9).

Ortega, L., & Byrnes, H. (2008). *The longitudinal study of advanced L2 capacities.* New York, NY: Routledge.

Pallotti, G. (2009). CAF: Defining, refining and differentiating constructs. *Applied Linguistics*, *30*(4), 590–601. doi:10.1093/applin/amp045

Pica, T., Kanagy, R., & Falodun, J. (1993). Choosing and using communication tasks for second language research and instruction. In S. M. Gass & G. Crookes (Eds.), *Task based learning in a second language* (pp. 9–34). Clevedon, England: Multilingual Matters.

Polat, B., & Kim, Y. (2014). Dynamics of complexity and accuracy: A longitudinal case study of advanced untutored development. *Applied Linguistics, 35*(2), 184–207. doi:10.1093/applin/amt013

Potowski, K. (2003). Chicago's *Heritage Language Teacher Corps*: A model for improving Spanish teacher development. *Hispania*, *86*(2), 302–311.

Potowski, K. (2004). Student Spanish use and investment in a dual immersion classroom: Implications for second language acquisition and heritage language maintenance. *Modern Language Journal*, *88*(1), 75–101.

Ritchey, F. (2008). *The statistical imagination* (2nd ed.). Boston, MA: McGraw Hill.

Robinson, P. (2001a). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22*(1), 27–57.

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

Robinson, P. (2003). The cognition hypothesis, task design and adult task-based language learning. *Second Language Studies, 21*(2), 45–107.

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

Robinson, P. (2007). Task complexity, theory of mind, and intentional reasoning: Effects on L2 speech production, interaction, uptake and perceptions of task difficulty. *International Review of Applied Linguistics in Language Teaching (IRAL), 45*(3), 193–213.

Robinson, P., Cadierno, T., & Shirai, Y. (2009). Time and motion: Measuring the effects of the conceptual demands of tasks on second language speech production. *Applied Linguistics, 30*(4), 533–554. doi:10.1093/applin/amp046

Robinson, P., & Gilabert, R. (2007). Introduction: Task complexity, the cognition hypothesis, language learning, and performance. *International Review of Applied Linguistics in Language Teaching (IRAL), 45*(3), 161–177.

Salaberry, M. R. (1999). The development of past tense verbal morphology in classroom L2 Spanish. *Applied Linguistics, 20*(2), 151–178.

Salaberry, R. (2002). Tense and aspect in the selection of Spanish past tense verbal morphology. In R. Salaberry & Y. Shirai (Eds.), *The L2 acquisition of tense-aspect morphology* (pp. 397–415). Amsterdam, Netherlands: John Benjamins.

Salaberry, R. (2003). Tense aspect in verbal morphology. *Hispania, 86*(3), 559–573.

Salaberry, M. R. (2011). Assessing the effect of lexical aspect and grounding on the acquisition of L2 Spanish past tense morphology among L1 English speakers. *Bilingualism: Language and Cognition, 14*(2): 184–202.

Salaberry, M. R. (2013). Contrasting preterit and imperfect use among advanced L2 learners: Judgments of iterated eventualities in Spanish. *International Review of Applied Linguistics*. *51*, 243–270.

Schmid, M., & Jarvis, S. (2014). Lexical access and lexical diversity in first language attrition. *Bilingualism: Language and Cognition, 17*(4), 729–748.

Seedhouse, P. (2004). *The interactional architecture of the language classroom: A conversation analysis perspective.* Malden, MA: Blackwell.

Seyyedi, K. (2012). Task-based instruction. *International Journal of Linguistics, 4*(3), 242–251.

Shrum, J. L., & Glisan, E. W. (2009). *Teacher's handbook: Contextualized language instruction*. Boston, MA: Heinle Cengage.

Silva-Corvalán, C. (1996). *Language contact and change*. Oxford, England: Oxford University Press.

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

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford, England: Oxford University Press.

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

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532. doi:10.1093/applin/amp047

Skehan, P., & Foster, P. (1997). Task type and task processing conditions as influences on foreign language performance. *Language Teaching Research, 1*(3), 185–211. doi:10.1177/136216889700100302

Skehan, P. & Foster, P. (2012). Complexity, accuracy and fluency and lexis in task-based performance. In A. Housen, F. Kuiken, & I. Vedder (Eds.), *Dimensions of L2 performance and proficiency: Complexity, accuracy and fluency in SLA* (pp. 199–220). Amsterdam, Netherlands: John Benjamins.

Spoelman, M., & Verspoor, M. (2010). Dynamic patterns in development of accuracy and complexity: A longitudinal case study in the acquisition of Finnish. *Applied Linguistics*, *31*(4), 532–553. doi:10.1093/applin/amq001

Swain, M. (1995). Three functions of output in second language learning. In. G. Cook & G. Seidhofer (Eds.), *Principles and practices in applied linguistics: Studies in honor of HG Widdowson* (pp. 125–144). Oxford, England: Oxford University Press.

van Geert, P. (2008). The dynamic systems approach in the study of L1 and L2 acquisition: An introduction. *The Modern Language Journal*, *92*(2), 179–199.

Varonis, E. M., & Gass, S. (1985). Non-native/non-native conversations: A model for negotiation of meaning. *Applied Linguistics*, *6*(1), 71–90.

Verspoor, M., & van Dijk, M. (2012). Variability in a dynamic systems theory approach to second language acquisition. *The Encyclopedia of Applied Linguistics*. doi:10.1002/9781405198431.wbeal1251

Verspoor, M., Lowie, W., & van Dijk, M. (2008). Variability in second language development from a dynamic systems perspective. *The Modern Language Journal*, *92*(2), 214–231. doi:10.1111/j.1540-4781.2008.00715.x

Willis, D., & Willis, J. (2007). *Doing task-based teaching.* Oxford, England: Oxford University Press.

Appendix A

Interview Questions

What is your favorite thing about learning Spanish? Please give me an example.

With whom do you normally speak Spanish? Please give me an example.

Where do you usually speak Spanish? Please give me an example.

Do you seek out opportunities to speak Spanish? Why/why not? Please give me an example.

Do you consider yourself a competitive person? How do you think that affects you when you are speaking Spanish in class or with friends? Please give me an example.

Do you like a challenge? Does that affect how you interact with your classmates in Spanish classes? Please give me an example.

How do you feel when you are speaking with a native Spanish speaker? Please give me an example.

With a high-proficiency non-native speaker? Please give me an example.

With someone who seems to be the same proficiency as you? Please give me an example.

With a lower proficiency speaker? Please give me an example.

Do you ever feel intimidated by other Spanish speakers, native or non-native? Why/Why not? Please give me an example.

When you are in class, with whom do you prefer to work? (i.e., native speakers, high proficiency classmates, good friends only, doesn't matter) Please explain/give me an example.

Appendix B

Background Questionnaire

Name_____ Age _____ Gender_____ email/x500_____

How long have you been learning Spanish? _____

With whom do you speak Spanish the most? _____

How many hours a week do you speak Spanish? _____

Where does most of this interaction occur? _____

Complete the following:

In general, native Spanish speakers are _____

In general, non-native Spanish speakers are _____

In general, I prefer (native/non-native—circle one) speaking partners because

_____

In general, I prefer (more proficient/less proficient/same proficiency) speaking partners because

_____

Appendix C

Arcsine Transformation

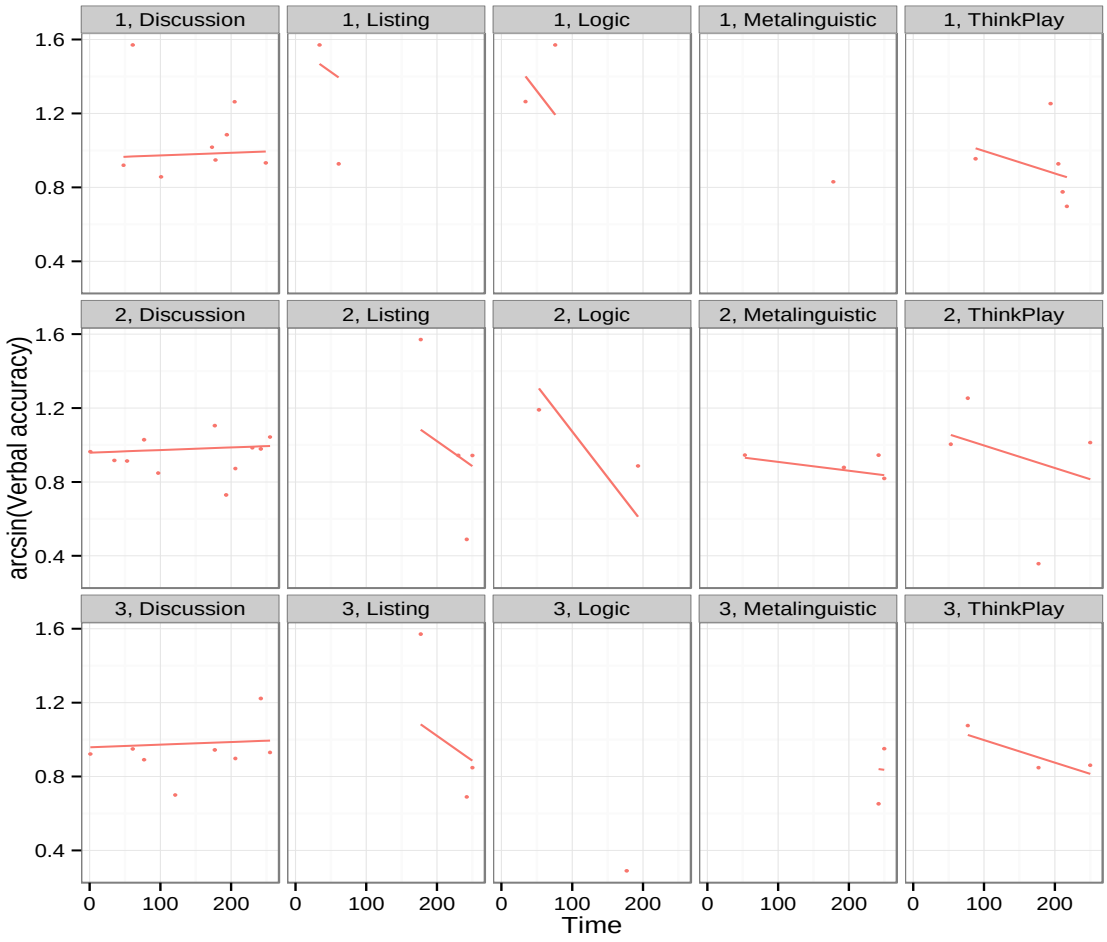

Figure C-1. Fitted values from restricted maximum likelihood estimated maximum likelihood models: Arcsine transformation of verbal accuracy.

Table C-1

Maximum Likelihood Parameter Estimates for a Linear Mixed-Effects Model Describing Changes in Verbal Accuracy (Arcsine Transformation), All Three Participants

Parameter	Estimate (s.e.)	*t*-value	*p*-value
Intercept	0.96 (0.09)	10.46	<.001
Time interval	0.00 (0.00)	0.26	0.797
Listing	0.60 (0.22)	2.74	0.009
Logic	0.61 (0.22)	2.78	0.008
Metalinguistic	0.00 (0.30)	-0.01	0.995
ThinkPlay	0.16 (0.20)	0.80	0.423
Time Interval * Listing	0.00 (0.00)	-2.50	0.016
Time Interval * Logic	-0.01 (0.00)	-3.03	0.004
Time Interval * Metalinguistic	0.00 (0.00)	-0.43	0.669
Time Interval * ThinkPlay	0.00 (0.00)	-1.20	0.235
Variance Components	Chi Square	*df*	*p*-value
Time Interval	6.27	1	0.012
Task Type	4.04	4	0.400
Time Interval * Task Type	16.284	4	0.002