Complexity, Accuracy and Fluency Development through Study Abroad Programmes Varying in Duration

Ann Rebecca Lara

TESI DOCTORAL UPF / 2014

DIRECTORS DE LA TESI

Dra. Carmen Peréz Vidal (Departament de Traducció i Ciències del Llenguatge, Universitat Pompeu Fabra),

Dr. Joan Carles Mora (Departament de Filologia Anglesa i Alemanya, Universitat de Barcelona)

DEPARTAMENT DE TRADUCCIÓ I CIÈNCIES DEL LLENGUATGE

ACKNOWLEDGEMENTS

I would first like to thank my thesis directors Dr. Carmen Pérez-Vidal and Dr. Joan Carles Mora for all of their support and patience with me during this time. Dr. Pérez-Vidal has always made an effort to be available, in spite of an extremely busy schedule. She is one of the most encouraging people I have run across and the best advocate for her students within the university system. I have benefitted greatly from watching her in her many roles, as researcher, professor and administrator. I am also very grateful for the expert insight she had brought to this dissertation. I am very grateful for Dr. Mora's professionalism and rigor, but also for his generosity when it comes to giving of his time, which I know is often scarce. My sincerest thanks go also to Dr. Pilar Avello, who has helped me through this process as much as anyone. I am especially appreciative of her help in data processing and in helping me through the early stages of this work. I would also like to thank Dr. Margalida Valls-Ferrer for her encouragement and helpful advice since day one, as well as Dr. Elisa Barquin, whose respective dissertations have served to guide my own writing to a great degree.

Many thanks also go to Dr. Isabel Tejada and Dr. Carmen del Río for the many times they have inspired me to carry on in this process. Seeing that they were able to persevere in the writing of their theses, always with a positive outlook and a sense of humour, has been a great encouragement to me. I would also like to express my gratitude to my former officemates, Estela García, Mònica Tarrés, Anna Denisenko, and Victoria Zaytseva for making the UPF such a lovely place to work, and for all their support and encouraging words that have meant so much to me.

Many thanks also to Dr. Raquel Serrano for her interest in this work in its very early stages, and for pointing out some very useful references that have added greatly to the dissertation. I would also like to thank Stephanie Cowell, Jessica Ballanti and Dorcas Gonzalez Bataller for their much needed help in the final revision of this work.

I am very thankful for the three-year study grant provided by *l'AGAUR, Generalitat de Catalunya*, which has made the development of this dissertation possible. I would also like to highlight the contribution of the UPF administrators in the *Departament de Traducció i Ciències del Llenguatge* in facilitating every step in the doctorate programme, for their eagerness to help and their expertise, and with special thanks to Susi Bolós and Lali Palet (who has moved on to a new post).

Of course, I am very much indebted to the informants who participated in the SALA research project, and would like to especially thank the students who made up the Long-stay group of subjects of the present work who I had the pleasure of working with personally.

Many thanks are also due to my family. I am very grateful to my parents for instilling in me the value of education, language learning, and a deep interest in other cultures. I am

especially thankful for the prayers and constant support of my mother, Maria Lara, whose own self-discipline and scholarship have inspired me my whole life, and have proven a much needed source of strength during this time. I would also like to thank my husband, Rubén, for all his generous support especially in the final stages of this process. Without his support, both financially and in the many extra hours of caring for our son in my absence, completing this dissertation would have been next to impossible. I am also very grateful to my son, Rubén David, for being as patient as a four-year-old can be with his very preoccupied mother. He has been my greatest inspiration and my most lovable distraction.

Finally, my most heartfelt thanks go to God, for the life-changing opportunity that my time at the UPF has been, and for his ever-present, sustaining hand in this time that has been extremely challenging for me. Without his grace, none of this would have been possible, and I am amazed at the ways in which he has facilitated the whole process, placed the right people at the right time in my path, and provided all the tools I have needed along the way. I am forever humbled and grateful.

Rebecca Lara

Castelldefels, August 2014

ABSTRACT

Study abroad (SA) as an integral part of university studies is often seen by students, educators and institutions alike as an important means of boosting foreign language skills; yet, still little is known about how long of a stay abroad is long enough to see considerable gains in L2 proficiency. The present study explores the L2 oral development of a group of 47 adult learners of English as a foreign language who participated in SA programmes in English-speaking countries that differ in Length of stay (LoS). We compare the progress made by learners who went abroad for a quarter (approximately 3 months) with that of students who spent a semester abroad (approximately 6 months). We also examine the role students' initial proficiency level, and experience abroad prior to the current SA had on linguistic outcomes. Learners' oral production was elicited through a role-play task just prior to, and upon return from SA, and was analysed through measures of syntactic and lexical complexity, accuracy and fluency (CAF). The same task was performed by 24 native speakers of English in order to provide a baseline reference with which to compare learner performances. Results indicate that learners' oral fluency increases considerably during SA. Lexical complexity moves toward more target-like use, and little change at all is observed in measures of syntactic complexity and accuracy as a result of SA. We did not find any compelling evidence to suggest that any one LoS was more beneficial than the other in terms of post-SA outcomes. Initial proficiency level was robustly impactful on post-test outcomes, and previous periods spent abroad only marginally influenced the outcome of the current SA period.

RESUM

Les estades a l'estranger integrades en els estudis universitaris són sovint considerades pels estudiants, docents i institucions acadèmiques com un mitjà per millorar la competència en una segona llengua. Tanmateix, a hores d'ara, es coneix poc sobre la durada mínima que hauria de tenir l'estada per obtenir una millora notable en el domini de la L2. Aquest estudi pretén examinar el desenvolupament oral de la L2 per part d'un grup de 47 adults que aprenen l'anglès com a llengua estrangera. Tots han participat en programes d'intercanvi acadèmic en països de parla anglesa, i de durada diferent. Comparem el progrés dels aprenents que han fet una estada durant un trimestre (3 mesos, aproximadament) amb el dels que han passat un semestre a l'estranger (prop de 6 mesos). També, avaluem quin paper tenen la competència lingüística inicial i

l’experiència a l’estranger prèvia a l’observada en els resultats lingüístics després de l’estada en qüestió. Mostres de la producció oral dels aprenents s’han obtingut mitjançant una tasca en forma de joc de rol que es va dur a terme abans i després de l’estada, i s’han analitzat aplicant mesures de complexitat sintàctica i lèxica, correcció i fluïdesa. 24 natius d’anglès han dut a terme la mateixa tasca que els aprenents per tal de poder comparar les dades de producció de tots dos grups. Els resultats indiquen que la fluïdesa oral en la L2 augmenta considerablement durant l’estada. La complexitat lèxica progressa cap a un ús més proper al dels natius, al contrari del que es destacaen les mesures de complexitat sintàctica i de correcció com a resultat de l’estada. No observem cap evidència ferma per assegurar que una durada sigui més efectiva que l’altra tenint en compte els resultats posteriors a l’estada. El nivell de competència inicial té un impacte destacable en els resultats i l’experiència prèvia al període estudiat influeix mínimament en els resultats de l’estada a l’estranger.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF ABBREVIATIONS AND ACRONYMS

ACTR: American Council of Teachers of Russian
AH: At home; within the learner's home country and/or educational institution.
ALLENCAM: *Grup d'Adquisició de Llengües a la Catalunya Multilingüe*
ASU: Analysis of speech unit
C/ASU: Clauses per AS-unit
CA: Communicative adequacy
CAF: Complexity, Accuracy, and Fluency
CEFR: Common European Framework of Reference for languages
CLAN: Computerized Language ANalysis (software)
COE: Council of Europe
CoordC/ASU: Coordinated clauses per AS-unit
DST: Dynamic Systems Theory
E/ASU: Errors per AS-unit
E/C: Error per clause
EC: European commission
EFL: English as a Foreign Language
EU: European Union
FI: Formal (language) instruction

GRAMSCOR: Composite scores on SALA Grammar and Cloze tests
GuirIndex: Guiraud's index of lexical richness
IDs: Individual (learner) differences
IM: Immersion, referring to a learning context where students use the target language continually during all instruction hours
L1: First/native language(s)
L2: Second/non-native language
MLAT: Modern Language Aptitude Test
NS: Native Speaker
NNS: Non-native Speaker
OPI: Oral Proficiency Interview
PAU: *Proves d'accés a la universitat* – Mandatory university entrance exam
Pruned SR: Pruned speech rate
SA: Study Abroad
SA-3m: Group of participants who went on study abroad programmes of three months
SA-6m: Group of participants who went on study abroad programmes of six months
SALA: Stay Abroad and Language Acquisition
SEM: Structural equation modelling
SLA: Second Language Acquisition
SR: Speech rate
SSLA: *Studies in Second Language Acquisition* (journal)
SubC/ASU: Subordinate clauses per AS-unit
TL: target language, language of study
TU: T-unit
UG: Universal grammar
VOT: voice onset time
WD/ASU: words per AS-unit
WiSP: Netherlands-based research project "What is Speaking Proficiency?"

INTRODUCTION[1]

During the course of each academic year hundreds of thousands of university students worldwide participate in Study Abroad (SA) programmes, temporary stays in a foreign country within the context of education, especially higher education. Young people who seek out an abroad experience do so for a variety of reasons ranging from concrete professional goals to the desire to travel and experience another culture firsthand. Not least among the motivations behind SA is foreign language learning.

The international movement of academics for the purpose of study has been a common practice since ancient times. Some of the most noteworthy intellectuals in history, pioneering thinkers in philosophy, medicine or theology were obliged to study outside of their local context in order to achieve their educational goals. Certainly, the scarcity of academic institutions or educational resources drove the first academic migrations. Of course this type of early cross-border education was reserved for those privileged few who had the time and resources to study in a formal setting far from their homeland. Today stays abroad are increasingly common in higher education. And although the idea of SA is nothing new, what was feasible for only the wealthiest of families or the most talented individuals has more recently been made available to a massive student population across the globe.

The practice of academic migration, that is, a sustained residence in a foreign country for the purpose of obtaining a degree, is still very much alive and well, although with very different motivations, and certain barriers for those who wish to do so. However, there is a particular modality of stay that allows the student to maintain ties with his or her home institution while participating in a temporary abroad experience. This idea of staying connected, even earning credit toward a degree from the home institution while away, is what best distinguishes what we refer to as SA (programmes) from other cross-border education pursuits.

[1] This research received financial support through HUM2004-05442-C02-01, HUM2007-66053-C02-01/02 and FFI2010-21483-C02-01/02 and ALLENCAM (SGR2005-01086/2009-140) from the Spanish Ministry of Education and the Catalan Government respectively.

Like many researchers who have developed an interest in the subfield of SA, my academic interests have been closely linked to my personal experience. Over the last thirteen years, SA has taken on a very significant role in my life, shaping both my personal and professional life in ways I could not have imagined a decade ago. In 2001 I first ventured abroad to study in Spain for a year. I had already spent a year at my home university in the US, and an unplanned hiatus from studies in my second year seemed a good opportunity to advance my language skills in a Spanish-speaking country; to become "fluent" in the language. I set out on my own; unlike many US based SA participants who tend move in packs, with a familiar professor from home to lead the way. I found an accredited programme that would accept unaccompanied students and leapt, unaware, into the unknown. Many of the assumptions and prejudices that accompany young people on their first sojourn abroad also went with me. I certainly expected some aspects of SA to be easier than they ended up being; nothing about SA that I can recall, proved easier than I expected.

Early into my stay I began to encounter many of the aspects of SA that have become well known general tendencies in SA contexts, and in some cases even specific lines of research within the subfield of SA and its effects on second language acquisition (SLA). For the first time, I found myself in a context that was completely foreign to me. My learning was at times intentional and at other times incidental, exciting and motivating but often at the risk of losing face in a land where I was an obvious novice in every way. Over the course of that year I was also introduced to the community of SA in Europe including the institution of Erasmus and the culture that surrounds it. I found that most of the first friendships that I made abroad were not with Spaniards but with Japanese, German or Brazilian exchange students, also new to the city, seeking out like-minded folks with whom to practice their Spanish. At first, I found locals to be uninterested in interacting with me, especially when they noticed my lack of competence in the language. Once the shock of realising that my ability to communicate in Spanish was not what I thought it was, my language skills advanced more slowly than I expected. I often noticed myself struggling in conversation to recall even basic vocabulary items, plugging them into some half-forgotten example phrase I had memorized. I stumbled over my words, mispronounced those phonemes that my vocal tract was not used to producing, committed every grammatical error imaginable and struggled to understand native speakers.

With all the “hardships” I endured during the first months of my stay abroad one might wonder at the fact that I did not catch the next available flight home at some point. Though I did encounter difficulties and disappointment, what I discovered in the process was invaluable to my personal development, not to mention my linguistic development. A new world was opened up to me. New ways of thinking and seeing the world amazed me. Novel ways of conceptualising the most commonplace events through another language fascinated me to no end. I was forever touched by my experience abroad to the point that upon my arrival home, all I could do was to devise ways of going back. Some years later a second stay abroad lead to my employment, marriage and the birth of my son – who has dual nationality. For me, the last decade has represented a change of spheres from novice newcomer to a fully integrated member of society in what is now my adopted country. During this same period, I was able to enter the university as a graduate student in applied linguistics, and began to formulate questions related to my own experience as a language learner abroad as well as that of other learners in similar situations I had observed through the years.

When I visited the US, many of my friends seemed to think of me as a sort of expert in language learning abroad, and though undeserving of such recognition, I often tried to answer their questions about SA and encourage them to venture out as well. A question that often reoccurred was “how long does it take to learn the language?” as if there was some kind of knowable point in SA at which one becomes a Spanish speaker (or French, German or a speaker of any foreign language). I then became interested in answering this question, backing it with empirical evidence that would point to an optimum period abroad that would maximize advancement in the target language (TL). Of course, this is no easy question to answer. If it were, someone other than myself would have already come up with a cut-and-dried approach to length of stay (LoS) in SA programme design. We can plainly see in the wide range of LoS in available programmes, often determined without consideration of the linguistic needs/goals of the individual learner, that the discovery of an optimum LoS for SA is likely a still distant reality. The present study, within its limitations, has the objective of shedding light upon the value of SA in the advancement of L2 linguistic competence, with a special look at LoS and its potential impact on the efficacy of SA.

The current study is developed against the backdrop of the exponential increase in popularity of SA programmes over the last three decades and across the globe. While developing countries continue to send students abroad on a more permanent, degree-seeking basis, SA is the stay of choice for students from wealthy nations who are more interested in going abroad for what it adds to their own personal enrichment rather than pursuing professional training during their stay. This is, however, only a tendency among SA participants, and individual students' motivations are diverse and often multifaceted.

Economically privileged countries in North America and Europe are top senders and receivers of SA participants. SA programme demand has increased by 50% since 1999 in Canadian universities (AUCC, 2011) and students going abroad from the US increased by 1.3% over a single school year (2009/2010-2010/2011). The numbers available from US institutions estimated 273,996 out-going SA participants in the 2010/2011 academic year (IIE 2012a). According to the OECD (2013) Asian students represent 53% of total foreign students with China, India and South Korea as the most represented countries of origin, although in 2004 Altbach and Basset estimated that about 80% of mobile students worldwide came from Asian countries[2]. Major world economies are overwhelmingly the destinations of choice, with 83% of all foreign students[3] enrolled in a G20 country (OECD 2013), that is, some representative nation of the top 20 world economies.

Europe has become a world leader in the promotion of SA within EU member states, but also in sending students to other parts of the world). The European Erasmus exchange programme facilitates the transfer of credits between partnering institutions as well as study grants to students who participate in the programme. Erasmus has been extremely successful, and has sent abroad an estimated three million participants since its beginnings in 1987. During the 2011/2012 school year alone 250,000 European students went abroad thanks to the Erasmus network (EC 2013). This relatively recent

[2] This discrepancy (53% vs. 80%) is likely due to a difference in the definition of student mobility neither source is explicit in their definition of "mobility". What is clear is that Asian countries are increasingly active senders of mobile students (foreign students and/or SA participants) worldwide.
[3] OECD (2013) makes no mention of a formal distinction between "foreign" students and SA participants; however the destination preferences of SA participants are similar to those pursuing degree programmes.

"boom" in the demand for SA in Europe is the setting for the research project in SLA which has provided the framework under which this dissertation has been developed. The practice of SA in the European context, and its impact on SLA is thus of particular interest throughout the discussion of SA contexts in the current work.

Language learning is one of the motivations most frequently mentioned among students wishing to study abroad. A survey of SA participants in Spain revealed that the target language for learning was the factor that best explained students' choice of destination country over any other factor (Pineda-Herrero, Moreno-Andrés, & Belvis-Pons, 2008). Institutions and individuals place a great deal of faith in the supposed efficacy of SA programmes when it comes to language learning. However, the deep seeded folk belief that the only way to become truly proficient in a foreign language is to spend time in the country where the language is spoken is not always supported by the research in every facet of second language development. While billions are spent each year on SA worldwide, in no small part in order to promote language learning, still relatively little is known about how consistently these programmes deliver positive results when it comes to specific areas of language acquisition. This gap in our knowledge of the true effectiveness of SA as a learning context has been noticed and addressed by an increasing number of researchers in the area of SLA. While some studies focus on the impact of the context on language acquisition, others look into the factors that best predict the success of such stays in furthering language learning.

While there is still much to be uncovered regarding SA and L2 acquisition, previous research indicates that the relatively naturalistic learning context associated with SA experiences tends to boost language acquisition in certain skills, while often leaving other areas of language ability untouched. SA has been seen to benefit fluency in terms of temporal measures and hesitation phenomena, but the same period abroad may have little to no measurable effect on more persistent interlanguage features such as the ability to avoid error in speech, or phonological accuracy, among others. Another observation that is repeated in the literature is the important role that individual differences play in the relative success of the learning experience. The uniqueness of each learner allows them to better (or less effectively) take advantage of the setting that SA provides. Individuals differ in their language learning aptitude, level of motivation, and demonstrate personality traits that affect to what degree the stay is beneficial to

their language ability. To further add to an already complex mix, SA programme design varies considerably across abroad endeavours, length of stay (LoS), pre-programme language requirements or previous experiences in the TL country, can equally influence outcomes, yet often prove difficult to control for in any rigorous manner.

A more complete understanding of the inner workings of SA, both at the individual and programme level will help educators and learners alike to better gauge the importance of SA in the course of higher education. Likewise, a more realistic picture of what SA can offer language learners will allow institutions to tailor programmes to the practical needs of their students at a variety of levels, adding to their efficacy and affordability. Also, a realistic vision of potential SA outcomes will allow sending institutions to better prepare students to take full advantage of the opportunity with which they are presented and avoid the disappointment that often accompanies returning students who did not reach their unrealistic learning goals.

The skill focus of the current work is oral production of L2 English. Indeed, oral abilities have been the most scrutinized in the SA literature, much more so than written or receptive skills. The assumption is that the naturalistic learning environment of SA – the enhanced access to native speakers and increased opportunities for practice of the spoken language – lead to gains in speaking ability and is likely the reason for the greater research interest in oral skills. Furthermore, the folk belief regarding the superiority of SA environments to at home (AH) settings is especially strong when it comes to speaking related skills. We consistently see in qualitative data collected through interviews and diary entries over the now ten-year span of the SALA project[4], that students often make reference to their hopes of improving their English accent or increasing their fluency while abroad. However, few, if any expectations are expressed regarding improved listening comprehension, reading or writing skills. With the current study, we thus seek to better link general expectations with the real possibilities for improvement among students at a relatively high level of proficiency, and at the same time, gauging the impact of the relative time spent in the TL at these higher levels of L2 oral competence.

[4] The SALA Project is discussed in further detail in section 4.2.

Most researchers in the area of SLA agree that the construct of proficiency in a second language is "multi-componential in nature" (Housen & Kuiken 2009:461). Many will also agree that while highly inter-dependent, components can be divided into the notions of complexity, accuracy and fluency (CAF). CAF are now commonly used as variables in L2 research to describe oral and written performance. They are also useful in assessing learner progress in the L2. Measures of CAF have enhanced research into language proficiency in that they provide objective measures of performance, while before the explicit use of the CAF triad; studies often relied on project-internal exam-like test scores or subjective ratings of performance. Here, we adopt measures of CAF as a manner of objectively assessing subjects' production before and after SA periods varying in length. Throughout the present work we refer the components of CAF, as "dimensions" of L2 performance and/or proficiency, reflecting the use of this term in the current related literature. The CAF dimensions are operationalised through concrete measures of each, namely, syntactic and lexical complexity, accuracy in terms of errors per unit, and fluency as measured through speech rate.

In the present study, our goal is to contribute to the growing body of research in SA contexts comparing SA periods that differ in LoS. The specific aim of the current work is to examine L2 proficiency as seen through the oral performance of Catalan/Spanish speaking adult learners of English. We examine L2 performance through the dimensions of complexity, accuracy and fluency applying specific measures of these dimensions to the oral data. We track the performance of 47 English learners who go on three (N=33) and six-month (N=14) stays respectively in English-speaking countries. We administered tests before SA and at the end of the SA period, the subjects having returned to the AH institution, in order to assess any progress learners had made over the treatment periods varying in length. Thus we seek to determine whether SA is beneficial, whether longer stays produce greater benefits, and if so, in what specific dimensions of linguistic performance included here. We take into consideration the development of the afore mentioned sample of English learners with a Spanish/Catalan L1 background, and additionally, measure learner results against those of native English speakers performing the same task. We formulate a series of research questions that are based on the overview of previous literature presented in chapters 1-3, and seek to

effectively answer those questions in the empirical study to which chapters 4-8 are devoted.

The chapters and subsections that compose this dissertation are developed with these key research objectives in mind. Topics revolve around SLA through SA contexts, LoS and its impact on the relative success of SA experiences, Methodological issues related to these areas, and finally, measures of linguistic complexity, accuracy and fluency as indicators of changes in the L2 learner's linguistic competence. The dissertation is divided in to two main parts (I and II). Part I provides an overview of the previous literature that is relevant to the empirical study contained in Part II. Part I is divided into three chapters and covers the following content.

Chapter 1 contains an introduction to SA research, its treatment in the literature as a learning context and its impact on SLA. In this chapter we also contextualize SA on a global level, placing it within the international trend toward increased student mobility worldwide. Here we also present previous findings dealing with the impact of SA on linguistic competence. We review the relevant studies by areas of linguistic ability with special attention given to oral skills. Several methodological issues are often discussed in the study of SA. These issues range from the difficulty in SA research in controlling certain programme variables to the influence of individual differences in SA outcomes. These and other methodological issues will be tackled in Chapter 1. Finally, the question of LoS is also first discussed in this chapter, presenting the still limited findings associated with this variable in SA programme design. We look at LoS as a programme feature of a wider framework for the characterization of SA programmes overall.

Chapter 2 focuses on the theoretical underpinnings of second language acquisition (SLA) and second language speech production presenting the principal approaches and models that have driven second language research. The second part of Chapter 2 turns to the speech dimensions of complexity, accuracy and fluency (CAF) as well as some of the theoretical takes on how these interact with one another. To do so, we review some of the underlying concepts related to the CAF constructs, as well as how they have been operationalised and applied in L2 research. We also discuss how hypotheses concerning the interaction of CAF elements have taken different directions according to whether

researchers embrace the “Trade-off hypothesis” or the “Cognitive hypothesis”. Finally, Chapter 3 concludes the theoretical background section (Part I), by summarizing the most recent innovations in the concept of second language proficiency, competence and development, as well as new conceptualisations of SLA through Dynamic Systems Theory.

The remaining chapters are dedicated to the empirical study (Part II), which is made up of Chapters 4-8. A brief introduction to the present study (Chapter 4) recapitulates the main concepts introduced in Part I, linking them to the current study. Also contained in this chapter are a description of the local context in which the current study was developed, and a more specific look at the wider research project, ‘Stay Abroad and Language Acquisition (SALA)’, which made the development of this dissertation possible. Also contained in Chapter 4 is a statement of the global research objectives and the research questions upon which the remainder of the dissertation is built.

Chapter 5 details the method behind the present study. The design that encompasses our study is discussed in more detail. The different profiles of the participant groups who acted as informants in the study are given here as well. Also in Chapter 5, the data collection procedure that was used throughout the course of the observation period is explained in terms of the task that was used for elicitation of the data, as well as the procedure that researchers and subjects followed during the data collection sessions. The processing and analysis of the data are described in Chapter 5 as well, as is the application of specific codes to the data. Finally, a section of this chapter is devoted to the CAF measures themselves, detailing how each of the indices was computed.

The results are reported in Chapter 6, and systematically unpacked in the discussion section contained in Chapter 7. The dissertation ends with a conclusion, in which we critically discuss our findings, suggesting possible implications and future research goals

PART I: BACKGROUND

CHAPTER 1

STUDY ABROAD AND SECOND LANGUAGE ACQUISITION

1.1 An introduction to SA research

An ever increasing body of research in Second Language Acquisition (SLA) is centred on SA – a context in which learners spend a predetermined amount of time in a foreign country where the TL enjoys an important social and functional status (Collentine 2009). Popular belief has long pointed to SA as the best possible scenario for foreign language learning, even considered by many as the only way to become fluent in a second language. This widespread confidence in SA settings is not only present in the general population. From the onset of the first SA programmes, second language researchers, practitioners and education administrators have placed a great deal of faith in the efficacy of SA in advancing L2 abilities. Folk belief in the power of SA persists to this day, even though empirical research often does not report the spectacular results one might expect from an abroad experience. Early work in the field often failed to rigorously examine SA contexts and research instruments were lacking. In response to the gap in our real knowledge of SA contexts and how they contribute to language acquisition, pioneering researchers in the field have sought to investigate SA, its impact and implications for language learning. Over the last several decades SA research as a whole has gained credibility, rigour and has been able to better characterize SA as a learning context. Furthermore, SA research has pointed to the benefits the context generally provides, has set straight common misconceptions about the context, and exposed the extremely complex nature of SA settings.

1.1.1. Milestones in the study of SA

By the mid 1990s a number of groundbreaking contributions to the early stages of SA research had appeared and set the course of academic investigation in the field even until the present day (Meara 1994; Freed 1995a; Coleman 1998). Further contributions to the growing body of research specific to SA appeared in the 2000s, confirming and further detailing many of the initial results seen in work done the prior decade and building upon earlier results (Collentine & Freed 2004; DuFon & Churchill 2006; Collentine 2009). In subsections that follow, we provide an overview SA as it has been

discussed in the literature as a context of learning, and how it has been seen to impact different aspects of SLA. We also discuss the characterization of SA basing ourselves in a specific framework and deal with common methodological issues in the study of SA and SLA.

The year 1995 marked the publication of Barbara Freed's hallmark attempt at capturing the state of study abroad (SA) as a context of second language acquisition. With the appearance of Freed (1995a), empirical findings were brought to the forefront for the first time in SA research, a new subfield of SLA that in the present day is becoming increasingly consolidated in the applied linguistics community. Freed edited a series of papers on SA that made up the volume of *Second Language Acquisition in a Study Abroad Context* and set the direction of the emerging field of SA research even to the present day. Collentine (2009) notes the impact of the 1995 volume describing it as a milestone that marks the beginning of a new period of rigorous empirical study in SA research. The main conclusions that were drawn by Freed (1995a) were that: first, in general terms, SA seems to be beneficial for most learners, and secondly, that research at the time was highly problematic for a number of reasons.

Up until the publication (Freed 1995a), the scant research that existed in the field, although generally optimistic about the effects of SA on linguistic ability, was limited in its validity due to several reasons. First, while previous studies declared SA as beneficial, many of these studies were based on standardized or institution internal test scores as a measure of improvement rather than empirical measures of linguistic performance. An early, noteworthy study was Carroll's (1967) large-scale study of US language majors. Although reliant on test scores alone and at only one testing time, the study was highly ambitious and relevant for the time with 2,782 participants at over 200 participating institutions. The study was not specific to SA as a learning context, but rather intended to be a "status study" of American language students upon graduation. SA was but one of many factors taken into account in relation to student attainment toward the end of their university training. However, Carroll expresses great confidence in the efficacy of SA noting that "even brief time spent abroad had a potent effect on a student's language skills" (p.1). He also sees SA as an opportunity for weaker students to catch up as he observed that "many low-aptitude students are able to compensate by

diligent study and practice or because of special opportunities such as *study abroad*" (p.1, emphasis added).

A second issue in early SA research was that many studies relied on subjective "oral proficiency ratings" or self reports to gage learners' improvement over a SA period, providing little concrete information about how linguistic ability might evolve as a result of SA. In spite of numerous shortcomings in the early SA research efforts, Collentine (2009) mentions some significant discoveries in the 1970s and 80s and into the 90s, such as superior gains in speaking and listening skills over reading skills (Willis, Doble, Sankarayya & Smithers 1977; Dyson 1988 as cited in Collentine 2009)[5], as well as the notions of the importance of learning thresholds in the relative efficacy of SA (Opper, Teichler & Carlson 1990 as cited in Collentine 2009)[6].

Yet another shortcoming in the literature on SA research prior to Freed (1995a) also tended to lack comparative data. For example, there was no comparison of SA participants to their AH counterparts nor were control groups such as native speaker subjects employed so as to provide baseline data to compare with SA experimental groups. Meara (1994), though large in scale, proves an example of some of these shortcomings. The study developed under the Nuffield Modern Languages Inquiry provided valuable information on residence abroad as undertaken by UK students and had a decidedly linguistic focus. Students completed questionnaires detailing how they spent their time abroad, and how this influenced their L2 linguistic competence. The study was indeed ambitious in scale, with 586 participants. However the data gathered were reliant on learners' impressions of how they had improved rather than empirical measures of language development. According to these self-report, the majority of students had the impression that their oral-aural skills had improved during their year abroad. However, less than half of the participants felt they had made similar progress in reading and writing abilities.

[5] Original sources unavailable, but as cited in Collentine 2009.
[6] Original source unavailable, but as cited in Collentine 2009.

What Freed and her colleagues were able to establish quite convincingly through the series of papers contained in Freed (1995a) was that SA, at least in general terms, is beneficial to most students, however SA cannot yet be named superior to AH contexts in many important areas of linguistic development. Morphosyntactic ability, for example, tends to be one of these areas. Development in this and other areas of linguistic ability will be further developed in section 1.2. Furthermore, it is not clear that all learners at all levels significantly benefit from a SA experience, but rather the relative success of learners while abroad may depend quite a bit on their level at the start of SA among other important individual factors that shall be further discussed in section 1.3. Many of these initial observations contained in Freed (1995a) were confirmed or further qualified in a series of papers published the 2004 special issue of *Studies in Second Language Acquisition (SSLA)* dealing with learning context, and in which SA was compared with other common contexts such as formal classroom instruction (FI) and immersion setting (IM).

The studies presented in the *SSLA* special issue represent another turning point in SA research. The issue came as a type of reprisal for the many years of focus on instruction type in the formal classroom setting (Chaudron 2001), as well as an overreliance on purely psycholinguistic approaches to SLA (Norris & Ortega 2001). These studies provided an alternative to then current views on L2 acquisition in that they adopted a context sensitive approach to SLA. On a very practical level, the papers compared the AH, IM and SA learning contexts either as a triad or comparing two of the three within a single study. The 2004 studies provided no firm evidence that any one context is superior to another across cases. Rather, the research indicated that depending upon learners' level and the specific skill(s) being analysed, some contexts may prove more beneficial than others in certain instances (Collentine & Freed 2004). The researchers who contributed to the special issue added greatly to the field in that their work rigorously compared those settings to which secondary students and undergraduates are most likely to have access. Also, more finely tuned methods of analysis were applied to data collected in these settings, reflecting the current practices in SLA studies in general, and further replacing internal test scores, or impressionistic evaluations of data in SA research.

As these timely publications appeared in North America, a growing interest in SA was emerging among European researchers as well. A widely cited article by James Coleman on the European take on SA was published in 1998. In it Coleman offers an amble historical review of the evolution of the practice of SA in Europe as well as a review of the impact of SA on foreign language skills and intercultural competence. A Few years before, Meara had published a large scale study on the effect of the year abroad. Although only based on information collected through a questionnaire, the large pool of subjects, 586, were asked to report specifically on the impact they felt the stay had had on their foreign language abilities. Milton and Meara (1995) was a study of vocabulary growth among English learners in the months prior to, and upon return from a 6-month SA experience. They found that the English as a foreign language (EFL) learners' vocabulary grew considerably more during SA in the UK than it had in the months prior to departure. Vocabulary growth during SA even reflected growth rates found in the L1 literature (Milton 2009:232). Milton and Meara's (1995) study was the first of its kind, groundbreaking both in the subject matter, vocabulary growth during SA, and in the application of assessment tools, thus representing an important contribution to early SA research in Europe.

Still another timely work in European based SA research appeared in 2006 with the volume edited by Margaret DuFon and Eton Churchill entitled "Language Learners in Study Abroad Contexts". This collection of studies has a decidedly sociolinguistic focus. The principal idea that the authors put forward is that of the individual language learner who deals with choices that are shaped by the local sociolinguistic context. Contributors to the book are especially concerned with the development of pragmatic competence, interaction with the host family and cultural beliefs as they are unpacked through such interactions, the development of negotiation strategies and classroom dynamics. Finally, the volume concludes with additional attention paid to programme and learner variables that influence the relative success of SA. More specifically, Isabelli (2006) reports on the importance of the construction of social networks while abroad and how these influence students' motivation, attitudes toward the host culture and consequently in their L2 linguistic ability during the SA experience. Adams (2006) analyses learning strategies and how they evolve over a SA period.

Overall, DuFon and Churchill bring together studies from a wide variety of SA settings raging from the more typical European destinations to Japan and Indonesia, and pose relevant questions about the very pertinent sociolinguistic aspect of SA. They also acknowledge the role individual and programme variables can have in facilitating or limiting opportunities for engagement with the host culture, improving learning strategies and determining personal investment in language learning.

Thanks to the groundbreaking work of pioneering researchers in SA contexts and their calls for improved academic work in this area, an increasing amount of attention in the SLA community is centred on SA. In subsections that follow, we provide an overview of the special context of SA as it has been discussed in the literature as a context of learning and how it has been seen to impact different aspects of SLA.

1.1.2. Student mobility in a global setting

Student mobility is often seen as a whole on a global level. SA, although varying considerably in practice and in its regional characterizations, can be separated out of the totality of student mobility and defined as a specific context, separate from other modes of cross-border education. In the following sections we deal with the ever strengthening trend toward greater student mobility, SA within the greater context of education related mobility, the motivations for its promotion as well as some often confused or overlapping terms to describe different modalities of mobility. First we will distinguish SA from other forms of cross-border education as such a distinction better contextualizes the empirical study that will follow in Part II. We also outline the current status of mobility and differing takes on mobility by region. Although it is not by any means an exhaustive review of mobility practices worldwide, the intention is to sketch an overall picture of SA in Europe and other regions where mobility has been of particular importance in recent decades.

1.1.2.1 SA within international student mobility

Early academics often encountered the need to travel great distances in order to be educated. History's intellectuals were certainly faced with the scarcity of at home institutions that could meet their needs and were often obliged to seek opportunities overseas, producing the first academic migrations. Education in and of itself was reserved for those privileged classes, or a few uncommonly talented individuals.

Aristotle, a Macedonian by birth (now northern Greece), travelled to Athens to study under Plato in 367 BC, and is later thought to have studied biology in Asia Minor (modern-day Turkey)[7]. The Apostle Paul originated from a small city in what is now Turkey, but travelled to Jerusalem as a youth to learn Judaic law from one of the most prestigious teachers of the day. After his conversion to Christianity, he travelled yet again for the purpose of study to an undetermined region of Arabia where he studied for three years before he began his missionary activities (teaching-related journeys) throughout the Mediterranean[8]. Oxford University claims to have taken on its first international student in 1190 with the arrival of Emo of Friesland, "setting in motion the University's tradition of international scholarly links"[9]. In North America, the first SA *programmes* that were comparable to what we might find today emerged in the mid-1800s. Indiana University, home to the oldest recorded "faculty-initiated study abroad programs," began offering European tours over the summer term in 1879[10]. While the University of Delaware reports the first US university-based year abroad programme in 1923, sending eight SA participants to France by ship. Today stays abroad during the course of one's high school or university career are becoming more and more common. The idea of cross-border education is nothing new, but what was only available to a select few, is now, more than ever, being offered to an immense global student population.

For our purposes, it is important at this point to distinguish between at least two present-day types of student mobility as the nature of the learning context, duration and objectives are not generally the same. International student mobility can be defined as "crossing country borders for the purpose of or in the context of tertiary education" (Richters & Teichler 2006) and encompasses practically any cross-border move at least loosely related to academic study. Parting from this rather general definition, here we seek to separate SA experiences from "foreign" or "international" student status and distinguish SA participants from those students who happen to have foreign student status in a given institution.

---

[7] Encyclopedia Britanica. Retrived from http://global.britannica.com/
[8] Acts 22:3; Galatians 1:16-18.
[9] 'A brief history of the University'. Retrieved from: http://www.ox.ac.uk/
[10] 'Early history'. Retrieved from: http://overseas.iu.edu/

In our characterization of SA we take into account Kinginger's observation that "While all forms of student mobility are available in principle to all students, in practice the meaning of 'study abroad' varies considerably depending on student origin, destination and ultimate goal" (Kinginger 2009:10). Thus SA, as it is dealt with in this dissertation, is to be set apart from mobility in general in that it is a markedly temporary international study situation. That is, the student's "origin" remains his or her principal academic institution throughout the stay. SA participants do not typically set off to spend their entire high school or undergraduate career in the host country, but rather go with a very definite timeframe in mind and typically ranging from a few weeks to a year. Few programmes offer SA terms beyond a year and many Europe-based SA initiatives send students on semester programmes. Short-term SA of a year or less is typical of sending countries such as the US – which often sends students for only a few weeks – the UK, European countries and Japan. In these cases, participants rarely intend to extend their stay beyond the predetermined study period. Destinations are often selected according to personal interest, the target language the individual wishes to study, or social/relational reasons (Pineda-Herrero, et al. 2008) rather than due to some concrete professional goal. The motivations behind students' participation in SA often have to do with self-enrichment, experiencing a foreign culture first hand, the pursuit of gains in language ability or a mixture these. The student who chooses to go on SA often comes from an economically privileged country. Kinginger (2009:11) characterizes this type of student as follows:

> ...these students may pursue intellectual growth, cultural enrichment, international awareness, and language learning through various forms of study abroad. The possibility of longer-term migration, while not excluded, is rarely the focus of such sojourns.

Finally, the "ultimate goal" of SA tends to be, as we have already touched upon, self-enrichment, language learning, or a disconnection from the home culture. Foreign or international students, on the other hand are typically degree-seeking students who spend the whole period of their tertiary education abroad. In contrast we can further isolate SA in that it is most typically undertaken within the academic framework of home institutions. SA students are thus internationally mobile *within* the context of their

tertiary education, earning credit toward the degree that they are pursuing AH, and are often only loosely connected to the host institution.

SA is infamous for its large array of methodologies and features across programmes, but a few examples may serve to at least broadly distinguish SA from other types of student mobility. An example of SA is when, for instance, an Italian engineering student spends a semester abroad in Germany through in interuniversity exchange programme and takes classes relevant to her major while maintaining her status at her home university. She knows going into the SA period that it is for a predetermined amount of time, and the credit she earns abroad will serve her in obtaining her degree from her AH university. In this case she is likely to go without FI in the local language unless she pursues instruction out of her own initiative. Although in a rather different vein, SA would also include American students who travel to Spain for several weeks during the summer term, attending Spanish literature, language and culture courses. Here the time abroad is reduced to a few weeks in many cases, and FI in the TL is very much a part of the itinerary. Students are unlikely to have any kind of link to the host institution that would resemble the registration process of local students, and their SA is clearly within the context of their US university curriculum.

This form of temporary mobility differs greatly from the experiences of the more than 30.000 Chinese students perusing university degrees in the UK, or the upwards of 27.000 students of Turkish nationality enrolled in German universities who are internationally mobile for the *purpose* of tertiary education (Lanzendorf 2006). Motivations are inherently different. This long-term form of mobility or "academic migration" (Kinginger 2009), is the most common form of student mobility for sending countries such as China, Latin American, and African countries. In this case the sending country has relatively few resources for scholars, and the alternative is that these students pursuit valued and highly recognised degrees abroad. This type of arrangement implies at least several years of residence in the host country and often leads to longer stays as work opportunities, marriages and other life commitments arise.

In many cases, reports of mobile student numbers are based on nationality alone. Thus, foreign nationals studying in what has become their home country, yet without having acquired official nationality, would also factor in, statistically speaking, as foreign

students. Likewise, nationals who have been schooled abroad and return to their country of origin to receive their higher education, or indeed, participate in SA, would fit the definition of mobile students, however, these students are registered as home students if nationality is the only criterion in determining who is "foreign". Some researchers who dedicate their work to tracking student mobility have noted some potentially misleading statistics associated with the movement of students across borders Kelo, Teichler, & Wachter (2006) note the following:

> A first glance at publications of national governments, specialised agencies, research institutes, and international organizations…conveys the impression that there is no shortage of quality data on international mobility. This impression is misleading: The available data are not (always) the data we need.

What is more, institutions and national governments are generally those with the task of compiling data on mobile students at their universities. While some statistics may accurately reflect the reality of student mobility, others may very well oversimplify student profiles grouping SA enrolment with that of foreign students. Numbers from these under-informed reports then make their way into reports elaborated by international institutions at the highest level. Although is it still unclear exactly how many of the millions of international students worldwide participate in SA as it is defined here, what is clear is that on all levels "more students are going abroad for a wider range of purposes, and within a broader selection of programs than ever before" (Kinginger 2009:7), a trend that shows no signs of relenting at the present time. Although the methodology associated with the task of classifying and tracking the direction of student mobility worldwide leaves much to be desired, for our purposes it will suffice to distinguish SA as a unique and separate institution within the phenomenon of mobility in general.

Government and educational institutions, in an ever more globalized world, are seeking ways of preparing young people for a highly international/intercultural future. SA has long been seen as an effective method of allowing students to discover the "other" and thus improve vital intercultural and foreign language skills that will serve them in their future pursuits.

> As national economies become more interconnected and participation in education expands, governments and individuals are looking to tertiary education to broaden students' horizons and help them to better understand the world's languages, cultures and business methods. One way for students to expand their knowledge of other societies and languages, and thus improve their prospects in globalised sectors of the labour market, is to study in tertiary institutions in countries other than their own. (OECD 2013)

How institutions go about initiating students in these areas differs considerably depending upon the country or region. In the following subsections we will review how mobility has been handled in those parts of the world that most often send or receive foreign and SA students. Europe is more thoroughly described as it is a remarkable case in terms of the promotion of SA, but also because it is the setting of the empirical study that is presented in part II of the current work. Attention will also be drawn to other parts of the world where mobility has had particular prominence or is gaining in popularity and in investment from local entities.

1.1.2.2. SA in Europe

The institution of SA has had perhaps no more transformed a scenario than that found in Europe. Coleman (1998) provides an enlightening perspective of what has been the backdrop to a modern, integrated Europe, taking us from the fallout of a post World-War II continent to what is now the very picture of inclusion and unification among neighbouring nations. The typical, day to day conflict we read about in the news may often cause us to lose perspective as to just how far Europe has come. In a few short decades European nations have demonstrated a clear determination to promote mutual understanding and respect and foment cooperation among member nations. This conviction is perceivable at nearly every level of European governance. Indeed, the sheer scale of the movement toward a peaceful multicultural, multilingual, multiethnic Europe is astounding. The practice and promotion of SA as an integral part of higher education are a direct result of such efforts. As Coleman puts it:

> Understanding study abroad in the European context means first appreciating, on the one hand, the political will to tolerance and integration respecting cultural

> and linguistic diversity, and, on the other hand, the geographical and ethno-cultural scale involved.

Mobility between countries that is now commonplace within the European community would have been unthinkable only a generation ago. Pérez-Vidal (Forthcoming) comments on the European strategy underlying mobility at the university level stating that "mobility has been seen as part of the EU's political strategy towards the goal of the construction of united Europe" (see also: Pérez-Vidal 2009; 2011). This unifying factor may seem trivial until we review just how far Europe has come in only the last few decades of its recent history, and the role that mobility, especially student mobility, has played in the process.

Coleman credits the work of the European Union (EU) and the Council of Europe (COE) as the principal advocates for the mobility of EU citizens across states, and for those social-cultural efforts that have so greatly influenced the exponential growth of SA in the European context. The EU, in its political and economic role, is the governing body responsible for the lifting of border formalities among member states, facilitating movement for work, education or residence abroad. The EU has also worked toward equality in the job market for newcomers to any given member state, and in the promotion of foreign language learning while, at the same time, preserving and protecting local customs and identities.

In terms of education, the EU has installed a number of mechanisms for cooperation among institutions on the continent. The most dynamic of which is the Erasmus student exchange programme. Here we will refer to these exchanges as Erasmus programmes, as this is the branch of exchanges associated with higher education. We refer to them as such with the understanding that at the institutional level, student exchanges fall under the direction of European Commission's (EC) Lifelong Learning Programme since 2007 (EC 2012). The programme provides grants to students who wish to study abroad temporarily anywhere from three months to full year. The system also allows for the transfer of academic credits between institutions, so that a student from one member nation can easily make academic progress toward his/her degree while away. What we know as Erasmus today was initiated in 1976 as the modest Joint Study Programme. This first effort to promote student mobility was succeeded in 1987 by the European

Community Action Scheme for the Mobility of University Students, and was already referred to as Erasmus. That year Erasmus exchanged 3,000 students among 300 universities. These exchanges were then brought under the well established umbrella of ERASMUS-SOCRATES in 1995. By 1997, 1,500 institutions exchanged 80,000 undergraduates (Coleman 1998). Erasmus has sent abroad an estimated three million participants since 1987. Over 250,000 went abroad during the 2011/2012 school year alone (EC 2013).

Currently, more than 4,000 higher education institutions in 33 countries participate, with more hoping to join (EC 2012). The higher education branch of exchanges draws from an annual budget in excess of €450 million. What is more, Erasmus conventions for the transfer of credits between institutions have largely inspired the implementation of the Bologna process: the re-visioning of higher education methodology and academic credentials in European universities. Indeed, the reach of the Erasmus programme has made it the most successful SA network in the world.

As publish in the European Commission's official summary of most resent Erasmus data (EC 2013), the 2011-12 school year saw 204,744 Erasmus participants sent abroad for the purpose of studying in a partner institution. An additional 48,083 students went on work placement programmes or traineeships, bringing the total number of mobile Erasmus students to 252,827. This is a 9% increase from the previous year. In 2011-12 the length of SA through Erasmus averaged 6.3 months, and student traineeships averaged 4.3 months. Thus, whether as students or in work placement, Erasmus participants spent an average of 5.9 months abroad. In the midst of a dramatic economic crisis, Spain managed to send the most students (34,103) followed by Germany (27,593), France (25,924), Italy (20,404), and Poland (12,106). Spain was also the top destination for incoming students followed by France, Germany, the UK and Italy. It should be noted that these are the European countries with the highest student populations. The average age of Erasmus participants was 22.5 years and 60.6% of abroad students were women.

While the EU operates at the political level, the Council of Europe (COE) takes on a social role in the zone. The COE is Europe's principal human rights organization and as such, works toward furthering the most fundamental rights of Europe's citizens,

democracy, and the rule of law. Currently there are 47 member nations serving approximately 820 million people. Its principal contribution to SA in Europe, and language learning specifically, has been the drafting of the Common European Framework of Reference for Languages (CEFR) (COE 2001). The first version of the document was published in 1996. Coleman (1998) highlights CEFR as one of the Council's more comprehensive publications on languages, an ever more widely used tool in establishing proficiency standards and in the preparation of curricula for language teaching. The CEFR document (EC 2003) sums up the purposes behind its elaboration as follows:

> The Common European Framework provides a common basis for the elaboration of language syllabuses, curriculum guidelines, examinations, textbooks, etc. across Europe. It describes in a comprehensive way what language learners have to learn to do in order to use a language for communication and what knowledge and skills they have to develop so as to be able to act effectively. The description also covers the cultural context in which language is set. The Framework also defines levels of proficiency which allow learners' progress to be measured at each stage of learning and on a life-long basis.

CEFR has become quite useful in efforts such as those put forth by the EC in action plan on language learning and linguistic diversity implemented between 2004 and 2005 (EC 2003). This proposal encourages the learning of two foreign languages other than one's own mother tongue during his/her lifetime. The aims of the action plan were to encourage language learning at all educational and life stages, improve language teaching, and promote a language friendly environment in the EU space. The reasoning behind such an initiative is as follows:

> In the European Union (EU) more than 500 million Europeans come from diverse ethnic, cultural and linguistic backgrounds and it is now more important than ever that citizens have the skills necessary to understand and communicate with their neighbours. All European citizens should be able to communicate in at least two languages other than their mother tongue (EC 2003).

Here, as in CEFR methodology, we see the grounding of language learning in the real use of language, as opposed to traditional text-book driven institutional learning, and the emphasis placed on the need to communicate with co-citizens. The realities of the European community begin to impact language legislation, furthering international connections. The role of SA through programmes like Erasmus and others has been fundamental in the changes in language policy that the EU has implemented in recent years.

To sum up, two major socio-political movements have brought about rapid and dramatic change across Europe. There are namely, heightened mobility across European borders, and the conscious promotion of multilingualism, and foreign language learning. The former, the will to open up the European space to all EU citizens, is made manifest in the lifting of restrictions of the mobility of citizens among member states. The latter, and equally important catalyst for change in Europe, has been the very conscious promotion of multilingualism in European society, especially in education. We repeatedly find evidence of the promotion of a multilingual profile in European society, the education system, and particularly in the area of student mobility. Efforts associated with this movement are perhaps best represented in large-scale investment in Erasmus and other language learning schemes, the elaboration of CEFR and in formal recommendations such as those contained in the action plan on language learning and linguistic diversity.

1.1.2.3. SA in other parts of the world

Studies of SA as a learning context have been until now most prolific in North America. There is a relatively long-standing interest in the impact of study abroad given the considerable increase in demand for such programmes. Even so, leaders in higher education are concerned that SA programmes are in need of further promotion. (Bradshaw 2012), highlighting the predominant popular and institutional belief in the value of SA in the overall development of students at the university level. The US and Canada, holding true to their wealthy-nation profile, are countries that tend to send students on brief SA sojourns within the context of their otherwise AH course of study. They are also top receiving countries for both short-term SA participants and more importantly, degree seeking internationals.

The demand for SA programmes among Canadian college students has increased by 50% since 1999 (AUCC, 2011). Students studying abroad originating from the United States increased by 1.3% between the 2009/2010 and 2010/2011 academic years. The world-wide total of US SA participants rose from 270,604 to 273,996 during that time (IIE 2012a). Europe is by far the most sought after destination for North American undergraduates with the United Kingdom, Italy Spain and France representing the top four destinations. These four countries host nearly half of the total number of SA participants, taking on 106,527 students during the 2010/2011 school year (IIE 2012b). Europe as a whole receives well over half of the total number of Americans studying abroad (54.6%). Top European destinations are followed by China which received nearly 14,000 US students during the 2010/2012 sessions (IIE 2012b).

The dramatic increase in the interest in SA in the United States is remarkable considering the traditional, widespread monolingualism in the country. Prior to World War I, the US was highly multilingual. Many immigrant families maintained their native languages, passing on their mother tongues to the subsequent generation. However, the societal view on mother tongue maintenance changed after the war and a new wave of immigrants arrived on US soil. Growing fears of appearing foreign motivated Americans to abandon their native languages in favour of English which had acquired a decidedly superior status. (Kinginger 2009:13). “The anti-German hysteria engendered by World War I, merged with anti-immigrant xenophobia, challenged the legitimacy of hyphenated – and multilingual – Americans” (Pavlenko 2002). English proficiency was linked to proper citizenship and intellectualism while native languages were associated with ignorance and even subversion. A generation later World War II, and more concretely, the attack on Pearl Harbor, reignited fear of foreigner disloyalty leading to systematic discrimination against Japanese-Americans. The practice was taken to such an extreme that even families who had resided in the US for generations were pre-emptively sent to internment camps across the Western US and Hawaii (Pavlenko 2002). The foreigner fear legacy lives on in the post-World Trade Center attacks era, this time directed at Arab-Americans.

As statistics clearly demonstrate, in recent decades, the traditionally insular US has demonstrated a growing curiosity toward the “other”. This is reflected in the ever increasing number of SA participants, programmes and modalities. Carroll (1967) put a

great deal of faith in SA early on in the initial exploration of the subfield. His findings in favour of SA were no doubt influential in the growth of such endeavours in the years following the publication of his large-scale study on language learning in the US.

Asia represents an emerging research base for SA as it is estimated that as much as 80% of international students across the globe originate from Asian countries (Altbach & Basset 2004). China has registered an increase of 23.1% in the number of students sent abroad between the 2010/2011 and 2011/2012 academic years. In fact, China, together with India and South Korea, round out the top three sending countries world-wide. Taiwan, Japan and Vietnam also fall within the top ten (IIE 2012c). As noted in Section 1.1.2.1., students originating from many Asian countries are not typically sojourners abroad, but rather longer-term, degree-seeking academics. The nature of these abroad experiences is quite different from that of a SA participant as we have defined here. Just how many of these "exchanges" are in line with the type of SA experience of interest in the present study is difficult to estimate. Nonetheless, it is undeniable that Asia-based students represent a considerable proportion of the world's students studying overseas, a population whose experiences have gone largely undocumented, not to mention their gains in language competence over the course of such stays. As globalization expands and emerging middle classes become wealthier, and Asian universities begin to sponsor short-term SA (Jackson 2006), it is likely that sojourns for the purposes of study among Asian students become more common place, providing more opportunities for researchers to tap into these experiences. Understanding SA from an Asian perspective will undoubtedly shed light on what, until now, has been a subfield of SLA with an almost entirely western representation. Wang (2009) and Jackson (2010) have contributed with studies of students on SA programmes originating from Taiwan and Hong Kong respectively, and point to a very interesting new research base for understanding both socio-cultural and linguistic development in the TL.

Having taken in something of the historical perspective in Europe and elsewhere, it is no wonder that researchers in the language sciences are increasingly interested in the impact of hugely popular mobility programs. In the following section we will briefly review some of the findings that have emerged from the study of SA and its impact on linguistic competence in a second language.

1.2. The Impact of SA on Linguistic Competence

We have already mentioned the widespread belief in the efficacy of SA in boosting foreign language skills. Since the advent of SA as part of modern higher education, a great deal of faith has been placed on SA's ability to accelerate that language learning process. Many even hold SA to be the only effective means of acquiring the TL. Qualitative evidence from within the SALA project repeatedly points to just how prevalent this idea of the superiority of SA over other learning environments is among students. The following extracts taken from the SALA main corpus provide insight into the common folk-linguistic beliefs about language learning that learners bring to their abroad experiences:

> "I think [SA] is very useful because **until you go abroad you can't speak the language properly**." (SALA participant, prior SA).

> "I think [SA] is really good 'cause **you don't learn a language being in your country** and going to class twice a week." (SALA participant, upon return from SA)

> "I think it's something we all look forward to, studying abroad. And especially **in the country of the language you should speak best**. I think **it's the best way of improving** [the language]." (SALA participant, prior to SA)

SALA participants also frequently express that they perceive considerable improvement in their L2 as a result of SA, especially in terms of progress in oral skills. An illustration is given in the following sample taken from a diary entry written by a SALA participant during his three-month SA in Nottingham, UK:

> I think that staying abroad has been a very useful experience. It has been a great chance to practice my speaking, to improve my style and my accent. […] I can speak much better than the day I arrived in Nottingham…I have also learnt how to pronounce words better because of some personal experiences.

In spite of these common assumptions, the question of how and what linguistic skills improve over a stay abroad is still very much an open one, when it comes to empirical

findings. Nonetheless, there are a few cautious generalizations to be made. Freed (1998), in her then updated review of the SA literature noted a number across-the-board benefits that are seen in those students who have the opportunity to study in the target language country as opposed to those whose language learning experience is limited to the AH, FI classroom. She found that students who participate in SA:

> ...appear to speak with greater ease and confidence, expressed in part by a greater abundance of speech, spoken at a faster rate and characterized by fewer dysfluent sounding pauses. As a group, they tend to reformulate their speech to express more complicated and abstract thoughts, display a wider range of communicative strategies and a broader repertoire of styles. It is equally clear that their linguistic identities extend beyond the expected acquisition of oral skills to new self-realization in the social world of literacy (Freed 1998:50).

Another milestone in the SA subfield confirmed many of tentative conclusions reached in Freed (1995a), namely, the publication of a special issue of *Studies in Second Language Acquisition* dealing with the role of learning context and its impact on SLA. The studies contained in this special issue further detailed the generalizations made by Freed (1995a; 1998) as to the benefits of SA that tend to been seen across SA situations. Along with SA, the papers dealt with two more learning contexts that are quite readily available to American students, especially during their secondary and higher education; namely, at home (AH) formal study and Immersion settings (IM). The series of studies published in 2004 provided no evidence that any one context is intrinsically superior to another. Rather, the research contained in the issue pointed to the notion that depending upon learners' level and the skill of interest, some contexts may be more beneficial than others (Collentine & Freed 2004). For example, morphosyntactic control seems to develop more readily in an AH setting especially for discrete grammatical forms (Collentine 2004). Likewise, development toward more native-like phonological systems seems to progress similarly in AH and SA settings. Diaz-Campos (2004) found that external factors such as initial level, hours spent speaking the TL outside of class, even gender were more reliable predictors of gain than was the learning context variable (AH vs. SA). SA did prove more beneficial than other contexts in terms of gains in lexical breadth (Segalowitz & Freed 2004) and narrative ability (Collentine 2004) and in the reduction of communication strategies due to some linguistic deficit (Lafford

2004). Fluency was benefited in SA over AH contexts (Segalowitz & Freed 2004; Freed, Segalowitz & Dewey 2004) but Freed et al. found that the IM context included in their study of the triad proved more of a boost to fluency that even the SA context. Overall, SA participants were made significant gains in oral fluency, especially when fluency is defined as "ease and smoothness of speech" (Collentine & Freed 2004). Similarly, students who participated in SA produced speech at speed that more closely resembled native speaker (NS) norms. Similar findings regarding fluency are reported in studies both prior to and since the publication of these papers.

To complement the basic generalizations mentioned above, we present a summary of findings related to language skills as they are affected by SA periods. In foreign language teaching-learning research, skills are often divided into productive and receptive categories. This is the breakdown we use here, recognizing that production and perception, oral and writing skills are often linking on a fundamental level. Here the division by skills is purely organizational in nature, and does not imply any firm divisions in cognitive processing or the like. In fact, a few studies present both oral and written data as they examine a particular skill set (e.g. Isabelli 2004). However, we will follow this scheme as closely as possible. It should also be noted that studies on production, especially oral production, are far more numerous than those of any of the other areas covered here, and thus occupy considerably more space in the following sections. Furthermore, oral production data is of greater interest to us in the present study as the data we draw upon in Part II are of this type.

First we address the case of productive skills in spoken and written production and their development in SA. Then we present the still relatively scant literature on receptive skills (listening and reading) as they are impacted by SA. In the sections that follow we introduce each section with a brief summary of findings by skill based on key studies of, or including, a particular aspect of linguistic competence (skill). Then we will discuss each study in greater detail as its particular relevance dictates.

1.2.1 Productive skills

An array of studies has appeared dealing with productive skills and how they are potentially impacted by SA. The more numerous instances of production related studies are perhaps due to a heightened interest in enhancing students' communicative

competence in the foreign language classroom that has been prominent since the 1980s (see Canale & Swain 1980). Similarly, in SA research productive skills have drawn great interest as compared to receptive skills. Here we divide production into oral skills and written skills, summarizing the findings of the most relevant work in these areas.

1.2.1.1. Oral skills

Grammatical ability in the L2 is one of the most examined areas of language acquisition in SA contexts. Nonetheless, results in this area of development are highly mixed. In oral production, some studies find that the more naturalistic setting of SA contexts does indeed benefit certain areas of grammatical development (Regan 1995; Guntermann 1995; Isabelli 2004). Other studies have found, however, that SA provides no additional benefits to grammatical ability when compared with AH contexts (Rothman & Iverson 2007), and in some cases AH FI even seems to be more conducive to gains in grammatical competence than SA when comparing the two contexts (Dekeyser 1991; Collentine 2004).

Regan (1995) and Guntermann (1995) contributed to Freed's 1995 volume on SA with studies dealing with the acquisition of specific grammatical aspects of French and Spanish, respectively. Both authors reported benefits associated with time spent abroad. Regan performed a study of the acquisition of French negation morphemes during a year abroad programme in France or Belgium, emphasising the sociolinguistic knowledge that is required in the correct use of negation in French. Six Irish subjects provided the L2 French data pre- and post-SA through an oral interview. Results showed a considerable strengthening of native-like criteria en French *ne* deletion as a result of the stay abroad, presumably in the students' "drive towards integration into the native speech community". Similarly, Guntermann (1995) found considerable gains in grammatical ability in L2 Spanish in a small sample of subjects after a year of foreign service in Central America. Peace Corps volunteers received intensive language training before entering the field, and testing times occurred at the end of training and then at about one year into their abroad experience. The data are then compared, although rather roughly, with SA data dealing with acquisition of the same structures, but in an academic abroad setting. The author found that proficiency ratings among the Peace Corps volunteers trained in the TL country were higher than those of the American College students who spent a semester abroad in Spain, as well as when compared to

volunteers whose training was undertaken in the US. It must be noted that this study is of a very specific context, foreign service, with initial intensive language training, and just how comparable it is with SA studies is somewhat questionable. However, the study does provide evidence that a period abroad paired with FI in the TL, does seem to yield considerable benefits in grammatical ability at the one-year mark.

A study by Isabelli (2004) also provided evidence for significant gains in the acquisition of the Spanish null-subject. A grammatically judgement task was administered and oral interview were performed among 31 American subjects before and after a year-long SA in Spain. Results showed that SA participants performed better on the grammatically judgment test at post-test in two of the three aspects of null-subject use: free verb-subject inversion and apparent "that-trace" sequence violations (grammatical and ungrammatical). However no significant improvement was found in students' ability to detect ungrammatical "that-trace" items. As for the oral interview task, similar improvement was seen. Subjects decreased the use of overt subject pronouns post-test and increased the number of "emphatic" post-verbal subjects characteristic of Spanish norms. Thus the oral data appear to support the evidence of improvement seen through the results of the grammaticality judgement test.

However, as we have already mentioned, results in the area of grammar development in SA do not always imply improvement. Dekeyser (1991) did not find that a semester-long SA group in Spain improved over a AH group in their use of specific grammatical structures covered in the second year Spanish curriculum at the home university, specifically: the copula, the subjunctive, conditional clauses and relative clauses. Nor did he find dramatic differences across groups in their monitoring of the grammar structures in those students who showed some mastery of the grammar elements analysed as determined through scores on a grammar test.

Collentine (2004) also reported on differing contexts and grammatical, as well as lexical development in L2 Spanish. This paper looked at the effect of learning context (SA vs. AH environments) on grammatical and lexical abilities in conversational discourse captured through oral proficiency interviews conducted among 46 participants before and after a semester-long treatment period in either SA or in FI at the home university. According to the results, students who stayed home during the semester and received FI

outperformed SA participants in some grammatical and lexical abilities. However, it should be pointed out that these students from the AH group performed better in those aspects of grammar/lexis that the curriculum is known to emphasize to a great extent—namely, verbs and subordinate conjunctions. The SA group however, was better able to produce more native-like narrative behaviour as well as more "semantically dense lexical types" than the AH group at the end of the same time period. Day-to-day interaction with the TL community may be the driving force behind this type of progress.

A more recent study in the acquisition of null-subject pronouns in Spanish (Rothman & Iverson 2007) also failed to detect any superior benefits for a SA context over FI AH. Like Isabelli (2004), the authors assume that UG related parameter resetting is at work in L2 grammatical development and investigate the role of increased positive input, available in a SA situation, and its potential beneficial effect on "triggering universals" and consequently, the acquisition of null-subject pronoun-related properties. They interpret the lack of gains in the SA group to mean that the "parameter resetting" that needs to take place in order to be able to consistently apply null-subject grammar rules in Spanish does not seem to be particularly enhanced during a five-month period of increased exposure to the TL environment in a way that FI in the classroom cannot.

While firm conclusions regarding the benefits of SA in specific areas of grammatical competence continue to be elusive, oral fluency is one of the domains of L2 production that seems to benefit most in SA contexts. Indeed fluency analysis quite consistently reveals gains in learner production post SA. In the 1990s SA was quickly gaining the reputation as a failsafe fluency booster. A series of studies done by Möehle and Raupach in the 1980s (as cited in Dekeyser 1991) were some early studies in SA contexts in which fluency was shown to benefit from a sojourn in the TL country. Lennon (1990a) and Towell, Hawkins, & Bazergui (1996) also found similar benefits to SA in the area of fluency. More recently, Segalowitz and Freed (2004), Trenchs-Parera (2009), Valls-Ferrer (2011), Mora and Valls-Ferrer (2012) Valls-Ferrer and Mora (2014) all find significant improvement in a number of areas and aspects of fluency upon return from SA. However, SA is not always found to be the best case scenario for the promotion of fluency in all its facets. Freed, Segalowitz and Dewey (2004) did find a fluency advantage in those students who went abroad over those AH who only

participated in FI. However, the L2 fluency gains in the SA group did not supersede those of the IM group also included in the study.

During the early 1980s, Möehle and Raupach published a number of case studies where L1 German students studied French and L1 French learners studied German for several months abroad (Möehle 1984; Möehle and Raupach 1983; Raupach 1983, 1984). Though the original sources are now difficult to access, these studies are widely cited. Dekeyser (1991) reports on their findings as do Freed and colleagues (2004) and Adams (2006). Though finding no significant improvements in syntactic complexity or accuracy, they did however find that subjects' speech rate, as measured in syllables per second, increased, the number of pauses decreased, and the length of stretches of speech between pauses increased as a result of having spent time in the TL (Möehle 1984 as cited in Dekeyser 1991). However, the German students of French well outperformed the French learners of German in quantitative measures of fluency. This was perhaps due to disparities in the initial level between groups. Raupach (1984 as cited in Dekeyser 1991) also posited a relationship between the acquisition of formulaic language and fluency. The researcher noticed that the more fluid speech of one L2 French subject was linked to her expanded use of formulas. Post SA French "fillers," "modifiers," and "organizers", were more readily available to the learner allowing her to avoid certain hesitation phenomena such as drawls, filled or unfilled pauses.

Lennon (1990a) did not set out to investigate SA as a learning context, but rather the construct of fluency itself. The motivation behind the study was to determine the quantitative elements of fluency that correlate with global evaluations of fluency by native-speaking teachers. More specifically, the main interest was "...to advance our knowledge of what constitutes fluency and especially what makes for perceived fluency differences among learners and how an individual learner improves in fluency over time". Nonetheless, the treatment in this study happened to be a six-month SA in the UK, almost under the assumption that the stay had every potential to boost subjects' fluency levels. The subjective judgments of listener raters did indeed detect an overall improvement in fluency at post-test, and of the 12 quantitative measures used to evaluate the learner speech samples only three (speech rate, filled pauses per T-Unit, and percentage of T-Units followed by pause) proved to be good predictors of improved perceived fluency. Self-corrections did not prove a good indicator. It should be noted

that the study was quite small in scale with only four learner participants. Even so, Lennon intuitively points to the role of individual variations among learners "in the precise areas in which fluency improvements may occur". He also postulates that further research would be able to identify "core" versus "peripheral" fluency variables. The study, thus, provides us with a rather early example of work with regard to the pairing of qualitative and quantitative evaluations of fluency in a SA context.

Segalowitz and Freed (2004) looked into the effects of SA on a number of measures of oral and cognitive fluency in American undergraduates studying Spanish in Spain as compared to students studying at home in a FI condition. Results showed that students who went abroad showed greater gains in temporal fluency measures, occurrences of hesitation phenomena were reduced, and overall oral proficiency increased as measured through an oral proficiency interview (OPI). That being said, the authors also note the presence of significant interaction effects and correlations that point to an important relationship between individual variables such as proficiency level, cognitive abilities and degree of language contact while abroad.

Valls-Ferrer (2011) examined the development of L2 fluency in advanced EFL learners from Catalonia (Catalan/Spanish bilinguals) studying in the TL country for three months. This study uniquely examines the relationship between utterance fluency, rhythm and listeners' perceived fluency ratings. Both fluency and rhythm, as measured through temporal fluency measures and rhythm metrics, were found to improve significantly as a result of the SA period. Furthermore, listener judges rated post-SA speech samples more favourably than pre-SA samples to a significant degree. A Study done by Mora and Valls-Ferrer (2012) followed an analogous population to that of Valls-Ferrer (2011), investigating the differential effects of formal instruction (FI) at home and a SA period on the oral production of Catalan/Spanish undergraduates learning English at an advanced level. Data were collected through oral interviews at three points in time over a two-year period and assessed for fluency, accuracy and complexity. Results revealed considerable fluency gains during the SA period and moderate improvement in accuracy. No significant gains were found for complexity, nor were any gains obtained during FI. A follow up study following the same design and by the same authors (Valls-Ferrer & Mora 2014) sought to explain gains in fluency over the SA period through initial level and contact patterns students maintained during

the stay. They found that a lower initial fluency level and increased levels of contact correlated with greater gains in fluency during the SA period.

The main findings of Freed, Segalowitz and Dewey (2004) seem to indicate an advantage for IM over and above either SA or FI. Results showed statistically significant gains in oral performance as measured through total number of words, length of longest turn, speech rate and other composite fluency indexes in the IM group. SA participants performed better than those who stayed home in the FI context only in terms of speech fluidity, but did not reach the degree of improvement that the IM group did, suggesting the superiority of the IM condition over SA at least in this instance. The AH group, in contrast, made no significant gains. Through multiple regression analysis, the authors noted that hours per week spent writing outside of the classroom were significantly associated with gains in composite oral fluency measures. The IM group reported more weekly hours writing in French outside of the class than did the SA group, suggesting a link between writing practice and gains in fluency.

While fluency is one of aspects of L2 oral production that most consistently benefits from an SA context, phonological development seems to be found on the other end of the spectrum. The existing studies suggest that phonological production in adult learners often resists significant development under SA conditions. Studies focusing on phonological development during SA are relatively few in number and results are mixed. Diaz-Campos (2004) compared phonological gains of Spanish learners who went abroad with those who stayed home finding no significant advantage for the SA context. Likewise, Avello (2013) found no improvement as a result of SA in terms of the objective measures of L2 phonology employed, while Mora (2008) found evidence to support the benefit of SA in this area of development.

In the first study of L2 phonology in SA and FI context mentioned here (Diaz-Campos 2004), 26 students of Spanish went on a SA programme to Spain while 20 of their classmates remained in the AH environment in the US following regular classroom instruction. Subjects read a text containing target words aloud and resulting speech samples were analysed according to the segments of interest: word initial voiceless stops, intervocalic fricatives, word-final laterals and palatal nasals. The segments in question are speech sounds that are particular to Spanish, some of which are notoriously

troublesome for Anglophones when acquiring Spanish as adults, namely word-initial non-aspirated voiceless stops, intervocalic voiced fricatives and the alveolar variant of the word-final, [l], and to a lesser degree the Spanish palatal nasal. Results showed similar gains for SA and AH groups in the realization of voiceless stops, word-initially, and word-final laterals. Neither group improved significantly in their pronunciation of Spanish intervocalic fricatives and in the case of palatal nasals, students already showed a high degree of accuracy at pre-test. The author notes that other factors independent of type of context such as time spent speaking the TL outside of class, age at onset of learning, gender and initial level at pre-test, proved better predictors of phonological development than context did.

While Diaz-Campos (2004) and Avello (2013) found no evidence for gains in L2 phonology over a SA period, Mora (2008) did find some evidence of gains after spending three months abroad. For this study, Mora looked at the voice onset time (VOT) associated with voiceless English stops as produced by advanced English L2 learners. Subjects were undergraduates participating in two quarters of FI at the home university before spending one quarter abroad, finally a fourth testing time was situated at approximately 15 months after arrival from SA, following the SALA project timing for data collection (see section 4.1 for further details). The author found consistent, albeit non-significant gains in VOT in the pronunciation of voiceless stops after the SA period. Additionally, he noted loss at the fourth testing time, that is, after a relatively long period of time in which no further exposure to English is reported, neither in the context nor through FI. It seems that once the intense exposure to English that occurs during SA is absent, the associated benefits also seem to wane.

A study investigating the impact of a three-month SA on second language (L2) phonological development in speech production drew from objective acoustic measures, as well as subjective listener judgments of perceived foreign accent (Avello 2013). In this study, 23 Spanish/Catalan learners of English provided speech samples before and after SA. Acoustic-phonetic measures tracked learners' progress as measured through VOT in voiceless plosives and vowel duration and quality. Pronunciation errors were also coded and reported as were error rate scores. Listener judges then provided ratings of perceived foreign accent. The objective acoustic measures did not indicate

improvement over the three months abroad but judges did detect a slight decrease in perceived foreign accent and there was a significant improvement in error rate after SA.

1.2.1.2. Written skills

Relatively little academic research has been done in the area of writing and the effect of SA in this skill. Writing has long been eclipsed by an interest in oral performance, but there are some interesting contributions to the literature. Sasaki has been quite prolific in this area in recent years publishing a series of studies on the subject (2004, 2007, 2011). Barquin (2012) contributed with a longitudinal study of a number of relevant measures of writing development where subjects were their own matched pairs. Finally Freed, So and Lazar (2003) did not find considerable improvement in their subjects L2 writing as a result of SA.

Sasaki (2004) noticed how Japanese students who had gone on SA to an English speaking country tended to attributed their improved L2 writing ability to these experiences while their fellow students who had stayed in Japan attributed their improved writing skills to their AH English classes. Students who went abroad had gone through similar coursework when at home, but still saw the SA period as more impactful in their L2 writing progress than the AH, FI experience. Sasaki then followed up on these observations investigating a number of external factors that influenced the learners L2 writing. In Sasaki (2007) it was seen that while both AH and SA groups improve their overall language proficiency over the same observation period, only SA participants improved their writing ability and became more motivated to improve their L2 writing. Sasaki (2011) confirmed improved L2 writing ability and motivation levels in those native Japanese students who had studied in an English-speaking country as compared to those AH students who had no such opportunity. Furthermore, she found that gains in these areas were a direct function of length of stay. Barquin (2012) looked into the writing development of 30 Spanish/Catalan learners of English. She compared progress in writing ability through sequential periods of FI and SA in the same group of subjects over a period of 15 months. Learner texts were evaluated in the domains of complexity, accuracy, fluency, lexical diversity, sophistication and cohesion. A group of native English speakers also performed the writing task so as to provide comparable data by which to gauge the L2 productions. Results showed that students made

considerably more progress while abroad in an English-speaking country than during the FI context at their home university.

While still little work has been done with respect to L2 writing and SA, a time spent abroad seems to be generally positive for this skill. However, this finding is not necessarily the same across studies. Freed and colleagues (2003) for example found that American undergraduates who stayed in the US for study improved their perceived writing fluency in French when rated by native French, non-teacher judges over and above those students who went on SA to France.

1.2.2. Receptive skills

Receptive skills are essentially the ways in which "people extract meaning from discourse they see or hear (Harmer 2001)" and here, refer to listening and reading skills. Receptive skills tend to be acquired at an earlier stage than do productive skills as they draw on the learner's previous knowledge of the world, types of discourse and literary genres. Although both listening and reading tend to be overlooked in studies of SA some notable exceptions are a study by Dewey (2004) on reading development and Beattie (2008) in listening comprehension, and Mora (2014) in phonological perception.

1.2.2.1. Listening

Surprisingly, very little attention has been given to listening-related skills in the SA literature. Some exceptions have been developed within the SALA project and are presented below. We note the particular contributions of Mora (2008) in the specific area of phonological discrimination and Beattie's paper on listening comprehension (2008), in both cases, as they are impacted by FI and a subsequent SA period.

Mora (2008) reported on a results obtained through an auditory perception test by Catalan EFL students who had reached an advanced level in the L2 at the time tests were administered. Students were tested at 4 different testing times following the SALA project design (see section 4.1). Students' development in phonological perception was measured through a categorical AX auditory discrimination test containing English word pairs representing nine phonemic contrasts presented to the listener as minimal pairs. Participants heard 135 English word pairs – a total of 108 minimal pairs, plus 27 same-word pairs that served as distracters). Results showed that the learners' ability to

discriminate between the English phonemes examined did indeed improve over time. However, it was only after FI that gains reached statistical significance, showing little evidence of an important effect for the SA condition on the overall gains that occurred during the observation period.

Another study accomplished under the auspices of the SALA project (Beattie 2008) looked into the effect of consecutive periods of FI and SA on listeners' comprehension of an English recording. Beattie found that only after SA were there significant improvements in listeners' comprehension of the recording as measured through an exam-like test.

1.2.2.2. Reading

Although studies of potential gains in literacy as a result of time spent in a SA context are rare, demonstrating a bias in general expectations in favour of aural/oral skills (DuFon & Churchill 2006), Dewey's 2004 study analysed reading development in two contexts of learning, one of them being SA and the other IM. Dewey's aim was to look into the role of context in reading development, comparing reading comprehension and processes in groups of American students of Japanese participating in SA and intensive immersion settings respectively. Significant differences between groups were found in only one measure: reading comprehension, showing a slight advantage for SA participants. The author notes, however, that this measure was based on a self-assessment. Thus, it seems that SA participants felt more confident when reading Japanese than the IM participants.

1.3. The interplay of context and individual differences: Methodological considerations.

The essence of research, at least from a quantitative perspective, is to control conditions and predict outcomes. Experimentation is obtrusive, in that independent variables are isolated and manipulated so as to produce differing, predictable outcomes in the dependent variables, while controlling conditions in such a way as to be able to attribute the result to a single (set of) variable(s). Such research is "objective, generalisable, outcome oriented and assumes the existence of 'facts' which are somehow external and independent of the observer..." (Nunan 1992:3). When applied to SA contexts, this type of research proves problematic in several ways. Rigorous SA research should be able to demonstrate that linguistic progress in the TL made during SA is significantly different

from that of AH learners who only have access to FI (Dekeyser 1990; Meara 1994). However, the complexity of SA and participant variability can make drawing clear conclusions about SA outcomes quite difficult. Here, we concentrate on two major issues in the field of SA research and the methodological considerations that should be taken into account when developing studies where SA settings are involved.

The first of these issues that we discuss here is the problem of comparability of groups. It has been argued that SA participants may differ from students who stay home, and that these differences may influence the relative language attainment in each group. Second we will look at the role of individual differences (IDs), which have rightfully received more pointed attention in recent SA studies. With regard to this issue, we consider at length the role of initial level as it is one of the IDs that has most consistently been seen as a reliable predictor of L2 gains during SA. Finally, we take on the context itself, considering the sheer complexity of SA settings, and the resultant difficulty in controlling for all impactful context variables within in a given research design. Here we dwell on the role of length of stay (LoS) as a programme design factor that is of special interest in the present study.

1.3.1. The problem of comparability of groups

A number of researchers within applied linguistics have dedicated much of their work to comparing two or more learning context that, at least in theory, differ considerably (e.g. FI vs. SA). Likewise, these researchers often desire to compare two or more groups of subjects. In SA research, this is typically set up as a comparison of the different learning contexts represented in SA and the AH instruction of the language classroom and the corresponding FI/SA groups of participants. This is especially the case in US based studies where SA tends to be optional[11] for university students, even in modern language majors. Thus, SA and more or less similar FI groups studying at the home institution are available for participation in these research projects. While this scenario drew quite a bit of interest in the 1990s as SA research became increasingly visible within the field of SLA, many who designed such studies failed to appropriately justify the comparability of groups.

[11] This is as opposed to the European modality of SA (especially that of the UK) that tends to be (semi)compulsory for language related studies (Coleman 1998; 2005).

In their critique of how learner progress has been measured in SA related studies, Rees and Klapper (2008) give the example of Lafford's (1995) study in which the comparability of groups is dealt with insufficiently. Lafford (1995) produced a comparative study in which data from two SA groups, studying Spanish in Mexico and Spain respectively, are compared with their AH colleagues studying in the FI context. Rees and Klapper chose Lafford's 1995 study to illustrate a shortcoming that many early SA studies demonstrate, namely the rather blatant overlooking of pre-programme learner variables that can potentially influence outcomes. SA participants are described as "heterogeneous", an attribute that the reader is to understand given that this is the case "in most universities" (pg. 98). The control group, (those who stayed home) were described as evidencing "this same type of heterogeneity" (pgs. 98-99). When describing learners' pre-programme experience in Spanish instruction, the author uses institution internal descriptors, including number of credits and course numbers used by the university (e.g. SPA-101), with little to no descriptive power whatsoever. The vagueness with which the author goes about describing groups takes away from the rigour of the comparison, and leaves the reader wondering if factors other than context alone might have influenced what Lafford reports as superior gains for the SA groups; in this case, with regard to communication strategies.

Students who seek to participate in SA may do so because of high levels of motivation to learn the language. They may feel a strong desire to learn about the target culture first hand, and mingle with native speakers. They may be naturally more outgoing and independent than students who opt out of SA experiences. All of these factors have been shown to impact linguistic outcomes to varying degrees (Freed 1990, 1998; Isabelli 2006). Yet, early on in the history of SA research, many studies ignored such crucial factors, assuming that comparisons of learners at home and abroad were comparisons of "like with like", only differing in treatment (FI vs. SA contexts).

Other factors such as socio-economic level, family and work commitments, and academic requirements such as obtaining a minimum grade are also likely to influence the student's ability to participate in an abroad programme, and thus the pool of potential subjects. The student's economic resources may dictate whether or not he or she can go abroad and for how long. Although not necessarily attested in the SA

literature, qualitative evidence from SALA interviews of SA participants point to the restrictive nature of SA for some. This can be further complicated within the European context, where SA is often a required element of foreign language degrees. To illustrate, we include below a selection of interview excerpts in which students, either preparing to leave on SA or having arrived home from SA, respond to the interview question "*What do you think of the requirement [at this university[12]] of having to study abroad?"*:

> "I think it's something that some people cannot afford, and making it compulsory it's not very helpful but...I think that if you come to this university you know it beforehand" (SALA participant, upon return from SA).

> "I think it's a great idea even though it's maybe a problem for, let's say, the less rich families...but maybe it would be better if they had a little more money, but I don't" (SALA participant, upon return from SA).

> "I like it but I don't think that with the marks of the first term [I will go]." (SALA participant, prior to SA).

These comments serve to illustrate what is likely happening during the grouping of SA/AH students for purposes of comparison. The likelihood that there are qualitative differences between students who are willing and able to study in the target language country and those who cannot (or do not choose to do so) is quite high. Springboards and barriers to SA in the form of economic and/or academic (dis)advantages that determine who goes abroad (and thus the make-up of the group) may influence attainment in the L2 in ways that are completely unrelated to the context itself.

More recent contributions to the body of research done in SA contexts have made great strides in controlling for factors such as demographic variables, pre-programme level, motivation, other affective variables and language learning aptitude, yet the problem of obtaining validly comparable groups remains a challenge. One way researchers have avoided this potentially confounding practice is by following learners through

[12] Here the name of the university, which was included in the original interview question, has been replaced with "at this university".

longitudinal studies in which different treatments are given to the same learners. Students are tested before and after a period of FI, and well as pre- and post-SA (Milton & Meara 1995, Pérez-Vidal, Trenchs, Juan-Garau, & Mora 2007). By adopting this design type, participants serve as their own control group (i.e. comparing initial performances with subsequent performances). This may prove to be a better way of monitoring progress in SA conditions as all other learner variables are presumably held constant.

1.3.2. Individual differences

If the essence of research is to control and predict, as we have already avowed, it logically follows that a considerable portion of the literature in SA research is devoted to predicting linguistic gains as a result of a sojourn abroad. However, early on in the development of the field of SA, researchers began to take note of the influence that individual differences exercised in linguistic outcomes (or gains). Thus, finding reliable predictors of gains in SA contexts has proven more challenging to apprehend than one might expect. Studies in this area have provided somewhat mixed results, but a growing set of studies are devoted to unravelling how individual differences (IDs) can affect degrees of successful learning in SA participants.

As already mentioned, a set of variables that plays into the efficacy of SA in post-test L2 attainment is made up of a variety of IDs in the learners themselves. Initial proficiency level at the beginning of SA, especially pre-programme metalinguistic knowledge (Golonka 2006), and one's own motivation level, and ability to establish quality contact with locals, producing language learning opportunities, may strongly influence the final result of an SA experience on L2 progress (Freed 1990; Isabelli-Garcia 2006; Juan-Garau & Pérez-Vidal 2007). More recently, researchers have looked into the role cognitive variables such as speed and efficiency of lexical access and attention control (Segalowitz & Freed 2004), or working memory resources (Sunderman & Kroll 2009; Anderson 2012) may have in facilitating learning when plentiful and impeding it when lacking.

Here we understand gains to be the difference between pre-programme and post-programme scores (or measures), and predictors to be those factors thought to affect gain, usually identified through regression analysis (Brecht, Davidson, & Ginsburg

1995). We briefly cover findings from studies that draw from the very large scale database on Russian learners compiled by the American Council of Teachers of Russian (ACTR). The series of studies using ACTR data have, by far, provided the most comprehensive study of predictors of success in language learning abroad (e.g. Brecht et al. 1995; Rivers 1998; Davidson 2010). We will briefly comment on the finding of these studies that proved robust predictors of gains then focus additional attention on the role of initial level as this factor is of particular interest in the current empirical study of Part II.

Brecht and colleagues (1995) was the first large scale study dedicated to statistically identifying predictors of foreign language gain during SA, drawing on data from a large sample of subjects (N=658) who contributed to the American Council of Teachers of Russian (ACTR) database. In this case, data from participants who studied for four months in Russia were analysed. Gender[13], knowledge of other foreign languages, general language aptitude as measured through the Modern Language Aptitude test (MLAT, Carroll & Sapon 1959), and pre-programme proficiency level in grammar and reading were the variables that proved significant predictors of gain.

A number of studies have yielded significant results for gender as a predictor for post-SA gains where men proved to be greater gainers over women (Carlson, Burn, Useem, & Yachimowicz 1990; Brecht et al. 1995). Others, unwilling to so quickly attribute gender differences to physiological traits, have made qualitative observations about gender differences in an unfamiliar culture, especially those that affect women's experience abroad (Twombly 1995; Polanyi 1995). These studies seek to qualify gender differences by describing the very different socio-cultural environment that young women come into, and which often proves daunting for them.

Knowledge of additional languages other than the TL contributes to the learner's status of what Brecht et al. call "expert language learners". The experts consistently gain over and above first time foreign language learners. It is assumed that as one learns

[13] It should be noted that gender was no longer a significant predictor of gains in a replication study by Davidson (2010) 15 years later. Davidson notes that this important detail reflects possible changes in gender roles in Russian society over the last few decades, but also improved pre-programme preparation, enabling students to better engage with the target culture.

additional languages the skills related to language learning are sharpened. These may include a heightened awareness of how to use learning and communication strategies, to make the most of learning situations. Their superior gains may also be due to more realistic expectations. Those who have already gone through the process of learning a second language may not be so easily frustrated when learning proves quite challenging as they already know what to expect and know that perseverance has a pay off. However, these are merely possible explanations for a phenomenon whose examination went beyond the scope of the 1995 study.

An aptitude for language learning in general seems to predict language gains during SA in only limited ways. In their large scale study of Russian, Brecht, Davidson, and Ginsberg (1995) found that language aptitude, as measured by two subsections of the then current version of the MLAT, had a strong positive correlation with gains in reading and listening. They did not, however, find any relationship between aptitude (as measured through these sections of the MLAT) and speaking related skills. Freed (1995b) obtained a similar result in the use of the MLAT for predicting success in SA settings. She found that the MLAT did not serve to predict gains in oral skills as measured through OPI and listener judges' ratings of global proficiency. The inability of the MLAT to predict students' gains in speaking related areas of proficiency has been thought to be due to biases in the design of the test. Specifically, it would seem that the MLAT measures one's ability to perform tasks that are typical of FI mediated learning in the language classroom rather than an innate ability to learn languages (Larsen-Freeman & Long 1991).

Perhaps the most interesting of the findings from Brecht et al. (1995) was that having to do with initial level. The authors found that greater control of basic grammar and reading ability prior to SA were significantly predictive of increased speaking, reading and listening proficiency during the time in country. However, it should be noted that gains were *negatively* related to general pre-programme level, that is the higher the students' overall initial level the lesser the gains. Indeed, the pre-programme level factor was such a robust predictor of gains that the analysis of the effects of other variables (gender, aptitude, specific skills, etc.) could only be seen having controlled for pre-program levels. In the same year, Lapkin, Hart and Swain (1995) reported similar findings in terms of initial level to those of Brecht et al., namely, that L2 French

speakers who came into SA with lower pre-test scores experienced greater gains at post-test. Self-assessments also indicated that French learners who reported lower self evaluations also had greater perceived proficiency gains at post-test.

Rivers (1998), however, obtained a seemingly contradictory result, drawing on L2 Russian data collected from ACTR, the same database as Brecht et.al (1995). Wanting to compare homestay and dormitory placements, he found that dormitory placed students outperformed homestay participants in speaking proficiency gains after SA. The result was surprising given that the homestay environment, in principle, provides increased quantity and quality of input that would be expected to lead to greater grains than those seen among students who opt for the dormitory stay together with other non-native speakers (NNSs). A similar finding was reported in a homestay study by Veguez (1984, as cited by Sutton & Rubin 2004). Rivers did find however, that the homestay participants who were better prepared in the TL before SA could significantly benefit from the added exposure to Russian provided by the host families. At first glance, this seems contrary to previous findings by Brecht and colleagues, who found: the lower the level the greater the gains. Rivers' study, on the other hand, found that students with higher pre-programme abilities benefited more from the homestay than lower level learners in the same living conditions. In1995, Brecht and colleagues had already mentioned an S-shaped curve, where low level SA participants initially experience superior gains to those of advanced students, followed by a decline in post-SA gains as initial-level increases. It seems, then, that learners reach a point at which they are better experienced in the language so as to take advantage, now at a higher level, of the intense input environment such as that offered by a homestay. It could be that the two studies represent snapshots of two different "peak periods" for availing oneself of the SA condition. (1) Low level learners have more to learn and thus experience greater gains over their more advanced peers. But, at the same time (2) High degrees of input can be too much for the learner. Should he have insufficient experience in the TL, it may be very difficult for him to maximally benefit from the additional input in terms of post-SA linguistic gains. Yet another, more recent study using ACTR Russian data, and to some extent a replication of Brecht et al. (1995), Davidson (2010), found that pre-program control of language structure and listening ability proved to be good predictors of gains in speaking. The 2010 study found pre-programme abilities to be *positively* correlated with gains in speaking. Again, these findings seem to contradict those initially reported

by Brecht and colleagues, yet Davidson (2010) looks at the performance of students who go abroad for periods of up to a year, while the Brecht et al. study examined performances of students who only spent four months in Russia. Davidson's findings seem to suggest that the role that initial level plays in predicting outcomes may change as LoS increases.

Results from "predictor" studies point to a highly complex interplay between a number of factors, here we focus on initial level, but other elements related to learner differences have been found to significantly influence SA outcomes. Affective features such as attitude and motivation as well as one's ability to create a strong social network while abroad (Isabelli 2006), or other learner dependent variables, such as the learner's ability to develop learning strategies (Adams 2006), have been seen to influence the eventual outcome of SA experiences.

1.3.2. Problems related to context.

Study abroad contexts have proven difficult to define in a narrow sense given the variety of schemes and the array of goals and priorities represented therein. Although describing SA contexts is no simple matter, an external/internal dichotomy of variables is often presented. That is, SA research often looks at variables that are intrinsically related to the situation of living and studying in the TL country for a limited amount of time (i.e. external or context variables). This angle on SA largely deals with quantity and quality of input (e.g. Collentine & Feed 2004; Segalowitz & Freed 2004; Dekeyser 2007; Pérez-Vidal & Juan-Garau 2011). Internal features, or individual differences, on the other hand, have to do with what the individual learner, or type of learner (e.g. highly motivated learners), brings to the SA scenario. Some examples of these elements are the individual's aptitude for language learning, knowledge of, or proficiency in the TL prior to departure (Davidson 2010), or the learner's level of anxiety when using the TL (Wang 2009). Together with internal/external features, some research has been dedicated to the study of programme design, and how concrete decisions about the specifics of a SA programme can influence the degree to which learners benefit from the experience in terms of linguistic gains (Rivers 1998; Davidson 2010; Avello & Lara 2014).

Beyond simply describing the context, and looking into how learning in the TL country is different from learning at home, many studies have sought to find a research based combination of optimal SA conditions for the improvement of L2 skills. Yet, the search for optimality is more problematic than one might think. As Huebner (1998) puts it, "...the range of experiences which fall under the rubric of 'study abroad' is so varied and complex that generalizations about optimal learning contexts need to be made with great caution". Furthermore, the complexity of SA does not reside only in differential study settings that fall under the category of SA, but also the human players' reactions to the opportunity of living together in the target/host culture and NS-NNS interaction further adds to the complexity of SA contexts. Kinginger notes:

> Students abroad potentially observe, participate, and communicate in classrooms, homes, personal relationships, service learning, or commercial interactions. Study abroad programs have varying objectives, academic foci, and expectations for student activities. Students abroad may be received with warmth, enthusiasm, and patient assistance, or they may find their presence noted with indifference or even with hostility (2009:5).

SA research should take into account these factors and how learner behaviour might be influenced by the local conditions. Indeed the resources learners bring to the situation, and certain personality traits have been the subject of examination in a number of studies set in SA contexts. Learner related variables have been shown to impact SA outcomes. Individual behaviour, and thus learning opportunities, may be mitigated by the student's L1 background and previous knowledge of the TL, the learning strategies he/she is able to implement while abroad, personal goals and motivation level, and cultural sensitivity, personality traits such as outgoingness, among other facets. The host culture and the individuals who happen to interact with the SA participant may facilitate the student's integration or impede it to a great extent, but how the learner copes with the experience seems to be an important part of explaining varying outcomes at the end of SA. Finally, The details of programme design, that is, practical arrangements such as type of housing, to what extent study of the TL is integrated into the hours of study, and other factors that influence the amount and type of contact with the target language the learner experiences make up another set of factors that have been postulated as influencing the efficacy of SA situations. The shape of programme design may increase

the occurrence of learning opportunities or, on the contrary, block access to the native speaking community to varying degrees. It should be noted that these different categories of external/internal/programme factors do not operate independently of one another, but rather make up a very complex web of variables that play off of one another, complement one another or indeed hinder one another.

Given the highly complex nature of SA contexts, some recent work in the field has responded to calls from the research community for alternative research designs drawing on qualitative data from L2 sojourners (e.g. Jackson 2008). While this type of research does not usurp the study of SA based on quantitative analysis alone, qualitative or mixed method, process-oriented approaches to SA research and programme evaluations certainly add to the effort of filling in the gaps in our knowledge of the context and examining its impact.

1.4. A framework for the characterization of SA

Here we briefly present a framework for a generalised characterization of SA programmes developed by Pérez-Vidal and Juan-Garau (2011), in which they effectively classify SA features at the context, individual and programme levels. The authors use the terms Macro- and Micro-level features to respectively denote those factors that depend on the context of SA, that is, following academic classes in the TL country, and factors that depend on the individual learner studying abroad. Programme design features make up the third parameter of the framework. We then focus on a particular aspect of programme design mentioned in Pérez-Vidal and Juan-Garau (ibid.), specifically the factor of LoS. At that point we discuss how LoS in SA has been researched in previous studies, a still largely under-investigated area of SA research. LoS is of special interest in the present study and thus receives special attention.

In order to better frame the current study, we extend the characterisation of SA along the same three parameters as Pérez-Vidal and Juan-Garau (2011). The authors identify the following distinctions: Macro-level context features, Micro-level learner features, and the Architecture of SA programmes.

At the Macro-level, context features are those elements which are common to nearly all SA programmes. These are the context features that are represented in what SA can

offer the learner. These include (1) rich and intense exposure to input, (2) varied, real life, situations in which interaction patterns can develop with many interlocutors who take on variable social roles, and lastly, (3) the facilitation of meaning-oriented, rather than form-oriented learning mechanisms.

At the Micro-level we see how individual learner variables come into play in the SA context. That is, to what extent individual learner differences allow students to take advantage of what a SA learning context offers. These factors may include learners' level in the target language at the start of SA, his or her natural level of outgoingness, the learner's specific cognitive abilities or aptitudes, and so forth.

Programme Architecture refers to the many issues to resolve when designing SA programmes. The practical matters of accommodation, whether or not students are allowed/encouraged to work, and the real possibility of doing so while abroad, and other particulars related to living conditions make up this area of concern. Furthermore, some stipulations may (or may not) be made as to the level of the target language students are required to have before leaving home. They may also undergo varying degrees of pre-departure orientation in order to prepare them for the experience. Pre SA preparation may range from a matter of a brief meeting with counsellors to a full academic subject for credit. Those details specific to the academics of SA also must be taken into account. At what point in their university career are students allowed/encouraged to to stay abroad? What type of academic work must they accomplish while on SA? Finally, what, if any, post SA follow up do students receive upon their return to their home country?

While any and all of these factors can potentially play an important role in the relative success of a given programme, we are especially interested in the impact of time spent abroad. Thus, in the current study, we centre on this area of the third parameter, within the architecture of SA programmes which is length of stay (LoS).

1.5 Length of Stay

The amount of time students spend abroad is subject to a number of opposing pressures. On the one hand, common sense dictates that longer periods abroad should lead to improved outcomes in the TL, and indeed, the few studies that report on LoS have

found, in general terms, that longer stays tend to yield greater benefits in different linguistic domains. For instance "the longer the better" conclusions have been drawn for progress in vocabulary size and lexical knowledge (Ife, Vives Boix & Meara 2000), listening comprehension, oral fluency and accuracy (Llanes & Muñoz 2009), complexity accuracy and fluency in written and spoken language (Serrano, Tragant & Llanes 2012), L2 writing ability and motivational factors (Sasaki 2011), interlanguage pragmatics, namely negotiation and politeness strategies (Félix-Brasdefer 2004), confidence in the TL and an increased frequency of use post-SA (Dwyer 2004). Yet, with all its potential benefits, SA has a cost, requiring a considerable investment of both time and money. The costliness of SA has lead to programme offers with reduced LoS especially in recent decades.

Length of Stay is a factor of vital importance in the design of SA programmes, as it constitutes an important variable of L2 contact and therefore a greater or lesser amount of L2 input. Within their three levels of parameters in the characterization of the SA context, Pérez-Vidal and Juan-Garau (2011) mention LoS first among the elements of what they call "the architecture of the SA programme".

Though LoS is often recognised as significant factor in the ultimate success of the SA programme in achieving linguistic gains, the research into this variable is scant. Indeed, the scarcity of studies into the impact of LoS in SA may be an indirect cause of the wide range of programme lengths available, varying anywhere from a few weeks to more than a year abroad. Some more recent and welcome contributions to the study of LoS in SA programmes are discussed below.

Ife and colleagues (2000) researched two groups of English speakers studying Spanish in Spain for one and two semesters respectively. The subjects in question also had different proficiency levels, "intermediate" and "advanced", at the onset of SA. They found improvement in the vocabulary of both level groups and in both LoS durations in terms of both the number of words known and the depth of their lexical knowledge. Both intermediate and advanced students experienced improvement during SA for which LoS was a significant contributor to relative gains. Along with LoS, students' level of integration and their degree of motivation for learning the TL were analysed as potential predictors for greater success in vocabulary acquisition; however, only LoS

showed a significant relationship to higher scores at post-test, that is, upon arrival from SA. The longer, two-semester stays were especially beneficial in achieving improved vocabulary organization, but also for the number of words known – results that clearly imply that longer stays are more beneficial in this area.

Llanes and Muñoz (2009) make a case for even the shortest stays abroad available. They found that short stays of only three to four weeks yielded significant gains for Spanish L1 learners of English in listening comprehension, four measures of temporal fluency (syllables per minute, ratio of L1 words, articulation rate and length of longest run), and two accuracy measures (errors/clause and percent of error-free clauses). This study also dealt with differing proficiency levels at the onset of SA, a variable which proved impactful as lower proficiency participants benefited the most from SA in a reduction of L1 words used when producing the L2, and in terms of accuracy and fluency. LoS, measured in weeks, was examined as a potential predictor of gains. Students whose LoS was greater by only one week (three vs. four weeks) yielded significantly superior results in two fluency measures (longest fluent run and silent pauses) and in the percent of error-free clauses (a 16,6% decrease) over learners who spent only three weeks abroad.

In 2011 Miyuki Sasaki published a longitudinal study that took place over a period of 3.5 years investigating the effects of SA LoS on Japanese learners' L2 English writing ability and motivation levels. During the lengthy observation period students were evaluated upon entering university and then on a yearly basis over the course of their studies. During this time 28 of the 37 participants spent some time abroad in an English-speaking country. LoS varied from 1.5 to 11 months. She found that those students who spent a period abroad significantly improved their L2 writing while those students who remained in Japan for the university training did not. She also goes on to observe that those students who spent four months abroad or more significantly improved their writing over those students who spent less time on SA. In terms of motivation, she found that students who went abroad were more motivated to improve their English writing than AH students. Also, interestingly, she found that only learners who went on stays of 8 months or more became intrinsically motivated to improve their L2 writing, that is, voluntarily seeking out ways of practising the written L2 and thus improving their writing ability.

In a cross-sectional study of university students who had resided in Spanish speaking countries, Félix-Brasdefer (2004) found that in comparing four different LoS durations: 1–1.5 months, 3–5 months, 9–13 months, and 18–30 months that longer stays enabled student to approach NS pragmatic norms. Students with longer LoS were better equipped to negotiate refusals in the L2 and patterns in their politeness behaviour more closely mirrored NS behaviour.

Some studies do not reveal any correlation between LoS and gains in linguistic abilities. Llanes and Serrano (2011) report on the non-impact of an additional month (two versus three months) spent in the TL country on subjects speaking and writing performances in the domains of fluency, accuracy, lexical richness and complexity. Avello and Lara (2014) found no added benefit to a 6-month stay in an English speaking country to that of a three-month stay in terms of pronunciation accuracy.

Finally a rare and valuable longitudinal study of an academic year abroad was undertaken by Serrano, Tragant and Llanes (2012) who followed 14 English learners over the course of a year studying in the UK. This study is of particular interest to us as there are a number of similarities between their study and our empirical study that follows in Part II in terms of the background of the informants and the speech measures used in the analysis. The authors found that the first semester abroad was enough time for significant progress to be made in fluency as measured through speech rate (syllables per minute) and lexical diversity as measured through Guiraud's Index. From the midpoint (Time 2) to the end of the academic year, students were not seen to progress any further, except in accuracy (Error per T-unit), yet when assessing improvement over the whole academic year (i.e. from Time 1 to Time 3), students were seen to have improved significantly in all linguistic measures although improvement in syntactic complexity lagged behind and with a relatively small effect size for *Time* as compared with fluency and accuracy.

In spite of some mixed results regarding LoS during SA, mounting evidence indicates that longer stays abroad yield greater benefits for the language learner. However, several real life factors are commonly at work against longer period options in SA. When the sending institution does place a high importance on the abroad experience as

a means of acquiring language skills, the immediate need is to fit the SA period into the home university's academic calendar. The task of meshing the home and host universities' timing may already place constraints on the amount of time students are able to spend abroad. Thus, in the interplay of seeking to provide the student with an optimal amount of time in the target language environment through SA and adhering to formal academic requirements, the very practical question of how to accumulate credits toward graduation may often win out in real life situations.

The affordability of SA is surely a second factor in determining LoS. When students are expected to finance their SA experience, shorter stays may become the only feasible option. Long stays may require that students work while abroad, or come up with alternative means of staying in a foreign country. Finally, the present day culture of "quick fixes" with regard to language learning, and the corresponding abroad programmes industry outside of higher education, may promise enormous benefits in language ability even after a very brief SA experience. The results of these intensive options are not necessarily supported by scientific research into the time it takes to progress in an L2, nor do they specify whether or not certain areas of language development could require longer periods spent in the target language environment. However, the propagation of intensive approaches to language learning may contribute to unrealistic expectations being attached to the SA context be it an intensive option or those university sponsored stays that last only a few weeks.

Dwyer (2004) reports on results from a publication by the Institute for the International Education of Students, a third party SA provider for hundreds of US universities and colleges. The trend in SA as practiced in North American universities consists in a steady move away from the year abroad and toward shorter LoS options. She notes that over the past two decades[14] this trend holds true even as the aggregate number of college students who go abroad has increase dramatically within the same time frame. Furthermore, she reports that since the 1990s the fastest growing programme type in the US are those with a LoS of less than one academic quarter. At the same time Dwyer (2004) reports on findings that point to the considerable benefits of stays of a year and

[14] At the time of publication, 2004, Dwyer make reference to data reports collected over the past 16 years. Currently this would encompass the last two decades

beyond, not only in linguistic ability and L2 maintenance post-stay, but also in sociocultural awareness and overall academic achievement.

As is the case in the SA context in general terms, stays abroad varying in length may be found to have different degrees of impact according to the linguistic skill(s) a given study seeks to address. Fluency, for example, may benefit much sooner than syntax or pronunciation at the segmental level. For instance, Mora & Valls-Ferrer (2012) found significant improvement in fluency measures after a stay of three months; while complexity and accuracy did not seem to improve at the same rate. Such may be the case in L2 phonology, where development towards native-like norms tends to be particularly resistant to change. Avello and Lara (2014) found no significant improvement in the pronunciation of English vowel segments among Spanish-Catalan learners who spent three and six months in an English-speaking country. Not only were there no significant effects of LoS, but neither were there significant effects of the time spent abroad. Nonetheless, Højen (2003) found a correlation between participants' perceived foreign accent scores and the duration of their stay in an English-speaking country, in that longer stays resulted in less accented speech. However, accuracy in the production of specific L2 sounds did not improve significantly after SA.

Certainly, further research into how programme design elements impact SA outcomes is much needed. As Kinginger puts it, "a clearer understanding of how study abroad functions in the development of students' language ability might assist in developing reasoned approaches to program design (2009:4)". Not least of these design elements is the amount of time students spend in the host country, a factor that is central to the original study developed in the second half of this dissertation.

1.6. Summary of Chapter 1

In this first chapter we presented a brief history of SA research highlighting several significant contributions to the field that have marked the course of research up until the present day. The collection of papers contained in Freed (1995a), Collentine and Freed (2004) as well as other contributors to the 2004 special issue of *Studies in Second Language Acquisition* on learning context, and Collentine (2009) remain some of the most noteworthy contributions to SA research based in North America. Meara (1994)

Coleman (1998), and the volume edited by DuFon and Churchill (2006) represent some of the most referenced works on SA contexts in Europe.

In order to better understand the global setting in which SA is found, we have looked into the worldwide practice of student mobility and specifically, where SA fits in this global phenomenon. We have emphasized that SA is not synonymous with student mobility in general, but rather is a specific modality of mobility. The European take on SA is of special interest in the present work, and thus has been more prominent in our presentation of SA by regions.

Also in this chapter, we have summarised the impact of SA on linguistic abilities as reported in SA related literature. More time has been spent on oral skills both because of the greater number of studies of oral skills over comprehension and literacy skills, and because oral skills are the focus of the empirical study to follow in Part II.

Finally, to conclude Chapter 1, we addressed a number of methodological issues that tend to arise in the study of SA contexts, namely the problem of adequate comparability of experimental and control groups in SA and AH contexts respectively, the role of individual differences, and issues related to the relatively uncontrollable environment that most SA situations present. LoS, another key issue for the original empirical work found in Part II, is dealt with as a subcomponent of programme design, one of the three parameters that are included in the framework for the characterization of SA that we subscribe to here (see Pérez-Vidal & Juan-Garau 2011).

We have seen that while much popular faith has been placed in SA as a failsafe foreign language learning opportunity, even seen by some as the only way to achieve proficiency in foreign language, conclusive evidence about the benefits of SA in obtaining L2 proficiency is still to be seen. It is clear that learners do not respond to SA situations in the same ways across the board, but rather show a great deal of individual variation. Similarly, we now know that not all areas of language ability are impacted in the same ways or at the same rate. On the contrary, some skills may be significantly enhanced after only a short stay abroad while others may remain unaffected indefinitely. In spite of the uncertainties surrounding SA, abroad programmes continue to attract enormous amounts of public and private investment, not to mention the confidence of

institutions, professionals in language pedagogy, and the language learners themselves for the improvement of language skills. Therefore, rigorous, scientific study of SA contexts continues to be extremely important if we are to correctly gauge the value of SA for language learning, and determine best practices at the institutional level.

CHAPTER 2

THEORETICAL APPROACHES TO SLA AND L2 SPEECH PRODUCTION

While Chapter 1 has introduced the main concepts and issues surrounding the study of SA as a language learning context, Chapter 2 has a broader theoretical focus. Here we review the essential concepts, theoretical frameworks and practices surrounding research into SLA in general and subsequently, in L2 oral production. We first dedicate a subsection to the approaches to SLA that have been particularly influential in the study of SA, namely, interactionist and context sensitive approaches. We go on to describe the most commonly accepted models of speech processing, namely Levelt's modular model of speech production (Levelt 1989) and its subsequent adaptation to the bilingual[15] speech production system (De Bot 1992), and its application in the study of SLA. We also present an alternative view of L2 processing described by spreading activation models. Finally, we will outline the study of complexity, accuracy and fluency in this field. Our aim is to both review the range of different conceptualizations of CAF and to clearly define CAF constructs that are an integral part of the study that is developed in Part II.

2.1. Approaches to SLA

The different theoretical approaches to SLA in use today continue to be quite numerous, so much so that the description of each would fall outside of the scope of the present literature review. This being the case, we make the effort to limit this review to a brief overview of SLA since its emergence in the 1960s and 70s to the present day, covering some notable paradigm shifts that have lead to the current movement toward more integrative theories of SLA.

2.1.1. Overview

[15] It has been suggested that the terms used to refer to the knowledge and use of two languages (bilingualism) and the knowledge and use of three or more languages (multilingualism) be collapsed into the term "plurilingualism" to refer to the knowledge and use of one or more language (Bhatia & Ritchie 2012:xxi). In this dissertation we do not stress a distinction between the three terms, but rather use them practically interchangeably to refer to the knowledge and use of one or more languages.

Over the last five decades the field of SLA has evolved considerably as researchers from varying points of view debate the fundamental questions underlying the acquisition of second languages. These areas include questions related to (1) how humans learn additional languages having already acquired their mother tongue, (2) in what ways learning additional languages is different from learning language(s) from birth, and conversely, how might it be similar to L1 acquisition, (3) what factors contribute to variability between learner outcomes in the L2, and (4) what it takes to achieve advanced language abilities in languages learned beyond childhood (Ortega 2011). Corder (1967) and Selinker (1972) were some of the first researches to contest the behaviourist model that dominated first and second language acquisition and teaching practices based on contrastive analysis. Behaviourism, the legacy of Skinner (1957) and others who saw language acquisition as nothing more than the process of establishing new habits, a process that could be influenced by providing positive and/or negative reinforcement. By the 1960s the behaviourist view of SLA was quickly losing credibility in the field of psychology, and particularly among linguists.

A new, alternative view of SLA looked into learner errors, not merely as a problem in need of correction, but as evidence into how languages are actually learned, and pointing to the existence of an interlanguage, the linguistic system created by the individual in the course of learning a foreign language, and that is both different from the speaker's L1 and the TL (Crystal 2008). Interlanguage reflects elements of the L1 system ('transfer'), contrastive interference from the TL, and the overgeneralization of newly learned rules or elements, all of which, it was argued, can be identified in learner production. Along with the notion of interlanguage, "for the first time, learners were viewed as active and rational agents who engaged in the discovery of underlying L2 rules" (Ortega 2011: 172), very much revising behaviourists' views of the language learner.

In the 1970s and 80s Krashen would offer the first attempt at an integrated theory of SLA (1977; 1976; 1985). Although often criticized as a set of ill defined concepts and a set of largely untestable hypotheses, Krashen's work provoked a considerable increase in the number of research endeavours into the mechanisms underlying SLA, and subsequently the proliferation of proposals as to how to fill the holes in his proposals. Others attempted to apply Chomskyan linguistic theory to SLA, most notably White

(1989; 1991; 2003). These research endeavours revolved around describing the relative pressures of two forces at work in the learner largely independent from other cognitive processes or the external context at hand. The first of these factors was the extent to which universal grammar (UG), an innate human endowment from birth, was (in)accessible to learners of additional languages and second, how specific knowledge of the L1, determined in the early years of life, worked its influence over the L2. Still another approach to SLA emphasized the cognitive nature of language learning (McLaughlin 1987). This vein of research thought of SLA as a similar process to that of learning any complex skill, although language is admittedly, extremely complex when compared to other skills.

By the 1990s the rather compatible approaches to SLA found in interactionist and information-processing based theories had more or less coalesced into a cognitive-interactionist approach. This vision of SLA was summed up as the union of learner-internal factors (attention and memory for example), and learner external factors offered by available interactions in the TL (Larsen-Freeman & Long 1991; Ortega 2011). SLA Research that took on a UG perspective, however, continued to maintain its distance from cognitive-interactionist approaches where theory was concerned, representing the second of the two dominant theoretical approaches to SLA that have persisted to this day.

Linguistics has undergone a series of paradigm shifts since the 1950s, and SLA as a subfield of linguistics has similarly adjusted its view of what second language is and how it is acquired. The first of these shifts occurred when behaviourism, in its view of language as a static system, was overtaken by the view that language was a generative system. Both behaviourism and generative linguistics continue to see language as an object that can be separated from its use. In general linguistics, the reductionist idea of language as essentially a sign system, impacted second language approaches first in the traditional practice of contrastive analysis, and then in modern linguistics through the abstraction that was needed to test the principles and parameters framework offered by generative grammar. Today we find ourselves perhaps still in the midst of yet another paradigm shift, in which language (including second languages) is seen more and more as intimately connected to its use within the social environment. Calls for a more holistic view of language and language use have accumulated over the past decades, and

have still only begun to be addressed in the SLA community. As Weigrand (2011) has put it.

> Linguistics in the 20th century sought to address language by way of the simple, i.e. of reduction and abstraction. In contrast, linguistics in the 21st century is called upon to take up the adventure of the complex and to develop a genuinely holistic approach for a complex natural object (p.549).

Since the 1980s an interactionist perspective on SLA has made great strides toward a more integrative approach to SLA, a view that is further developed in the following subsection (2.1.2.). More recently, Dynamic Systems Theory (DST), has offered new methodological approaches to the study of the L2, positing that language (and SLA) exhibits all the characteristics of what researchers in the hard sciences have termed complex, dynamic systems, and thus cannot be thoroughly measured through traditional methods. DST, as an attempt at a fully integrative theory of SLA is further discussed in Chapter 3, Section 3.3.

2.1.2. Interactionist approach

Partly in response to the abstraction of Chomskyan linguistic theory, largely divorcing language from real life communication, partly in response to the insufficiencies of Krashen's hypotheses composing his monitor model, in the 1990s a new approach to SLA emerged in what has come to be known as the cognitive-interactionist (or simply, 'interactionist') approach.

As early as the 1970s researchers became interested in the discourse patterns observed between native speakers and learners of the language. L1 research had described modified language on the part of adults toward young children "caregiver talk", and Ferguson (1971) noted similar modifications in the speech directed toward language learners with the purpose of making speech more comprehensible, terming it "foreigner talk". Long, the author of the Interaction Hypothesis (1981), considered how native speakers (NSs) modify their speech when speaking with non-natives (NNSs) altering their use of two linguistic phenomena: input (the linguistic forms used) and interaction (the functions served by those forms such as expansion, repetition, clarification). It was thought that this type of simplification made for more comprehensible input thus

facilitating acquisition of the L2. Early interactionist studies were largely only descriptive, reports on frequency of input modifications or characterizations of NNS directed speech (Gass & Varonis 1985; Long 1983; Pica, Young, & Doughty 1987), or how NNS negotiate meaning amongst themselves (Porter 1986; Varonis & Gass 1985) however, researchers soon made strides in explaining how input and interaction affect SLA, and what factors influence individual variation among learners (Larson-Freeman 1991)

Proponents of the interactionist approach drew from Krashen's work on the role of input, and particularly his input hypothesis that claimed that L2 acquisition would occur automatically (or subconsciously) in environments where there was a pairing of comprehensible input and a positive affective environment (Krashen 1977; 1985). However, the input hypothesis also claimed that as long as these two criteria were in place, comprehensible input, only slightly above the learner's current level of proficiency, would be sufficient to drive learners' acquisition on a subconscious level. In this sense, interactionists would later disagree with Krashen's claims (Swain 1985). While comprehensible input was clearly necessary, growing evidence would discount its *sufficiency* for acquisition of the TL. The wider collection of Krashen's hypotheses known as the Monitor Model gleaned criticism as being overly metaphorical and often failing to include precise definitions of key elements of his hypotheses (McLaughlin 1987; Greg 1984). From an interactionist perspective, comprehensible input is essential but not sufficient for acquisition to take place. (Swain 1985; 2005), and researches from the UG perspective would second the notion of the insufficiency of comprehensible input, or in their terms "positive input". It was argued that the type of environment where the learner was exposed to comprehensible input would not provide the learner with information about non-acceptable utterances in the L2 as he/she draws from what is acceptable in the L1 system (White 1991; 2003).

Neither Chomsky nor Krashen ever pronounced themselves on the role of interaction, a phenomenon that interactionists took to have a similarly vital role to that of comprehensible input in order for acquisition to take place, even thought to be the site of acquisition (Hatch 1978; Gass & Selinker 1994). Advancements in the field of SLA introduced important notions to the growing body of research on interaction. Some of these are Schmidt's (1990, 1993) work on noticing, claiming that new TL material

(=linguistic forms) can only be incorporated into the learner's developing L2 system as "intake" if they are consciously noticed. Discoveries having to do with the limited attentional resources of the L2 learner also contributed to overall picture (VanPatten 1989). Long's (1996) revision of the Interaction Hypothesis drew from current research in these areas, describing how negotiation of meaning between native speakers (or relative experts) and novices in the TL prompted interactional adjustments on the part of the more competent interlocutor, thus connecting input with the learners internal capacities, effectively channelling his/her selective attention and providing an opportunity for the NNS to produce output. Output, in itself became a focal point in a number of interaction centred studies.

Swain (1985; 1995; 2005) further developed the concept of output and how it contributes to acquisition. Producing output, she claimed, is crucial for L2 development in that it (1) gives learners the opportunity to practice and thus automatise elements of the L2, (2) allows students to test hypotheses concerning target forms, (3) forces focus on the structure of the TL and (4) draws the learner's attention to gaps in his/her knowledge of the TL. Swain's Output Hypothesis (2005) rounded out the interactionist perspective on SLA, closing a full circle from input to output.

Today in the field of SLA, it is generally accepted that there is a "robust connection between interaction and learning (Gass & Mackey 2007:176). Thus, in the last two decades research has shifted from asking "if" interaction impacts L2 outcomes to determining which aspects of the L2 benefit most from interaction, explaining how individual differences mediate the relationship between interaction and development in the L2 and how types of interaction as seen through different types of feedback impact the development of specific L2 forms (Ortega 2011).

In the end, it is the potential for receiving input and participating in interaction that makes SA such an interesting context for SLA research. When compared to the home environment, a SA period will generally offer massive amounts of input, as well as ample opportunities for interaction with competent and NSs of the TL. We have already addressed in section 1.3 some of the reasons why SA programmes may fail to deliver the maximum contact with the target culture (and thus fewer interaction opportunities), and we have also looked into learner dependent variables that can lead to little or no

improvement post-SA, many related to the inability to maintain significant levels of interaction in the TL community. Nonetheless, engaging in meaningful interaction in the TL culture remains the most valuable resource that SA contexts have to offer the language learner. Thus, more and more applied research is dedicated to maximizing the SA experience (e.g. Cohen, Paige, Shively, Emert, & Hoff 2005) in the hope of enabling students to best take advantage of what SA has to offer in terms of learning opportunities. In much the same vein, the body of research into the type and quantity of input and interaction within specific learning contexts has received a considerable amount of attention in recent years. We now turn our focus to such research and how it has contributed to what we know about SA and its impact.

2.1.3. Context sensitive approaches

First language and bilingual acquisition research is nearly always developed within a naturalistic setting. Babies and young children acquire their first language(s) within the family unit, or a slightly wider language community. However, SLA research into how older children, adolescents and adults learn additional language other than their mother tongue takes place in a wide range of contexts and settings. These range from purely naturalistic acquisition that takes place in informal settings, that is non-instructional settings (such as the learner's TL-speaking neighbourhood or workplace), to the traditional formal instruction classroom. Furthermore, we find many combinations of formal and naturalistic settings in between (e.g. L2 immersion) (Ortega 2011). In Secondary and higher education, SA is one of several contexts of learning that combines formal classroom and naturalistic exposure to the TL.

In this section we introduce the idea of learning context as it pertains to SLA. As SA is the principal context of interest in the present work, the descriptions of other contexts mainly serve to contextualize SA in the realm of secondary and higher education. We begin by characterizing educational and natural learning contexts as they have been dealt with in the previous literature, including a brief mention of the origin of context sensitive approaches to SLA, and how they have influenced studies of SA. We then go on to contrast educational and natural contexts in more detail while noting that clear borders where one context ends and another begins are seldom identifiable. Finally, we discuss what aspects of language acquisition these contexts purportedly facilitate and why that may be the case.

In research done among students of secondary and higher education, SA is often mentioned as one of three principal second language learning contexts, with formal instruction (FI), and immersion (IM) settings rounding out the triad (Collentine & Freed 2004; Collentine 2009; Pérez-Vidal 2011). As these are the main study options available at these levels, the comparison of two types of context has been the focus of a number of studies of the impact of context on SLA (Segalowitz, Freed, Collentine, Lafford, Lazar, & Diaz-Campos 2004; Segalowitz & Freed 2004; Trenchs-Parera 2009). Comparisons of all three settings: FI, IM and SA, are fewer in number, but some examples of such endeavours are Freed et al. (2004) and Dewey (2004; 2008). Context descriptions of this type fit well with the movement in recent decades as researches have taken on context-sensitive theoretical approaches to SLA.

In the 1960s and early 70s Dell Hymes contributed greatly to what was then the just emerging field of sociolinguistics. His work on the ethnography of communication equipped researchers in the field with the tools needed to describe speech situations and the players that shape discourse (Johnstone & Marcellino 2010). His work marked a clear alternative to generative linguistics that was the principal line of research at the time, and drastically shifted focus from a purely theoretical concept of language to the context in which language takes place. Context-sensitive approaches consider linguistic knowledge as more of an extension of culture and experience rather than some separate entity, divorced from everyday living and communication (Atkinson 2002; Collentine & Freed 2004), the nature of which must be taken into account together with cognitive approaches to SLA if we are to represent a more complete picture of SLA.

2.1.3.1. Educational versus natural settings

Returning to our discussion of learning context, we can think of these different settings, FI, IM and SA, as three somewhat blurry points on a continuum where two manners of engagement with the TL, one form focused and the other meaning focused, are measured out in different proportions and leading to more formal or more informal learning. Batstone (2002) calls these approaches to the TL, ways of "contextual engagement," and uses the terms "learning context" and "communicative context" to distinguish formal contexts from more naturalistic ones. Batstone's ways of engagement roughly coincide with Krashen's notion of contexts that lead to learning versus those

that lead to acquisition (Krashen 1976), and the same contrast is made in Ellis's (1994) description of natural versus educational contexts. Here, for clarity's sake we follow Ellis's terminology, referring to these as educational and natural contexts/settings respectively.

The principal difference between educational and natural settings arises from the type of contact that learners have with other speakers of the L2 and with the TL itself. Educational context is the type of engagement that typically characterizes the second/foreign language classroom. The teacher manages the classroom dynamic with the goal being that students attend to form. In other words, this context involves conscious attention to linguistic forms. The natural context, on the other hand, is best represented by those interactions where the use of language is but a means of exchanging information and is not consciously form focused. Here learning is the result of the learner's direct observation and participation without any explicit mention of underlying principles or rules. Examples of natural settings can be drawn from an infinite array of every-day situations and experiences outside the classroom such as ordering lunch at a restaurant, asking for directions on the street or chatting with a friend. Some learners will receive no formal language instruction and thus will experience the L2 in a completely natural setting. This is the typical case of adult emigrants from Asian countries who arrive in Europe. Many will never set foot in a language classroom; their contact with the L2 occurs in their workplace, through the local media and in a diversity of other settings wholly naturalistic in nature. Other learners' experience with the TL will be nearly entirely in the educational setting. For instance, US high school students whose first and last significant contact with German occurs in the language classroom experience the TL within the educational setting only. These examples are hypothetical, and purely formal or purely informal learning may be quite rare occurrences. Most learners will experience a mixture of both contexts in varying proportions.

It is assumed that the learning that takes place in natural settings and that which takes place in educational settings is very different. Natural settings provide informal learning opportunities; this is understood to be learning as a result of participation and observation rather than through the explicit articulation of linguistic rules. Conversely, formal learning takes place through conscious attention to linguistic rules, and there is a

greater emphasis on the language as subject matter and its mastery (Ellis 1994) That being said, it should be noted that cut-off points where educational contexts end and natural contexts begin (and vice versa) are seldom clear. Ellis (1994:215) puts it as follows:

> Learners in natural settings often resort to conscious learning and may deliberately seek out opportunities to practise specific linguistic items they have studied[....] Conversely, learners in classrooms may not be required to treat the language as 'subject matter', but instead be given opportunities for acquisition[...] [...]The correlation between educational settings and formal language learning depends on the pedagogic approach.

Nonetheless, in general terms, educational context is the manner of engagement that best characterizes the foreign language classroom, and natural contexts more often lead to informal, often spontaneous, out-of-class learning.

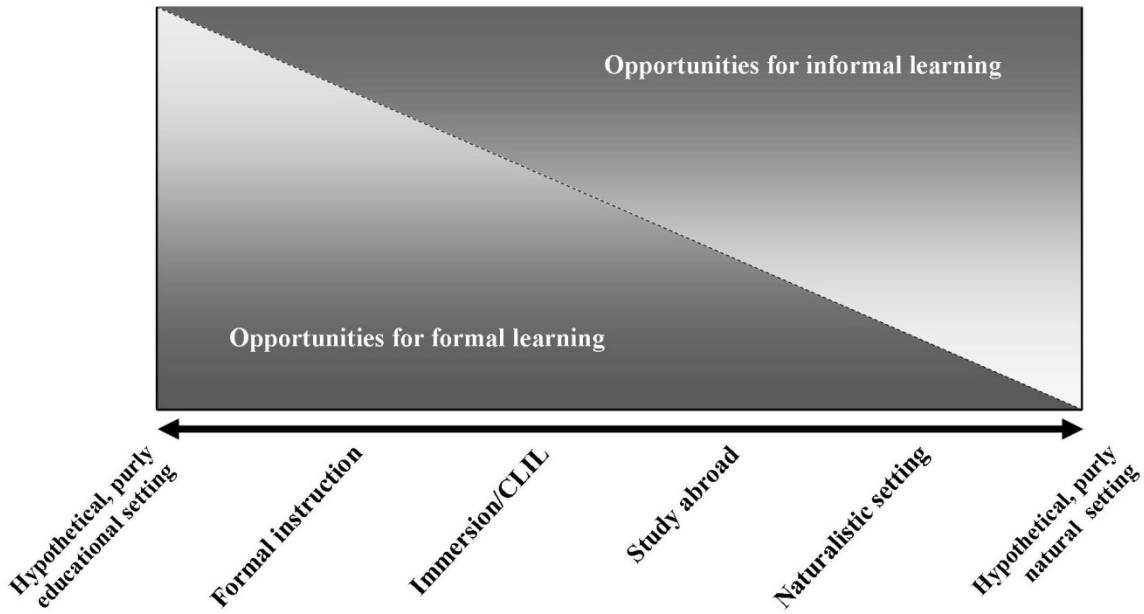


Figure 2.1 Formal and informal learning opportunities by learning context

As one moves along the continuum from a more typically formal context, such as the language classroom, toward a purely naturalistic setting, likely free of any FI whatsoever. We see that the types and proportions of opportunities for learning vary depending upon the learning context at hand. In Figure 2.1 we offer a visual summary

of these types of learning settings in terms of amount of exposure to each type of context (educational to natural) and thus, the (in)formal opportunities associates with each. In figure 2.1 the relative proportions of learning opportunities are represented by the triangular figures that coincide with the scale below. The point on the scale representing FI thus coincides with a high degree of form focused input and educational context and a very small degree of natural input/interaction and informal learning opportunities. At the other extreme we see that the naturalistic setting is dominated by the natural setting with little to no educational context involved. Different points represent different "mixtures" of the two settings. The centre of the figure represents those contexts where learning and communicative contexts maintain a relative balance as in the case of IM and SA to varying degrees.

2.1.3.2. What different contexts facilitate

It is also thought that each type of engagement facilitates different types of language learning. Educational contexts tend to facilitate access to new linguistic forms as the interaction is most typically directed by a teacher who explicitly guides students' attention toward form. Here the discourse includes explicit focus on the language as the object of learning, thus intake of new forms is very much one of the main purposes of these encounters and the ultimate expectation of both teacher and learner. Along with this type of support, learning contexts tend to deemphasise the social consequences of any errors made while trying out new forms in production. After all, the language classroom is a community of relative novices in the TL, and does not demand the performance of its members that out-of-class, real-world interaction often would. Ideally, the learning context thus facilitates both the intake of new forms and the risk-taking associated with the use of newly acquired forms. The latter may be especially important in "pushing out" those forms that are not yet automatised for the learner and thus more effortful and challenging. The concepts of intake and pushing out output have been briefly discussed in section 2.1.2, for a comprehensive discussion see Swain (1985).

While in educational contexts access to new forms is generally quite deliberate and interactions are contrived to meet the learners next-level needs for growth, FI requires little of the learners in terms of the meaning driven communication, or social appropriateness of their production. The language classroom may allow students to

freely experiment with new forms without the risk of losing face though frequent or grievous error; however these uses of the TL are made outside of the community of native speakers or habitual users of the language. The artificial FI speech community may lack socio-cultural information that real-world interaction can provide. Socio-cultural approaches to SLA see the wider language community as a society of language experts, that is, native speaking adults, who corporately transmit socially constructed linguistic meaning and appropriateness to novices (children and L2 learners) allowing them to eventually achieve competence in the language (Lantolf 1994; Atkinson 2002). Thus situations where the natural context predominates may be better suited for connecting form and function in the mind of the learner (Collentine & Freed 2004).

Researchers with an interest in SA learning contexts have been draw by the offer of both educational contexts and natural contexts within the same setting. Collentine and Freed (2004) refer to this mix of educational and natural settings when they note that "studying abroad heavily involves both communicative and learning contexts which may entail a hybrid communicative-learning context". Similarly, Miller and Ginsburg (1995), describe how students draw from what they have learned in one context in order to negotiate the other. The potential for this contextual "hybridization" is very attractive, for researchers and educators alike, and while the truth behind the efficacy of SA contexts, as we have seen in Chapter 1, is not as straightforward as all that, SA continues to be a rich context for research due to the coexistence of educational and natural modalities. Thus far, we have characterized the two extremes of the learning contexts presented in Figure 2.1, FI and naturalistic settings. We will now look into two midpoint scenarios that are common in secondary and higher education environments.

Starting at the point making FI in Figure 2.1, and moving along our imaginary continuum from formal to informal situations we find the IM/CLIL context. The IM setting is constructed by a community of instructors and language learners who use the L2 in all aspects of study, and perhaps when interacting outside the classroom to some extent, but without leaving the wider AH, L1 context (Collentine & Freed 2004). Programmes of this type have enjoyed popularity in Canada, and a number of studies have arisen from the growth of French IM in North America. For a comprehensive review of literature associated with French IM and SLA see Swain (2000). In IM we find that educational and natural contexts come more into balance than in FI and

naturalistic contexts in terms of time spent in either educational or natural modalities. In IM, the TL may be the object of learning in a particular subject, but it is also used as a communicative means to an end when tackling other areas of study, and not always as the explicit focus of study. In figure 2.1 we also include a sub-grouping of programmes that have drawn increasing interest in recent years, namely Content and Language Integrated Learning (CLIL) classrooms. Although differing from IM to some degree, CLIL is largely based on the same premise as IM, and like IM, takes place in the home country. While CLIL programmes can be found in other parts of the world, they have been most widely embraced in Europe. This is due to the fact that it is seen as beneficial to the international movement toward multilingualism (Ruiz de Zarobe & Jiménez-Catalán 2009: xi) a tendency that is discussed at length in Section 1.1.2.2 of this dissertation. In this method, courses are imparted in the TL and students are expected to perform through this medium. CLIL methodologies are diverse, and control of "who gets what" in terms of quality input is still quite unclear across studies, yet for a thorough review of the current CLIL literature see Dalton-Puffer (2011). In Figure 2.1 CLIL is placed on the continuum between with IM, but this placement could vary depending upon the nature of the CLIL setting itself. For our purposes, it is sufficient that we understand IM and CLIL as other contexts of learning that serve to give us perspective on instruction settings and current modalities of language learning overall.

Moving still further down the line in Figure 2.1 we have SA contexts. In this case, any learning context specific to the TL itself is reduced. Instead, the communicative context increasingly dominates the learners' interactions in the TL. It should be noted that, depending upon programme design, students may or may not receive specific language training while abroad. It is at this point that we begin to see the enormous assortment of programme features across SA endeavours, and how elusive a concrete characterisation of SA can be. In those cases where no instruction in the TL is provided while abroad, and barring any own initiative on the part of the learner to receive instruction, the context is practically purely naturalistic in nature. In such cases the TL is the medium through which content learning takes place and through which objectives are accomplished, but not the object of study. This would be true of many Erasmus exchanges in Europe, where students are often enrolled in content courses having little or nothing explicitly to do with language learning. The variation in programme features makes the questions of how much of each type of learning opportunities learners

experience a complex question to answer across the board. Nonetheless, a generalization to be taken with some caution is that SA contexts provide more opportunities for use in communicative contexts, thus an input rich environment and enormous amounts of practice in the TL (Pérez-Vidal, 2011). Still we must keep in mind a caveat that holds true for SA "It is rarely, if ever, possible to describe the quality and extent of social contact and linguistic interaction." (Freed 1995a). Nearly two decades later, little conclusive evidence to support popular intuitions about the superiority of the SA "hybrid" learning environment. However, some research into how social networks and language contact (e.g. Mitchell, Tracy-Ventura, McManus, Richard, & Romero de Mills 2013) seek to compensate for what has been until now extremely difficult to determine, just what the SA context looks like in terms of input and interaction, and how learners engage with the host environment.

2.2. Principal speech production models

In this section we briefly describe the principal models of speech production that have most notably gained acceptance in cognitive linguistics and SLA research. The principal model discussed here, namely Levelt's (1989) modular model has been fundamental in research related to oral production for decades. The empirical study contained in Part II of this dissertation, while not directly investigating the mechanisms underlying speech production, does assume, at least to some degree, that the processes described here are those at work in what we can observe about fluency, accuracy and complexity in subjects' speech productions. De Bot (1992) initially applied Levelt's model to bilingual speakers and language learners and has relevance for understanding the speech production of the multilingual subjects who participated in the study.

Speech production modelling seeks to describe what is considered to be the subsystem of the whole of human cognition dealing with language production. Levelt (1989) refers to speaker as an "information processor," and language production is commonly seen as a specific way of taking an initial intention through a series of hierarchical steps on its way to an appropriate, meaningful utterance. As each cognitive step takes place, the output more closely resembles actual speech until the final product, perceivable speech, is produced.

Traditionally, speech errors have been our only window into the realities of speech processing. Depending upon what types of errors are observed, or indeed, those which are not produced, certain assertions can be made about the existence of stages or levels of processing (Crookes 1991). However, more recently, error analysis has been increasingly replaced by techniques such as, newer ways of measuring fluency, studying reaction time in picture naming tasks for instance, or even computer driven simulations, or connectionist modelling, of how the human brain may actually produce spoken language (Harley 2001).

Models of speech production generally fall under two categories, spreading activation models, and modular models. Here we summarize the main characteristics of each, how they are similar and how they differ. Likewise, we discuss the implications of being bilingual or multilingual, and how that condition is thought to influence speech processing in the speaker of more than one language.

By far the most influential model of speech production in recent history is a model of L1 language production proposed by Levelt (1989) who divided his "blueprint for the speaker" into three principal processing components: conceptualization, formulation and articulation (or encoding). These components function interdependently, yet are largely autonomous and able to work in a parallel fashion to produce speech.

The Conceptualizer is where a communicational goal is set along with accompanying subgoals. The retrieval of all necessary conceptual information to be expressed, and thus meet the set goals, is also performed at this stage. The product of this phase is what is called the preverbal message, the intention of the speaker prior to specifying the linguistic forms to go along with that message. It is also thought that self-monitoring takes place here, where the speaker attends to his own speech, controlling for the correctness and content of his message. Conceptualization involves stages of both macro-planning and micro-planning. The macro-planning phase is where intentions are organized into a goal and subgoals and the retrieval of all necessary information takes place. A second phase is characterized by a micro-planning stage. Micro-planning involves structuring the necessary information into the propositions to be expressed, and making decisions such as what the focus of the utterance will be. Thus, micro-planning

allows the speaker to guide the addressee's attention in accordance with his intended message.

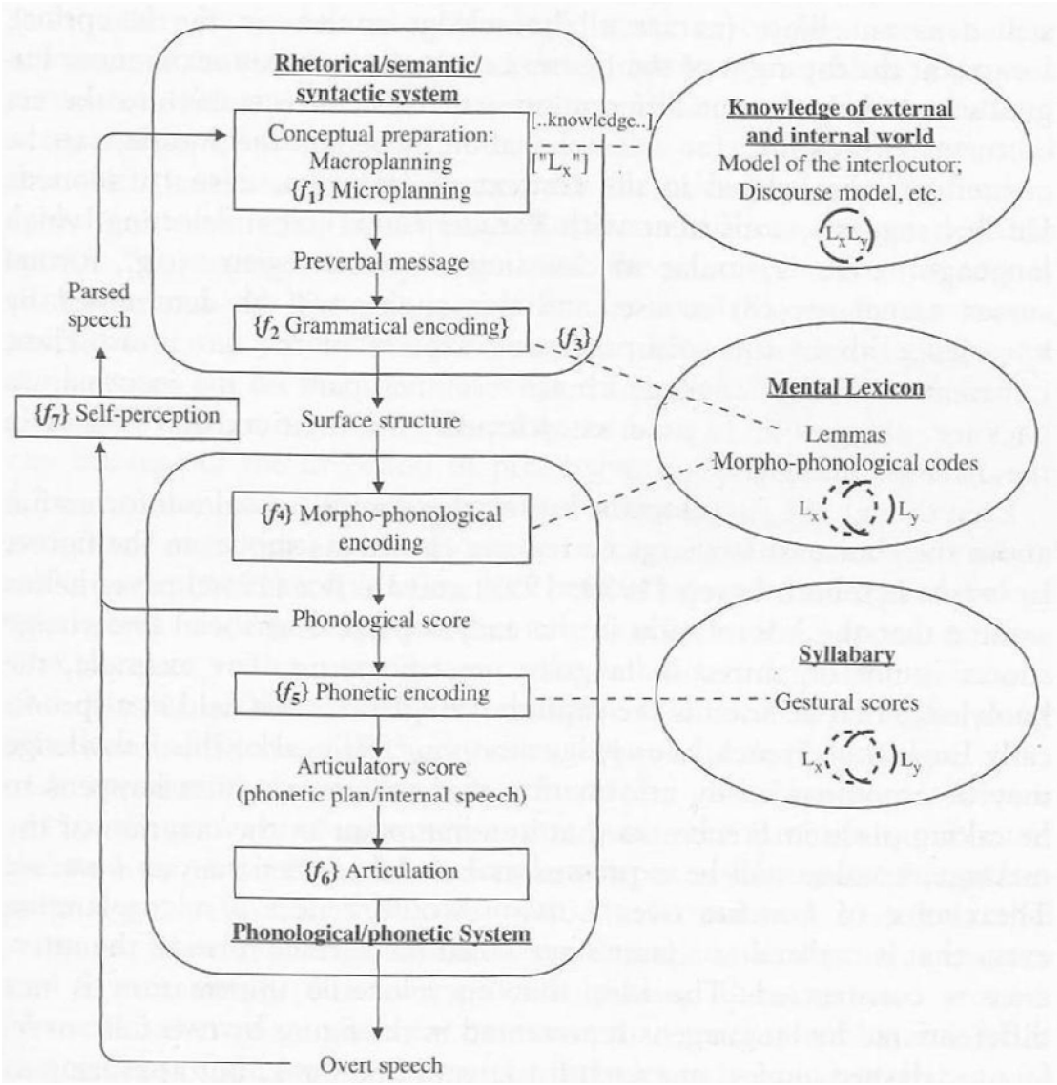
Moving from conceptualization to formulation, we find that the Formulator, in simple terms, "translates a conceptual structure into a linguistic structure (Levelt 1989:11). This subcomponent of the model draws from the preverbal message produced through the process of conceptualization taking in the preverbal message as input to be further processed. It is here, in the formulation phase, that specific linguistic forms are applied to the preverbal message. Formulation can be broken down into two major functions according to modular models: lexicalization, where lemmas are selected, and syntactic planning, where the selections are arranged in appropriate order. A final phase where a phonological plan is prepared is also within the domain of the Formulator (Harley 2001), while Levelt (1989) would lump lexical retrieval and syntactic planning together under those processes performed by the Grammatical Encoder, its product being the surface structure of the message to be expressed. The message represented by the surface structure then passes to the Phonological Encoder were a phonetic plan (or internal speech) is produced. In moments where the phonetic plan may not be immediately executed, it may be temporarily stored in an Articulatory Buffer for later use. Whether immediately or postponed, the message represented in the phonetic plan becomes the input for the next subcomponent, namely, the Articulator.

The Articulator functions at the end of the speech production line. Here the phonetic plan is executed through the use of the muscles that control the vocal tract. The Articulator is fed "chunks of information" in the form of internal speech reserved in the Articulatory Buffer and produces overt speech. At this state the process comes full circle in the monitoring function that the Conceptualizer is thought to perform. As speech is monitored by the Conceptualizer, the message has completed the speech production circuit. That is, the Conceptualizer acts as both a generator of messages and an editor for internal speech and overt, spoken messages either on the part of the speaker himself or his interlocutor.

Figure 2.2 illustrates Levelt's updated, 1999 Model (as adapted by Segalowitz 2010:9). In this figure, Segalowitz provides an adapted framework which accounts for both monolingual and bilingual (L2) processing, incorporating De Bot's (1992) proposals on

second language variations on the original monolingual model. Dotted and dashed circles that surround L_x and L_y (i.e. L1 and an additional language) represent how information regarding the speaker's languages x and y are considered to be related to one another. Where circles partially overlap, it is thought that systems are partially distinct. Fully overlapping systems, represented by circles that fully overlap, are considered to be undifferentiated in the speaker's mind. The symbol {*f*} denotes "fluency vulnerability points", referring to those moments along the production process where underlying processing difficulties are associated with dysfluenies in overt L2 speech.

Figure 2.2 Levelt's updated "blueprint for the speaker" taken from Segalowitz (2010).



Levelt's and other similar characterizations of speech processing are commonly referred to as modular models, where each subcomponent works autonomously, taking in a

specific type of input and producing a different type of output. This model also assumes that the process of speech production is lexically driven. That is, the retrieval of words from the mental lexicon activates syntactic building procedures already specified as part of the information associated with each lemma. The words themselves and their meanings are the basis for syntactic construction. Furthermore, modular models only allow for error correction through self-monitoring after erroneous speech has been produced as internal or overt speech. It is then that the message can be reformulated and correctly delivered. Thus, the speaker monitors his own speech much in the same way that he does that of his interlocutor. As a message moves from intention to articulation there is no feedback along the line of production. If this is the case, an error produced at any point in the series of steps toward overt speech will not be corrected until phonetically encoded or even audibly perceived by the author of the message himself.

An alternative to the modular view of Speech production is a spreading activation Model (Dell 1986). Spreading activation models also arose out of a desire to explain common speech errors and discover clues to the inner workings of the speech production system through the observation of those errors. Indeed, spreading activation models assume many of the same processes as modular models. For example, like modular models, spreading activation models also assume varying levels of speaker knowledge about their language. Semantic, syntactic, morphological and phonological levels are often assumed and explainable through generative linguistic theory (Dell 1986) and what neurologists know about brain function (Kormos 2006). Also, as in modular models spreading activation models assume conceptualization, formulation, articulation, and self-monitoring constitute the principal language processing levels, and that processing happens in that same hierarchical order. Likewise, at least in L1 speech production, planning requires attention, while formulation and articulation are automatic and information is transmitted by activation spreading. The speech production system is thought of as a network containing “nodes” for linguistic units. These can be in the form of concepts, words, morphemes, phonemes, and phonemic features, and in some cases syllables and syllabic constituents are also included in the network. A conceptual node may activate a word node, defining the word to be used. A word node may connect with a morphological node, which connects with a phoneme, and on down the line toward the production of overt speech.

Although fundamentally quite similar, spreading activation models differ from strictly modular models in two ways. First, instead of assuming that production is lexically driven spreading activation models propose that syntax is set in place before lexical items are selected and inserted into appropriate "slots". Thus, spreading activation is a syntactically driven model inspired in generative theory as in Chomsky's (1965) treatment of syntax. Modular models, in contrast, are thought to be lexically driven. Second, spreading activation theories assume that feedback regarding error can occur at any stage of processing and allowing for a backward flow of activation. When an error occurs a type of warning signal is issued and the problem is immediately attended to, no matter the processing phase in which the message is found at the moment the error occurs. Modular models do not include this feedback between levels of encoding, as modules are thought to be independent, feeding off of a unique type of input at any given stage. For the modular modal proponent, error is dealt with in the final stages of encoding through self-monitoring, and if detected would be sent back through the whole process starting with the Conceptualizer.

De Bot (1992) was the first to adapt Levelt's model to the bilingual/multilingual condition. De Bot does not seek to reinvent the model, but rather considers what differences in the bilingual system might be postulated given that the speaker is managing two (or more) languages. An effort is made to hold true to the monolingual model as much as possible while also seeking to explain the uniqueness of bilingual speech production.

A fully agreed upon definition of bilingualism is still unsettled in the relevant literature. Whether second languages learned later in life (beyond early childhood), by an otherwise monolingual speaker, make him or her bilingual, or whether the term is reserved for only well balanced "ambilingual" speakers, is also a point of some debate (Dewaele, Housen, & Wei 2003). Some would consider only the most proficient L2 speakers to be bilinguals (Bloomfield 1933), while others opt for a more inclusive definition of bilingualism such as Grosjean's (2010), who sees bilinguals as people who "use two or more languages in their everyday lives" (p. 22). Indeed, the trend in the field is to use an inclusive definition, allowing for a wide range of degrees of proficiency in all the user's languages, acknowledging the fact that perfect bilinguals probably do not exist, and even balanced bilinguals are quite rare (Dewaele et al. 2003).

Furthermore, a broader definition of bilingualism allows for the inclusion of L2s that the speaker is in the process of acquiring. Inclusive definitions of bilingualism are in line with a more recent focus in SLA on communicative competence (as opposed to formal linguistic rules) (Butler 2012). Thus, broad definitions include not only balanced bilinguals having been exposed to his/her languages from birth, but also "various 'imperfect' and 'unstable' forms of bilingualism, in which one language takes over from the other(s) on at least some occasions and for some instances of language use" (Dewaele et al. 2003)

Where many would draw the line in defining bilingualism, however, would be in considering bilingualism as equal to multilingualism. Some studies have pointed to differences between bilinguals and multilinguals that should not be ignored. These differences may include the influence of having (an) additional language(s), while bilinguals only draw on one additional language, cognitive differences, greater metalinguistic awareness and communication strategies (Butler 2012). Therefore, there is some debate as to whether or not to lump bilingualism and multilingualism together into a third term designating the knowledge and use of more than one language, some preferring the term "plurilingualism" to refer to the grouping of bilinguals and multilinguals (e.g. Mackey 2012), while still taking into account their differences.

For the bilingual speaker, as in monolingual processing, the model is incremental and parallel. Each component works rather autonomously at any given level of message representation, and lower level processing is more automatised than higher level processing. The main differences in the bilingual system, as cited by De Bot (1992; 2005), are as follows. Firstly, it is thought that the highest level component, the Conceptualizer, is probably language independent in the speaker of more than one language. However, there appear to be distinct Formulators for each language as the Formulator deals with grammatical information, and grammar varies across languages. It is theorized that the Formulator draws from one lexicon (De Bot 1992) where items from both/all the speaker's languages are stored. The message then moves to the Articulator which seems to have a wide range of speech motor plans which are not necessarily language specific. In De Bot's later work (2002; 2005) the idea of a single lexicon comes into question, especially when dealing with learners of a second language with varying levels of proficiency.

The modular theory of speech production has clearly dominated both L1 and L2 research in the field. Modular theory is more systematic and detailed in its representations of verbal processing, and a great body of empirical studies based on its precepts have pushed modular models to the forefront in the language sciences.

2.3. Complexity, accuracy and fluency in L2 oral production

Here we briefly review how the dimensions of CAF have been viewed in the relevant literature. We will first define what the CAF triad represents in language research overall. After that, we will discuss each element of CAF separately, noting how they have been characterised in previous research. Finally we will expound upon current theories of how the three dimensions interact in speech production and under specific conditions.

Although some references to a fluency-accuracy dichotomy had already been seen in related literature (Hammerly 1991) in the 1980s and 90s, Skehan (1998) seems to have been the first to unify the three dimensions as the three principal elements of proficiency.

In recent years the research domain of CAF has drawn increased attention among academics in the field of SLA and applied linguistics in general. The definitions and nature of CAF components are, as a result, becoming ever better defined. However, the very broad scope CAF is quite difficult to encompass in a few simple definitions, and the multiple takes on the triad have generated a great deal of overlap of terms which may change in meaning considerably from one study to the next.

Complexity is by far the most troublesome of the three when it comes to characterising the CAF constructs. For example, complexity may refer to how difficult a given task is to perform (relative complexity, or in plain terms, “difficulty”), or it could refer to the absolute complexity of the language itself (to what extent the morphosyntactic structure of a sentence, for example, is “complex”). Complexity in the L2 can be dealt with at the linguistic, discourse-interactional or propositional level. We can still further filter out structure complexity and contrast it with system complexity and at the lexical, morphological, syntactic or phonological level. For a very complete taxonomy of

complexity see Bulté and Housen (2012). Most of these facets of complexity will not be dealt with here for brevity's sake, however, anyone seeking to build on an overview of complexity (not to mention accuracy and fluency), must understand the tremendous breadth of the topic that is often channelled according to a specific scientific approach be it linguistic, cognitive or pedagogic (Housen, Kuiken and Vedder 2012b). Therefore, before isolating a particular branch on which to focus attention, one must, albeit briefly, take in the enormous scope of CAF in all its varieties and functions.

When applied to L2 performance and proficiency, there is a wide consensus among researchers in language sciences that affirms CAF constructs as "multi-componential in nature". There is also a broad agreement that the main facets of performance/proficiency can be (roughly) divided into the principal dimensions of complexity, accuracy and fluency (Housen & Kuiken 2009). In recent years CAF have frequently been utilised as research variables in SLA. When applied to L2 performance, CAF are used as descriptors when assessing oral and/or written language. When applied to proficiency, CAF become indicators of a deeper linguistic knowledge underlying performance. Finally, CAF have been used as measure progress in L2 acquisition and language learning (Housen & Kuiken 2009.).

The CAF triad is in essence, an aid to learning about the nature of language through experimentation. As Housen and colleagues (2012b) point out, CAF do not represent a specific type of analysis nor methodological approach at present. To date, Research efforts dealing with CAF constructs also fall short of providing a theory of aspects of the L2 such as language learning, development or use. CAF have consistently been found to be dimensions of language that allow for systematic inquiry and observation of second language performance/proficiency. Yet, it is premature to attempt to fit CAF into a specific theory of language production or acquisition. Therefore, in the present work, we treat CAF for what they are, a three-fold tool for observation, and do not seek to assign them further theoretical dimensions.

2.3.1 CAF as speech dimensions

Pallotti (2009) notes that, while CAF do not represent any theory or even any specific line of research in and of themselves. They do serve as dimensions for describing language performance. CAF are most frequently used as dependent variables so as to

examine variation in independent variables such as learner level or task features (e.g. task complexity or planning time). CAF measures have been operationalised in a number of ways and although many find them useful, many would also agree that further clarification regarding their validity is needed and that CAF, as it is used today, does not yet provide an exhaustive description of language performance.

Complexity is "the extent to which the language produced in performing a task is elaborate and varied" (Ellis 2003). Although this definition may seem straight forward, as already mentioned, complexity is by far the most problematic dimension of the CAF triad in terms of settling on a clear definition, and meaningful operationalisation of specific measures. The term "complexity" in language sciences has been used in a myriad of contexts and with reference to very different language related phenomena. "Language complexity has cognitive and linguistic dimensions, and performance and developmental facets, and can manifest itself at all levels of language structure and use" (Bulté & Housen 2012). Though the scope of language complexity is immense, recent efforts to make sense of the multiple definitions and uses of the term "complexity" have shed light on the otherwise daunting task of deciding on a particular type of complexity within language complexity as a whole. Here we briefly review language complexity in all its facets, then hone in on the specific type of complexity that is dealt with in the empirical study of Part II.

In a comprehensive taxonomic survey of L2 complexity, Bulté and Housen (2012: 22-27) make a number of important distinctions in the classifications of types of second language complexity from which we draw from here. At the most basic level complexity is defined as a property (or quality) of a phenomenon or entity in terms of (1) the number and nature of its discrete components and (2) the number and nature of the relationship(s) between these. That is, some entity that is made up of (multiple) component parts that are distinguishable one from another, yet interact in some sort of interconnectedness.

In some branches of language sciences, such as psycholinguistics, we run across a type of complexity that deals with the relative complexity of certain language components. Cognitive complexity, sometimes called simply "difficulty", refers to the costliness of a particular language feature. If a feature is relatively complex it is more taxing on the

mental processing of the language user or learner, in the case of L2, than a less relatively complex feature. This type of complexity is thus user dependent. That is, while some learners may find a certain feature complex (taxing on their processing resources), another learner may find the same feature less costly to produce, and could even produce the feature with ease. Examples of studies that take on this relative approach to complexity can be found in Byrnes and Sinicrope (2008) and Diessel (2004). While it is important to be aware of this use of the term "complexity", the use of the term we adopt in the study contained in Part II refers to absolute complexity. We now take a closer look at what we mean by absolute complexity.

An absolute (or inherent) take on complexity considers complexity in objective terms. Instead of complexity being determined by the user of the language as he/she experiences more or less difficulty, absolute complexity is defined in quantitative terms. Thus the degree of complexity dealt with in this case depends on the number of components that a particular feature consists of, and the number of connections that exist between its discrete components. In current study in Part II, when we make reference to complexity, we refer to this second, objective approach to the notion of this dimension.

Under "absolute complexity", Bulté and Housen (2012) move further down their L2 complexity taxonomy, to "linguistic complexity". Although other forms of absolute complexity have emerged in recent years, namely discourse-interactional and propositional approaches, linguistic complexity remains the most scrutinized in L2 research and can deal with either the learner's global (and dynamic) L2 system as a whole, or can refer to a property of individual linguistic forms such as specific structures or rules that make up the learner's L2 system. While global complexity describes the extent of the elaboration, size, breadth or richness of the learner's repertoire as a whole, local (also call structural) complexity is focused on the depth of the system as reflected in the manifestation of specific linguistic items. The later is the definition that we take on in the empirical study contained in this dissertation. Figure 2.3 is a simplified version of Bulté and Housen's (2012) taxonomy of L2 complexity. For brevity's sake we adapted the wider taxonomy to start at the linguistic level, excluding higher levels of L2 complexity.

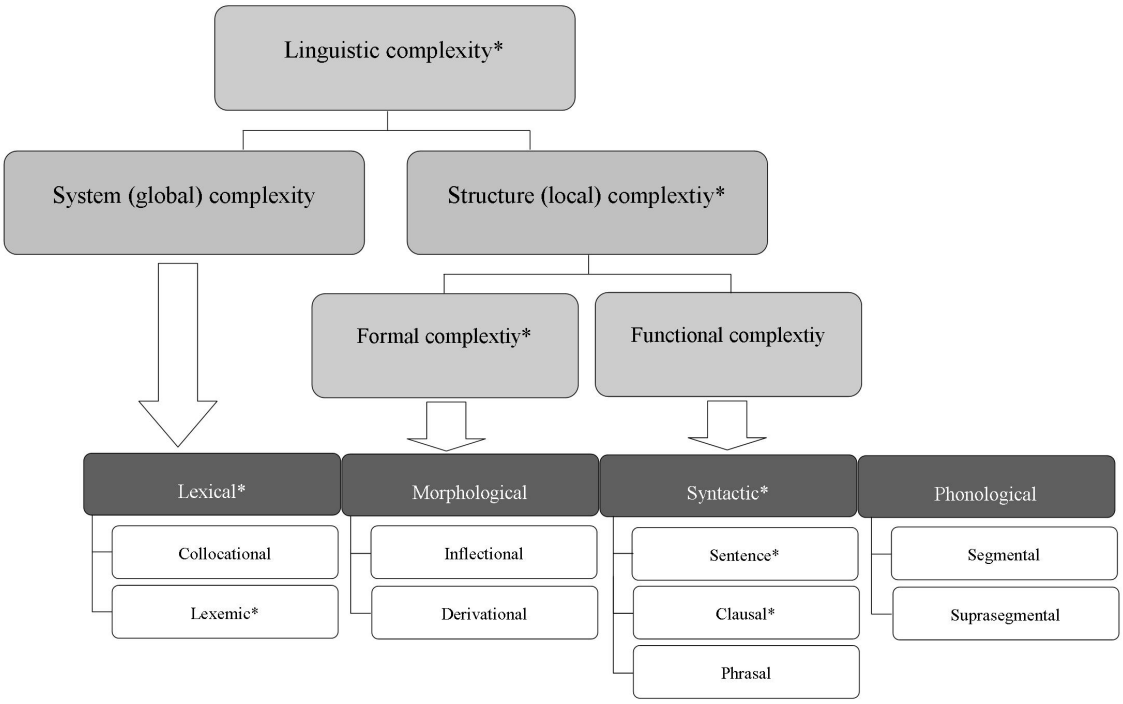


Figure 2.3 Taxonomy of complexity constructs (adapted from Bulté and Housen 2012)

Levels marked with (*) are relevant to the current study found in Part II, and measured through indices of lexical (Guiraud's Index) and syntactic complexity (words, clauses, subordinate clauses, and coordinated clauses per AS-unit[16]).

In the L2, accuracy is always considered with reference to the norm, that is, the target language system and presumably the intuitions and productions of NSs of the target language. Accuracy has been defined as "how well the target language is produced in relation to the rule system of the target language" (Skehan 1996:23). Of the CAF triad, Accuracy is the dimension of speech that authors tend to dwell on the least, and when compared to complexity and fluency, accuracy stirs up little discussion at present. However, there was a time when accuracy (or error rather) was the central discussion of what was to become the study of SLA. Behaviourists, up until the 1960s, thought that error was a settled question in second/foreign language acquisition. Inaccuracies in learner production, it was thought, were due simply to negative transfer from the L1 and thus learning a new language was really nothing more than forming new language

[16] In the taxonomy in Figure 2.3 the sentence level is included. The study in Part II the Unit most closely resembling the sentence is AS-unit, the denominator in the complexity measures mentioned here.

habits. Contrastive analysis, the process by which L1 and L2 systems were systematically compared, would in the behaviourist's view, lead to a list of discrepancies or problematic features between the native language and TL systems that would predictably correspond to the areas in which learners would have difficulty (Ellis & Barkhuizen 2005). Thus, for learning to take place, negative transfer from the L1 would have to be overcome by the learner.

Selinker was one of the first to take issue with behaviourism and the practice of contrastive analysis, and brought the analysis of learner error to the forefront in language learning research, turning language studies toward the observation of real learner production for the first time. This gave rise to an array of theories in the then still emerging field of SLA.

While the predictive power of contrastive analysis in second language learning was postulated based on anecdotal evidence and theorists' intuition about the effects of transfer, Tarone (2006) describes the advent of error analysis as follows:

> Error analysis was an enterprise born of the attempt to validate the predictions of contrastive analysis by systematically gathering and analyzing the speech and writing of second-language learners. For perhaps the first time in history, the focus moved from teaching materials and hypotheses about second-language learning problems, to the systematic observation of learner language.

Of course, some decades later, we know that the outcome was not the confirmation of contrastive analysis, but rather the development of the notion of interlanguage (Selinker 1972), a concept that would shape SLA research considerably over the long term. In his observation of real life L2 speech samples and the (in)accuracies contained therein, Selinker comes to the conclusion that "one would be completely justified in hypothesizing…the existence of a separate linguistic system based on the observable

output which results from a learner's attempted production of a TL[17] norm" (Selinker 1972:214).

Since the onset of the practice of error analysis and the development of the Interlanguage hypothesis, much of the work done in SLA with regard to accuracy/error has serve to defined what accuracy is not, or how it cannot be applied to data in order to gauge development over time, or postulate mental process of L2 speech production. Some have attempted to over systematize inaccuracies present in the interlanguage (Tarone, Frauenfelder & Selinker 1976, as cited by Bley-Vroman 1983), while others have confounded accuracy with other constructs (e.g. comprehensibility).

Bley-Vroman (1983) published a critique of Tarone, Fruenfelder and Selinker's 1976 work that for Bley-Vroman, was a regression to the ways of contrastive analysis, drawing too heavily from the comparison of L1 and target systems, and missing the original point of the Interlanguage Hypothesis, that is, the observation and description of the IL itself. The conclusion that can still be drawn from Bley-Vroman's criticisms is that "accuracy *per se* is not a direct indicator of interlanguage development" (Pallotti 2009). A similar "dead end" in research related to accuracy was the idea of assigning relative weights to errors, with the notion that some errors are more grievous than others as they impede comprehension to a greater degree thus compromising communication (e.g. Humburg 1984). While error gravity studies were getting at a very real problem in L2 communication, that of comprehensibility, the approach of assigning weights to errors confounds the two constructs (Pallotti 2009). Pallotti also illustrates how a given speech sample may be perfectly accurate, yet utterly meaningless and uncommunicative evoking Chomsky's famous nonsense phrase "Colorless green ideas sleep furiously", perfectly grammatical, error free, yet nonsensical. Likewise, if a given speech sample contains 7 errors, yet those errors do not impede comprehension and thus are assigned low gravity scores, then that speech sample would not be more *accurate* than a speech sample of the same length containing three grievous errors, although depending on the scoring system, the numbers may indicate just that. The second sample with fewer errors could indeed be the *less* accurate of the two should the errors be weighty enough.

[17] With "TL" Selinker (1972:214) refers to the "target language" just as we have done throughout the present work.

Studies of intelligibility/comprehensibility as a construct in its own right and utilizing a very different methodology have replaced this technique of error analysis.

Error analysis also lead to the discovery and subsequent discussion of developmental sequences in learner language (Corder 1967). Given the observation that certain types of errors tend to appear at certain stages of development independent of the L1 and TLs, accuracy was postulated as a way of determining advancedness or at least as an indicator of what language forms should be taught at what stage of development. However, what should be avoided is the collapsing of the constructs of accuracy and development or advancedness or even proficiency as Pallotti (2009) explains: "we may find texts that are very accurate but scarcely developed…and texts containing many errors but exhibiting several traits of evolution".

Thus, the straightforwardness of accuracy has not always been evident in SLA. Progress in other areas of study such as intelligibility/comprehensibility and L2 development have better clarified what accuracy does not point to in learner language. Yet, accuracy remains an integral part of the CAF triad and indispensible in the description of performance and proficiency in the L2. In the empirical study that follows, accuracy is measured through error ratios at the unit and clausal level (errors per AS-unit and errors per clause).

Varying definitions of fluency assume some normative standard of the dimension that is either explicitly or implicitly identified in the NS. Lennon (1990a) speaks of fluency as "native-like rapidity", making direct reference to the native norm. Skehan (2009) offers: "the capacity to produce speech at normal rate and without interruption", and Ellis and Barkhuizen provide: "the production of language in real time without undue pausing or hesitation' (Ellis and Barkhuizen 2005: 139). Again, as when defining accuracy, the terms "normal" and "undue" imply the existence of a standard and appropriate fluency. Alternatively, fluency can be relativised as "the extent to which the language produced in performing a task manifests pausing, hesitation, or reformulation" (Ellis 2003: 342).

At this point we can make the distinction recommended by Segalowitz (2010) between different notions of fluency. These are cognitive fluency, utterance fluency and perceived fluency. Cognitive fluency has to do with the speaker's ability and relative

efficiency in planning and producing speech (i.e. the fluency of speech processing as it occurs in the mind). Utterance fluency is the type or level of fluency that can be objectively measured in a speech sample. Utterance fluency has been further decomposed into breakdown and repair fluency and speed fluency (Travakoli & Skehan 2005). Breakdown fluency, as the term suggests, measures the amount of "breakdowns" within the flow of speech, that is, the number and/or length of pauses (filled and unfilled). Similarly, repair fluency measures the frequency with which speakers produce false starts, self-corrections and repetitions within the speech sample. Speed fluency, measures the speed at which the speech is produced and is calculated with temporal measures (e.g. speech rate in syllables per minute).

De Jong, Steinel, Florijn, Schoonen and Hulstijn (2013) point out that quite a few fluency measures confound these categories of fluency in that one measure may capture elements of two or all three types. They provide the example of SR. As SR is generally calculated by taking the number of units (i.e. syllables or words) and dividing the count by total time –including pause time – the resulting measure of SR (a measure of speed fluency) includes material that is representative of both breakdown fluency, as (un)filled pause time is counted, as well as repair fluency where restarts, repairs and repetitions are included. Therefore, careful attention to what is actually being measured is essential in justifying the use of certain measures of fluency over, or in combinations with others and in the interpretation of the results.

Finally, fluency can be conceived of as a perceivable quality of speech. Perceived fluency is the impression the listener has when listening to a given speech sample. In an experimental setting the listener is also a rater or judge, who assigns a score to the utterance/subject usually making use of a likert scale. Thus, perceived fluency is a subjective and global rating of the speaker's performance. Fluency studies have frequently attempted to link perceived fluency with objective measures of utterance fluency. Like Lennon (1990a), the authors of such studies see perceived fluency as the impression that (in)efficient speech planning and processing at the psychological level makes on the listener. Of course, the medium through which the listener can make judgments about a certain speaker's ability to efficiently (fluently) process speech is the utterance. It follows then that certain aspects of fluency (e.g. speed fluency measures, or dysfluency phenomena) may contribute differentially to perceived fluency. Studies by

Lennon (1990) and Riggenbach (1991) recruited expert (teacher) judges to rate NNSs speech samples, and tested for correlations between global scores provided by the judges and objective measures of fluency. Derwing, Rossiter, Munro, and Thomson (2004) performed a similar study using non-expert judges and both Rossiter (2009) and Valls-Ferrer (2011) compared several groups of raters with differing (expert/non-expert) profiles.

Although the measures used vary considerably across studies of this type and the results from one study to another may be mixed, in general terms utterance fluency is correlated with perceived fluency ratings in measures of both speed fluency and breakdown/repair fluency. To provide some examples of findings, Lennon (1990a) determined that speech-pause relationships in performance and frequency of occurrences of certain dysfluency phenomena (i.e. filled pauses and repetitions) proved significant predictors of global fluency scores, while Derwing et al. (2004) found pruned speech rate to be the best predictor of subjective fluency scores. They calculated pruned speech rate by dividing syllables by total time (in seconds) but after excluding all repair phenomena (false starts, reformulations, and repetitions) from the syllable count, thus distinguishing Pruned speech rate from regular speech rate in syllables per second. Pruned speech rate was found to be a particularly robust predictor of perceived fluency judgments, accounting for 65% or more of the variance. In determining which aspects of utterance fluency more greatly influence listener perceptions of fluency, research in the field hoped to both better understand what constitutes L2 fluency, allowing for objective assessment of spoken learner language and improve pedagogy where fluency was concerned. In the current work fluency is tapped through both speech and pruned speech rate, measured in syllables per minute.

CAF has been, and continues to be a very useful way of gauging language performance, and some widely accepted conclusions about language processing and how L2 speakers use their second language(s) to perform language related tasks have been developed thanks to evidence gathered through CAF measures. These conclusions, however, are continually subject to revision and refinement as the field of SLA continues to improve upon methodologies. While CAF constructs are important tools to better understand language proficiency and development over time, CAF can only refer to the corresponding properties of language performance as they are observed as a product at a

given point in time. Proficiency and development should be taken as separate dimensions of language that CAF measures can be related to empirically. However, as Pallotti (2009) stresses, CAF should not be considered to be a sub-dimension of development.

2.3.2. Trade-off versus Cognitive hypotheses

Two competing hypotheses are present in the literature involving tasks and often utilize CAF dimensions in their analysis. These are Skehan's Trade-off hypothesis and Robinson's Cognition hypothesis. In essence, both approaches seek to better understand and predict how task demands affect the allocation of memory and attention (Robinson 2005a), and seek to explore how performance is affected by task characteristics and conditions (Skehan 2003). In both approaches, this is typically done by looking into how the manipulation of task elements and conditions affects change in the different dimensions of CAF as seen in the production of the language learner.

Both the Trade-off and Cognition hypotheses deal with the allocation of attentional resources; indeed there are many similarities between the two approaches. Robinson (2005a) sums up the principal phenomenon of interest in such studies noting that "allocation of attentional capacity to task demands is a control process, and as task components and demands proliferate, so does the difficulty of managing allocation policy, with consequent lapses in perception and production" (p.647).

Where these two accounts differ however, is in how they view the limitations of attentional recourses, and what facets of CAF are correlated with one another as task complexity (i.e. demandingness) increases. Drawing from knowledge of cognitive psychology and long-held beliefs in that field about the limited capacity of attention and working memory, Skehan considers attentional recourses to be in short supply as a speaker is performing in the L2. Manifestations of limited attention capacity will show themselves in 'trade-offs' reflected in the operationalised dimensions of CAF. According to the Trade-off hypothesis, one can expect to find tension between attention paid to form (i.e. accuracy and complexity) on the one hand, and fluency on the other (Skehan 2009). Thus, a typical outcome would be to find that greater fluency may be accompanied by greater accuracy or greater complexity, but not both greater accuracy and greater complexity (Skehan 2003).

Thus in Skehan's view, if successful performance in the L2 is to be achieved, each component of the CAF triad requires attention and the involvement of working memory. Where there are not enough resources to go around, as may be the case in the L2 speaker, committing attentional resources to one (e.g. accuracy) may have a negative impact on others (e.g. complexity/fluency) resulting in lower performance.

While the two approaches are very similar, the Cognition hypothesis differs from the Trade-off hypothesis in at least two significant ways. First, the Cognition hypothesis does not hold that the scarcity of resources is the main influence in the allocation of what Robinson would still consider to be *limited* attentional resources, but rather that learners have access to multiple 'attentional pools' that are not necessarily in competition with one another (Robinson 2005a). Thus, the main influence at work in performance variation is how task elements and conditions direct or disperse attention. The terms resource-dispersing and resource-directing task dimensions refer to this idea of channelling attentional resources, or on the contrary, dispersing them. While resource-directing elements increase performance levels (e.g. tasks which differ along the Here-and-Now versus There-and-Then dimension), attention-dispersing elements lower them (Robinson 2003).

Also within this view, more demanding tasks will generally lead to more complex and accurate performances, pushing learners to produce more accurate and complex speech so as to meet the greater functional and communicative demands that the more difficult task requires of them (Robinson 2003). Thus, the Cognition hypothesis predicts that there is a F versus CA opposition where C and A correlate with each other. On the other hand the Trade-off hypothesis would expect F to be correlated with either A or C, but not both and in the best case scenario (Skehan 2003). While no clear evidence supports one view over the other to any decisive degree. Fine tuning has been done to both hypotheses in recent years, and surely further research into the nuances of task design will demand further revisions in both cases.

2.4. Summary of Chapter 2

In this chapter we shifted our attention away from SA specifically, and toward a more general look at theoretical approaches to SLA, models of speech production that have

been applied to the bilingual/L2 user/L2 learner conditions, and covered the basic notions of CAF, and how these constructs have been used in SLA studies.

We began this chapter by providing a brief overview of some of the most prominent approaches to SLA since its emergence in the 1960s and 70s. Researchers in the field have witnessed a number of significant paradigm shifts in how language, and therefore second language, is viewed as an object of study. The nature of language has been seen as a mere sign system, first by behaviourists who considered language to be an essentially static system, but also by generative linguists, who have resorted to the abstraction of language in order to tap its underlying forms, yet ignoring how it is used in everyday communication. We have also seen how later on purely cognitive approaches to SLA and input/interaction approaches have coalesced in many senses into a widely accepted interactionist view of language acquisition. In this view, language is seen as inseparable from its use and social significance, marking a trend that persists to the present day. As we have noted, the study of learning context, namely FI, IM and SA fits well with the interactionist perspective, and we have discussed the roles of educational versus natural settings and what each may contribute to language learning.

Some principal speech production models were presented in this chapter as well. In section 2.2 we discussed Levelt's "blueprint for the speaker" model (1989) and how others, most notably De Bot (1992), have applied this model to bilingual speech production. These modular models have been contrasted with a spreading activation model. While the two ways of modelling speech production are quite similar in many ways, we have discussed several discrepancies between the two.

Finally, as the dimensions of production represented in the CAF triad as of particular importance in the present study, we have dedicated the final section of Chapter 2 to the discussion of these constructs, their definitions and how they have been operationalised in the relevant literature. We have also discussed how CAF dimensions may enter into competition with one another when the speaker is presented with a particular linguistic task. Likewise, we have seen how some authors have attempted to account for how this competition is played out when task elements vary. Skehan and Robinson have observed patterns of differential competition between CAF dimensions across tasks and have attempted to provide an explanation through the Trade-off hypothesis (Skehan

1998) and the Cognition hypothesis (Robinson 2005a) respectively. With this discussion we conclude Chapter 2 and turn our attention to some novel concepts in SLA that have begun to take root in recent years. These innovative concepts are further discussed in the following chapter.

CHAPTER 3

RECENT SHIFTS IN THE CONCEPT OF SECOND LANGUAGE PROFICIENCY AND DEVELOPMENT

In the introduction to their volume on L2 performance and proficiency and CAF, Housen, Kuiken and Vedder (2012) very succinctly identify the two main objectives of not only SLA, when they pose the questions: "What makes a second language (L2) learner a proficient language user? And how can L2 proficiency be most adequately (i.e. validly, reliably and feasibly) measured?" Acceptable answers that satisfy academic inquiry related to these questions have evolved greatly over the past decades. Here, we would like to highlight three topics that represent considerable advancements in the study of SLA since the turn of the present century, and in particular, how each applies to the study of oral proficiency and development. First we mention the contributions of the Netherlands based project "What is Speaking Proficiency?" (WiSP), (De Jong, Steinel, Florijn, Schoonen & Hulstijn 2012a; De Jong, Groenhout, Schoonen & Hulstijn 2013). The series of studies produced through WiSP have made impactful advances in our understanding of what speaking proficiency really is as well as how to relate proficiency to evaluation tools such as CEFR. Perhaps most notably, the WiSP methodology avoids the circular definitions that have been used in studies of proficiency until now far. Here we examine several studies produced by the WiSP research project summarizing their results and potential applications.

Then, we turn to a trend in the study of proficiency, also taken into account in WiSP publications, but developed by other authors as well, which is a focus on communicative adequacy (also called 'functional adequacy') as an important element of L2 proficiency (Pallotti 2009; Hulstijn, Schoonen, De Jong, Steinel & Florijn 2012). We discuss some recently evolving views on communicative adequacy and some proposals as to how it can be separated out of more traditional CAF constructs as a separate dimension of L2 performance, yet considered of equal importance in assessing learner proficiency (Pallotti 2009). Finally, Chapter 3 concludes with a discussion of the influential Dynamic Systems Theory (DST) (Larsen-Freeman 2002; De Bot, Lowie & Verspoor 2007), its application to the study of SLA and how this approach fits with L2 research involving CAF dimensions. Still in its initial stages in terms of its application

to real learner data, DST promises to shed new light on L2 development using innovative methodologies that recognise the complex and dynamic nature of language and language development.

3.1. The characterization of speaking proficiency

An ambitious research project dealing with the nature of speaking proficiency has emerged out of the University of Amsterdam and partner universities, headed by Dutch researchers Jan Hulstijn and Rob Schoonen. The project has yielded a series of publications on varying aspects of the topic (De Jong, Steinel & Florijn Hulstijn J. H. 2012a; Hulstijn, Schoonen, De Jong, Steinel & Florijn 2012; De Jong, Groenhout, Schoonen, & Hulstijn 2013). "What is speaking proficiency?" or "WiSP" for short, set out in 2004 to investigate the componential structure of L2 oral proficiency, gauging the weightiness of the different components that seem to determine varying levels of L2 speaking proficiency in adult learners of Dutch. Here we would like to highlight two areas in which these studies have contributed to the field, first by providing an innovative methodology in determining components of proficiency anchoring their approach in language processing theory. And second, how these components of proficiency can be practically applied to a language assessment tool such as CEFR.

The first study in the series aimed to better define facets of speaking proficiency and assigned each relative weight according to the outcomes. While this practice is nothing new in SLA, the breadth and depth of this study shed new light on the componential nature of proficiency in a second language in ways that previous studies had not. The authors offer an alternative to often limited methodologies that have been employed in the "decomposition" of speaking proficiency up until the present day. Prior studies have often relied on purely subjective measures of language proficiency; for example judges will rate overall proficiency, then rate performance in specific subskills that appear to contribute to proficiency. Even when subjective measures have been complemented with objective measures (for example: measures of syntactic complexity, or repair fluency of a given speech sample), the authors argue that, as both types of analysis (subjective and objective) use the same speech samples, there is the potential problem of circularity in that, the truth of the observations about proficiency cannot be established independently of the conclusion. To provide an example, one possible circular conclusion might be that reduced instances of repair phenomena indicate better

performance because those speech samples with fewer instances of repair phenomena are considered better performances (and thus indicate greater proficiency). While De Jong and colleagues acknowledge that this type of measurement can lead to perfectly valid conclusions in terms of language assessment, they offer what they consider to be a superior methodology where the real breakdown of proficiency is concerned.

In order to better determine how individual differences in those subskills hypothesized to be components of proficiency are related to individual differences in successfully conveying information through speech, subskills and functional adequacy (relative success of conveying information) where measured separately. In this way, the problem of circularity is avoided. Specifically, functional adequacy was measured by non-expert raters, judging the speech performances of 181 learner subjects. The relative success of conveying information thus became the dependent variable in the study. Conversely, the independent variables that were tested as predictor variables were taken from subjects' performance in a number of linguistic skill tasks assessing "knowledge skills (testing their declarative knowledge about the language)" and "processing skills (testing the rapidity and accuracy with which they process linguistic information)". Knowledge and processing skills are roughly related to Anderson's (2009) concept of declarative and procedural knowledge and are rooted in the premises of Levelt's speech production model (1989).

Structural equation modelling (SEM) was used to determine the relative "weight" of each predictor variable. Of the seven skills tested, knowledge of vocabulary and correct intonation proved to be the best predictors of functional adequacy. In fact, these two variables together explained 75.3% of the variance in subjects' functional adequacy. Of course, vocabulary knowledge and intonation alone were not the only factors contributing to proficiency; grammar knowledge, lexical retrieval and sentence building speeds, correct pronunciation and stress patterns were also strongly associated with proficiency. Thus proficiency is a balance between declarative knowledge, processing knowledge and pronunciation related skills.

The novel approach to the study of proficiency adopted in De Jong et al. (2012a) reinforces the widely held belief of the componential nature of L2 speaking proficiency. That is, the idea that proficiency consists of knowledge about the language as well as

procedural knowledge, or as the authors put it "language-processing components". What is more, the componential make up of proficiency held true for the learners at both the top and bottom 40% of proficiency ratings, a finding that provides evidence against the relative contribution model proposed by Higgs and Clifford (1982, as cited in De Jong & Van Ginkel 1992) who suggest that as learners' proficiency progresses over time, the relative weights of each facet as they contribute to overall proficiency may change as well. Following the relative contribution model, vocabulary and grammar, for instance were hypothesized be the most important factors across all levels of proficiency, but would diminish in importance at higher levels in favour of skills related to pronunciation, fluency and sociolinguistic competence. Thus at early stages of learning, speakers would differ more in knowledge related factors (vocabulary and grammar), while at advanced stages, learners would distinguish themselves from one another in their ability in skills such as pronunciation and fluency. These conclusions, however, were based on subjective ratings of the importance of each skill by teachers who were familiar with specific internal criteria for rating language proficiency, namely the guidelines provided by the Foreign Service Institute School of Language Studies (1968, as cited by De Jong & Van Ginkel 1992). The study by De Jong and colleagues (2012a) does not support this model, but rather provides concrete evidence that suggests that specific facets of proficiency account for perceived functional adequacy in L2 production in similar proportions independent of the learner's level.

The second study in the WiSP series resumed here was performed by Hulstijn, Schoonen, De Jong, Steinel and Florijn (2012), and examined the association between speaking proficiency and linguistic competencies based on the Common European Framework of Reference for languages (CEFR) classifications. The tasks were the same 8 tasks used in De Jong et al. (2012a), varying in formality and task complexity in order to capture learners' performance in an array of communicative situations. Also, just as in De Jong et al. (2012a), the subjective measure of proficiency was rated on a scale of communicative adequacy, the difference being that this time expert raters were recruited to perform the listener judgement portion of the experiment. Learners who were rated as falling into either the B1 or B2 CEFR levels were selected for comparison in order to determine whether or not knowledge (productive vocabulary and grammar knowledge), processing skills (speed of lexical retrieval, articulation and sentence building) and/or pronunciation skills as measured through separate, non-communicative tests would

prove robust predictors of assigned CEFR level (discriminating between B1 and B2 levels).

Discriminate analysis indicated that all the linguistic competencies, save one (speed of articulation), discriminated L2 levels of oral production. Further analysis showed that the distance between learners found at B1 and B2 levels was smaller in terms of knowledge of high-frequency words than in knowledge of medium to low-frequency words. Thus, extrapolation from the resulting scores estimated productive vocabulary sizes of 4000 words for the B1 level and 7000 words for the B2 level. Learners at the B2 level were seen to outperform those at the B1 level on all aspects of grammatical knowledge tested. Furthermore, the difference between groups seems to be a matter of degree of grammatical knowledge rather than based on categories or domains.

The contribution of this paper is a first empirical look into the linguistic underpinnings speaking proficiency as gauged through CEFR, a daunting task, yet a much needed validation of the use of CEFR in this area of linguistic proficiency.

3.2. Communicative adequacy (CA)

Measures of CAF have frequently been applied together, as they are thought to "jointly encompass overall performance" (De Jong, Steinel, Florijn, Schoonen & Hulstijn. 2012b:122). In language testing research and especially in task related research, we can assess performance in at least two ways, first by determining whether or not a L2 user has successfully accomplished a given language related task in that he/she has accomplished the communicative or pragmatic goal set before him/her. And second, we can indirectly analyse the performance by measuring linguistic complexity, accuracy and fluency (Robinson 2001; De Jong et al. 2012b).

In Hulstijn, Schoonen, De Jong, Steinel and Florijn (2012) Participants' speaking proficiency was measured utilizing CA as a basis for the evaluation by trained raters of L2 speech production. In order to capture listeners' perceptions of adequacy, raters judged speech samples along a numeric scale in accordance with CA criteria in terms of "(a) the amount and detail of information conveyed, relevant to the topic, setting (formal/informal) and discourse type (descriptive/argumentative) and (b) the intelligibility of the response" Hulstijn et al. 2012:207). It should be noted that in the

rating task no reference was made to linguistic quality so as to better isolate CA as the means of evaluation.

Both Hulstijn et al. (2012) and Kuiken, Vedder and Gilabert (2010) are studies that are interested in relating adequacy to CEFR oral proficiency scales (COE 2001). Furthermore, Hulstijn et al. relate linguistic knowledge (productive vocabulary knowledge, productive knowledge of grammar), access speed measures (speed of lexical retrieval, speed of articulation, speed of sentence building) and pronunciation skills to CEFR levels, showing that all except speed of articulation were able to discriminate between groups at B1 and B2 levels.

3.3. Dynamic Systems Theory (DST)

Promoted as a solution to long-standing and competing stances in SLA, namely cognitive/individualist approaches and social/interactionist approaches, Dynamic Systems theorist claim to have sufficiently widened the scope of SLA theory to include both cognitive and interactionist views, as well as taking into account how the L2 is actually used in real life situations (Larsen-Freeman 2002). DST has been applied in the field of developmental psychology (van Greet 1994), and carried over into SLA chiefly by Diane Larson-Freeman (1991; 2002; 2009) and the group of colleagues made up of Kees De Bot, Marjolijn Verspoor and Wander Lowie, who have frequently appeared as co-authors of DST centred work (e.g. De Bot, Lowie & Verspoor 2005; De Bot Lowie & Verspoor 2007; Verspoor, Lowie & De Bot 2009).

DST framework offers an alternative view of language development whether it be L1 or L2, assuming that language is an ever evolving system. The study of SLA, therefore, is very much the study of a ‘moving target’. As language is not a stable system, it logically follows that traditional analytical techniques have not been able to make exact predictions with respect to a number of issues in SLA (e.g. studies of input have not been able to precisely predict what and how the learner will acquire TL forms, Verspoor, Lowie & De Bot 2009).

Larsen-Freeman (2002) comments on a heated debate between opposing schools of thought in the SLA community that took place in the 1990s in which proponents of a purely cognitive (often highly abstract) approaches to SLA clashed with those who

promoted a context-focused approach; a debate which continues to have repercussions to this day. Indeed, the controversy has contributed to the advent of DST in various sectors of linguistics, and has lead to the rethinking of some of the most basic concepts in SLA held until now. As we have already mentioned, in the 1990s, Firth and Wagner (1997) initiated a debate in the SLA community revolving around what they perceived as an imbalance between cognitive approaches and context sensitive approaches to SLA. The imbalance, in their view, was toward the cognitive view, and the research inspired in Chomskyan linguistics and the focus on the individual and the mind, ignoring the surrounding social context, forms of interaction and what language is used for in the first place. They called for a correction of this imbalance and for nothing less than the reconceptualisation of SLA as "an enterprise that endeavours to attend to, explicate, and explore, in more equal measures and, where possible, in integrated ways, both the *social* and *cognitive* dimensions of [second/foreign language] use and acquisition" (Firth & Wagner 1997:286).

Along with this global objective of reconceptualisation, Firth and Wagner also more specifically sought to challenge a number of key concepts in SLA that had become assumed as truth yet were unacceptable for these authors for a variety of reasons. These include such fundamental concepts as "learner," "(non-)native speaker," and "interlanguage". For Firth and Wagner, the terms "native" versus "non-native", and "learner" emphasized the insufficiencies of the L2 speaker (as compared to the NS) while ignoring his/her role as a user of the language within an interactional context. Similarly "(non-)native speaker" served as blanket terms that assume homogeneity in each categorical group (NS vs. NNS). NSs are considered to be wholly unproblematic speakers of the language, which has been acquired from birth, and thus the ideal yardstick by which NNS performances should be measured. NNSs, on the other hand, are wrongly considered (they argue) to be substandard users of the language by nature. These characterisations of NS versus NNS ignore the bilingual condition, multilingualism and cases of L1 attrition, not to mention the various sources of identity that the speaker draws upon, other than that of being a (non-)native speaker. NSs are seen as the norm, while NNSs are seen as the anomalous variety of speaker, a view that world demographics do not uphold. With some 7000 living languages spoken in the world today across only about 200 countries (Lewis, Simons, & Fennig 2014), it is

evident that a large portion of the world's population speaks more than one language and a large percentage of plurilinguals have one or more of their language(s) as a L2.

In a similar way that NS and NNS have been separated categorically, the concepts of "learner" and "interlanguage" in SLA have often wrongly lumped everyday users of the language who happen to be NNSs of that language into the same group as those who are actively learning the language (e.g. within the language classroom). First, like in a class curriculum, NNSs as a collective are seen on the road towards NS competence and any lack of progress along this continuum is deemed "fossilization", a failure to progress toward the idealised NS status which is always considered to be the destination. Second, the transitional nature of interlanguage is taken as systematic and predictable. This view of the supposed 'learner' and 'interlanguage' ignores the fact that some speakers may have a L2 that completely satisfies their communicative needs even though it varies from NS norms in some sense. Such L2 users may not be on the same path toward nativelike abilities that say undergraduate modern language majors might be, however, in the research, they are treated as on a similar trajectory. Indirectly, this view also implies that the "target" (i.e. NS competence) is "constant, fully developed, and complete" (Firth & Wagner 1997:292), which is, of course, probably not the case.

Proponents of DST respond to the concerns of Firth and Wagner, and those who would corroborate their arguments by offering an alternative, integrative view of SLA. The main tenants of DST as they are applied to language in general and SLA in particular, are that language, be it a L1 or a L2 is by its very nature a complex system. DST theorists define complex systems as sets of interacting variables (De Bot, Verspoor & Lowie 2005). That is, complex systems are made up of a multitude of variables which function in complete interconnectedness such that a change in one variable will impact all other variables that are members of the system (De Bot et al. 2005; De Bot et al. 2007). Second, DST treats language, both at the societal and individual level, as a dynamic system. That is, language changes over time; the set of interacting variables that make up the system are in constant flux. Whether we refer to language of a given speech community or the language of the individual as it is represented and processed in his/her mind, the dynamics of language are always at work.

Due to the complex and dynamic nature of language, and especially language development, linguistic outcomes over time are often unpredictable. In the words of Larsen-Freeman: "In a dynamic system, not only do the components change with time, but the ways in which the components interact with each other and with the environment also change with time" (Larsen-Freeman 2009:583). This is thought to explain how some subfields of SLA, despite considerable experimentation, continue to produce very mixed results.

De Bot, Verspoor and Lowie (2005) go about summing up DST in the following characteristics. Dynamic systems are always part of another, larger system. In the hard sciences they understand this to be any system from the molecular level to the universe. As they develop over time, dynamic sub-systems appear to settle in specific states, which are preferred but unpredictable, termed 'attractor states'. States which are not preferred nor settled in are called 'repeller states'. The development of a dynamic system is a process based on iterations – the repetition of a sequence that grows ever closer to the desired result – and appears to be highly dependent on the beginning state such that even minor differences at the beginning can have dramatic consequences at the end of the observation period. This is presumably due to the fact that in dynamic systems, changes in any one variable have an impact on all other variables within the system. In this sense they are considered to be fully interconnected. In natural systems, development is dependent on the resources at hand. Language development, like all natural systems, will tend toward entropy when no additional energy is added to the system. Furthermore, Systems do not develop in isolation, but through interaction with the environment and through self-reorganisation. Because systems are constantly changing, they will show variation, which makes them sensitive to specific input at a given point in time and some other input at another point in time.

Though still very much in preliminary stages of application to the study of SLA, DST has been shown to be useful in a variety of studies of dynamic subsystems in the L2. DST seems to lend itself well to the study of CAF as can be seen in a handful of rather recent studies. Spolman and Verspoor (2010) use DST techniques to examine accuracy and complexity development in L2 writing samples provided by a Dutch learner of Finnish. Polat and Kim (2014) look into the relationship between accuracy, complexity and lexical diversity over time in oral production. Spolman and Verpoor (2010), Polat

and Kim (2014), and Verspoor, Lowie and De Bot (2008) are case studies of a single language learner and Larsen-Freeman (2006) examines the development of five subjects. Indeed, studies that apply DTS to real learner data tend to be limited to the case study modality. While DTS promises to be an all inclusive framework for the study of L2 development, much work is still to be done in terms of applying the framework to the breadth of data sources and phenomena that proponents claim that DST encompasses.

3.4. Summary of Chapter 3.

To conclude Part I of this dissertation, Chapter three has introduced three areas of SLA research that have proposed innovative ways at looking at proficiency and development in the L2. WiSP collaborators have contributed to our understanding of what proficiency is, working toward a theory of language proficiency founded in empirical evidence. They have developed a methodology to investigate the componential nature of proficiency which avoids the problem of circularity when it comes to defining proficiency constructs. Their contributions provide a theoretical basis for language assessment tools such as CEFR, which was developed in 2001 without reference to any particular theory of proficiency or SLA.

Communicative adequacy has received attention in recent years both as an aspect of the WiSP project (De Jong et al. 2012b) and in publications such as those by Pallotti (2009) and Kuiken et al. (2010). Few CAF studies include a measure of adequacy or any kind of communicative success in their design, yet there is a great emphasis on communication, and getting one's message across effectively in language teaching-learning methods and evaluation. Efforts to quantify adequacy seek to bridge the gap between what CAF measures can tell us about learner language objectively, and how real-live listeners perceive the relative success of their communication. In doing so, we may be able to better understand and assess proficiency in all its facets.

Dynamic Systems Theory offers an integrative view of SLA in which the main tenants are that language (L1 or L2) is by nature a complex system made up of a multitude of variables which are fully interconnected such that a change in one variable will impact all other variables that are members of the system. DST seeks to account for the complexity of language at all its levels, societal to individual, assuming that it is a

dynamic system. That is, the set of interacting variables that make up the system are constantly changing. Whether referring to language in society or the language of the individual, the dynamics of language are always at work.

These innovations do not represent, by any means, the only advances in the field of SLA in recent years, but all of these emergent concepts have been included here due to the fact that the author considers them to be important examples of progress in our treatment of proficiency and finds them of particular interest for future research both in the field of SLA as a whole and in the study of SA as a learning context.

PART II: THE CURRENT STUDY

CHAPTER 4

INTRODUCTION TO THE CURRENT STUDY

The chapters that make up Part I of this dissertation have provided a backdrop for the current study of L2 English development through SA programmes varying in length. In Chapter 1 we presented a brief history of SA research, highlighting several significant contributions to the field that marked the course of research up until the present day (Meara 1994; Freed 1995a; Coleman 1998; Collentine & Freed 2004; DuFon & Churchill 2006; Collentine 2009). These and other works point out that, although SA is a very interesting context to examine for L2 development in terms of the quantity and quality of input a stay abroad potentially provides, there are still some very open questions regarding its efficacy, and the literature does not consistently demonstrate across-the-board benefits for language learning. To illustrate, in section 1.2 we surveyed the relevant literature, reporting by skill types on the results of previous studies dealing with SA contexts. We saw that results have been very mixed, but that certain skills, or aspects of language performance seem to benefit more and sooner over others. Fluency, for example, is likely the facet of performance that most consistently benefits from SA (Lennon 1990; Towell et al. 1996; Segalowitz & Freed 2004; Trenchs-Perera 2009; Valls-Ferrer 2011; Mora & Valls-Ferrer 2012; Valls-Ferrer & Mora 2014), while L2 phonology, on the other end of the spectrum, often remains unaltered upon return from SA (Diaz-Campos 2004, Avello 2013, Avello & Lara 2014). We also noted that while SA is often seen as the optimal language learning environment by institutions and the general public, research does not always support this idea. AH and IM contexts may even yield superior results to those of SA in some linguistic areas (Collentine & Freed 2004).

We have also looked into the worldwide practice of student mobility and how SA fits into this global phenomenon as a specific modality of mobility. Clearly, there is no one way of doing SA and differing curricula, itineraries and programme specifics such as LoS vary considerably from one sending region to the next and from one programme to the next. Participants also vary greatly in how they approach their period abroad and in what resources they bring to the experience (both in terms of linguistic skills and

affective variables such as outgoingness and motivation) that can significantly impact the end result upon return.

In Chapter 2 we shifted our focus from previous research surrounding SA specifically to the broader climate of SLA in general, L2 speech production modelling and the speech dimensions of CAF. Section 2.1 provided us with a brief overview of the highly varied field of SLA, different theoretical approaches as they have evolved throughout the years and how they have influenced the course of SLA inquiry. We have dedicated time to two notions within SLA that are of particular interest when studying the impact of SA, the role of interaction as it has been hypothesized as a catalysis of language acquisition and the role of context, especially in terms of quality and quantity of input/output through the unique "hybrid" nature of SA (Collentine & Freed 2004; Miller & Ginsburg 1995). With this term we refer to the combination of educational and natural settings characteristic of most SA programmes that lead to both formal and informal language learning opportunities, all within the TL environment. Also in Chapter 2, the speech dimensions represented in the CAF triad were defined and discussed within the context of the relevant literature.

Finally, we have reviewed three areas of innovation that have arisen in recent years and either serve to better interpret the results that the reader will find in the coming chapters, or provide alternative views to be explored in future research. First, we have acknowledged the steps the WiSP project has taken toward unravelling what constitutes speaking proficiency apart from the potentially circular definitions that have been used up until now. Next, we addressed the proposed addition of communicative adequacy (or functional adequacy as some prefer) to the traditional grouping of speech dimensions that is CAF. Adequacy, it is argued, should be considered a separate dimension of proficiency, so as not to confound this aspect with other, more traditional elements of CAF (Pallotti 2009). Finally, we discussed the proposals brought forth by the principal proponents of DST as an all encompassing theory of SLA, its origins and its potential applications to learner data.

4.1. The SALA project

The present study utilizes oral data draw from two groups of learners and one group of NSs whose contributions make up part of the greater Study Abroad and Language

Acquisition (SALA) corpus. For a full description of the SALA project, its goals and design, see Pérez-Vidal et al. (2007), Pérez-Vidal (2014); see also Beattie (2014) for a detailed description of the mobility programme that inspired the project. SALA is a large-scale, research project recognised and funded by the state and based out of a public university located in Barcelona, Spain. SALA began in 2004 as a joint research project between the aforementioned university in Barcelona and a second public institution in the Balearic Islands. SALA was developed with the aim of tracking EFL acquisition during SA, contrasting the outcomes of this context with acquisition that takes place in the FI condition AH. In the beginning, SALA sought to take advantage of a curriculum requirement in the Translation and Interpretation department at the Barcelona based university that required second year students to go on a one-quarter (three-month) academic exchange to an English-speaking country. SALA successfully compiled data from such exchanges over the course of three years, taping a wide range of linguistics skills through written, oral, listening and perception tasks and collecting baseline data from native speakers of English. The original SALA design elicited data from a battery of tests that were performed at 4 different points over a two year period: about 6 months prior to SA and at the beginning of a two term period of formal instruction in English, immediately prior to SA, upon return home from SA and at approximately 15 months post-SA, the students having received no further English instruction (See diagram taken from Pérez-Vidal & Juan-Garau 2011, Appendix C.3). Now in its 10th year of operation, SALA has expanded in scope with new research questions requiring the incorporation of new participant profiles. One such profile is that of students who spend two quarters away from their home campus (Approximately 6 months).

In the present study we analyse data from both a cohort that participated in the original SALA data collection sessions, going on SA for three months and from students who spent two quarters abroad. In order to provide baseline comparisons we also utilized data from a group of NSs of English who were on an exchange programme in the Balearic Islands at the time. It should be noted that while the original three cohorts that provided data were tested at up to four testing times, the data we use here is restricted to only two testing times; pre-SA and post-SA (corresponding to T2 and T3 in the diagram in Appendix C.3 for the three-month group). The decision was made to only include these testing times so as to be able to compare the SA experiences of three-month SA

and 6-month SA participants. As the 6-month participants had not been tested prior to the testing time immediately before departure, there was no way of comparing three-month participant data that had been elicited 6 months prior to the pre-departure testing time as in the case of the three-month participant group. Thus, although both the original SALA design and the present study are longitudinal in nature, here we limit our pre-test/post-test testing times to those that are immediately prior to SA and upon arrival from SA.

The same tasks and procedures were performed by all the participants at the consecutive testing times, with the exception of the NS group who performed the tasks only once. Among the skills tested through the SALA battery of tests are speaking, lexical and grammatical knowledge, listening comprehension, phonological perception and writing ability. Students were also asked to complete a questionnaire about their language background and other key demographic information. Upon return from SA, they completed an additional questionnaire on the SA stay conditions, their level of contact with the TL and similar information specific to their SA experiences (See Appendix C.2 for a complete list of SALA tests). In the study that follows we utilize a role-play task that was included among the three oral tasks that participants performed. Two written tests were included in the analysis as well: grammar (Grammar I) and cloze tests. These tests helped us in gauging initial level in students' control of English grammar and vocabulary prior to SA. Chapter 5 provides further details regarding the SALA design as it pertains to the current study, participant profiles and the materials and procedures used during data collection.

A number of publications have been produced drawing from the SALA corpus. Like the present study, several of these publications analyse oral data, and have reported gains in oral fluency post-SA, pointing to an advantage of SA over FI, at least in this area (Trenchs-Parera 2009; Pérez-Vidal Juan-Garau 2011; Valls-Ferrer 2011; Mora & Valls-Ferrer 2012), while complexity and accuracy show mixed results. Pérez-Vidal and Juan Garau (2011) report significant gains in accuracy and a non-significant tendency toward greater complexity, while Mora and Valls-Ferrer (2012) found robust gains in fluency after SA; only moderate gains were seen for accuracy and no gains were found for complexity. Valls-Ferrer and Mora (2014) confirmed the significant gains in fluency found previously (Mora & Valls-Ferrer 2012), and found that lower initial levels at

onset and higher degrees of contact when abroad could, at least in part, explain differential effects of the sequential FI and SA periods. Juan-Garau (2014) reports no significant gains in accuracy during the FI period, but did find significant gains in two measures of accuracy after SA. Readers interested in the unique design of SALA should consult the very recent volume *Language Acquisition in Study Abroad and Formal Instruction Contexts* (Pérez-Vidal 2014). The book provides a collection of studies, many of which have already been mentioned, encompassing the whole breadth of SALA research to date and is a self-contained window into SALA findings over the past decade.

Now having contextualised the current study within the academic surroundings in which SALA research has taken place, we now turn to the wider social context of an ever more diverse Catalonia. The region provides a rich linguistic environment that merits due discussion as it is home to the NNSs who contributed to the current study.

4.2. The local context

Catalonia is one of three autonomous communities in Spain, together with the Basque Country and Galicia that have well established autochthonous languages spoken throughout their territories. The official language of the wider Spanish state, Castilian Spanish, is the most widely used language in Catalonia. According to a report published by the Catalan government in 2012, *Informe de Política Lingüística 2012* (Generalitat 2012), 100% of the adult population understands Spanish. 99.8% can el speak it and 99.3% can write in Castilian Spanish. Although not included in the most recent report as such, in the 2010 report the number of proficient Spanish speakers in all language abilities (understanding, speaking reading and writing) and over age 15 in Catalonia reached 5.400.000 inhabitants (Generalitat 2010:238).

Catalans who identify themselves as native and proficient speakers of the autochthonous language do so for a variety of reasons. One of the principal determining factors in what language they prefer is the geographical origin of their family and the transmission of the preferred language inter-generationally. Catalonia was subject to several waves of immigration in the twentieth century. The first flood of newcomers originated from other communities of Spain following the civil war as Spaniards from other regions fled the poverty of their post-war homelands with the promise of work and prosperity in the

more industrialized Catalonia. This brought with it an influx of Spanish speakers who subsequently influenced the linguistic identity of the next generations. A second and more recent wave of immigration occurred in the 1990s and during the first years of the 2000s. In this instance, those settling in Catalonia came from Latin-America, Africa and Asia. For these new comers, Spanish is either their native tongue or the default language for day to day interaction for speakers of other languages (Generalitat 2010).

In spite of the recent and massive flow of Spanish speakers to the region, Catalan has enjoyed resurgence in all areas of society over the last few decades, a tendency which has been catalyzed by regional legislation and supported by European language policy. For example, public servants are required to be proficient in co-official Catalan. Likewise, School age children and adolescents studying in Catalonia receive all classroom instruction in Catalan, with exceptions being Spanish and foreign language classes. The participants in the current study have grown up in midst of this resurgence in which Catalan enjoys a special status in public life.

Catalan universities are the preservers and promoters of the Catalan language, and at the same time, the force behind the multilingual identity of young professionals. The local governing body, *La Generalitat de Catalunya*, partners with the competent university departments and administrations in the elaboration of linguistic policies that "guarantee the presence of Catalan in the university environment and promoting its use among faculty and student body in universities of Catalonia" (Generalitat 2010). While the promotion of Catalan is very evident in Catalonian higher education, many institutions have also worked to introduce third languages into the classroom setting. English's particular status as an international language makes it an ever more popular medium of instruction. On average Catalan is the language of choice in 77.2% of undergraduate courses taught at Catalonia's 7 public higher education institutions. The remaining percentage of the course work is done through Spanish, and third languages such as English and French. The 2010-2011 school year saw a considerable rise in the use of both Catalan and third languages as vehicular languages in undergraduate courses (Generalitat 2012).

In these sections we have provided a brief account of both the academic setting in which SALA research has taken place over the last decade. We have also taken a look into the

complex social situation that is present in Catalonia today, and to which the subjects of the present study are exposed on a daily basis. We now turn out attention to the specific objectives of the current study, as well as the research questions that we will address in coming chapters.

4.3. Objectives and Research Questions

The sections that follow describe in greater detail the specific objectives of the empirical study. Section 4.3.1 provides an overview of the global goals of this piece of research and points out three specific areas of inquiry that we will tackle in the remaining chapters of this dissertation (Chapters 6-8). In section 4.3.2 we state our research questions with corresponding sub-questions.

4.3.1. Objectives

Two primary objectives are the focus of the current study. First, we seek to analyse the potential benefits of SA on the production of English learners over the observation period and measured through the domains of complexity, accuracy and fluency. Second, we look at the potentially differential impact of learners' participation in stays of three versus six months, and in particular, whether or not any one modality (3 or 6-month stays) prove more beneficial to these facets of L2 linguistic competence.

Along with the primary objectives described above, we include two secondary goals. First, we seek to determine if, and to what extent proficiency level at the beginning of SA was influential in post-SA outcomes among learners. Second, we seek to find out how stays abroad prior to the current SA may have influenced the outcomes we analyse here. Many students reported having spent as least brief periods in an English-speaking country prior to participating in the study. Information about informants' previous stay(s) in the TL country was gathered through questionnaires and used to determine whether or not these experiences would influence the relative success of the SA period (3 or 6 months) that is the treatment in the present study.

In order to achieve our research goals we made use of a role-play task to elicit student speech productions longitudinally at two different testing times, pre-SA and upon their arrival home from their SA experiences. The productions were evaluated in terms of syntactic and lexical complexity, accuracy and fluency. We also collected NS data,

recruiting native English speakers to perform the same task. Having NS data allowed for comparisons of native and learner performances and also enabled us to see if learner productions approached native productions in any significant way as a result of SA and/or LoS. Furthermore, with NS data in hand, we were better equipped to understand the quantitative measures selected for analysis, and interpret divergences in NS and NNS performances.

Along with CAF scores recorded at pre- and post-test, we included two other approaches to assessing learner level and development. At several points in the analysis we use mean gains in CAF scores that were calculated in order to examine changes in performance over the SA period independent of initial level. We also drew from test scores obtained from other SALA tests, namely a grammar test and a phrase-completion test (cloze) that subjects completed as part of the SALA battery of tests. Gains and composite grammar scores were then related to the measures of CAF to see if any relationship could be determined.

In sum, the overall objectives are listed as follows:

1. To determine the impact of SA on the oral production of English learners in terms of specific CAF measures.
2. To determine whether or not LoS has an effect on the outcomes of SA in these measures.
3. To identify any influence initial level may have on SA outcomes
4. To identify whether or not previous periods spent in the TL country affect current SA outcomes.

4.3.2. Research questions

In order to address the objectives presented above, we formulated the research questions that are outlined in the present section. Here we present three principal research questions that later guide the analysis and discussion developed in Chapters 6-8. Each addresses a single research question and can be broken down into sub-questions that immediately follow the main question. While the first question (RQ1) addresses the impact of SA and LoS on linguistic outcomes, questions RQ2 and RQ3 focus on the additional independent variables of initial level and previous time spend abroad that could have influenced the outcomes of the current SA period

Research question 1 (RQ1)

Which modality of SA (LoS of 3 or 6 months) is more beneficial in the development of L2 speaking performance as measured through CAF?

RQ1a. In what measure(s) do NNS performances change significantly across testing times?

RQ1b. In what measure(s) do SA-3m and SA-6m participants' production differ across testing times?

RQ1c. Do changes in NNS performances from pre- to post-test represent a movement toward NS performances?

RQ1d. Do learner groups differ significantly in terms of overall gains?

Research question 2 (RQ2)

Is linguistic development in the L2 as measured through CAF different for learners with different initial levels of proficiency?

RQ 2a. To what extent is initial proficiency level as measured through pre-test CAF scores related to post-test outcomes in terms of post-test CAF scores and gains that occurred during the observation period?

RQ2b. To what extent is initial proficiency level, as assessed through SALA lexico-grammatical testing instruments, related to post-test gains during SA?

RQ 2c. Do low initial level participants obtain significantly greater gains than high initial level participants during SA?

Research question 3 (RQ3)

To what extent do previous periods abroad in the target language country (previous SA and previous abroad experiences in general) affect the outcome of the present SA experience?

RQ3a. Do first time SA participants and those who have studied abroad before benefit differently from the current SA?

RQ3b. Does previous experience abroad impact the outcomes of the current SA?

RQ3c. Do previous SA and previous experience abroad impact current SA outcomes in terms of gains from pre- to post-test?

CHAPTER 5

METHODS

We will now turn our attention to the methods involved in the current study. This chapter is divided into four main sections with corresponding subsections where indicated. Section 5.1 covers the design of the study. Participant profiles are discussed in section 5.2, where we describe the three groups of subjects who provided us with the data for our study (two groups of NNSs and one group of NSs). Next we describe the process of data collection in detail (5.3), discussing the main task that provided us with the oral data for analysis to which we applied measures of CAF. This task was a role-play between peers (5.3.1). We also describe the grammar and cloze tests that are included among the SALA tests, and here serve to compare initial levels between groups (5.3.2). In Section 5.4, we detail the transcription and coding process in which CLAN software was used to prepare transcriptions and apply codes of syntactic complexity and accuracy to the transcribed role-plays. In this section, we also explain the procedure for capturing measures of lexical complexity (5.4.3) and fluency (5.4.4). Chapter 5 ends with a summary of the specific measures used in the CAF analysis, and how each was selected and calculated.

5.1. Design

As we have already mentioned above, the SALA project was developed in order to take advantage of an obligatory stay abroad that occurred in translation students' second year at university and following a period of two consecutive quarters of FI. In order to assess students' progress over the different treatment periods (FI followed by SA), a repeated measures design was set up (See Appendix C.3 for a diagram of testing times). In this way, students' performances were measured against their own subsequent performances[18]. Although this type of design does not include a control group of learners that did not participate in SA, it has its advantages in that it avoids the problem

[18] It should be noted that while the students who make up the three-month SA group did undergo testing at time 1 in the original SALA design (i.e. approximately 6 months prior to departure on SA), the testing times that are taken into account here correspond to those recorded immediately before leaving for SA and upon return. Therefore, here we only deal with one treatment type: SA.

of between subject variability and oversteps the question of whether or not groups (control and experimental) are really comparable in the first place. It can be assumed that since students' performances are measured against subsequent performances that all other factors will generally remain stable within subjects. However, we acknowledge that some criticisms can arise as to the validity of using the same instruments over two or more testing times as there may be an effect for task repetition that has nothing to do with real linguistic gains.

Coinciding with a widening of the scope of the SALA project, additional data was collected from students who participated in exchanges over the course of two quarters (one semester). These data were collected during the academic years 2010/2011 and 2011/2012. This new group of students was not limited to those studying translation, but rather subjects came from a wide range of degree programmes. The most frequently reported major for the 6-month group was economics and included representatives from a number of other departments ranging from biology to humanities. As linguistic background questionnaires and pre-SA lexico-grammatical tests suggest, the initial levels in English between the two test groups did not seem to be remarkably different. One can assume that due to their choice of studies, students choosing to study translation may vary somewhat in their metalinguistic knowledge to those to following other degree programmes; however, the testing instruments here did not detect any strong differences between groups at pre-test, suggesting that these are indeed comparable groups.

Regardless of degree programme, at the home university academic terms are organized into quarters. Each quarter, 10 weeks of in class instruction are followed by a two-week period for revision and exams. The original SALA cohorts, one of which makes up the three-month group here, was tested on up to 4 occasions, yet, the testing times of interest for the current study were conducted just before departure, and upon arrival from SA. Further details about how data collection was conducted are contained in section 5.3.

5.2. Participants

The oral production of 47 NNSs and 24 NSs was taken into account in the present study. Learner subjects performed a role-play task at pre-test and post-test in which they

were partnered with one of their fellow subjects[19]. NS subjects also performed the role-play with one of their NS peers, yet performed the task only once. In the end, we analysed 94 elicitations from 47 NNS individuals and 24 elicitations from NSs.

Three groups of participants were formed by the duration of their respective SA programmes and by their L1s. Two groups of NNSs who participated in SA periods varying in length (three months vs. approximately 6 months) are described in further detail below, as are a group of NS participants who provided baseline data. NNS Subjects were selected based on the fact that they had participated in the study at both data collection times used in the current study (i.e. just prior to, and upon return from SA). A single learner cohort was selected from the SALA cohorts of subjects who participated in a three-month SA. The decision was made to select participants from only one cohort for two reasons; first, the need to compare a three-month group with a 6-month group required that testing conditions be as uniform as possible. The final cohort of the SALA main corpus was selected as the procedures were better documented than those of the previous cohorts, and the same researchers who collected these data also oversaw the data collection of the 6-month group. Second, oral data from the selected cohort had been recorded under more favourable acoustic conditions; conditions that also more closely resemble those of the 6-month group as compared to the other two cohorts of the greater SALA corpus. This greatly facilitated the transcription process and the capture of measures of fluency.

As mentioned above, learner subjects within the selected cohorts of students who participated in 3- and 6-month SA exchanges were selected on the basis of their having participated in both relevant testing times. The number of 3-month SA participants based on these criteria came to 34. Upon further examining background questionnaire data, one longitudinal subject was excluded from the analysis due to his divergent L1 profile (the subject did not have Spanish or Catalan as a L1). Thus the number of shorter stay participants dropped to 33. Of the total participants who spent 6 months abroad (N=22), subject loss at post-test left 14 learners who had gone on a 6-month stay and participated at the two testing times. Of a total of 28 NS participants who performed the

[19] In cases where there were an odd number of subjects, and thus no peer partner available to perform the role-play, a research assistant performed the role of the missing partner.

task, 4 subjects (2 role-plays) were eliminated from the analysis due to either dissimilarities in profile or some issue with how the task was performed. One NS subject's age varied considerably from the rest (42 at the time the tests were performed). Excluding this subject, NS students' ages ranged from 20-22 years, and NNS from 17-23 years. Learners' average age was 18.8. The other NS role-play pair was eliminated from the analysis due to the fact that in the during the task one participant dominated nearly all of the speaking time, while the partner barely intervened, this particular role-play was considered anomalous and misleading for the analysis, and thus excluded. The selection process left a total of 47 NNS and 24 NS whose group profiles are discussed further in the following sections.

All learner participants reported native or native-like command of Spanish. Most (63.8%) reported being bilingual in Spanish and Catalan, having been exposed to both languages since birth. Another 31.9% considered themselves native speakers of only one language (N=8 Catalan L1 speakers and N=7 Spanish L1 speakers), yet remain early sequential bilinguals due to the wider linguistic context of Catalonia[20]. Two students (4.3%) were originally from other regions of Spain and had not been exposed to Catalan during childhood. One of these reported being bilingual in Spanish and Basque. Learner participants all studied English as their primary foreign language during their education. Participants reported their age of onset of English study (*M*=7.0 years) yet the reported onset ages ranged from 2 to 12, most if not all, were likely exposed to English through the education system before age 12 suggesting a misinterpretation of this questionnaire item. The most frequently reported age, at which participants reported initiating English study was 8 years-old. This age also marks the onset of a new cycle of primary education in the country and likely coincided with the beginning of more intensive English language curriculum at school. Many had previously studied French (68.1%) or German (21.2%) for an average of 3.7 years. All participants (NS and NNSs) were comparable in age and education level. Woman outnumbered men across groups (76.6% Female), and most were enrolled in either their second or third year at university.

[20] These bilingual/single L1 distinctions are based on self-reports and likely a reflection of participants' personal linguistic identity rather than any imbalance in their linguistic competence in Spanish and Catalan.

Experience abroad prior to the current SA period was common among participants. These previous experiences were SA experiences and non-academic abroad experiences. Information gathered by profile and SA questionnaires told us which students were first-time SA participants and which were repeat participants. About half (N=23, 48.9%) of the learner participants had travelled abroad for the purpose of study, this was either language study, or content courses taught in English. The remaining participants had never gone abroad with the express purpose of study (N=24, 51.1%). Participants reported that these SA stays were of three months or less. While no information on the exact duration of prior SA periods was available through the questionnaires, it seems that most of these stays lasted approximately 4 weeks during the summer. While only about half of the subjects were repeat SA participants, the majority of students had had abroad experiences with no formal study involved (N=27, 57.4%). These experiences varied considerably: holidays abroad, English-language camps, or brief stays for summertime work. In contrast, only N=17 (36.2%) had never been abroad in an English-speaking country before.

Thus far we have discussed what participants had in common. In the following subsections we will further define the profiles of the two learner groups, as well as NS participants.

5.2.1. Three-month SA participants

The first learner group included in our analysis (SA-3m from this point on) was made up of undergraduates (N=33) enrolled at a public research university in Barcelona between the academic years of 2007 and 2010. All SA-3m subjects were working toward an undergraduate degree in Translation and Interpretation at the time. All participants had English as their primary foreign language of study, but also studied either French or German as part of their studies and many had previously studied French (71.4%) or German (15.2%) with an average length of study of 2.8 years. Prior to entering university students had substantial exposure to English through FI in primary and secondary/pre-university education. The vast majority also received at least some extracurricular English language instruction in language academies or private lessons during secondary school (84%).

Before entering university subjects passed an entrance exam that assesses language competence in Catalan and Spanish as well as English as a foreign language, *Prova d'accès a la Universitat* (PAU). In 2007 students needed a score of 6.4 (out of 10)[21] in order to enter the degree programme in Translation and Interpretation. Furthermore, before enrolling in the English Translation stream, students had to accredit having a level of B2[22] (following CEFR criteria) or higher. This was done through an English language level test administered by the university. Furthermore, SALA publications have described participants as upper-intermediate language learners at the start of the observation period (Beattie 2014).

According to self-reports, 66,7% of SA-3m subjects were bilingual from birth or early sequential bilinguals (Catalan/Spanish)[23] while 30,4% reported to be native speakers of only one language, or having a strong preference for one of the languages they commonly use for communication (i.e. Spanish or Catalan). At the time the first testing session took place[24] subjects ages ranged between 17 and 21 (*M*=17.97).

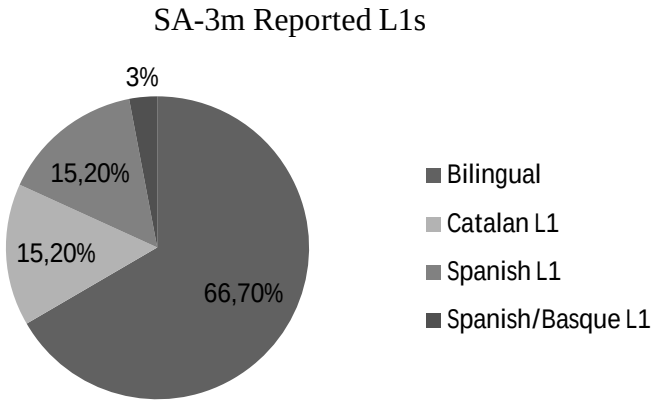


Figure 5.1 SA-3m L1 backgrounds

Other experiences abroad in an English-speaking country prior to the SA period in question were frequent among SA-3m subjects (48,5% of participants) and nearly a third (28,8%) reported having previously been abroad for the purpose of studying

[21] Information retrieved from http://www.gencat.cat/
[22] The requirement has since been reduced to a B1 level on CEFR.
[23] In all cases, we are referring to Catalan and Spanish when we mention NNS participants' bilingualism.
[24] SA-3m subject reported their age at the beginning of the academic year, yet did not perform the pre-SA testing time until approximately 6 months later, thus the difference in age between SA-3m and SA-6m would likely decrease had SA-3m reported their ages immediately before SA.

English – or study through the medium of English – yet for a period of no longer than three months. The exception was one participant who had spent a semester abroad. Only 27.3% reported never having set foot in an English-speaking country prior to the current SA.

Participants went on a short stay abroad for a period of three months during their second year at university. SA-3m destinations included a variety of English-speaking countries. Host universities in the UK, USA, and Australia received SA-3m students as part of exchange programmes partnering with the department of Translation and Interpretation at students' home university and their own equivalent departments.

During their second year at university SA-3m participants were required to spend the first quarter (mid-September to mid-December) abroad, in an English-speaking country. The exchanges were facilitated through the Erasmus scheme to allow for credit transfer from the host institution. While most subjects went to the UK (89.3%), a small percentage (7.5%) went to the US, and one student went to Australia. During this three-month period, students attended lectures and seminars at the host university. Students took a minimum of two courses while abroad; however, no official, across-the-board requirements were specified. For this reason, the number of hours students had to attend class while abroad varied according to the host university. SA-3m subjects lived in university residence halls (76%), shared apartments with other students (16%), or chose to stay with a host family (8%). After their time abroad, students returned to the UPF and followed the regular Translation and Interpreting Studies curriculum.

5.2.2. Six-month SA participants

The second NNS group consisted of undergraduates (N=14) from other departments of the same home university (SA-6m from this point on). Most subjects came from the economics department (61,5%; N=8), with representatives from Human Biology, Humanities, Business Administration and Marketing majors making up the remaining 38,5% (N =5). These data were collected between the academic years of 2010 and 2012, in their second, third or fourth year of study. We confirmed that in all cases, English was the primary foreign language they had studied in school. The mean age at which students began to learn English was 6.7 years. The majority (76.9%) also reported

having studied French for at least 2 years. The mean age at which they began their study of French was 12.8 years.

Along with the requirement of passing the *PAU* entrance exam, the department of economics recommends that incoming students majoring in business-economics degrees have at least a B2 level of English on the CEFR scale of proficiency[25] or the equivalent of First Certificate[26]. There is also a requirement that at least 30 of 240 total academic credits be taken through the medium of English. The number of courses with English as the vehicular language increases incrementally over the four years needed to complete degree programmes in the economics department. The departments that offer human biology and humanities majors also have similar language requirements. Additionally all three departments offer English language courses specific to their disciplines such as 'Business English', 'English for Biological Sciences', or 'English for the Humanities'[27].

Self-reports revealed that 84,6% of SA-6m subjects considered themselves to be bilingual[28] in Catalan and Spanish, while 2 subjects (15.4%) reported having only Spanish as their native language or a strong preference toward one of their languages they commonly use for communication. Their Ages ranged between 20 and 23 (*M*=20.6) at the time the tests were given.

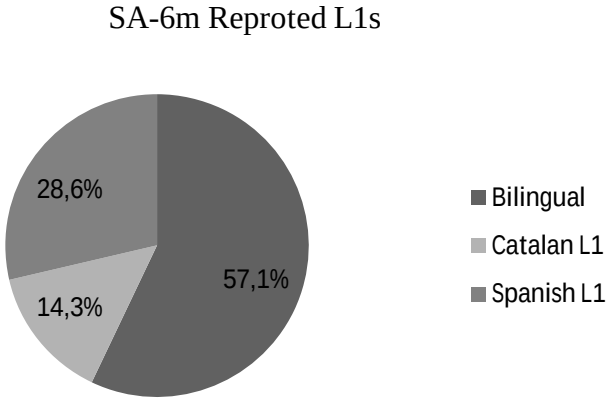


Figure 5.2 SA-6m L1 backgrounds

[25] No explicit department requirements are given, but see http://www.upf.edu/facecon/estudis/graus/ for recommendations.
[26] see http://www.cambridgeenglish.org/
[27] See http://www.upf.edu/estudiants/
[28] See note 18.

SA-6m participants went on longer stays abroad as compared to the length of stay of their SA-3m counterparts. The SA periods for SA-6m ranged from 19 to 26 weeks during their second (N=2, 14.3%), third (N=9, 64.3%) or fourth (N=2, 14.3%) year at university. One student was in the second year of his second four-year degree, reporting being a 6th year student. Their SA destinations included host universities in the UK, USA and Canada. Students were received as part of exchange agreements between their respective departments at the home university and those of their host institutions. Students from the economics department could earn up to 40 transferable credits toward their degree during SA[29] and the department reports that approximately 50% of students working toward the degree participate in an exchange programme during their university studies. The department of Health and Life Sciences reports that nearly all biology students study either in another Spanish university or abroad during the course of their studies, with the majority choosing to go abroad. Without specifying any limit to the credits students can earn abroad, the department states that along with Erasmus exchanges, "the course curriculum provides for a term of mobility so that students who undertake a study period at another university (be it in Spain or abroad) can gain automatic academic recognition for a significant number of credits"[30]. The Humanities department does not mention any specific English language requirement upon enrolment but does state that some required courses are taught in English. Abroad periods are actively promoted as well, and a list of partner institutions is given, but no further details are provided with reference to SA programmes linked to the department[31].

Like their SA-3m counterparts, members of the SA-6m group had studied English throughout their years of schooling within the Catalan education system with the exception of one subject who attended primary and secondary/college preparation in the Canary Islands. Extracurricular English classes in language academies or through private lessons were somewhat frequent among these participants. SA-6m subjects

[29] See http://www.upf.edu/facecon/estudis/graus/
[30] See http://www.upf.edu/estudiants/titulacions/ciencies-salut-vida/grau-biologia/presentacio/
[31] See http://www.upf.edu/estudiants/en/titulacions/humanitats/grau-humanitats/presentacio/

reported having attended private lessons during secondary and/or university preparatory school in 21.4% of cases.

Previous experience abroad, in an English speaking country, was reported to have been three months or less in 61.5% of cases. Two participants (15.4%) had been abroad for a period of more than three months (4 months and 9 months respectively), and three (23.1%) reported never having been to an English speaking country prior to the stay in question here.

Like those of the short-stay learner group, SA-6m exchanges were facilitated through the Erasmus scheme. Most SA-6m subjects went to the US (64.3%). The remaining 35.7% went to the UK, with the exception of one student who studied in Canada. During their time abroad, students attended lectures and seminars at the host university. No official, academic workload requirements were specified. Thus, the number of hours students attended class while abroad varied according to destination and host institution. SA-6m Students lived in university residence halls (28.6%), or shared apartments with other students (71.4%). None chose to stay with a host family.

5.2.3. Native Speaker Participants

Baseline data was collected from a group of 24 highly comparable participants, both in age and education level. Also, like the NNSs female participants outnumber the men (79.2% female). They were all NSs of English. At the time the data were collected, these students were participating in an international exchange program at a public university in Mallorca (Spain). NS participants came from a variety of English speaking countries including the UK (N=8), USA (N=13) and Ireland (N=3).

5.3 Data collection

At each data collection time NNS participants performed a series of tests designed to assess general English language proficiency. Each test captured students' performance in a particular skill. Paper and pen tests were administered that tested listening comprehension, phonological perception (both with audio support), grammar and lexical knowledge and writing ability. Students also completed questionnaires on linguistic background, attitude toward language learning, and upon arrival from SA, a questionnaire on SA conditions. These tests and questionnaires were administered in an

exam-like setting in a classroom or lecture hall on the university campus. The second half of the battery of tests was dedicated to capturing data on oral performance through a reading aloud task, a semi-guided interview, and a role-play in pairs. These tasks were performed in pairs in the presence of a research assistant and digitally recorded. Unlike the paper tests, the oral tasks took place in a small space such as a classroom with only the pair of participants and the researcher present so as to ensure the quality of the digital recordings. When available, the research team made use of soundproof spaces on the university campus ceded by the audio-visual department. The tests were timed, and the whole battery of tests was completed in 2 to 2.5 hours depending upon the availability of timeslots for recording the oral tasks. NS participants performed the same test following the same procedures, yet on only one occasion. A list and brief descriptions of the SALA tests can be consulted in Appendix C.2.

The current study takes into account three of the SALA tests mentioned above. The main task from which learner and NS oral data were compiled was the role-play. Grammar and cloze tests were also included so as to provide additional information about participants' level of English at pre-test in the areas of grammar and lexical knowledge. We also made use of two questionnaires, namely the linguistic background and the SA conditions questionnaires. The sections that follow provide further details about the tasks themselves.

5.3.1. The task

The principal data collection instrument was designed as a problem solving style role-play task to be performed in pairs. Students were presented with an initial scenario, but the task was otherwise open-ended. The roles consisted of a professional decorator and a potential client who were assigned opposing goals in the role-play instructions. The decorator was to try to sell an expensive, minimalist design, while the homeowner was to act as she preferred any design other than the minimalist design, and wanted to stay within a limited budget. Participants were to discuss four different living room decorations, and were given a visual support in the form of a print out with the different decoration styles available (for full instructions see Appendix A.1). In order to accomplish the task, participants had to provide arguments for their proposals as well as initiate and maintain the flow of conversation. The opposing task objectives were given

in this manner in order to elicit high levels of negotiation behaviour while replicating an everyday interaction situation.

The role-play task was chosen over the other speaking task, the interview task, in order to analyse the highly dialogic and highly interactional language that the task elicits. Furthermore, in addition to the factors *–Monologic* and *+Interaction* the role-play task also exhibits other certain elements of task complexity, hypothesised to affect learner performance as observed through the dimensions of CAF (Robinson 2005). The role-play requires increased reasoning demands as compared to the interview task. Moreover, in performing the role-play, students are drawing less from their own prior knowledge than in the interview task, where the object is to explain personal (well known) information. These characteristics simply serve to highlight the differences between the two tasks, justify the choice of one over the other, and roughly place the role-play on a task complexity scale. The role of task complexity itself as it impacts learner language is beyond the scope of the present work.

In further qualifying the choice of the role-play task over the interview we note that the SALA-based publication by Mora and Valls-Ferrer (2012), applied CAF measures to the oral data collected through the interview task. They found significant gains in fluency after a SA period of three months, but marginal gains in accuracy and no gains in complexity. By substituting the interview with the role-play task as the elicitation instrument in the present work, we seek to expand the findings of this and other previous studies (Juan-Garau & Pérez-Vidal 2007; Juan-Garau 2014 for accuracy) and shed further light on the L2 outcomes seen after SA with analyses applied across task types.

5.3.2. Other SALA internal tests (Grammar test and Cloze)

Included in the written portion of the battery of tests were a sentence-rephrasing task and a cloze task. Together, these tests were intended to assess subjects' knowledge of grammar and lexis. The rephrasing task presented subjects with a sentence and a blank line immediately underneath it. Students were to rephrase the sentence provided into a grammatically correct alternative in which the original idea remained unchanged. An example sentence transformation was provided. Subjects were presented with 20 such sentences and were given 15 minutes to complete the test (see Appendix C.1).

The cloze task consisted of 286-word text entitled "The lady who liked adventure". The story was divided into three paragraphs with 20 gaps in the text. Students were to fill in the gaps with a single suitable word that fit the immediate context. Some spaces could have more than one appropriate answer. As with the rephrasing task, students were given 15 minutes to complete the cloze test (See Appendix C.1). Participants' scores on the sentence rephrasing and cloze tests were combined into a global score of grammar and lexical knowledge. Students' scores were used in gauging initial level prior to SA as it has been used in SALA studies in the past (e.g. Barquin 2012).

Once considered to be an indicator of overall L2 proficiency, the validity of the cloze test for gauging overall proficiency has since been revised. Alderson (1979) suggests that cloze tests are adequate for taping low-order, or core proficiency in grammatical and lexical knowledge, but fall short of capturing high-level proficiency skills, such as reading comprehension. Others corroborate his suggestions, claiming that cloze tests, while useful in tapping lexico-grammatical knowledge at the sentential/suprasentential level do not capture global proficiency (Saito 2003). Nonetheless cloze tests continue to be used today in language teaching and assessment and in evaluation through standardized tests such as the Examination for the Certificate of English Proficiency (ECEP)[32]. Here, sentence rephrase and cloze tests administered at the pre-SA testing time, help to fill the gaps in our knowledge of learners' initial level and allow us to better demonstrate the comparability of the two learner groups. Furthermore, composite grammar-cloze scores provide us with an additional variable of pre-programme lexico-grammatical knowledge that allows us to further postulate a differential impact of SA as related to initial level. We do not suggest however, that these tests comprehensively capture learners' global L2 proficiency at pre-test.

5.3.3. Procedure

The procedure for administering the role-play was followed closely in the data collection sessions. Once organized into pairs, Student A was assigned the role of a homeowner seeking advice from a professional decorator. Student B was assigned the

[32] See Examination for the Certificate of English Proficiency (ECEP) Technical Review, available at http://www.cambridgemichigan.org/

role of decorator, and given a print out with photos representing four different decoration options to be used as a prop during the interaction. Students were assigned roles at random, and were given opposing objectives according to their role. While the homeowner sought a more traditional style, and to spend as little money as possible, the decorator sought to sell a minimalist design that would yield the highest commission for himself. Students were given written instructions and asked to read them silently to themselves before initiating the task. Any questions the students had about the task were answered prior to the start of the recording. For the exact instructions and visual support given to the subjects see Appendix A.1.

At the beginning of each recording students were asked to identify themselves by saying a code they had been previously assigned. Having the code recorded at the beginning of each recording allowed transcribers to identify each subject while maintaining the participants' anonymity in the compilation of the corpus and at all stages of the subsequent data analysis. After identifying themselves, students also stated the task name (i.e. "role-play") and the role they had been assigned ("decorator" or "homeowner").

Once the subjects were properly identified on the recording the interaction was initiated. As mentioned in the written instructions (See Appendix A.1), Student A (the homeowner) was to initiate the interaction by reading a line provided them in the written instructions: *I'm tired of this old decoration. I want a big change. How can you help me?* The majority of student pairs began the task in this way, with some exceptions being those who paraphrased the introduction, or where student B (the decorator) pre-empted the initiating turn of Student A with a greeting or an introduction of some sort. Once the take had begun subjects were allowed to negotiate from the standpoint of their assigned role for up to 7 minutes, however, most did not take the whole allotted time (M=3,03 minutes at pre-test and M=3,57 minutes at post-test). The interaction was free in the sense that the researcher intervened only to give instructions or inform them of the approaching time limit, leaving the students to interact spontaneously with one another.

When an odd number of subjects were present at a given data session, thus preventing student-student pairing, a member of the research team performed the opposing role.

This was the case on 13 occasions (13.8% of the recorded interactions). Admittedly, this solution introduces issues of inconsistency in task conditions in that, in these cases, the interlocutor is not a member of the peer group but rather a stranger, an elder and likely more proficient in English. Although not an ideal solution, the decision was made to have the researcher stand in so as to prevent the loss of valuable longitudinal data.

At the second testing time, roles where again assigned at random, thus, while some students repeated the same role from time T1, others took on the opposite role at post-test. Native speaker subjects, who provided baseline data in this task, performed the role-play only once with the same instructions and procedures as given to NNS subjects.

5.4. Analysis

In this section we detail the procedures followed in the transcription, coding and other analyses of the data included in the present study. We then present how each measure was calculated, giving us the 9 dependent variables we examine through statistical analysis. In the analysis and operationalisation of CAF dimensions, we closely follow the criteria offered by Mora and Valls-Ferrer (2012).

5.4.1. Transcription

Research assistants, including the author, transcribed the role-plays by hand creating text files while listening to the audio recording of the performances. We used Computerized Language Analysis (CLAN) software and produced the transcriptions with the CLAN editor in text mode. This way, the resulting CHAT files would be coding ready.

The basic CHAT transcription conventions described in MacWhinney (2000) where applied to the role-play data. Standard punctuation was not generally applied to the transcribed data except for an utterance terminator such as a period or question mark. Capitalization was applied only to the pronoun ‘I’ and proper nouns. The expression ‘okay’ was written as such, and titles such as ‘Mr.’ were written out in words: ‘mister’. Any numbers mentioned during the interaction were also written in words. Compound nouns composed of two words in conventional spelling but with a specific semantic distinction were written together. For example ‘dining room’ was written as ‘diningroom’. Incomprehensible speech was marked with ‘xxx’. The bullets *HOU: for

the homeowner's turns and *DEC: for the decorator's turns begin each new turn denoting the speaker's role. A fully transcribed sample role-play is included in Appendix B.1.

Beyond the basic CHAT conventions, other transcription norms were established. An effort was made to represent the actual speech sample as faithfully as possible. Thus, filled pauses (i.e. 'um', 'er', 'uh', etc.) were included in the transcription as were repetitions, false starts and reformulations. False starts and reformulations are contained in brackets {...}. Partial words were represented with the missing portion in parenthesis such as 'so(fa)' and were determined by the context whenever possible. Overlapping speech was set between angle brackets <...> in both adjacent turns. Pauses that potentially determine unit and/or clausal boundaries are recorded within parenthesis and in seconds, rounded to two decimal places. Silent pauses of (>0,5 seconds) were recorded as this is used as a cut-off mark in the criteria for identifying separate AS-Units (See Foster, Tonkyn, & Wigglesworth 2000).

5.4.2. Coding procedures

After the transcription process was complete, syntactic complexity measures and coding for errors could be performed. CLAN software allows the user to enter codes in a new line of text immediately underneath the transcribed text to be coded then automatically counts the entered codes presenting them in a generated output summary. A raw numbers count can then be recorded based on the CLAN summary. Syntactic complexity measures were taken by creating "codes" for AS-unit, clause, and subordinate clause and coordinated clauses. As the researcher analysed each syntactic element, codes were entered to record the count. At the same time, the researcher entered all error codes using the same coding line in CLAN. Errors were coded by type including grammatical, lexical, pragmatic and phonological errors, however, these error types were not discussed as part of the analysis, rather, they aided in forming criteria to identify errors in the data.

5.4.2.1. Syntactic units

The transcribed role-play dialogues were divided into AS-units. We chose to use AS-units due to the fact that the criteria for this unit were developed with the very often

fragmented nature of oral language in mind, and taking into account those aspects of speech which have proven to be problematic when applying other types of units to spontaneous speech production units (e.g. T-unit, C-unit). T-units, for example have been employed in a wide range of studies analysing the L2, and was originally established within studies of written language, however their application to oral data has proven awkward in some cases and requires some degree of criteria modification in order to be adequately applied to spoken language (Kuiken & Vedder 2012). Since the publication of Foster, Tonkyn and Wigglesworth (2000) in which the AS-unit is presented in detail and with specific examples of how it should be applied to real data, researchers dealing with L2 speech have increasingly opted for the use of the AS-unit over other alternative units.

Foster et al. (2000) define the AS-unit as "a single speaker utterance consisting of an independent clause, or sub-clausal unit, together with any subordinate clause(s)" (p.365). In Appendix B.2, we use extracts from our data, to provide examples of how transcribed speech samples were divided into AS-units. In the examples an upright slash (|) represents an AS-unit boundary. An independent clause is defined minimally as "a clause including a finite verb" (p.365). In our examples the symbol (::) represents a clausal boundary within the same AS-unit. Appendix B.2, Examples 1 and 2 illustrate these unit and clausal divisions. Our analysis, in following Foster and colleagues (2000), independent sub-clausal units are taken into account in the unit count and are considered to be any sub-clausal unit consisting of "...one or more phrases which can be elaborated to a full clause by means of the recovery of ellipted elements from the context of the discourse" (p.366). For an example see Appendix B.2, Example 3. This definition is helpful in the analysis of oral discourse due to its often fragmented quality.

Subordination has been established as a measure of complexity in many studies of L2 complexity (Wolfe-Quintero, Inagaki, & Kim 1998). See Bulté & Housen (2012) for a lengthy up-to-date list of studies in which subordination measures appear as complexity variables. Therefore, it is essential that we clearly define what is considered a subordinate clause. We define a subordinate clause as consisting minimally of "...a finite or non-finite verb element plus at least one other element" (Foster et al.

2000:366). The "additional elements" referred to here can be in the form of a subject, object, complement, or an adverbial element (See Example 4, Appendix B.2).

We limited the unit analysis of the role-plays to "level two" analysis, described in Foster et al (2000). This level is intended to lend itself to highly interactional data where a considerable portion of the data results in only minimal units such as one-word, minor utterances and echoed responses. For examples of the application of level two criteria see Appendix B.2, Examples 5-7. We further specify here the exclusion from the final AS-unit count of a string of one-word, minor elements (see Appendix B.2, Examples 5-7). Self-corrections, false starts, repetitions attributable to dysfluency were not taken into account unless these met the criteria to be counted as an AS-unit. As mentioned above, in the transcripts, these dysfluency phenomena are enclosed in brackets {...} (See Example 8, Appendix B.2).

Pausing behaviour and intonation were taken into account when making decisions as to where to draw the unit boundary. In runs of speech where the speaker uses elements of coordination or subordination to hold the floor, but where the connectors have no other clear grammatical or cohesive function, again, the criteria from Foster et al (2000) were applied to determine where the unit boundary lies. Pauses of 0.50 seconds or longer were considered to mark an AS-unit boundary when accompanied by either rising or falling intonation. The excerpts where this type of unit division was required were isolated on a digitally generated waveform and pauses were measured manually in seconds using PRAAT phonetic analysis software (Boersma & Weenink 2014). Measures of pause time were rounded to the nearest hundredth of a second 0.50. During the analysis, the presence of falling and rising intonation was determined impressionistically in most cases due to the fact that these variations in pitch are easily perceived by the naked ear; however, pitch contours automatically generated in PRAAT were also consulted whenever necessary. See Appendix B.2, Example 9.

The data in question are highly interactive and interruptions are frequent. We dealt with speakers' interruptions of one another by applying our global criteria on a case by case basis, taking into consideration that a unit may be produced across turns when an interruption has taken place (see Example 10, Appendix B.2). In cases where scaffolding occurred in the data, the speaker who completed the previous turn (i.e.

exhibited scaffolding behaviour) was credited with having produced a complete AS-unit (See Appendix B.2, Example 11).

In the majority of cases the AS-unit and clausal boundaries could be clearly determined according to our criteria, however, in (6.4%) of cases the correct application of units and/or subordination counts were uncertain.

AS-units thus served as a denominator for the 4 structural complexity measures used in our analysis (words per AS-unit, clauses per AS-unit, subordinated clauses per AS-unit and coordinated clauses per AS-unit). We also calculated one of two error ratios at the AS-unit level (i.e. Errors per AS-unit).

5.4.2.2. Coding for errors

The transcribed role-plays were coded for accuracy using the same technique mentioned in the syntactic units count. The researcher identified linguistic errors which were recorded though a system of codes, and applied directly to the transcripts using CLAN software. Linguistic elements were identified, labelled by type, and coded as incorrect if they exhibited one or more errors of any type. Errors were classified by type: lexical, semantic, morphosyntactic and phonological (i.e. phonemic substitutions, deletions or insertions that deviated from the norm).

Once all the syntactic unit codes and error codes were entered in a given set of CHAT transcripts and the relevant files selected, a frequency command was run using CLAN's commands window, automatically generating a raw numbers count of syntactic units and each error by type for each participant. Raw numbers were recorded on a spreadsheet for subsequent analysis.

5.4.3. Procedure for measuring lexical complexity

CLAN software automatically generates a type/token count for selected CHAT files when a frequency code is entered in the commands window. The type and number of tokens (WDs) for each subject's performance at each testing time was recorded in a spreadsheet for subsequent calculations of a measure of linguistic complexity, namely Guiraud's index of lexical richness.

5.4.4. Procedure for measuring fluency (Speech rate)

The measures of fluency used in this study were based on temporal fluency measures of speech rate and pruned speech rate in syllables per minute. To calculate these measures, a syllable count and a measure of total time are required. Computerized techniques for automatically detecting syllable nuclei in a recorded speech sample have been developed in recent years (see De Jong & Wempe 2009). Such techniques greatly facilitate the process of measuring speech rate, especially when dealing with large quantities of data. However, automatic syllable counts require that recordings be quite clear in terms of the signal to noise ratio. Unfortunately, many of the SALA role-plays were recorded in less than ideal acoustic conditions preventing us from using automatic techniques. In order to obtain the speech rate measures a research assistant listened to each role-play, hand counting syllables and entering the tally in a spread sheet. Similarly, a second, pruned syllable count was taken. In the pruned count repetitions, reformulations, false starts, and asides in the L1 were omitted from the tally following (Gilabert 2006). Along with the syllable and pruned syllable counts, a measure of total time (including pause time) was taken from the speech samples from the onset to the end of each run of speech. Research assistants recorded speaking time with the help of a Praat TextGrid. A script was then run that calculated total time (including pause time)[33] by summing the duration of all intervals marked as speaking intervals and for each participant. The total duration was then recorded in a spreadsheet for each subject and testing time. Syllables per minute and pruned syllables per minute were calculated to give us pre- and post-test scores.

5.4.5. Reliability

We have already mentioned that in (6.4%) of cases, the correct application of AS-units, clauses and subordinate clauses was doubtful. Accuracy coding proved more problematic as the application of error codes was doubtful in 17.2% of cases.

Syntactic units and errors were analysed by only one rater, the author, while syllables per minute, pruned syllables per minute and GRAMSCOR were analysed by two raters en each case. We begin by calculating intra-rater reliability for complexity and accuracy

[33] This script is attributed to Mietta Lennes (2002) and available at (http://www.helsinki.fi/~lennes/praat-scripts/public/total_duration_of_labeled_segments.praat)

scorings. Intra-rater reliability was assessed by randomly selecting a sub-sample of 10 role-plays (20 learner performances) that were evaluated a second time by the researcher. Using the Intra Class Correlation (ICC) method in SPSS we selected a two-way fixed-effects model and set the confidence interval at .95. The results revealed that intra-rater reliability was very acceptable (ICC >.90). This indicated that the criteria for identifying AS-units, clauses, subordinate clauses, coordinated clauses, and errors were consistently applied throughout the analysis in spite of some very minor discrepancies in the application of codes (see Table 5.1). Coding criteria were quite straight forward and did not leave room for a great deal of interpretation, this lead to highly consistent ratings overall.

Table 5.1 Intra-rater reliability

Syntactic unit	ICC (Single measures)
AS-unit	.952
Clause	.954
Subordinate clause	.927
Coordinated clauses	.941

Inter-rater reliability was calculated for those measures in which more than one rater had performed the analysis (Fluency, GRAMSCOR). We calculated the total scores reported for the complete set of role-plays, then the mean scores by rater were calculated for the same 10 role-plays (that is, 20 cases) that had been scored a second time. Inter-rater reliability was very acceptable, indicating that raters applied the criteria similarly (see Table 5.2).

Table 5.2 Inter-rater reliability

Measure	ICC (Average measures)
Syllables/min	.998
Pruned Syllables/min	.996
GRAMSCOR	.940

We can assume then, that fluency and GRAMSCOR scores reported in our study had been evaluated reliably across raters, following the established criteria, and that these differences in scores do indeed represent true differences in performance.

5.4.6. Measures

Here we provide a summary of how measures of CAF were chosen and the calculation of each. The analysis of role-play data gave us four measures of syntactic complexity, one measure of lexical complexity, two measures of accuracy, and two measures of fluency. Further specifications are included in the following subsections.

5.4.6.1. Syntactic complexity

Measures of syntactic complexity are used to gage the degree of complexification of learner speech. Norris and Ortega (2009) identify three major sub-constructs of syntactic complexity: subordination, general complexity and subclausal complexity. In the current study, we include measures that tap the first two of these classes of syntactic complexity.

We examine complexity through subordination; that is, the degree of embedding in the language sample. Syntactic complexity metrics can be considered as such when clause or subordinate clause appears in the numerator. The measures used here that are related to subordination are clauses per AS-unit and subordinate clauses per clause. Both these measures have been shown to distinguish proficiency levels at least in terms of programme level (Wolfe-Quintero 1998)[34] Measures of subordination have been widely used in studies of L2 production, often under the assumption that higher degrees of subordination reflect more elaborate or advanced language and more automatised structures on the part of the learner. This association of subordination with advancedness comes from findings that suggest that subordination is more difficult (more cognitively taxing) than other means of complexification such as coordination (e.g. Lord 2002). Others are more sceptical about the existence of a direct link between subordination and proficiency and urge that the relationship between these constructs be further tested empirically, and that it not merely be taken for granted (Bulté & Housen 2012).

We also look into general complexity, through a length measure. It is argued, that general complexity can be measured through practically "any length-based metric with a potentially multiple-clausal unit of production in the denominator (Norris & Ortega

[34] That is, the equivalent of these measured reported on in 1998, with T-unit in the denominator.

2009:561). Therefore, in order to tap general or overall complexity in the current work, we include the length measure of words per AS-unit.

Finally, we include a complementary complexity measure which taps a different source of complexification: coordination. Coordination has generally been seen as the primary means of complexification at lower levels of proficiency (Bardovi-Harlig 1992) and therefore, is often ignored in studies of more advanced learner language (Norris & Ortega 2009). Nonetheless, we include a measure of coordinated clauses per AS-unit so as to unassumingly assess learners' use of coordination at their current level and how it may evolve over the SA period and/or within each learner group. We have opted for the measure of coordinated clauses per AS-unit as this measure quite effectively isolates complexity through coordination independent of subordination[35]. The instances of coordination were counted when linking two clauses or material that fit the criteria of a subclausal unit. The criteria for determining what counts as a clause is detailed in section 5.4.2.1 above.

Raw-numbers count for AS-units, clauses, subordinate clauses and coordinated clauses ware obtained though CLAN outputs in terms of the total number produced per individual participant and separately at each testing time (pre- and post-SA in the case of NNSs). Once raw numbers were recorded by participant in a spreadsheet, the calculation of complexity indices could be performed. The resulting measures of syntactic complexity were: Words per AS-unit (WD/ASU), Clauses per AS-unit (C/ASU), subordinate clauses per AS-unit (SUBC/ASU), and coordinated clauses per AS-unit (CoordC/ASU).

5.4.6.2. Lexical complexity

The measure of lexical complexity included in the current study was Guiraud's Index of lexical richness (GuirIndex). This index measures lexical diversity within a given language sample. Guiraud's Index has come into use in studies of L2 vocabulary as an alternative to a simple type/token ratio as the latter has been seen to be highly sensitive

[35] The use of Norris and Ortega's (2009) recommendation, i.e. a *Coordination Index* proposed by (Bardovi-Harlig 1992), has been criticized as confounding complexity sources, that is coordination and subordination (Bulté & Housen 2012:38)

to text length (in words). As text length increases (i.e. total amount of words) TTR logically decreases thus potentially invalidating comparisons between samples in which text length varies considerably (Vermeer 2000). Guiraud's Index avoids the effect of text length by dividing the number of word types by the square root of tokens. Thus, GuirIndx compensates for variations in text length and is considered to be a more reliable measure of lexical complexity than the TTR in the analysis of oral productions where text length varies, as is the case with the role-play data used here.

The number of types and the number of tokens for each performance was automatically generated in the CLAN output and recorded in a spreadsheet organized by test subject. Guiraud's Index was calculated by dividing the number of word types by the square root of the number of tokens.

5.4.6.3. Accuracy

Accuracy ratios were calculated at two levels in order to capture any potential variations at the unit or clausal level. Thus, the two measures of accuracy used here are errors per AS-unit (E/ASU), and errors per clause (E/C), both of which have been found to correlate with holistic ratings of L2 proficiency (Wolfe-Quintero et al. 1998)[36]. The calculation of accuracy ratios consisted in dividing the total number of errors by the total number of AS-units, and dividing the total number of errors by the total number of clauses respectively (Wolfe-Quintero et al. 1998).

5.4.6.4. Fluency

Speech rate (the number of words or syllables per minute including pause time) has proven to be a very robust and reliable predictor of perceived oral fluency in a wide variety of studies (Lennon 1990a; Riggenbach 1991; Towell et al. 1996; Freed 2000; Kormos & Dénes 2004; De Jong & Wempre 2009). Speech rate continues to be a robust measure even in the very dysfluent speech of low proficiency learners produced by low-proficiency speakers (Mora 2006). Two measures of fluency were included here, namely speech rate (SR) and pruned speech rate (Pruned SR) both measured in syllables per minute following Griffiths (1991). The advantage of including these measures of

[36] Wolfe-Quintero et al. (1998) refers to T-units, the predominate unit of measure at the time.

speech rate (SR and Pruned SR) is that they include both the amount of speech and the length of pauses. They differ in that Pruned SR, as opposed to regular SR, omits repetitions, reformulations, false starts and utterances in the L1 from the syllable count (Gilabert 2006; Yuan & Ellis 2003; Lennon 1990b).

CHAPTER

6 RESULTS

The principal aim of the remaining chapters is to better understand the role of LoS in the SA outcomes observed in the current study. We also look into the role initial level plays in SA outcomes as well as how previous periods spent abroad may affect outcomes of the current SA. In this chapter, the results of the analyses pertaining to the present study are presented so as to address the three principal research questions that we outlined in Chapter 4. First, we report on preliminary analyses having to do with initial lexico-grammatical knowledge and the impact that the assigned role had on the data elicited through the role-play task. The remainder of the chapter is organised following the order of the research questions and presenting the results in each case. In each section we will restate the RQ in question and then present the relevant statistical analysis and results. The chapter concludes with a summary of our results.

6.1. Preliminary statistical analyses

We used statistical software (SPSS, Version 20) to generate summaries of descriptive statistics for each dependent variable, and by group (e.g. SA-3m, SA-6m and NSs, Groups with/without previous abroad experience, etc.). Before each round of analysis, tests of normality were performed in order to see whether or not each dependent variable was normally distributed, and thus allowing us to best meet the assumptions of the appropriate statistical test. The data were also screened for the presence of outliers and extreme cases, and decisions were made in each case as to how to treat these data.

6.1.1. Treatment of violations of the assumption of normality and outliers

Normality tests run in SPSS revealed through Kolmogorov-Smirnov and Shapiro-Wilk statistics which variables were considered (non)normally distributed by the subject group(s) to be tested. Whenever it was deemed advantageous for the analysis, nonnormally distributed variables were mathematically transformed following the steps described in the following subsection.

6.1.2. Variable transformation

Parametric statistical techniques are often applied under the assumption of multivariate normality. Although some controversy exists among statisticians as to how to treat nonnormal distributions in the event that the research design warrants the use of a parametric test, Tabachnick and Fidell (2006) argue for the transformation of nonnormal distributions in order to enhance the robustness of those parametric statistical techniques that assume multivariate normality. The decision was made to follow this recommendation; transforming the values of nonnormally distributed variables where needed in order to better meet these underlying assumptions.

6.1.3. Initial level through Grammar scores

In order to better assess the pre-programme English level of students arranged by LoS, an independent-samples t-test was conducted in order to compare composite scores on SALA grammar rephrase and cloze tests for the two learner groups. There was no significant difference in scores between SA-3m (M=13.99, SD=7.63) and SA-6m (M=34.02, SD=6.71) groups, $t(45)$=-1.206, p=.234, indicating that there was no significant difference between how the two learner groups performed on pre-SA grammar and cloze tests. This suggests, as does background information gathered through a questionnaire, that NNS student groups are highly comparable in terms linguistic knowledge prior to departure.

6.1.4. The effect of *Role*

We suspected that the role students took on during the role-play (i.e. either decorator or homeowner) influenced the amount of speech elicited through the role-play task. These different effects for role were confirmed by an independent-samples t-test. The t-test was conducted in order to compare mean number of turns, AS-units (ASU)[37] and words (WD) elicited through each role and across all subject groups and testing times. While no significant differences were found in the number of turns elicited (due to the fact that the interactions were dyadic and practically required the same amount of turns from each participant), significant differences in the mean number of ASUs and WDs elicited were found between decorator and homeowner roles. The results of independent-

[37]Mean number of AS-units occurred in nonnormal distributions, and was submitted to variable transformation (SQRT) prior to running a independent samples t-test. Therefore, the reported mean for this variable is not based on the true number of AS-units found in the role-play data.

samples t-tests are summarized in Table 6.1. We found that the decorator's role elicited significantly more speech in terms of ASUs (*M*=4.95, *SD*=1.12*)* and WDs (*M*=87.72, *SD*=36.85) than did the homeowner's role (ASU: *M*=4.41, *SD*=1.10; WD: *M*=63.6, *SD*=31.76).

Table 6.1 Independent-Samples t-tests comparing amount of speech (in AS-units and words) elicited through Decorator and Homeowner roles

Length measure	*t*(116)	*p*	Eta sq.
ASU†	2.589	.011*	.055
WD	3.797	.000*	.117

*Significant *p* values at the 0.05 level. Shaded values indicate significance.
Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT)

In order to determine the size effect for "role" we calculated eta squared[38]. Eta squared came to .055 for ASU and .117 for WD, indicating that, the role performed by the subjects did have a significant effect on the amount of speech elicited; and this effect was moderate in the number of AS-units the task elicited and was somewhat large in the number of words, following Cohen (1988).

6.2. Addressing RQ1: The effects of SA and LoS

In this section we address our first research question and four sub-questions, which were previously presented in Chapter 4 as follows:

Research question 1 (RQ1)

Which modality of SA (LoS of 3 or 6 months) is more beneficial in the development of L2 speaking performance as measured through CAF?

RQ1a. In what measure(s) do NNS performances change significantly across testing times?

RQ1b. In what measure(s) do SA-3m and SA-6m participants' production differ across testing times?

RQ1c. Do changes in NNS performances from pre- to post-test represent a movement toward NS performances?

[38] Eta squared was calculated using the following formula $eta^2=t^2/t^2+(N1+N2-2)$

RQ1d. Do learner groups differ significantly in terms of overall gains?

In order to address this first research question, we approached the data in several different ways. First, in order to determine in what ways learner performances changed over the SA period (RQ1a), and whether or not LoS had a differential effect on SA outcomes (RQ1b), we performed a series of mixed between-within ANOVAs where *Time* (i.e. SA period) represented the within subjects factor and *Group* (i.e. LoS) represented the between-subjects factor. CAF scores at pre- and post-test served as dependent variables. These analyses are detailed in sections 6.2.1 and 6.2.2.

Next, in order to further explore the factors of *Time* and *Group* separately and at the same time, further qualify the previous ANOVA results, we performed a series of t-tests on the same data. Section 6.2.3 contains a report of independent-samples t-tests that were run in order to determine any pre-SA differences in initial level between groups, followed up with another set of t-tests for post-test scores. Also in section 6.2.3, we examine how SA impacted each group independently of the other using a Paired samples t-test for each participant group and testing pre- and post-test scores against each other.

Section 6.2.4 addresses whether or not NNS performances approached those of NS as a result of SA (RQ1c). We have included NS data at various points in this section addressing RQ1 for comparative purposes, but in 6.2.4 we first apply statistical analysis to the NNS/NS comparison. Each LoS group was compared to NS baseline data in terms of pre-test, and then post-test CAF scores.

Section 6.3 introduces analyses of the linguistic gains obtained during SA, and thus providing a different perspective on how learner groups' L2 performance changed over time and independent of initial level. Independent-samples t-tests were performed where learner groups were compared in terms of gains. We also examine individual data in terms of the percentage of students who actually gained to some degree during SA. A report on the proportion of 'gainers' can be found in Section 6.3.2.

To begin, we first give an overall picture of the data that were used for analysis. A summary of mean scores on all speech measures and their standard deviations is given

in Table 6.2 and corresponding Bar graphs are provided in Figure 6.1. We see that learners' syntactic complexity remains quite stable over the SA period Lexical diversity and accuracy dropped by post-test and fluency scores increased; changes in learner performance that approach NS values.

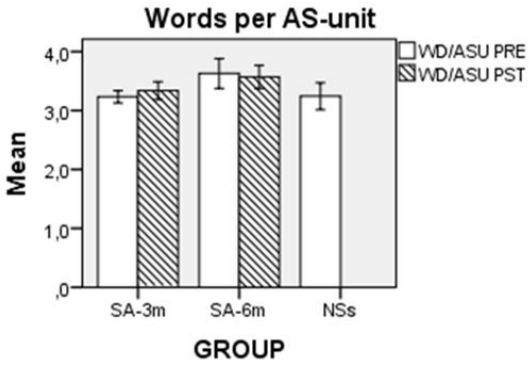
Table 6.2 Mean learner CAF scores, SA-3m (N=33), SA-6m (N=14), and NS (N=24) (*SD* in parentheses).

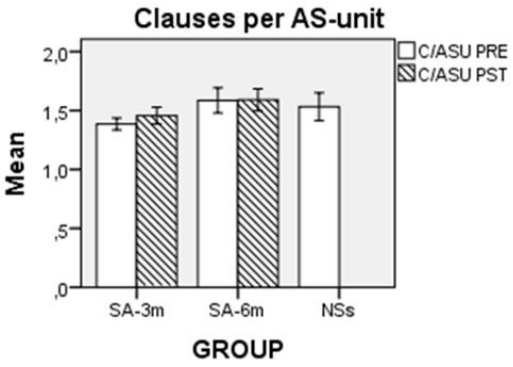
	NNSs				
	SA-3m		SA-6m		NSs
Oral measures	PRE-SA	POST -SA	PRE-SA	POST -SA	
Complexity					
WD/ASU	3.23 (0.60)	3.34 (0.87)	3.63 (0.94)	3.57 (0.73)	3.24 (1.12)
C/ASU	1.39 (0.29)	1.46 (0.41)	1.59 (0.40)	1.59 (0.35)	1.53 (0.58)
SUBC/ASU	0.74(0.36)	0.85 (0.54)	1.08 (0.66)	1.09 (0.53)	1.14 (0.94)
CoordC/ASU	0.29 (0.22)	0.36 (0.31)	0.30 (0.20)	0.30 (0.17)	0.22 (0.18)
GuirIndex	0.83 (0.17)	0.75 (0.15)	0.76 (0.16)	0.70 (0.14)	0.60 (0.17)
Accuracy					
E/ASU	0.30 (0.24)	0.23 (0.14)	0.34 (0.26)	0.29 (0.15)	0.03 (0.46)
E/C	0.22 (0.16)	0.19 (0.13)	0.20 (0.14)	0.18 (0.07)	0.02 (0.28)
Fluency					
SR	163.7 (33.0)	177.5 (32.5)	190.8 (26.2)	198.6 (31.2)	231.9 (44.9)
PRUNED SR	154.5 (32.6)	167.9 (32.1)	177.5 (26.3)	187.4 (29.1)	226.3 (45.3)

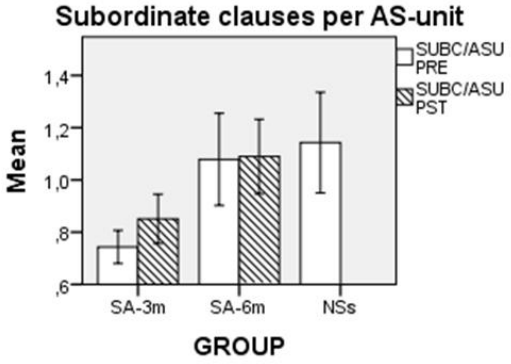
Note: means and standard deviations for SUBC/ASU, E/ASU and E/C appear here in their original form, although for further statistical analysis these variables were transformed in order to better meet the assumption of normality, and reduce the influence of outliers.

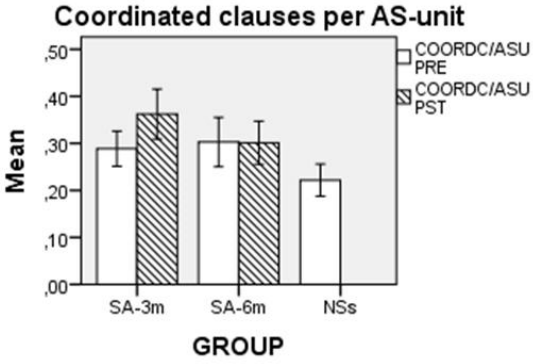
Those variables included in the analysis of LoS that were related to syntactic and lexical complexity made up reasonably normal distributions across groups and testing times with the exception of WD/ASU, SUBC/ASU and CoordC/ASU in at least one subject group and at pre- and/or post-test. Error ratios at the unit and clausal level were not distributed normally according to normality test statistics and also proved more problematic in terms of the presence of outliers. Upon transformation all variable distributions were found to be normal according to Shapiro-Wilk and/or Kolmogorov-Smirnov statistics. One SA-3m subject produced extreme values for SR and Pruned SR at pre-test. Upon the elimination of these two extreme values, mean fluency measures were distributed normally across groups.

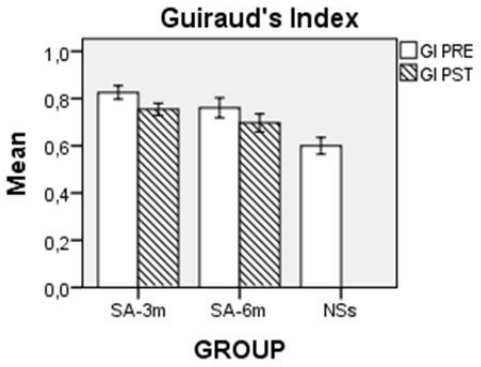
Figure 6.1 Bar graphs of mean pre- and post-test CAF scores of learner groups arranged by LoS. NS were tested only once (NS means shown as pre-test scores). Plots represent means prior to transformation. Error bars: +/- 1 SE

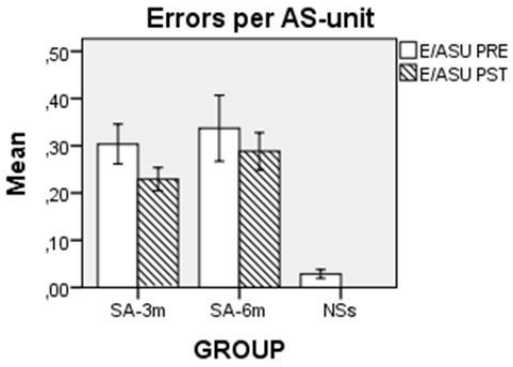

Words per AS-unit
Mean
4,0
3,0
2,0
1,0
,0
SA-3m
SA-6m
NSs
GROUP
WD/ASU PRE
WD/ASU PST

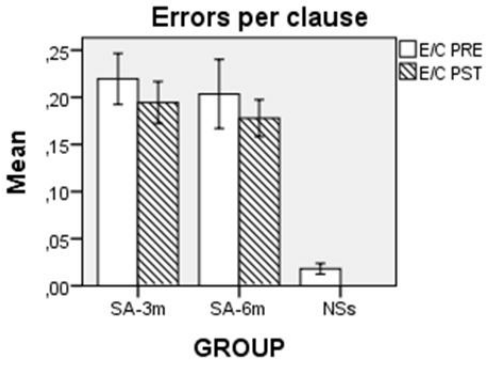

Clauses per AS-unit
Mean
2,0
1,5
1,0
,5
,0
SA-3m
SA-6m
NSs
GROUP
C/ASU PRE
C/ASU PST

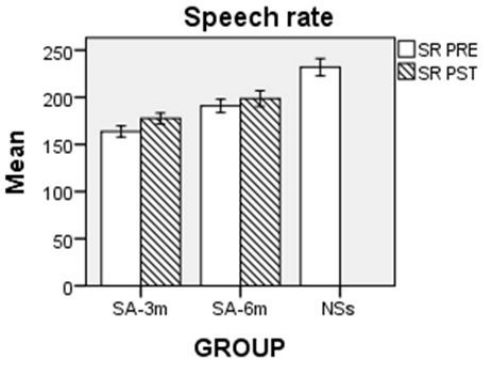

Subordinate clauses per AS-unit
Mean
1,4
1,2
1,0
,8
,6
SA-3m
SA-6m
NSs
GROUP
SUBC/ASU PRE
SUBC/ASU PST

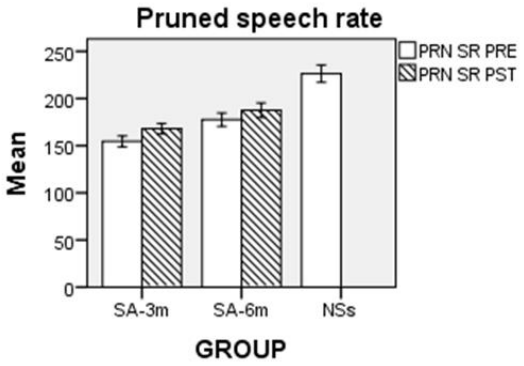

Coordinated clauses per AS-unit
Mean
,50
,40
,30
,20
,10
,00
SA-3m
SA-6m
NSs
GROUP
COORDC/ASU PRE
COORDC/ASU PST


Guiraud's Index
Mean
1,0
0,8
0,6
0,4
0,2
0,0
SA-3m
SA-6m
NSs
GROUP
GI PRE
GI PST


Errors per AS-unit
Mean
,50
,40
,30
,20
,10
,00
SA-3m
SA-6m
NSs
GROUP
E/ASU PRE
E/ASU PST


Errors per clause
Mean
,25
,20
,15
,10
,05
,00
SA-3m
SA-6m
NSs
GROUP
E/C PRE
E/C PST


Speech rate
Mean
250
200
150
100
50
0
SA-3m
SA-6m
NSs
GROUP
SR PRE
SR PST



In order to begin to provide an answer to RQ1, a series of mixed between-within ANOVAs was performed in order to explore the impact of SA and any differential effects of programme duration. Two learner groups were arranged by LoS, categorized into short-stay (SA-3m) and long-stay (SA-6m) participants. CAF scores were submitted to the ANOVA with *Time* and *Group* (each with two levels) as independent factors. Significant main effects of *Time* were seen through measures of fluency, with quite large effect sizes, and nearly reached significance in the lexical complexity measure, GuirIndex. *Group* effects were found to be significant in the fluency measures as well, also with large effect sizes, yet participants went into SA with already considerable differences in fluency level. These differences in initial level are examined further in later sections. We did not find any significant *Time x Group* interactions.

6.2.1. Main effects of SA

We first present the main effects of *Time*, i.e. the SA experiences overall. One of the most notable points to take in here and throughout the report of the results is the generalised lack of change in learner production across testing times in the complexity and accuracy measures analysed. As can be seen in Table 6.3, the analysis of 4 structural complexity measures (WD/ASU, C/ASU, SUBC/ASU and CoordC/ASU), one measures of lexical complexity (GuirIndex), and two measures of accuracy (E/ASU and E/C) did not indicate any significant changes in learner production over the SA period. On the other hand, fluency measures of SR, and Pruned SR did change significantly as a result of the SA experience. Fluency scores also consistently show changes over time, and toward NS levels throughout this report of our results.

Table 6.3 Summary table main effects for *Time* (SA)

Mixed between-	Partial eta

within ANOVAs	*F*	*df*	*p*	squared
Complexity				
WD/ASU†	.017	1, 45	.898	.000
C/ASU	.264	1, 45	.610	.006
SUBC/ASU†	338	1, 45	.564	.007
CoordC/ASU††	.155	1, 37	.680	.003
GuirIndex	3.821	1, 45	.057	.078
Accuracy				
E/ASU†	2.315	1, 45	.135	.049
E/C†	.938	1, 45	.338	.020
Fluency				
SR	5.522	1, 44	.023*	.112
Pruned SR	6.201	1, 44	.017*	.124

*Significant *p* values at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT), and variables marked with (††) were transformed using Log10. Shaded values indicate significance

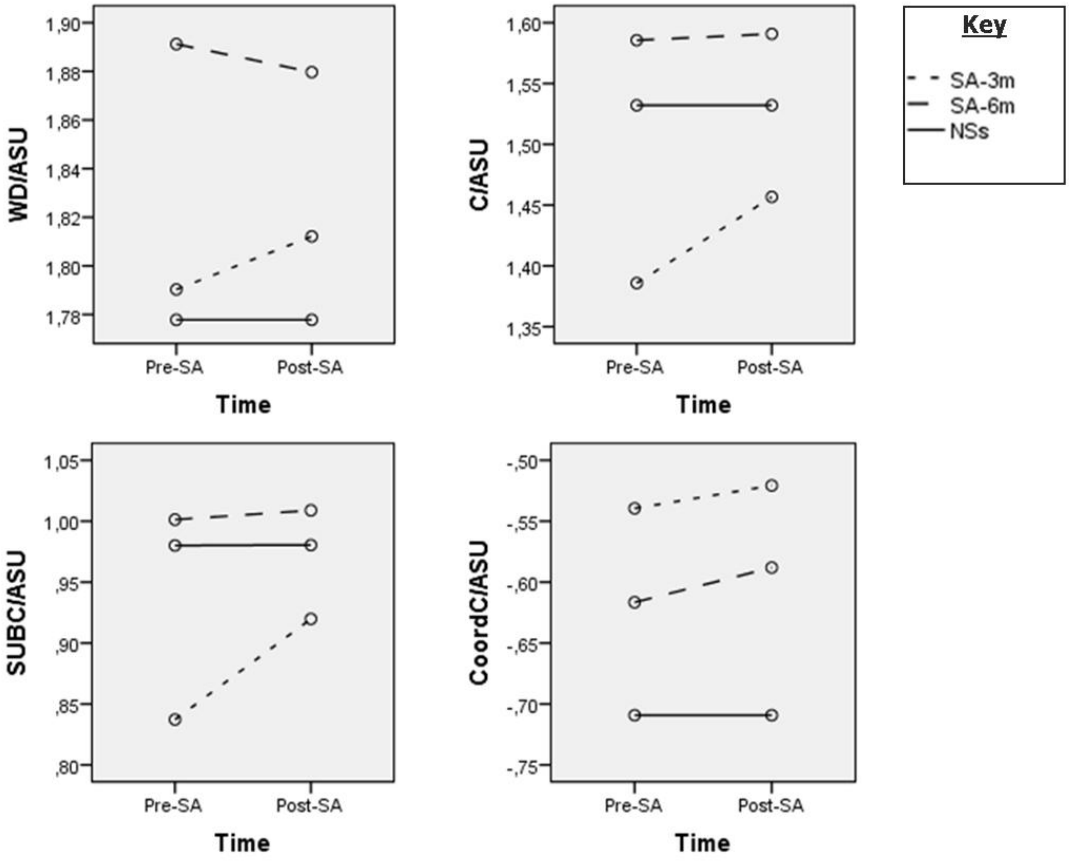
We then went on to examine partial eta squared values in order to determine the effect size for SA as seen through the two variables that had changed significantly over time. Fluency measures showed a partial eta squared for SR of .112 and for Pruned SR partial eta squared came to .124. These values imply a rather large effect size for SA. The remainder of the complexity and accuracy variables were not significantly impacted as a result of the SA experience.

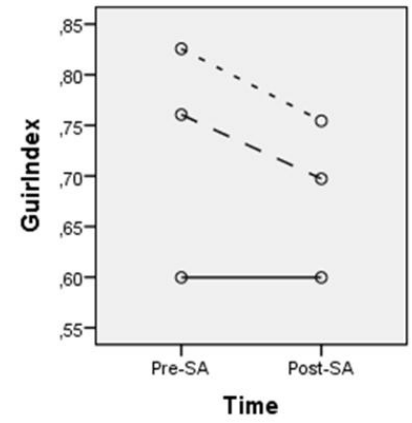
Changes in GuirIndex fell short of reaching significance at the 0.05 confidence level (p=.057). Nonetheless, it is interesting to note that the changes in GuirIndex over time represent negative gains, that is, learner production became less complex in terms of lexical diversity from pre- to post-test. At first glance, this change seems to be counterintuitive as we would expect learner language to become more lexically complex over time, and presumably, more diverse, if indeed progress in the L2 is being made. This change would be difficult to interpret had we not also included NS data as a baseline reference. With the inclusion of baseline data, we see that while learners' lexical complexity does indeed decrease during SA, this represents a shift toward NS scores as can be seen in Figure 6.2. Fluency, as measured through SR and Pruned SR reflects mean increases over time during SA. Like the results for GuirIndex, these changes in fluency scores represent a movement toward NS rates, this time reaching statistical significance. With these results we can begin to provide an answer to our first sub-question (RQ1a):

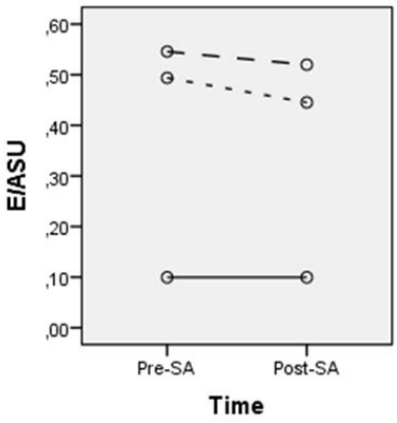
RQ1a. In what measure(s) do NNS performances change significantly across testing times?

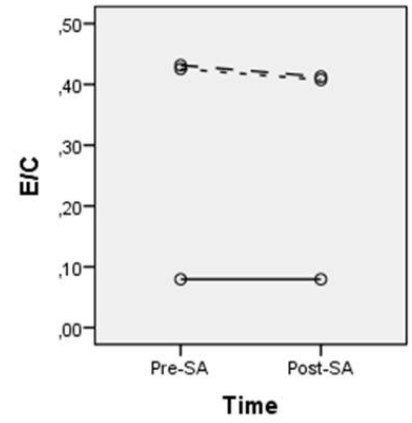
We see a rather sharp drop in learners' GuirIndex scores from pre-test to post-test as they more closely resemble NS values. This indicates that learners used less-rich language at post-test in that it was less diverse, yet this usage more closely reflects native-like behaviour for this task. Similarly, fluency appears to increase as a result of SA as SR and Pruned SR scores increase by post-test, also moving toward more native-like values. Figure 6.2 illustrates changes in learners' CAF scores from pre- to post-test, with NS values represented by a horizontal like as a point of reference.

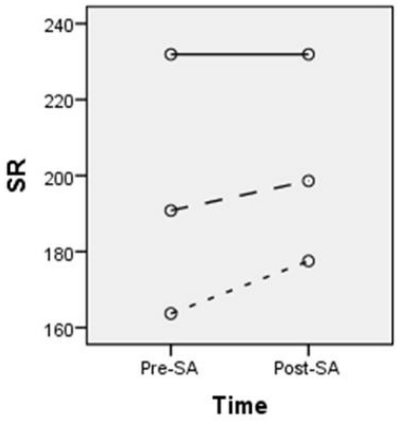
Figure 6.2 Mean pre- and post-test CAF scores with learners arranged by LoS and NSs. Plots represent transformed values.

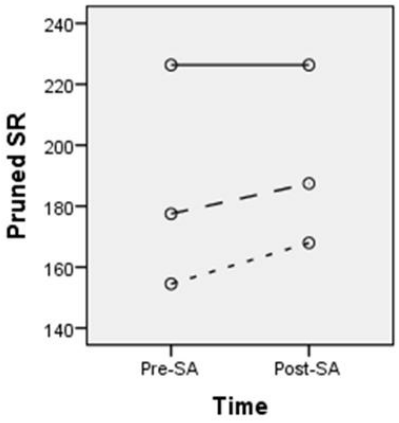












At this point in the analysis, we can provide a preliminary answer to RQ1a, in that we see that fluency measures change significantly from pre- to post test, and moved toward NS baseline rates. GuirIndex, while falling short of statistical significance, also changed notably over time, and approached baseline data values. These changes were presumably a result of the SA periods these learners experienced. We did not see any changes in learners' syntactic complexity and accuracy scores from pre- to post-test.

6.2.2. Between Subjects effects: LoS

We examined the between-subjects effects so as to better understand the role of LoS in learners' performance upon return from SA, and thus addressing the second sub-question from RQ1, that is RQ1b:

> *RQ1b In what measure(s) do SA-3m and SA-6m participants' production differ across testing times?*

Again, these analyses include the two learner groups whose SA experiences differed in LoS. ANOVAs, where the between-subjects factor *Group* represented learners arranged by SA duration, revealing significant group differences reflected in the two fluency measures: SR and Pruned SR. None of the mean complexity and accuracy scores indicated significant between-groups effects. Table 6.4 contains a summary of the results including the effect size through partial eta squared. For both significant between-groups effects outcomes (SR and Pruned SR) the effect size, as seen through partial eta squared, is large following Cohen (1988).

Table 6.4 Summary table between-subjects effects: *Group* (LoS)

Mixed between-within ANOVAs	*F*	*df*	*p*	Partial eta squared
Complexity				
WD/ASU†	2.529	1, 45	.119	.053
C/ASU	3.608	1, 45	.064	.074
SUBC/ASU†	3.595	1, 45	.064	.074
CoordC/ASU††	1.089	1, 37	.303	.029
GuirIndex	2.901	1, 45	.095	.061
Accuracy				
E/ASU†	.953	1, 45	.334	.021
E/C†	.044	1, 45	.835	.001
Fluency				
SR	7.127	1, 44	.011*	.139
PRUNED SR	5.872	1, 44	.020*	.118

*significant at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT). Variables marked with (††) were transformed using Log10. Shaded values indicate significance

Having determined the main effect of *Time* and *Group*, we then looked to see if any significant interactions had occurred. As can be seen in Table 6.5, no interactions reached significance.

Table 6.5. Interaction effects for *Time x Group*

CAF scores	*F*	*df*	*p*	Partial Eta Sq
WD/ASU†	.173	1, 45	.679	.000
C/ASU	.196	1, 45	.660	.004
SUBC/ASU†	.234	1, 45	.631	.005
CoordC/ASU††	.007	1, 37	.936	.000
GuirIndex	.013	1, 45	.909	.000
E/ASU†	.332	1, 45	567	.007
E/C†	.003	1, 45	.958	.000
SR	.431	1, 44	.515	.010
SR Pruned	.997	1, 44	.710	.003

*Significant *p* values at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT) and variables marked with (††) were transformed using Log10.

Given these main effects results and finding no significant interactions, we can now provide an initial response to RQ1b. We have found that the only measures included here in which learner groups performed differently to a significant degree were fluency measures SR and Pruned SR.

Here we conclude our report of the mixed between-within ANOVAs examining the role of SA and LoS. In the following sections, we look to other analyses to further shed light on these outcomes.

6.2.3. Further SA and LoS results

Significant main effects of factors *Time* and *Group* resulting from the ANOVAs were only seen through measures of fluency (SR and Pruned SR). Furthermore, plots in Figure 6.2, appear to indicate a number of notable discrepancies between SA-3m and SA-6m onset levels in CAF variables (see especially WD/ASU, C/ASU, SUBC/ASU, CoordC/AS, SR and Pruned SR in Figure 6.2). These pre-test group differences could potentially be the source of the significant between-subjects effects seen in the ANOVA results. Therefore, In order to test for any further effects of *Time* and *Group*, not seen when these factors were analysed together, and to determine whether or not learner groups already varied significantly at pre-test, paired-samples t-tests were conducted so as to test for any further effects of *Time*, while between-subjects differences were tested through independent-samples comparing mean pre- and post-test CAF scores.

First, we examined how SA had impacted each learner group separately, without the influence of the between-groups factor that had been in place for the two-by-two ANOVAs. In order to test each LoS independently of the other, paired-samples t-tests were run for SA-3m and SA-6m groups comparing CAF scores at pre- and post-test in each case. Results for these tests can be seen in Table 6.6

Table 6.6 Paired-Samples t-tests comparing CAF scores of learner groups arranged by LoS

	SA-3m				SA-6m			
Pre-/post-test CAF scores	*t*	*df*	*p*	Eta sq.	*t*	*df*	*p*	Eta sq.
WD/ASU†	-.538	32	.594	--	.148	13	.885	--
C/ASU	-.911	32	.369	--	-.039	13	.970	--
SUBC/ASU†	-.915	32	.367	--	-.072	13	.944	--
CoordC/ASU††	-.271	24	.789	--	-.277	13	.786	--
GuirIndex	1.841	32	.075	--	1.190	13	.255	--
E/ASU†	2.135	32	.041*	.125	.466	13	.649	--
E/C†	.962	32	.343	--	.513	13	.616	--
SR	-2.565	31	.015*	.171	-1.219	13	.244	--
PRUNED SR	-2.463	31	.020*	.159	-1.479	13	.163	--

*significant at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT), and variables marked with (††) were transformed using Log10. Shaded values indicate significance

The results of the paired-samples t-tests revealed significant changes over a three-month SA period for SA-3m subject in three measures, E/ASU, SR and Pruned SR. We calculated eta squared[39] to determine the effect sizes of the three-month SA period. The effect size of the three-month stay in these variables was quite large in E/ASU, and in the fluency variables. SA-6m, in contrast did not register significant differences between pre- and post-test scores in this test.

Independent-samples t-tests were performed where LoS acted as the grouping variable (SA-3m versus SA-6m) and pre-test and post-test scores served as the dependent variable. Tables 6.7 and 6.8 summarize the results.

Table 6.7 Independent-Samples t-tests comparing mean pre-test CAF scores of learner groups arranged by LoS

[39] Eta squared was calculated using the following formula $eta^2=t^2 /t^2 + N-1$

Pre-test CAF scores	*t*	*df*	*p*	Eta sq.
WD/ASU†	-1.661	45	.104	--
C/ASU	-1.936	45	.059	--
SUBC/ASU†	-2.201	45	.033*	.089
CoordC/ASU††	.939	42	.353	--
GuirIndex	1.249.	45	.218	--
E/ASU†	-.488	45	.628	--
E/C†	.192	45	.849	--
SR	-2.718	44	.009*	.141
SR PRUNED	-2.326	44	.025*	.107

*significant at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT), and variables marked with (††) were transformed using Log10. Shaded values indicate significance

As can be seen in Table 6.6, we found that SA-3m and SA-6m did differ significantly at pre-test in SUBC/ASU, SR and Pruned SR measures. These results point to SA-6m as the learner group with a significantly higher level prior to going on SA, as compared with their SA-3m peers. In SUBC/ASU, SA-6m (*M*=1.001; *SD*=.286) outperformed SA-3m (*M*=.837; *SD*=.209). Similarly, in fluency scores overall, SA-6m began SA with higher initial levels of fluency in SR (*M*=190.822; *SD*=26.157) and Pruned SR (*M*=177.509; *SD*=26.310) than did SA-3m (*M*=163.697; *SD*=33.006) and (*M*=154.506; *SD*=32.586). The mean differences between groups in these scores were *M*=.164 SUBC/ASU, *M*=27.124 syllables per minute, and *M*=23.004 pruned syllables per minute. The effect size in each case is between moderate and large.

We then went on to perform a second round of independent-samples t-tests, this time swapping pre-test scores for post-test scores in the comparisons. In so doing, we test whether or not these differences in level also held true at post-test. In those variables that had revealed significant between-groups differences at pre-test (SUBC/ASU, SR, Pruned SR), we found that only group differences in SR remained marginally significant at post-test. Table 6.8 contains a results summary.

Table 6.8 Independent-Samples t-tests comparing mean post-test CAF scores of learner groups arranged by LoS

Post-test CAF scores	*t*	*df*	*p*
WD/ASU†	-.942	45	.351
C/ASU	-1.067	45	.292
SUBC/ASU†	-.718	45	.477
CoordC/ASU††	1.126	42	.266
GuirIndex	1.209	45	.233

E/ASU[†]	-1.766	45	.084
E/C[†]	.064	44	.949
SR	-2.050	44	.046*
SR PRUNED	-1.949	44	.058

*significant at the 0.05 level. Note: measures marked with ([†]) were subjected to variable transformation prior to analysis (SQRT), and variables marked with ([††]) were transformed using Log10. Shaded values indicate significance

The results from the t-tests where learner groups' mean post-test scores were compared indicated that SA-3m caught up to SA-6m during SA in their use of subordination as well as in Pruned SR. Only SR remained significant, although barely reaching significance (p=.046) and with a reduced mean difference between groups (M=21.087 syllables per minute). Furthermore, the effect size for *Group* dropped in this measure, from .141 at pre-test to .078 upon return from SA. Therefore, we see that group differences that were seen at pre-test did not hold true at post-test. In Table 6.7, CAF scores that had reached significance at pre-test are boxed in, showing the results for post-test scores. This suggests that, on average, the lower level SA-3m participants were able to improve to the point of SA-6m after only three months abroad.

These results suggest that taken separately, SA-3m and SA-6m's production changed over their respective SA periods in ways that were not seen through the series of ANOVAs performed earlier. We see that SA-3m changed significantly over the three months in terms of E/ASU; a decrease in error rate with a mean difference of M=.058 E/ASU (SD=.157) between testing times. Similarly, SR scores had a mean difference of M=13.828 (SD=30.50) and Pruned SR scores M=13.442 (SD=30.87) representing a significant increase in syllables produced per minute (full count and pruned). Paired-samples t-tests for SA-6m, on the other hand, did not register any significant changes over their SA period (See Table 6.8).

Given these results we can now provide a more definitive answer to RQ1a and b. When learner groups were analysed together, we found that SA was significantly impactful on learner production only in fluency measures of SR and Pruned SR. When tested separately, however, SA-3m and SA-6m were seen to progress differently over their respective SA periods. While SA-3m registered significant changes in E/ASU and fluency scores from pre- to post-test, SA-6m's scores did not change significantly during SA. This apparent lack of change on the part of SA-6m could be due to the

already high proficiency level of this group relative to SA-3m. LoS did seem to result in significant group differences in SR and Pruned SR, but it is likely that these significant distinctions were due to the already notable gap in fluency levels between these groups at pre-test.

6.2.4. NNS-NS comparisons

In an effort to provide a response to RQ1c: "Do changes in NNS performances from pre- to post-test represent a movement toward NS performances?" we set up the following analyses. We compared learner groups to NS subjects who provided baseline data by maintaining our NNS groups that differed in LoS and comparing mean pre-test scores from each NNS group with mean NS scores. These comparisons were made separately through independent-samples t-tests. Results can be seen in Table 6.9

Table 6.9 Independent-Samples t-tests comparing the pre-test CAF scores of learner groups arranged by LoS and NS groups

	SA-3m / NS comparison				SA-6m / NS comparison			
Pre-test CAF scores	*t*	*df*	*p*	Eta sq.	*t*	*df*	*p*	Eta sq.
WD/ASU†	.187	55	.853	--	1.230	36	.227	--
C/ASU	-1.140	55	.263	--	.305	36	.762	--
SUBC/ASU†	-1.485	55	.148	--	.162	36	.872	--
CoordC/ASU††	2.314	48	.025*	.089	.887	34	.381	--
GuirIndex	5.010	55	.000*	.313	2.856	36	.007*	.184
E/ASU†	8.742	55	.000*	.582	7.987	36	.000*	.639
E/C†	9.496	55	.000*	.621	8.694	36	.000*	.677
SR	-6.557	54	.000*	.439	-3.118	36	.009*	.213
PRUNED SR	-6.904	54	.000*	.464	-3.673	36	.025*	.273

*significant at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT), and variables marked with (††) were transformed using Log10. Shaded values indicate significance

Results for the SA-3m/NS comparison resulted in 6 variables where NNSs and NSs differed significantly, while NNSs' mean syntactic complexity scores for WD/ASU. C/ASU and SUBC/ASU did not diverge significantly from NS values. CoordC/ASU, GuirIndex, the two accuracy measures and the two fluency measures did indicate significant NNS/NS differences. The mean differences between groups in those variables that revealed a significant divergence between the SA-3m and NS were: in CoordC/ASU *M*=.17592 (SE=.0760), GuirIndex *M*=.2259 (SE=.0451), E/ASU *M*=.4150 (SE=.0475), E/C *M*=.3616 (SE=.0381), SR *M*=68.2186 (SE=10.8662), and Pruned SR

M=71.8418 (SE=10.900). The largest effect sizes were seen through the accuracy measures. The rest of the effect sizes for membership in NNS/NS groups can be seen through eta squared in Table 6.9

Next we pitted SA-6m performances at pre-test against NS scores. Just as in the SA-3m/NS comparison, WD/ASU, C/ASU and SUBC/ASU did not distinguish NNS and NS groups. To these we add CoordC/ASU to the list of variables that were not significantly different for NNSs from SA-6m and NSs. We did see significant mean differences between SA-6m and NS groups in GuirIndex, E/ASU, E/C, SR and Pruned SR, all measures that also distinguished SA-3m from NSs (See Table 6.9). The mean group differences in those variables that revealed significant differences between the SA-6m and NS were in GuirIndex *M*=.1609 (SE=.0551), E/ASU *M*=.4463 (SE=.0617), E/C *M*=.3521 (SE=.0429), SR *M*=41.0945 (SE=11.5282) and Pruned SR *M*=48.8379 (SE=11.6221). The effect size for subjects' condition as NNSs or NSs is reported in Table 6.9 for those variables that indicated significant differences between groups. Again, we see that the accuracy measures showed the largest effect sizes, as was the case for the SA-3m/NS comparison.

We have already referred to in Figure 6.2, in pointing to the change and directionality of learner performances relative to NS data. In order to test whether learner outcomes approached NS scores significantly at post-test, we again performed paired-samples t-tests, this time with the post-test scores as the dependent variables. Results can be seen in Table 6.10.

Table 6.10 Independent-Samples t-tests comparing post-test CAF scores of learner groups arranged by LoS and NS groups

	SA-3m / NS comparison				SA-6m / NS comparison			
Post-test CAF scores	*t*	*df*	*p*	Eta sq.	*t*	*df*	*p*	Eta sq.
WD/ASU†	.492	55	.625	--	1.150	36	.258	--
C/ASU	-.575	55	.567	--	.344	36	.733	--
SUBC/ASU†	-.525	55	.602	--	.219	36	.828	--
CoordC/ASU††	2.744	50	.008*	.120	1.266	34	.214	--
GuirIndex	3.606	55	.001*	.191	1.787	36	.082	.081
E/ASU†	9.423	54	.000*	.618	8.967	36	.000*	.691
E/C†	10.568	54	.000*	.670	9.585	36	.000*	.718
SR	-5.265	54	.000*	.335	-2.445	36	.020*	.142

PRUNED SR	-5.650	54	.000*	.367	-2.876	36	.007*	.187

*significant at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT). Variables marked with (††) were transformed using Log10. Shaded values indicate significance

We see that upon arrival from SA, SA-3m scores as compared with NS scores are unchanged in terms of group differences having reached significance. WD/ASU, C/ASU and SUBC/ASU measures, as at pre-test did not distinguish between SA-3m and NS groups. Likewise, CoordC/ASU, GuirIndex, accuracy and fluency measures continued to show significant differences between groups.

If we look to the columns corresponding to the SA-6m/NS comparison however, we see that GuirIndex no longer distinguishes between the two groups (*mean difference* =.0975, SE=.0546). This outcome would suggest that SA-6m had approached NS values at post-test to the degree that they are no longer statistically distinguishable from NS performances in this measure. In Tables 6.9 and 6.10 we have boxed in the rows corresponding to GuirIndex to more easily compare results. Given this new evidence, we are further inclined to interpret the drop in GuirIndex that we saw in Figure 6.2 as a significant change toward NS norms for this task and at least in the case of SA-6m. No other post-test scores for SA-6m showed any further growth toward NS values in terms of group differences no longer reaching significance at post-test

Therefore, we can answer RQ1c, in that we have seen that a number of CAF variables seem to indicate a trend toward NS norms in both learner groups (e.g. GuirIndex, E/ASU, SR, Pruned SR). However, only GuirIndex can be confirmed to no longer differ significantly between SA-6m and NS groups at post-test when differences between these groups had reached significance prior to SA.

So far we have studied learner production in terms of CAF scores from the two testing times. We have found some differences, but these are closely tied to initial level at pre-test. To further look into how learner groups differed over the course of SA, independent of initial level, we analyse learner gains in the following section.

6.2.5. Analysis of learner gains

Mixed between-within ANOVAs were unable to detect any significant group differences in complexity and accuracy scores between learner groups whose SA

experiences differed in LoS by three months. Fluency measures did discriminate between learner groups; however, it was also confirmed through independent-samples t-tests of initial CAF scores that this distinction was already in place at pre-test.

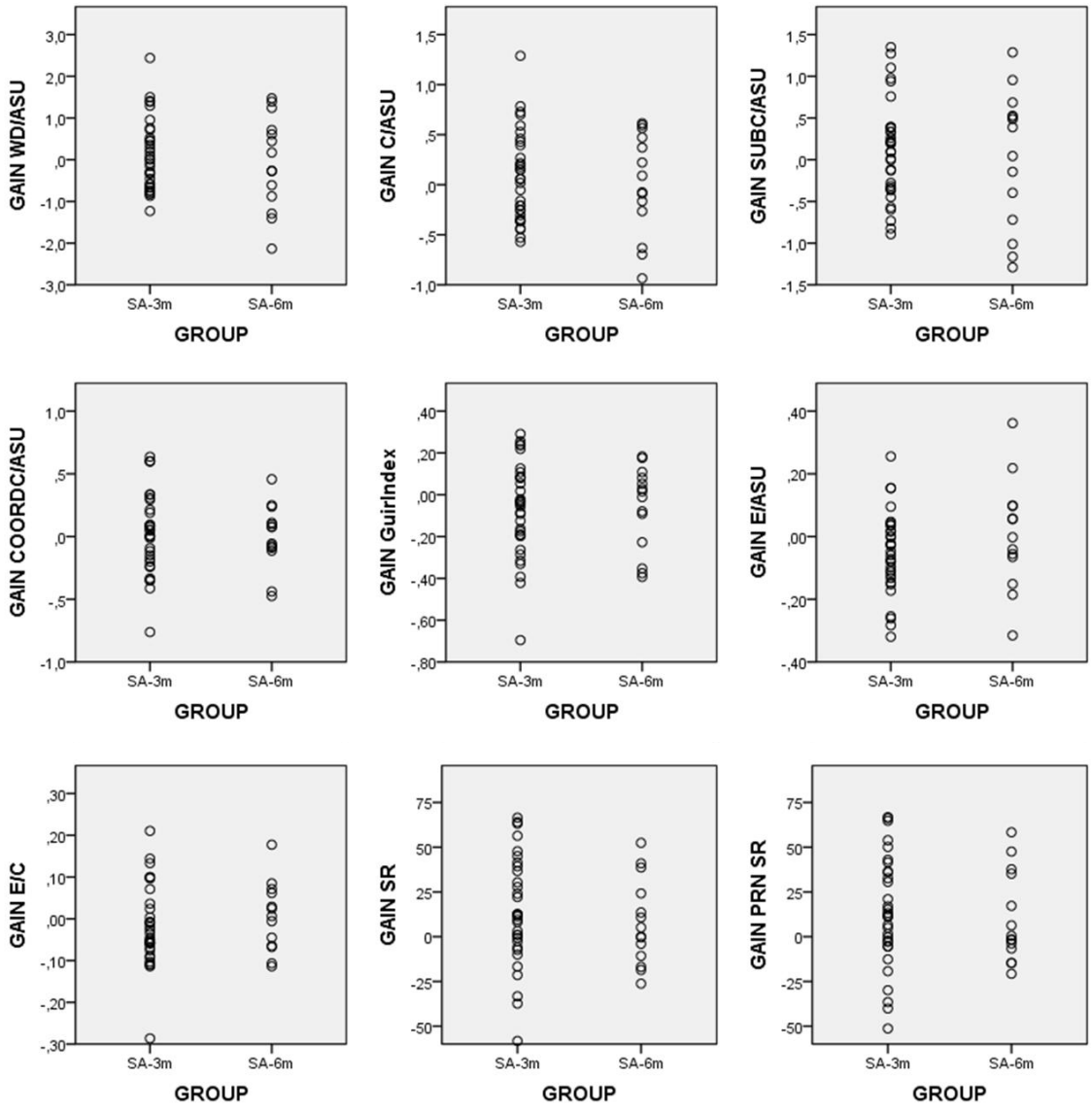
In order to test for group difference in terms of how much learners gained over the SA period, regardless of initial level, we calculated CAF gains upon return from SA by subtracting pre-test scores from post-test scores for all dependent variables. This left us with the gains experienced during the SA period. See Table 6.11 for a report of means and standard deviations for CAF gains at post-test and for each of the two learner groups. Gains for all CAF variables were found to be in normal distribution according to Kolmogorov-Smirnov tests of normality.

Table 6.11 Mean CAF gains for SA-3m (N=33) and SA-6m (N=14) groups (*SD* in parentheses)

CAF Gains	Group	Mean (*SD*)
GAIN WD/ASU	SA-3m	.104 (.85)
	SA-6m	-.058 (1.12)
GAIN C/ASU	SA-3m	.071 (.45)
	SA-6m	005 (.51)
GAIN SUB C/ASU	SA-3m	.108 (.59)
	SA-6m	.011 (.82)
GAIN COORDC/ASU	SA-3m	.074 (.39)
	SA-6m	-.001 (.25)
GAIN GuirIndex	SA-3m	-.071 (.22)
	SA-6m	-.063 (.20)
GAIN E/ASU	SA-3m	.074 (.17)
	SA-6m	-.049 (.27)
GAIN E/C	SA-3m	-.025 (.10)
	SA-6m	-.26 (.07)
GAIN SR	SA-3m	13.828 (30.50)
	SA-6m	7.791 (23.91)
GAIN PRN SR	SA-3m	13.442 (30.87)
	SA-6m	9.931 (25.13)

RQ1d references learner gains: “do learner groups differ significantly in terms of overall gains?” We addressed this remaining RQ1 sub-question through another round of analyses, this time, using learner gains as the dependent variable for the same learner groups (SA-3m and SA-6m). Individual gains in CAF scores are illustrated in Figure 6.3

Figure 6.3 Scatter plots of individual CAF gains scores, learners arranged by LoS.



6.2.5.1. Group differences in terms of gains

The gains we calculated were then submitted to independent-samples t-tests to determine to what extent groups SA-3m and SA-6m gained differently. T-test results for complexity and accuracy, which can be consulted in Table 6.12., indicated no significant differences between groups in terms of gains made during SA, corroborating the lack of group differences between SA-3m and SA-6m participants that ANOVAs had already revealed through mean scores related to these dimensions.

Table 6.12 Independent-Samples t-tests comparing post-SA gains for learner groups arranged by LoS

Post-test gains	*t*(45)	*p*
WD/ASU	.546	.588

C/ASU	.443	.660
SUBC/ASU	.456	.651
CoordC/ASU	.665	.510
GuirIndex	-.114	.909
E/ASU	-.390	.698
E/C	.016	.987
SR	.656	.515
SR PRUNED	.374	.710

*Significant *p* values at the 0.05 level. Shaded values indicate significance.

With the results of these independent-samples t-tests, we conclude our statistical analysis of LoS, finding no remarkable group differences between SA-3m and SA-6m in terms of gains. Given that we did not find significant differences in fluency gains between learners we can further confirm that significant group differences were already present before SA.

6.2.5.2. Analysis of individual data.

Given the generalised lack of change over the different SA periods, we sought to wrap up our look at SA and LoS by examining individual data in terms of those participants who experienced gains as a result of SA, and those who did not. In so doing, we hope to obtain a better picture of what the proportion of students who benefited from the experience in terms of linguistic performance really was.

In a spreadsheet we grouped learner data by whether or not their performance in each CAF measure at post-test reflected gains or not. 'Gainers', were those who improved CAF scores to some degree during SA, while 'non-gainers' were those who had not gained, or whose scored at post-test indicated a movement away from target L2 use. In determining what changes in scores represented gains, we took into account that 'improvement' in learners' production may be reflected in positive gains in some measures (e.g. SR), while at other time, negative gains may be indicators of more target-like language (e.g. E/ASU). We also considered the effects of this particular task on NS performance. Knowing how NSs handled the task in terms of the linguistic resources they applied in the interaction, we could better gauge whether or not changes in NNSs' performances at post-test actually indicated a more target-like use of the L2 on this task or not. Measures such as GuirIndex, where more diverse language would typically indicate greater proficiency, was shown earlier to decrease considerably by post-test in

learner production; yet it was also demonstrated that NSs performed this role-play task with levels of GuirIndex well below those of the NNS participants. Therefore, in this measure we suggest that negative gains in GuirIndex, actually indicate change toward target-like use in terms of lexical diversity.

In order to report on the proportion of gainers/non-gainers that resulted from the two modalities of SA, we examined the individual CAF gains scores for each participant, grouped them into gainer and non-gainer categories, and did a manual count of those performances that indicated gains, calculating a percentage of 'gainers' for each CAF measure and by LoS group. After obtaining the results for each LoS group, we then did the same for learners as a whole, calculating the percentage of participants who had experienced gains during SA regardless of their LoS.

Table 6.13 Percentage of 'Gainers' as measured through gains in CAF variables with learners arranged by LoS.

CAF Gains	SA-3m (% Gainers)	SA-6m (% Gainers)	All NNSs (% Gainers)
WD/ASU	51.52	42.86	48.94
C/ASU	51.52	57.14	51.06
SUBC/ASU	54.55	57.45	57.45
CoordC/ASU	45.45	42.86	44.68
GuirIndex	66.67	42.86	59.57
E/ASU	66.67	57.14	63.83
E/C	69.70	50.00	62.22
SR	63.64	64.29	63.83
PRUNED SR	60.61	57.14	59.57

As can be seen in Table 6.13, well over half of SA-3m participants benefited from their SA experiences as seen through these linguistic measures, with the exception of CoordC/ASU, in which only 45.5% experienced gains during SA. Where lexical diversity, accuracy and fluency are concerned, gainers represented well over half of SA-3m participants. In fact, in these measures 60% or more of participants benefited from their stay abroad. The most notable percentage of gainers for this group was seen in E/C where nearly 7 out of 10 students were able to reduce the frequency with which they made linguistic errors at the clausal level.

In the case of SA-6m, gainers outnumbered non-gainers in 5 of the 9 CAF variables (C/ASU, SUBC/ASU, E/ASU, SR, and Pruned SR). SR was the measure in which the

greatest number of participants benefited, while in the measures WD/ASU, CoordC/ASU, GuirIndex and E/C, 50% or less of SA-6m participants registered gains.

When grouped together, we found that the proportion of learners who gained in syntactic complexity measures as a result of SA hovered around the 50% mark, suggesting that those who improved their L2 proficiency in these measures were more or less on par with those who did not in terms of the number of individual subjects. On the other hand, in lexical diversity, accuracy and fluency measures, the students who improved during SA outnumbered those who did not. In these last measures 62.8% of learner participants improved to some degree during SA. Although not fully reflected in the statistical analysis, we see that the majority of SA participants did in fact benefit linguistically from their time abroad, even if this improvement did not generally reach statistical significance.

In order to determine whether or not there was a significant relationship between LoS group membership and learners' post-test status as gainers or nongainers, chi-square tests were run where *Group* (LoS) and *Gainer (yes/no)* in each CAF measure, served as the two variables to be tested. Our results by the variables in which we found participants to be (non)gainers, are summarized in Table 6.14.

Table 6.14 Chi-square tests of learner groups arranged by LoS and (non)gainers

	$\chi^2(1, N = 47)$	*p*
WD/ASU	.009	.924
C/ASU	.082	.775
SUBC/ASU	.001	.978
CoordC/ASU	.228	.633
GuirIndex	.759	.384
E/ASU	.386	.534
E/C	1.652	.199
SR	.578	.447
SR PRUNED	.998	.318

*Significant *p* values at the 0.05 level. Shaded values indicate significance.

In spite of the percentage differences we saw between groups in the previous analysis, chi-square results were not significant, meaning that the proportion of SA-3m learners who gained during SA was not significantly different from the proportion of SA-6m learners who gained.

Our results so far, taken as a whole, have shown that SA significantly impacted fluency as measured through SR and Pruned SR measures, and LoS was not seen to be impactful in terms of post-SA outcomes. The lack of impact of LoS was true for both CAF scores over the SA period as well as net gains accrued. Nonetheless, we saw that the majority of SA participants did benefit from their time in the TL country. In some measures (e.g. error ratios), nearly 70% of learners experienced some degree of gain as a result of SA, while in syntactic complexity measures there were approximately the same percentage of gainers as non-gainers.

6.3. Addressing RQ2: The impact of initial level

Initial level has been widely discussed in SA related literature as it has been seen to impact outcomes in a number of studies that we have already discussed in Part I (e.g. Brecht et al. 1995; Lapkin et al. 1995; Golonka 2006; Davidson 2010; Valls-Ferrer & Mora 2014). Given the range of studies that point to subjects' initial level as a robust predictor of differential SA outcomes, we examined this factor through several approaches in order to determine whether or not previous findings also hold true for the SA situations studied here. The approaches we adopted so as to examine the role of initial level were based in the set of questions contained in RQ2 which appeared in Chapter 4 as follows:

Research question 2 (RQ2)

Is linguistic development in the L2 as measured through CAF different for learners with different initial levels of proficiency?

RQ 2a. To what extent is initial proficiency level as measured through pre-test CAF scores related to post-test outcomes in terms of post-test CAF scores and gains that occurred during the observation period?

RQ2b. To what extent is initial proficiency level, as assessed through SALA lexico-grammatical testing instruments, related to post-test gains during SA?

RQ 2c. Do low initial level participants obtain significantly greater gains than high initial level participants during SA?

We first tackled sub-question RQ2a by examining the relationship between initial level and post-test outcomes across the CAF variables. We did this first by running correlations between learners' CAF scores recorded at pre- and post-test (Section 6.4.1.1). We then ran correlations between pre-test CAF scores (representing initial level) and gains obtained during SA (See section 6.3 for an explanation of how gains were calculated and interpreted as more or less target-like). These results are reported in Section 6.4.1.2.

We then went on to examine to what extent initial level in lexico-grammatical knowledge was related to gains (RQ2b). We did this by running *Pearson's* correlations to determine any significant relationships between pre-SA GRAMSCOR and gains during SA. We report these results in Section 6.4.2.

The final step we took in addressing RQ2 consisted in determining whether or not participants obtained significantly different outcomes during SA in terms of CAF score and gains (RQ2c). To perform these analyses we used median splits to group learners into high and low initial level groups based on their pre-test CAF scores. High/low groupings resulted for each CAF variable (See Table 6.17). Wilcoxon signed rank tests were run for CAF scores in order to test for significant effects of *Time*. Furthermore, CAF gains were submitted to independent-samples t-tests in order to determine whether or not high/low level groups gained differently to a significant degree. In our analysis of gains, we also include groups arranged by high/low grammar level that were grouped using K-means cluster analysis. Results are presented in Section 6.4.3.

6.3.1. The relationship between initial level and post-test outcomes

In order to address to what extent initial level was related to post-test outcomes in terms of CAF scores and gains that occurred during SA we performed two sets of *Pearson's* Correlations: (1) between learners' pre-test and post-test CAF scores and (2) between learners pre-test scores and gains.

6.3.1.1. Correlations between pre-test and post-test CAF scores

When NNS scores were collapsed into a single group, normality tests revealed nonnormally distributed mean scores in all the same variables that had been found to be

in nonnormal distribution in previous analyses. Therefore, all the variables used here have been transformed using the procedure already discussed in section 6.1.2.

Pearson's correlations were performed for pre- and post-test scores, thus allowing us to determine the direction and strength of the relationship between initial level and SA outcomes. Table 6.15 provides a summary of the results for *Pearson r* correlations.

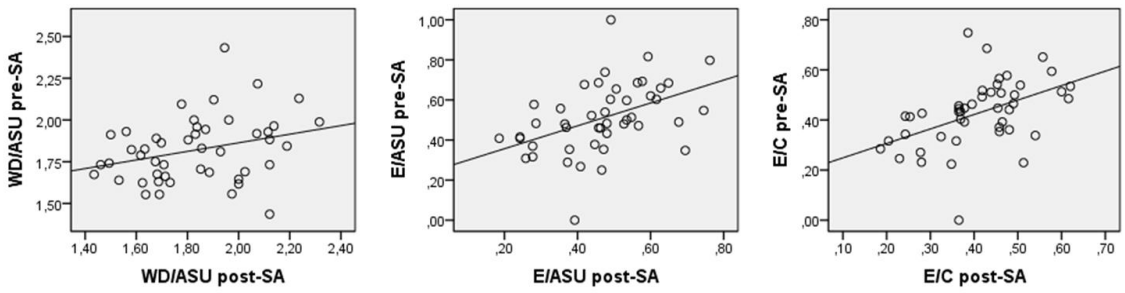
Table 6.15 *Pearson* Correlations between learners' pre- and post-test scores (N=47)

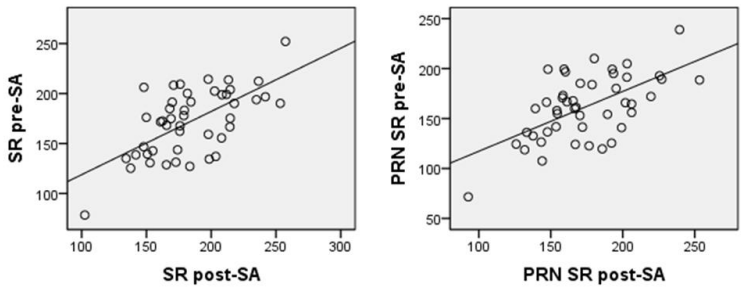
	r	*p*
Complexity		
WD/ASU†	.297*	.042
C/ASU	.207	.163
SUBC/ASU†	-.127	.395
CoordC/ASU††	.154	.349
GuirIndex	.072	.629
Accuracy		
E/ASU†	.537**	.000
E/C†	.593**	.000
Fluency		
SR	.632**	.000
PRUNED SR	.596**	.000

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).
(†) variable subjected to variable transformation (SQRT).
(††) variable subjected to variable transformation (Log10)
Shaded values indicate significance

Results showed significant correlations for one complexity variable (WD/ASU), the error ratios (E/ASU and E/C) and the fluency measures (SR and Pruned SR). All of these were positive correlations. Scatterplots in Figure 6.4 illustrate these relationships.

Figure 6.4 Scatterplots of learners' pre- and post-test scores in WD/ASU, E/ASU, E/C, SR and Pruned SR.





As can be seen in Table 6.15, a general complexity measure of WD/ASU at post-test correlated with pre-test scores. *Pearson's r*=.297 reflects a medium correlation and a moderate relationship between pre- and post-test scores, Thus, students who produced more words per units at pre-test were more likely to continue producing more words per unit at post-test. Similarly, post-SA error ratios (i.e. E/ASU and E/C) were significantly, positively correlated with pre-SA scores. E/ASU at pre- and post-test, and E/C at pre- and post test had large correlations of *r*=.537 and *r*=.593, respectively, suggesting a strong relationship between pre- and post-test scores following Cohen (1988). Therefore those who made frequent errors before SA were also more likely overall to make errors at post-test. Fluency scores at pre-test were also correlated positively with those at post-test. SR and Pruned SR at post-test correlated with pre-test scores, indicating a strong relationship (*r*=.632. and *r*=.596, respectively). Therefore, fluency levels prior to SA are closely related to fluency outcomes upon return from SA in that, high scorers at pre-test remained high scorers at post-test overall.

6.3.1.2. Correlations between initial level and gains

Where we can best see the impact of initial level upon SA outcomes is in correlating initial CAF scores with the amount of gains obtained during SA. Where correlations between pre- and post-test scores resulted in only five significantly correlated pairs, if we examine the correlations between initial CAF scores and gains from pre- to post test, we see highly significant, very strong correlations across CAF variables (See table 6.16). Given that all these relationships were negative correlations, we see that subjects who had relatively low scores at pre-test gained most, while those who already scored on the higher end of the spectrum at pre-test did not experience considerable gains during SA.

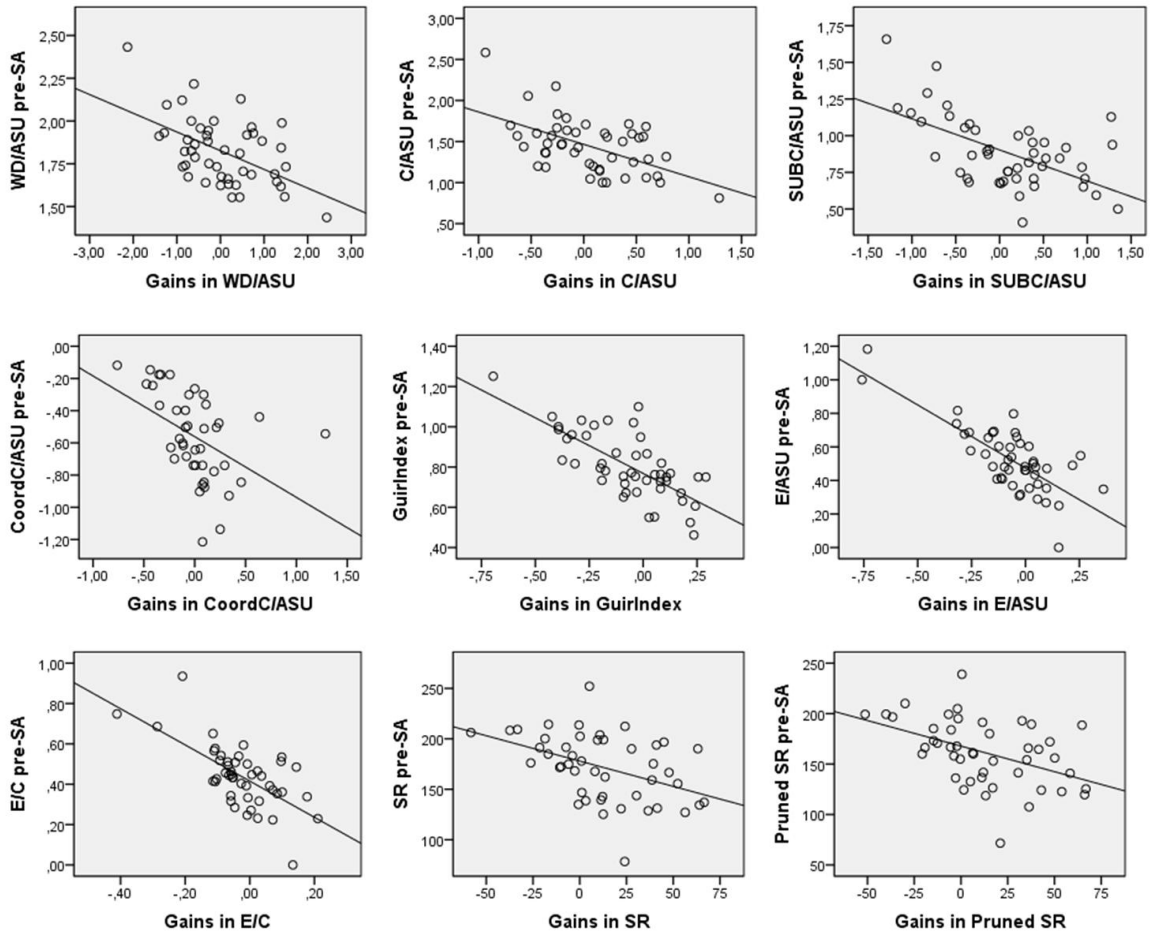
Table 6.16 *Pearson* Correlations between learners' pre-test scores and gains (N=47)

CAF Gains	*r*	*p*
Complexity		
WD/ASU†	-.523**	.000
C/ASU	-.556**	.000
SUBC/ASU†	-573**	.000
CoordC/ASU††	-.463**	.002
GuirIndex	-.718**	.000
Accuracy		
E/ASU†	-.764**	.000
E/C†	-.652**	.000
Fluency		
SR	-.430**	.003
PRUNED SR	-.455**	.002

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).
(†) variable subjected to variable transformation (SQRT).
(††) variable subjected to variable transformation (Log10)
Shaded values indicate significance

Scatterplots of pre-test scores and gains in Figure 6.5 illustrate the statistically significant correlations mentioned above.

Figure 6.5 Scatterplots of pre-SA scores and post-SA gains



6.3.2. Initial lexico-grammatical knowledge and gains

Along with pre-SA scores, our study incorporated an additional measure of initial level, that is, initial lexico-grammatical knowledge as measured through a composite score on a sentence rephrasing task and a cloze test. *Pearson's* correlations were run to determine any significant relationships between pre-SA GRAMSCOR and gains during SA, thus further exploring the role of initial level mentioned in RQ2. Table 6.17 summarizes our findings.

Table 6.17 *Pearson* Correlations between learners' pre-test Grammar scores (GRAMSCOR) and gains (N=47)

	r	*p*
Complexity		
WD/ASU	-.043	.776
C/ASU	-.050	.740
SUBC/ASU	-.073	.627
CoordC/ASU	-.067	.653
GuirIndex	-.143	.339
Accuracy		
E/ASU	.224	.130
E/C	.184	.216
Fluency		
SR	-.181	.229
PRUNED SR	-171	.257

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).
Shaded values indicate significance

In this instance we found no significant relationship between pre-programme GRAMSCOR and gains occurring over the observation period. It should be noted that as in the significant correlations between pre-test CAF scores and gains, the results of the GRAMSCOR/gains pairing also indicated negative correlations for the pairs made up of complexity and fluency variables. However, these did not even approach significance at the 0.05 level. This could indicate a very loose relationship between grammar level at pre-test and gains across testing times where subjects with lower levels at pre-test gained more during SA.

6.3.3. High/low initial level groups and SA outcomes

Here we address the third sub-question for this research question: RQ2c which read:

RQ 2c. Do low initial level participants obtain significantly greater gains than high initial level participants during SA?

In SPSS, we created high and low level learner groups for each CAF variable by performing a median split of pre-test CAF scores. Taking NS baseline data as a point of reference for performance on this task, it can be noted that higher scores in certain CAF variables indicate more native like behaviour (e.g. increased SR) and on the contrary, lower scores in other variables we have seen, may better represent movement toward NS usage (e.g. GuirIndex). See figure of mean high/low group CAF scores with NS scores for comparison in Appendix D.1. Taking this directionality into account, we assigned high/low level labels to the two new groups.

Descriptive statistics revealed that nonnormal distributions occurred in all high/low initial level learner groupings. Many of these nonnormal distributions, when plotted as histograms, were seen to be skewed in opposite directions, for example, low-level learner fluency means were positively skewed, as these learners tended to score at the lower end of the scale. High-level learners tended to score at the higher end of the scale producing negatively skewed distributions. This pattern was particularly evident in measures of accuracy and fluency, but was also seen in GuirIndex, and CoordC/ASU. These distributions, skewed in opposite directions, made variable transformation difficult. The transformation of mean scores in one group would often alter the normality of the distribution of scores in the other. Therefore, we chose a non-parametric test to analyse these data. We present median CAF scores for these high/low groups in Table 6.18

Table 6.18 Median CAF scores of high/low initial level groups (*Range* in parentheses).

Measures	N (high/low)	High initial level PRE-SA	High initial level POST -SA	Low initial level PRE-SA	Low initial level POST -SA
Complexity					
WD/ASU	24, 23	3.72 (2.60)	3.47 (3.11)	2.80 (1.22)	2.92 (2.44)
C/ASU	24, 23	1.61 (1.12)	1.20 (.63)	1.60 (1.34)	1.29 (1.34)
SUBC/ASU	24, 23	.34 (.28)	.29 (.56)	.20 (.20)	.26 (.38)
CoordC/ASU	24, 23	.14 (.24)	.25 (.60)	.43 (.51)	.32 (1.57)
GuirIndex	24, 23	.72 (.31)	.76 (.50)	.95 (.48)	.69 (.62)
Accuracy					

E/ASU	23, 23	.17 (.23)	.15 (.48)	.38 (.76)	.29 (.50)
E/C	23, 23	.12 (.19)	.13 (.25)	.26 (.37)	.21 (.25)
Fluency					
SR	23, 23	198.82 (76.87)	202.62 (109.29)	142.70 (96.52)	165.56 (111.99)
PRUNED SR	23, 23	188.59 (74.36)	192.67 (106.28)	137.56 (89.58)	154.56 (113.53)

CAF scores of high and low level learners were submitted to Wilcoxon signed ranks tests in order to determine whether any significant effect of *Time* (the current SA) could be detected. Results are presented in Table 6.19.

Table 6.19 Wilcoxon signed rank tests comparing CAF scores of high/low initial level groups

	High			Low		
CAF scores	*z*	*p*	*r*	*z*	*p*	*r*
WD/ASU	-1.257	.209	--	-.1825	.068	--
C/ASU	-.886	.376	--	-1.977	.048	.41
SUBC/ASU	-1.686	.092	--	-2.737	.006	.57
CoordC/ASU	-2.772	.006	.57	-1.494	.135	--
GuirIndex	-2.314	.021	.57	-3.924	.000	.82
E/ASU†	-.304	.761	--	-2.920	.004	.61
E/C†	-.487	.627	--	-2.220	.026	.46
SR	-.243	.808	--	-3.467	.001	.72
PRUNED SR	-.112	.903	--	3.559	.000	.74

*significant at the 0.05 level.

The Wilcoxon test results showed that there was a significant effect of SA for the low initial level group in 7 of the 9 CAF variables included here. Effect sizes were calculated[40] for those measures that had reached significance. These effect sizes ranged from medium to large. The effect size for SA in the GuirIndex measure was quite large, *r*=.82, and the smallest effect size, that corresponded to C/ASU, was still medium-large, *r*=.41. The high initial level group, on the other hand, demonstrated little change over the SA period. With the exceptions of CoordC/ASU and GuirIndex, no significant differences between pre- and post-test scores were found. Furthermore, the changes in CoordC/ASU, experienced by the high-level group represent increased rates of coordination as a source of complexification in their performances. This actually

[40] Effect size (*r*) was calculated using the following formula: $r=z/\sqrt{N}$

represents a movement away from NS values for this task. The low-level group did reduce their use of coordination by post-test, but this change did not reach significance. GuirIndex was the only measures in which both high and low-level groups changed significantly over the observation period. The high-level group significantly increased their lexical diversity, according to median GuirIndex scores (see Table 6.18); an increase that was also reflected in the mean difference in pre- and post-test scores, *M*=.072 (*SD*=.13). Again, this represents a shift away from nativelike behaviour in this task for the high-level group. The low-level group, on the other hand, did successfully reduce their lexical diversity so as to more closely represent target-like levels in GuirIndex at post-test. On average, low initial level participants reduced their scores in GuirIndex, *M*=-.216 (*SD*=.18), adjusting their language to more closely resemble NS performance.

We next tested for between-groups differences in high/low initial level groups submitting pre- and post-test scores to Mann-Whitney U-tests. Results are summarized in Table 6.20.

Table 6.20 Mann-Whitney U-tests comparing CAF scores of learner groups arranged by high and low initial level groups at pre- and post-test High (N=24), Low (N=23).

	Pre-SA				Post-SA			
CAF scores	*U*	*Z*	*p*	*r*	*U*	*Z*	*p*	*r*
WD/ASU	.000	-5.875	.000*	.86	189.0	-1.852	.064	--
C/ASU	.000	-5.875	.000*	.86	192.0	-1.788	.074	--
SUBC/ASU	.000	-5.876	.000*	.86	246.0	-.639	.523	--
CoordC/ASU	.000	-5.880	.000*	.86	245.0	-.660	.509	--
GuirIndex	.000	-5.874	.000*	.86	237.5	-.819	.413	--
E/ASU	.000	-5.812	.000*	.85	96.0	-3.702	.000*	.54
E/C	.000	-5.812	.000*	.85	95.5	-3.715	.000*	.54
SR	.000	-5.811	.000*	.85	107.0	-3.460	.001*	.50
PRUNED SR	.000	-5.811	.000*	.85	122.0	-3.131	.002*	.46

*significant at the 0.05 level. Shaded values indicate significance

Through these results, we found that at pre-test, high and low initial level groups differed significantly across all CAF variables. Effect sizes were also quite large with *r* values of .86 and .85. By post-test, high and low level groups had converged in complexity scores, including GuirIndex scores, yet remained significantly different in accuracy and fluency variables. The effect size of *Group* in these variables that continued to be significantly different between high/low groups remained large ($r > .50$)

but was down from pre-test values. These results suggest that low-level students were able to improve their performances to the level of high initial level participants, at least in measures of complexity.

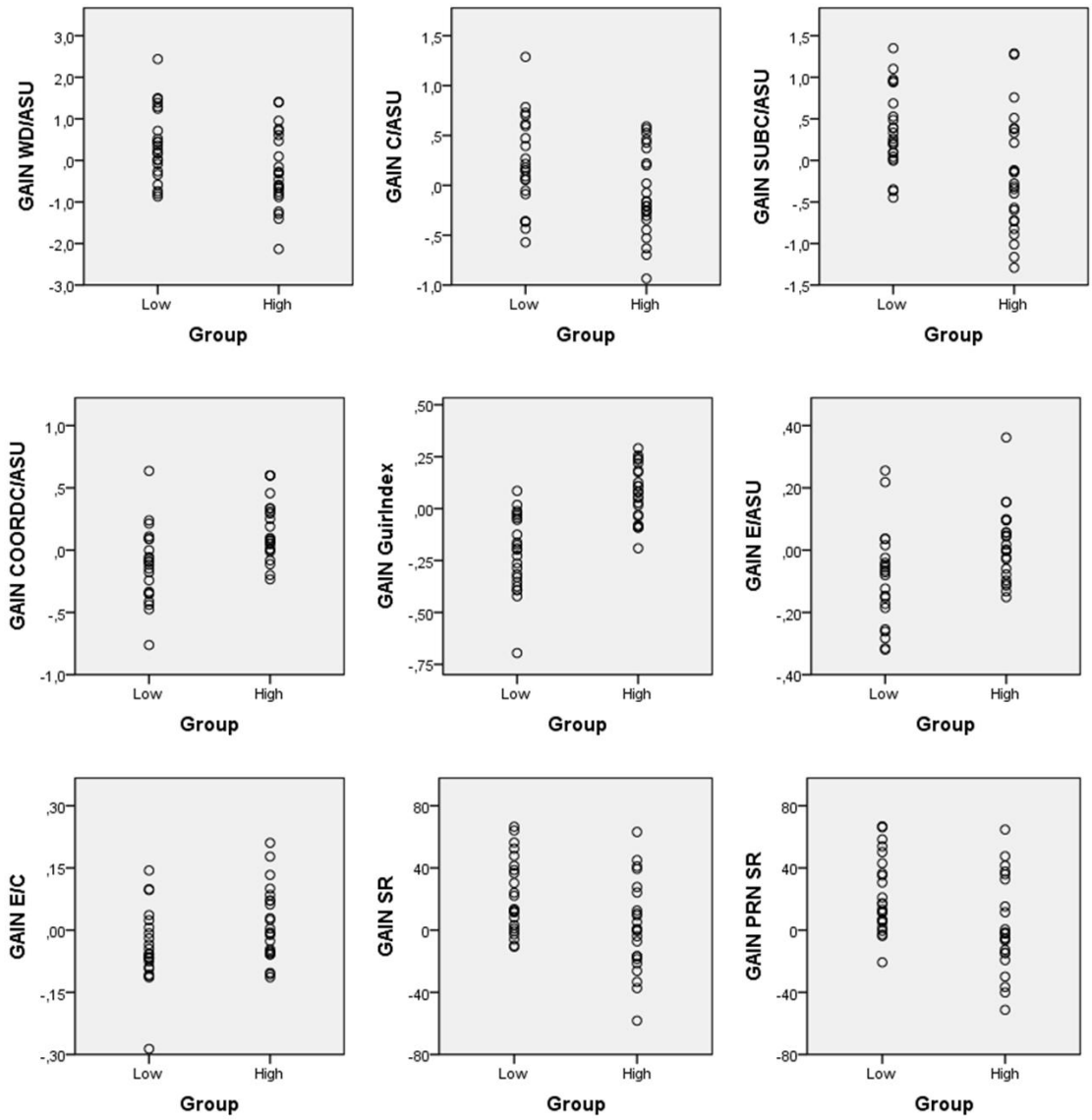
Having analysed CAF scores for high/low level groups, we went on to test mean gains. CAF gains were compared where learners were grouped according to initial level based on their pre-SA CAF scores, just as in the previous analysis. Mean gains and standard deviations for these groups are summarized in Table 6.21.

Table 6.21 Mean gains for high and low initial level groups (*SD* in parentheses)

CAF Gains	Group	N	Mean (*SD*)
GAIN WD/ASU	High	24	-.256 (.90)
	Low	23	.381 (.85)
GAIN C/ASU	High	24	-.087 (.43)
	Low	23	.196 (.46)
GAIN SUB C/ASU	High	24	-.183 (.70)
	Low	23	.352 (.48)
GAIN COORDC/ASU	High	24	.161 (.24)
	Low	23	-.125 (.30)
GAIN GuirIndex	High	24	.072 (.13)
	Low	23	-.216 (.18)
GAIN E/ASU	High	23	.017 (.11)
	Low	23	-.093 (.15)
GAIN E/C	High	23	.014 (.09)
	Low	23	-.041 (.09)
GAIN SR	High	23	1.704 (29.35)
	Low	23	24.276 (24.13)
GAIN PRN SR	High	23	1.970 (29.99)
	Low	23	22.777 (24.42)

Scatterplots of individual gains scores by high/low initial level groups are provided in Figure 6.6 so as to better visualise how learners, arranged by high/low initial level, had progressed upon arrival from their SA periods.

Figure 6.6 Scatter plots of individual CAF gains scores, learners arranged by high/low initial level.



All of these variables were found to be normally distributed across learner groups. Thus, Mean CAF gains were submitted to independent-samples t-tests comparing high and low initial level groups. T-tests revealed significant differences between groups in all CAF variables. Results can be seen in Table 6.21

Table 6.22 Independent-Samples t-tests comparing gains for high/low initial level groups.

CAF Gains	*t*	*df*	*p*	Eta sq.
WD/ASU	2.483	45	.017*	.120
C/ASU	2.192	45	.034*	.096
SUBC/ASU	3.026	45	.004*	.169
CoordC/ASU	-3.587	44	.001*	.222
GuirIndex	-6.255	45	.000*	.465
E/ASU	-2.717	43	.009*	.141
E/C	-2.053	43	.046*	.086
SR	2.597	44	.013*	.130

Pruned SR	2.581	44	.013*	.129

*Significant *p* values at the 0.05 level. Shaded values indicate significance.

Statistically significant results were seen across CAF gains scores, indicating that high and low learner groups gained differently to a significant degree in all 9 CAF measures. These results also serve to confirm that the strong relationship between pre-test level and gains seen through *Pearson* correlations in Section 6.4.1.2 were based on scores that also proved significantly different for students classified as high and low initial level subjects prior to SA. When we calculated eta squared to determine effect size for initial level we found quite large effect sizes ranging from .086 to .465 (see Table 6.22) Overall, we find that subjects who start out with a lower initial level tend to have a greater propensity toward gain than their high-level peers.

We also tested how knowledge of grammar and lexis might impact post-SA gains. To do so, we grouped learners by their score on lexico-grammatical tests at pre-test (GRAMSCOR). See Table 6.23 for a summary of these groups. These high/low groups were created using K-means cluster analysis of the GRAMSCOR variable in SPSS, this gave us two new groups to which we assigned high initial grammar level and low initial grammar level labels. As this variable is based on a combined score on two exam-like tests, determining which groups were in the high or low category was very straightforward. Those who scored higher numerically were assigned to the high-level category and the low scorers to the low-level category.

Normality tests revealed a number of nonnormal distributions for these variables when grouped by high/low GRAMSCOR, therefore two variables were submitted to transformation (Gain WD/ASU and Gain C/ASU). Two additional variables were pruned of extreme values (E/ASU and E/C) and upon doing so, all variables became normally distributed. Mean CAF gains scores and standard deviations are reported in Table 6.23.

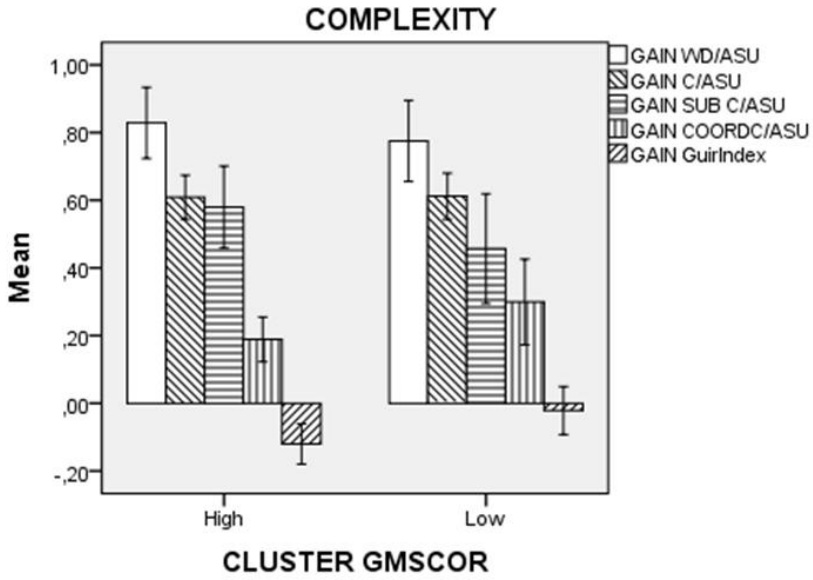
Table 6.23 Mean gains for high and low initial GRAMSCOR groups (*SD* in parentheses)

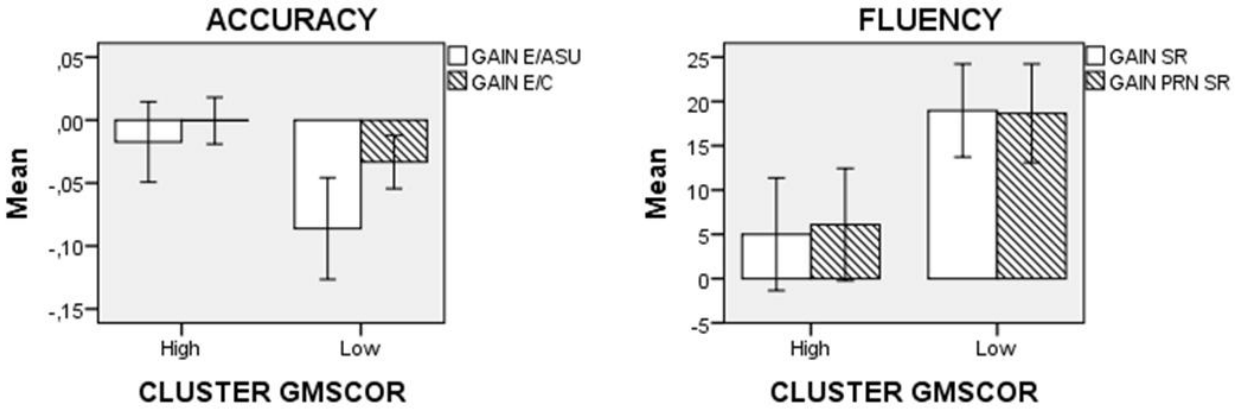
CAF Gains	Group	N	Mean (*SD*)
GAIN WD/ASU†	High	13	.829 (.38)
	Low	11	.775 (.40)
GAIN C/ASU†	High	14	.575 (.26)

GAIN SUB C/ASU	Low	11	.612 (.23)
	High	23	.076 (.75)
	Low	24	.081 (.57)
GAIN COORDC/ASU	High	23	.060 (.28)
	Low	24	.043 (.41)
GAIN GuirIndex	High	23	-.075 (.19)
	Low	24	-.063 (.24)
GAIN E/ASU	High	23	-.018 (.15)
	Low	23	-.086 (.19)
GAIN E/C	High	23	-.001 (.09)
	Low	23	-.033 (.10)
GAIN SR	High	23	5.006 (30.49)
	Low	23	18.974 (25.16)
GAIN PRN SR	High	23	6.088 (30.39)
	Low	23	18.660 (26.75)

Bar graphs representing mean gains are provided in Figure 6.7.

Figure 6.7 Bar graphs of mean gains for high and low initial GRAMSCOR level groups. Error bars: +/- 1 SE





Accuracy and fluency bar graphs seem to indicate a certain degree of between-groups differences based on pre-test GRAMSCOR results, however, none of these outcomes reached significance. Indeed, we found no significant differences between high and low scorers on the grammar and cloze tests in terms of post-test CAF gains (See Table 6.24).

Table 6.24 Independent-Samples t-tests comparing post-SA gains for high/low initial GRAMSCOR groups.

CAF Gains	*t*	*df*	*p*
WD/ASU	.336	22	.740
C/ASU	-.374	23	.712
SUBC/ASU	-.025	45	.980
CoordC/ASU	.157	45	.876
GuirIndex	-.201	45	.842
E/ASU	1.334	44	.189
E/C	1.164	44	.251
SR	-1.695	44	.097
Pruned SR	-1.489	44	.144

*Significant *p* values at the 0.05 level. Shaded values indicate significance.

Given the nonsignificant results for these tests, we cannot claim that initial level according to the GRAMSCOR measure predicts any considerable difference in post-SA gains.

6.4. Addressing RQ3: the effect of previous periods abroad

Drawing from questionnaire information, we included the following research question assessing the impact of previous SA experiences on the outcome of the SA period studied here. We also looked into how previous experience in an English speaking country, or abroad experiences where English was the means of communication,

impacted the SA in question. To guide the last round of analyses we formulated RQ3 and corresponding sub-questions:

To what extent do previous periods abroad in the target language country (previous SA and previous abroad experiences in general) affect the outcome of the present SA experience?

RQ3a. Do first time SA participants and those who have studied abroad before benefit differently from the current SA?
RQ3b. Does other previous experience abroad impact the outcomes of the current SA?
RQ3c. Do previous SA and previous experience abroad impact current SA outcomes in terms of gains from pre- to post-test?

First, it is important to remember what we mean by previous SA and previous abroad experience. Thanks to information gathered by profile and SA questionnaires, we know which students were first-time SA participants and which were repeat participants. About half (N=23, 48.9%) of the learner participants had travelled abroad for the purpose of study at some point. leaving the rest of participants who had never gone abroad for language or content study (N=24, 51.1%).

On the other hand, most students had had abroad experiences in English-speaking countries with no formal study involved (N=27, 57.4%). while only 17 (36.2%) had never been abroad in an Anglophone country before. Therefore, for these analyses we grouped students in two ways: first, by previous SA (yes/no), and then by previous experience in English-speaking countries or through the medium of English (yes/no: no= no prior experience/ yes= one or more stays abroad).

In order to find out if first-time SA participants and those who had studied abroad before benefit differently from the current SA (RQ3a), we submitted syntactic complexity and accuracy scores to mixed between-within ANOVAs with *Time* (i.e. the current SA period) as the within-subjects factor and previous participation in an SA programme (yes/no) as the between-subjects factor (See Section 6.5.1). As distributions of mean GuirIndex, SR and Pruned SR scores violated the assumption of normality

underlying ANOVA, median lexical diversity and fluency scores were submitted to non-parametric tests to determine the effects of *Time* and *Group* separately (See Section 6.4.1.3).

We then performed a similar set of analyses examining previous experience abroad as it impacted the current SA (RQ3b). Thus previous experience abroad acted as the between-subjects factor. *Time* continued as the within subjects factor, just as CAF scores served as the dependent variables. Results of these ANOVAs can be seen in Section 6.5.2. Just as in the Previous SA analysis, GuirIndex, SR and Pruned SR scores were left out of the ANOVA analysis due to the fact that these mean scores occurred in nonnormal distributions and transformations did not result in improved normality. Median GuirIndex, SR and Pruned SR scores were submitted to non-parametric tests so as to test for effects of *Time* and *Group* on these variables (See Section 6.5.2.3).

After this initial look at previous periods abroad, we went on the test how these stays impacted the current SA outcomes in terms of gains from pre- to post-test (RQ3c). To do so, we performed independent-samples t-tests with the previous period abroad (SA or previous experience) as the grouping factor, and gains as the dependent variable. These tests and results are covered in Section 6.5.3.

6.4.1. Previous SA and current SA outcomes

First we present the results we obtained when grouping participants by previous SA experiences. See Table 6.25 for a summary of mean scores. NS mean scores are included for purposes of comparison. Bar graphs representing CAF scores for these groups can be seen in Appendix D.2.

Table 6.25 Mean CAF scores of learners grouped by previous SA (yes/no) Yes (N=23), No (N=24) and NS (N=24) (*SD* in parentheses).

	NNSs				
	Prev. SA yes		Prev. SA no		NSs
Oral measures	Pre-test	Post-test	Pre-test	Post-test	
Complexity					
WD/ASU	3.40 (0.86)	3.47 (0.95)	3.30 (0.61)	3.35 (0.72)	3.24 (1.12)
C/ASU	1.49 (0.37)	1.52 (0.44)	1.40 (0.30)	1.47 (0.36)	1.53 (0.58)
SUBC/ASU	0.94(0.58)	0.92 (0.59)	0.75(0.37)	0.93 (0.51)	1.14 (0.94)

CoordC/ASU	0.27 (0.19)	0.42 (0.33)	0.32 (0.23)	0.27(0.17)	0.22 (0.18)
GuirIndex	0.77 (0.18)	0.73 (0.15)	0.84 (0.14)	0.75 (0.03)	0.60 (0.17)
Accuracy					
E/ASU	0.30 (0.28)	0.24(0.15)	0.33 (0.22)	0.25 (0.15)	0.03 (0.46)
E/C	0.19(0.16)	0.19 (0.13)	0.24 (0.14)	0.19 (0.10)	.018 (0.28)
Fluency					
SR	178.44(34.37)	193.59(38.74)	165.46(31.56)	174.30(23.73)	231.92(44.91)
PRUNED SR	167.22(33.49)	182.92(38.09)	155.79(30.81)	164.84(22.23)	226.35(45.33)

Note: means and standard deviations appear here in their original form, although for further statistical analysis certain variables were transformed in order to better meet the assumption of normality and reduce the influence of outliers.

A number of variables were not normally distributed for these groupings of subjects according to Kolmogorov-Smirnov and/or Shapiro-Wilk tests. Transformations were performed on variables: WD/ASU, SUBC/ASU, CoordC/ASU, E/ASU, and E/C. One extreme case was pruned from the error ratios. After these operations, these variables were found to be normally distributed. GuirIndex, SR and Pruned SR were also found to be in nonnormal distribution but several attempts to transform the data did not improve normality. We decided to leave these variables in their original form, leaving them out of the ANOVA analysis and testing effects of *Time* and *Group* separately using non-parametric tests. Upon transformation WD/ASU, SUBC/ASU, CoordC/ASU and the accuracy measures all became normally distributed.

When mean syntactic complexity and accuracy scores of students arranged by whether or not they had studied abroad before were submitted to mixed between-within ANOVAs no significant main effects were found for *Time* or *Group*, nor did we find any significant interactions.

6.4.1.1. Main effects for current SA

The first sub-question in this round of analyses (RQ3a) asked:

> *RQ3a. Do first time SA participants and those who have studied abroad before benefit differently from the current SA?*

In an attempt to provide an answer to this question, syntactic complexity and accuracy scores were submitted to mixed between-within subjects ANOVAs with *Time* (the

current SA) as the within-subjects factor, and Previous SA (yes/no) as the between-subjects factor. A summary of main effects for *Time* can be consulted in Table 6.26.

Table 6.26 Summary table main effects for *Time* where subjects are grouped by previous SA (yes/no).

Mixed between-within ANOVAs	*F*	*df*	*p*	Partial eta squared
Complexity				
WD/ASU†	.106	1, 45	.747	.002
C/ASU	.562	1, 45	.457	.012
SUBC/ASU†	.697	1, 45	.408	.015
CoordC/ASU††	.074	1, 37	.787	.002
Accuracy				
E/ASU†	2.717	1, 44	.106	.058
E/C†	.844	1, 44	.363	.019

*Significant *p* values at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT). Variables marked with (††) were transformed using Log10. Shaded values indicate significance

Main effects for *Time* did not reach significance in these syntactic complexity and accuracy variables. See the nonsignificant results for main effects of *Time* in Table 6.26.

6.4.1.2. Between-subjects effects: *Previous SA*

In grouping participants by previous SA, we did not find significant between-subjects effects in any of the syntactic complexity and accuracy variables submitted to mixed ANOVAs. Results are presented in Table 6.27

Table 6.27 Summary table between-subjects effects: *Group* (Previous SA, yes/no)

Mixed between-within ANOVAs	*F*	*df*	*p*	Partial eta squared
Complexity				
WD/ASU†	.216	1, 45	.645	.005
C/ASU	.656	1, 45	.422	.014
SUBC/ASU†	.055	1, 45	.816	.001
CoordC/ASU††	.037	1, 37	.848	.001
Accuracy				
E/ASU†	1.004	1, 44	.322	.022
E/C†	1.886	1, 44	.177	.041

*significant at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT), and variables marked with (††) were transformed using Log10. Shaded values indicate significance

We also discard any impactful interaction effects in these factors as no significant interaction effects were found for *Time x Group* (previous SA: yes/no). Nonsignificant interactions are shown in Table 6.28

Table 6.28 Interaction effects for *Time x Group*

Oral measure	*F*	*df*	*p*	Partial Eta Sq
WD/ASU†	.003	1, 45	.954	.002
C/ASU	.078	1, 45	.782	.012
SUBC/ASU†	1.086	1, 45	.303	.024
CoordC/ASU††	2.665	1, 37	.111	.067
E/ASU†	.304	1, 44	.584	.007
E/C†	1.126	1, 44	.363	.025

*Significant *p* values at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT) and variables marked with (††) were transformed using Log10.

Given our results, we cannot yet claim any benefit for any one group over the other. Learners grouped as first-timers and repeaters (in SA) did not demonstrate any significant group differences in terms of syntactic complexity and accuracy scores over the two testing times.

6.4.1.3. Non-parametric tests of GuirIndex, SR and Pruned SR variables

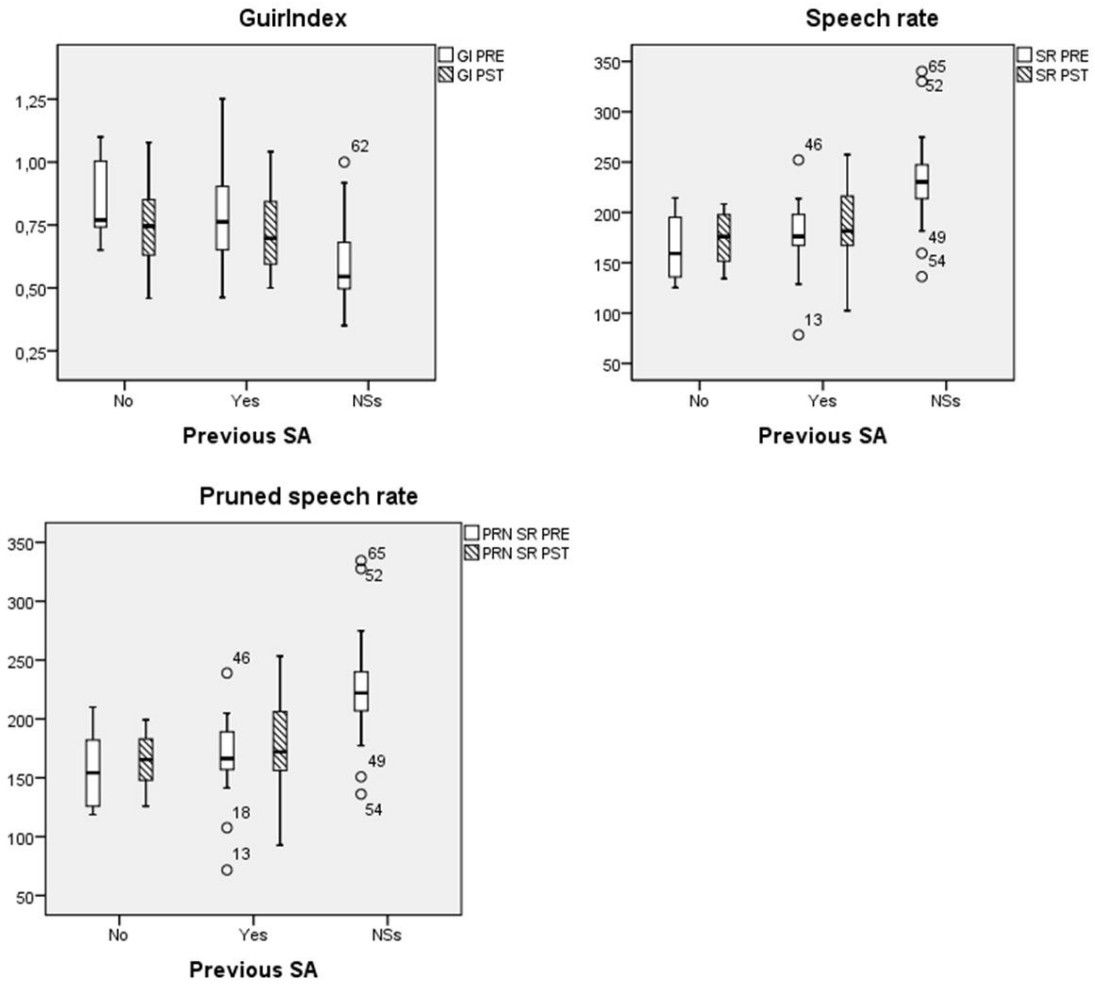
GuirIndex and fluency measures were left out of the ANOVA analyses as variable transformation of nonnormal distributions of GuirIndex, SR and Pruned SR variables did not improve normality. Therefore, lexical diversity and fluency scores were submitted to Wilcoxon Signed Rank tests, in order to determine any significant effects of *Time*, Similarly, pre- and post-test scores in GuirIndex, SR and Pruned SR were submitted separately to Mann-Whitney U-tests of independent samples in order to determine whether or not any significant effects for *Group* could be seen through these measures at the different testing times.

Wilcoxon test results indicated that those who had not studied abroad before significantly reduced GuirIndex scores as a result of the current SA, Z=-2.400, p=.016, r=.49 However, this group did not demonstrate significant changes over the SA period in SR, Z=-1.186, p=.236, and Pruned SR, Z=-1.247, p=.212. Those who had studied abroad prior to the current SA did not show significant changes over time in GuirIndex, Z=-.487, p=.627, but, repeaters did change significantly over time in SR, Z=-2.585,

p=.010, *r*=.54 and in Pruned SR, *Z*=-2.403, *p*=.016, *r*=.50. Effects of *Time* for this second group had rather large effects sizes, where *r* was equal to or greater than .50 in both SR and Pruned SR.

When median pre-test lexical diversity and fluency scores for first-timers and repeaters were submitted to Mann-Whitney U-tests we found that differences in pre-test GuirIndex scores did not reach significance, *U*=208, *Z*=-1.437, *p*=.151; nor were any significant differences found at post-test, *U*=247, *Z*=-.617, *p*=.537. Likewise, differences in SR scores between first-timers and repeaters did not reach significance at pre-test, *U*=204, *Z*=-1.329, *p*=.184, nor were they significant at post-test, *U*=178, *Z*=-1.900, *p*=.057, falling just short of significance. The results for Pruned SR were very similar to those of SR at pre-test, *U*=203, *Z*=-1.351, *p*=.177, and at post-test, *U*=178, *Z*=-1.900, *p*=.057. Boxplots of these variables can be seen in Figure 6.8

Figure 6.8 Boxplots of GuirIndex, SR and Pruned SR scores with subjects arranged by previous SA (yes/no), NSs.



Our results suggests that both first-timers and repeaters progressed in lexical diversity and fluency over the SA period, approaching NS levels, but only first-timers progressed to a significant degree in GuirIndex, and only repeaters progressed to a statistically significant degree in SR and Pruned SR. Furthermore, no significant differences between groups were found in GuirIndex at either testing time, while in the case of the fluency measures, no significant group differences in SR and Pruned SR could be detected at pre-test, but between-groups differences in these measures very nearly reached significance at post-test, suggesting a trend toward further progress on the part of the repeaters group over and above the first-timers.

In summing up our look at the impact of prior SA periods on the current one, we were unable to determine any general benefits to having studied abroad before, with the exception of GuirIndex. This measure did indicate group differences in that, first-time SA participants significantly reduced scores in GuirIndex by post-test while repeaters did not. However, at the same time, there were no between-groups differences revealed by Mann-Whitney tests at pre- or post-test in any of these measures.

6.4.2. Previous experience abroad and current SA outcomes

We next went on to perform a similar set of analyses; this time grouping learners by previous experience abroad that did not involve classroom study. This was done so as to provide an answer to the second sub-question included in this set of analyses:

RQ3b. Does previous experience abroad impact the outcomes of the current SA?

In the rearranging of the learner data we found that first-timers were in the minority (N=17) while repeaters came to 27 subjects. A summary of mean scores by group is available in Table 6.29. Mean NS scores are included as a point of reference. Bar graphs of CAF scores can be seen in Appendix D.3

Table 6.29 Mean complexity, accuracy and fluency scores of NNSs grouped by previous experience abroad (yes/no) Yes (N=27), No (N=17), NS (N=24) (*SD* in parentheses).

	NNSs

Oral measures	Prev. abroad exp. yes		Prev. abroad exp. no		NSs
	Pre-test	Post-test	Pre-test	Post-test	
Complexity					
WD/ASU	3.28(0.84)	3.40 (0.87)	3.41(0.50)	3.37 (0.76)	3.24 (1.12)
C/ASU	1.43(0.39)	1.51 (0.42)	1.47(0.22)	1.48 (0.39)	1.53 (0.58)
SUBC/ASU	0.88(0.60)	0.93 (0.60)	0.80(0.30)	0.91 (0.48)	1.14 (0.94)
CoordC/ASU	0.26 (0.17)	0.35(0.30)	0.34 (0.24)	0.33(0.26)	0.22 (0.18)
GuirIndex	0.77 (0.16)	0.74 (0.17)	0.83 (0.12)	0.75 (0.12	0.60 (0.17)
Accuracy					
E/ASU	0.24 (0.15)	0.24(0.15)	0.36 (0.23)	0.26 (0.10)	0.03 (0.46)
E/C	0.17(0.08)	0.17 (0.09)	0.25 (0.15)	0.20(0.10)	.018 (0.28)
Fluency					
SR	181.31(33.02)	192.45(37.17)	158.40(30.18)	168.79(21.61)	231.92(44.91)
PRUNED SR	169.92(32.10)	181.68(36.38)	150.54(29.75)	160.55(20.10)	226.35(45.33)

Note: means and standard deviations for SUBC/ASU. E/ASU and E/C appear here in their original form, although for further statistical analysis these variables were transformed in order to better meet the assumption of normality and reduce the influence of outliers.

According to Kolmogorov-Smirnov and/or Shapiro-Wilk tests, a number of variables were not normally distributed when subjects were regrouped by whether or not they had been abroad before. Transformations were performed on variables WD/ASU, SUBC/ASU, CoordC/ASU and E/ASU. One extreme case was pruned from the error ratios. Upon performing these operations these distributions became normal. GuirIndex, SR and Pruned SR were also found to be nonnormally distributed and as in the Previous SA arrangement, several different transformations were unable to improve the normality of these distributions. For this reason, we left lexical complexity and fluency measures in their original form and separated them out of the ANOVA analyses. Separate analyses using non-parametric tests were performed on GuirIndex and fluency data in Section 6.5.2.3.

Again, mixed between-within subject ANOVAs were performed where *Time* was the within-subjects factor, representing the current SA period, and previous abroad experience (yes/no) served as the between-subjects factor. Syntactic complexity and accuracy scores served as the dependent variables.

When mean syntactic complexity and accuracy scores of students arranged by previous experience abroad (yes/no) were submitted to mixed between-within ANOVAs no

significant main effects were found for *Time* or *Group*, nor did we find any significant interactions.

6.4.2.1. Main effects for current SA

The main effects for *Time* (i.e. the SA period) were not found to be statistically significant in these syntactic complexity and accuracy measures. The remaining non-significant results can be seen in Table 6.30.

Table 6.30 Summary table main effects of *Time* (Previous abroad experience, yes/no).

Mixed between-within ANOVAs	*F*	*df*	*p*	Partial eta squared
Complexity				
WD/ASU†	.035	1, 42	.852	.001
C/ASU	.411	1, 42	.525	.010
SUBC/ASU†	.1.137	1, 42	.292	.026
CoordC/ASU††	.001	1, 35	.981	.000
Accuracy				
E/ASU†	2.761	1, 42	.104	.062
E/C	2.079	1, 42	.157	.047

*Significant *p* values at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT). Variables marked with (††) were transformed using Log10. Shaded values indicate significance

6.4.2.2. Between-subjects effects: *Previous experience abroad*

The between-subjects effects summarized in Table 6.31 showed no significant results for syntactic complexity and accuracy variables. It should be pointed out that E/C very nearly reached significance with a *p*-value of .052. The remainder of the accuracy and complexity variables did not point to significant group differences for previous experience abroad.

Table 6.31 Summary table of between-subjects effects: Previous experience abroad (yes/no)

Mixed between-within ANOVAs	*F*	*df*	*p*	Partial eta squared
Complexity				
WD/ASU†	.164	1, 42	.688	.004
C/ASU	.004	1, 42	.953	.000
SUBC/ASU†	.114	1, 42	.737	.003
CoordC/ASU††	2.059	1, 35	.160	.056

GuirIndex	.966	1, 42	.331	.022
Accuracy				
E/ASU†	1.414	1, 42	.241	.033
E/C	4.003	1, 42	.052	.087

*significant at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT). Variables marked with (††) were transformed using Log10. Shaded values indicate significance

We did not find any noteworthy interactions for factors *Time x Previous experience abroad*. Nonsignificant interactions are reported in Table 6.32.

Table 6.32 Interaction effects for *Time x Group*

Oral measure	*F*	*df*	*p*	Partial Eta Sq
WD/ASU†	.369	1, 42	.547	.009
C/ASU	.179	1, 42	.674	.004
SUBC/ASU†	.007	1, 42	.935	.000
CoordC/ASU††	1.980	1, 35	.168	.054
GuirIndex†	1.033	1, 42	.315	.024
E/ASU†	1.704	1, 42	.199	.039
E/C	2.375	1, 42	.131	.054
SR	.007	1, 41	.935	.000
SR Pruned	.036	1, 41	.851	.001

*Significant *p* values at the 0.05 level. Note: measures marked with (†) were subjected to variable transformation prior to analysis (SQRT) and variables marked with (††) were transformed using Log10.

In light of the ANOVA results we cannot point to any significant effects for previous experience abroad (first-timers versus repeaters in non-study related periods abroad) in syntactic complexity and accuracy measures. Learners grouped as first-timers and repeaters did not demonstrate any significant group differences in these measures over the observation period.

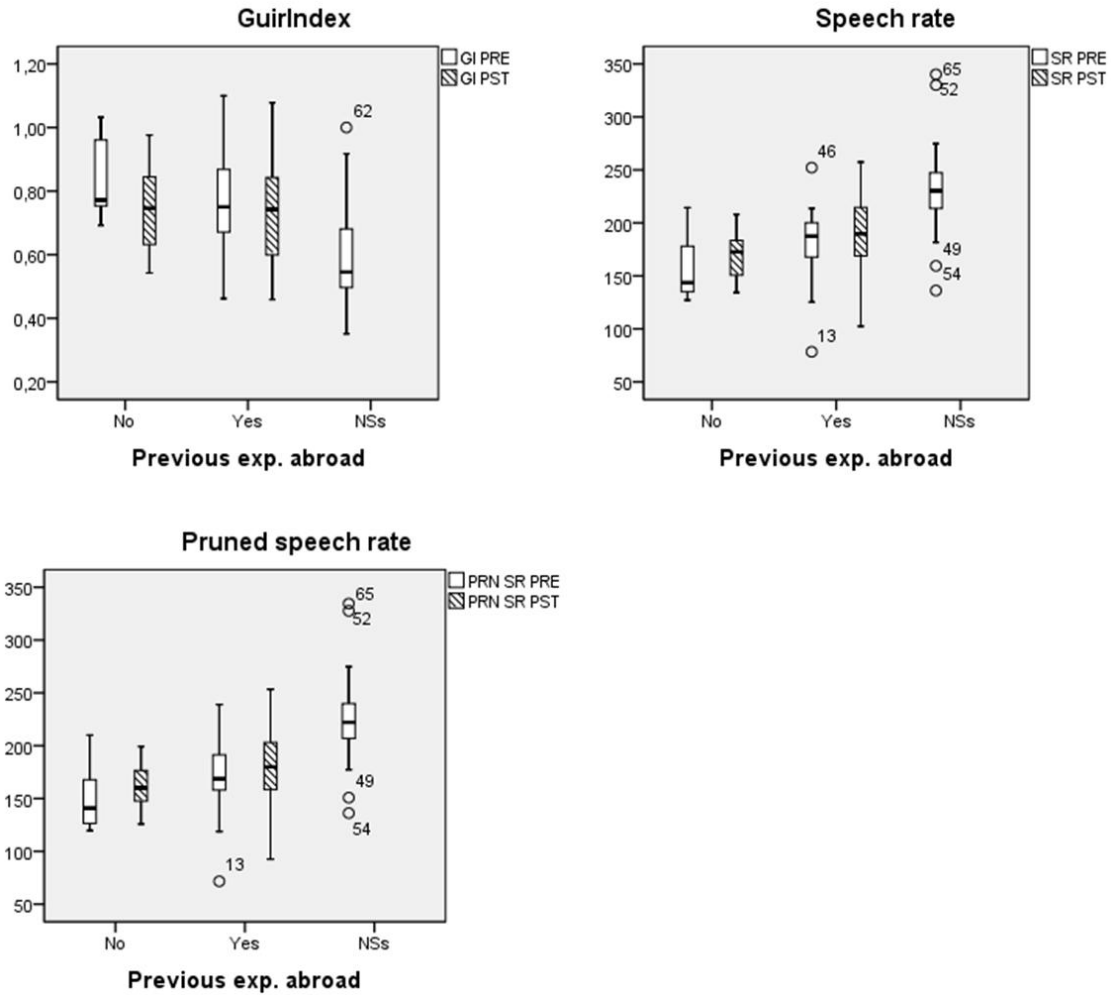
6.4.2.3. Non-parametric tests of GuirIndex, SR and Pruned SR variables

GuirIndex and fluency measures had been left out of the ANOVA analyses as variable transformation of these nonnormal distributions did not achieve acceptable normality. Thus, to see if any significant effects of *Time* had impacted lexical diversity and fluency outcomes, we submitted GuirIndex, SR and pruned SR scores to Wilcoxon Signed Rank tests. To test for significant effects for *Group,* pre- and post-test scores in GuirIndex, SR and Pruned SR were submitted to Mann-Whitney U-tests of independent samples.

Wilcoxon test results indicated that those who had never been abroad to an English-speaking country before significantly reduced GuirIndex scores as a result of the current SA, *Z*=-2.201, *p*=.028, *r*=.53. However, the first-timers group did not demonstrate significant changes over the SA period in SR, *Z*=-1.349, *p*=.177, or Pruned SR, *Z*=-1.254, *p*=.210. Those who had been abroad prior to the current SA did not show any significant changes over time in GuirIndex, *Z*=-.480, *p*=.631, nor did they change significantly over time in SR, *Z*=-1.791, *p*=.073 and in Pruned SR, *Z*=-1.664 *p*=.096. Therefore, first-timers differed from repeaters in that they did demonstrate significant drops in GuirIndex as a result of the SA period, while repeaters did not show any statistically significant changes during SA in GuirIndex, SR or Pruned SR.

When median pre-test lexical diversity and fluency scores for first-timers and repeaters were submitted to Mann-Whitney U-tests we found that between-groups differences in pre-test GuirIndex scores did not reach significance, *U*=168, *Z*=-1.483, *p*=.138; nor were any significant differences found at post-test, *U*=211, *Z*=-.446, *p*=.656. Differences in SR scores between first-timers and repeaters did reach significance at pre-test, *U*=132, *Z*=-2.211, *p*=.027, *r*=.33. Group differences in SR also held true at post-test, *U*=138, *Z*=-2.260, *p*=.024, *r*=.34. The results for Pruned SR were similar to those of SR, showing significant between-groups differences at pre-test, *U*=133, *Z*=-2.186, *p*=.029, *r*=.33., and at post-test, *U*=130, *Z*=-2.161, *p*=.029, *r*=.33. Effect sizes for *Group* were all very similar, with *r* values of around .33, indicating a medium effect size in all three variables tested here. Boxplots of GuirIndex, SR and Pruned SR scores and can be seen in Figure 6.9

Figure 6.9 Mean GuirIndex, SR and Pruned SR scores with subjects arranged by previous experience abroad (yes/no), NSs.



Our results suggest that only first-timers progressed in lexical diversity over the SA period, approaching NS levels, but only repeaters progressed to a significant degree in fluency. No significant differences between groups were found in GuirIndex at either testing time, fluency measures, on the other hand, were found to indicate significant group differences in SR and Pruned SR at both pre- and post-test, suggesting that repeaters already performed with higher levels of fluency at pre-test and maintained their advantage over the first-timers over the SA period examined here.

In summing up our look at the impact of prior abroad experiences on the current SA, we were unable to point to any generalised benefits to having gone abroad before or not; the only exception being that first-timers' performances changed significantly in GuirIndex over the current SA period, moving toward NS levels for this task, while repeaters did not. However, between-groups differences in this measure did not reach significance at pre- or post-test.

As can be seen through the report of mean scores (Table 6.29), participants who had been abroad before had higher mean fluency scores at pre -test than did their first-timer counterparts. Repeaters, performing with a faster SR than did first-timers prior to SA, continued to outperform first-timers to a significant degree at post-test. The non-parametric tests we report on here confirmed significant group differences across testing times. We can plausibly attribute the differences at pre-test to the fact that repeaters had had a considerable amount of experience in English-speaking countries and thus had already benefitted notably in terms of SR. Furthermore, repeaters maintained an advantage over first-timers at post-test, suggesting that fluency may be accumulated from one abroad experience to another, at least for these participants.

6.4.3. Previous periods abroad and gains

To conclude our look at previous periods abroad we sought to provide an answer to the final sub-question in this section. RQ3c which asked:

RQ3c. Do previous SA and previous experience abroad impact current SA outcomes in terms of gains from pre- to post-test?

In order to answer this question we performed independent-samples t-tests where students were grouped as first-timers (N=24) and repeaters (N=23) with respect to SA and then by previous experience abroad (yes: N=27, no: N=17). The CAF gains we calculated in section 6.4 served as dependent variables. (See mean gains by previous periods abroad groups, Table 6.33)

Table 6.33 Mean CAF gains with learners arranged by previous SA and Previous Experience Abroad (*SD* in parentheses).

	Previous SA		Previous exp. abroad	
CAF gains	Yes (N=23)	No (N=24)	Yes (N=27)	No (N=17)
Complexity				
WD/ASU	.067 (.97)	.045 (.91)	.121 (1.02)	-.038 (.85)
C/ASU	.032 (.47)	.007 (.46)	.079 (.50)	.016 (.44)
SUBC/ASU	.108 (.59)	.172 (.64)	.041 (.74)	.113 (.59)
CoordC/ASU	.155 (.37)	-.477 (.31)	.085 (.33)	-.009 (.39)
GuirIndex	-.046 (.25)	-.091(.17)	-.031 (.21)	-.090 (.14)
Accuracy				
E/ASU	-.058 (.20)	-.075 (.21)	-.004 (.14)	-.115 (.21)

E/C	-.004 (.10)	-.046 (.12)	-.002 (.08)	-.050 (.14)
Fluency				
SR	15.145 (24.20)	8.836 (32.53)	11.133 (27.39)	10.388 (32.06)
PRUNED SR	15.698 (24.68)	9.049 (33.01)	11.763 (27.63)	10.006 (32.83)

Three variables were found to be in nonnormal distribution when learners were grouped by previous SA periods (CoordC/ASU, E/ASU, C/ASU). We found that one extreme value in the gains in CoordC/ASU data was altering the normality of the distribution. Upon excluding this value, normality was achieved for this distribution according to Kolmogorov-Smirnov and Shapiro-Wilk tests. Similarly, E/ASU and E/C contained two extreme values each. Once these were pruned from the data, these distributions became normal and no variable transformation was required.

We ran Independent-samples t-tests, where Previous SA (yes/no) served as the grouping variable and CAF gains acted as the dependent variables. Results for previous SA and previous experience abroad can be seen in Tables 6.34 and 6.35.

Table 6.34 Independent-Samples t-tests comparing post-test gains for first time and repeat SA participants

Post-test gains	*t*	*df*	*p*
WD/ASU†	-.079	45	.937
C/ASU†	.279	45	.782
SUBC/ASU	.987	45	.329
CoordC/ASU	-1.724	44	.092
GuirIndex	-.706	45	.485
E/ASU	-.286	43	.776
E/C	-1.318	43	.194
SR	-.746	45	.459
SR PRUNED	-.774	45	.443

*Significant *p* values at the 0.05 level. Shaded values indicate significance.

In this instance we found no significant results reflected in CAF gains. This suggests that first-timers and repeaters gained similarly over the current SA period.

The same tests were performed with students grouped by previous experience abroad. Two variables were not normally distributed according to normality tests (gains in WD/ASU and C/ASU). We tried transforming both variables before proceeding with the analysis, but upon transforming WD/ASU we found that the calculation left us with many missing cases. Therefore, we left the WD/ASU variable as it was and substituted

the independent-samples t-test with its no-parametric alternative. C/ASU was transformed by calculating the square root of the original values. This made the distribution normal. Independent-samples t-tests were performed where previous experience in the TL country (yes/no) acted as the grouping variable and CAF gains were the dependent variables. Again no significant p-values resulted (see Table 6.35).

Table 6.35 Independent-Samples t-tests comparing gains for previous experience abroad (yes/no)

Post-test gains	*t*	*df*	*p*
C/ASU†	-.179	42	.859
SUBC/ASU	.341	42	.735
CoordC/ASU	-.868	41	.390
GuirIndex	-1.017	42	.315
E/ASU	-1.582	41	.121
E/C	-.993	41	.326
SR	-.829	42	.935
SR PRUNED	-.182	42	.851

*Significant *p* values at the 0.05 level. Shaded values indicate significance.

As mean Gains in WD/ASU could not be adequately transformed, we submitted these scores to Mann-Whitney tests in order to determine the effects of *Group*. Results did not indicate any between-groups effect, U=199, Z=-.735, p=.462. Therefore, no significant between-groups effects for Previous experience abroad (yes/no) cold be seen here.

In these tests, just as with the Previous SA (yes/no) arrangement, we found no significant results in terms of gains when comparing groups with no prior abroad experience with those who had travelled to English-speaking countries.

Having completed these final analyses, we can now provide an answer for RQ3c, in that periods abroad prior to the current SA do not seem to impact gains accrued during the current SA. These results suggest that previous stays abroad for this demographic do not add to gain, but rather seem to slow the acquisition of additional gains to a marginal degree. First-timers, on the other hand, were able to reduce GuirIndex in their handling of this task, which more closely resembled NS performances. Repeaters did not obtain the same result.

6.5. Summary of Results.

To summarize our findings we provide a brief overview of what we have reported at length in Chapter 6. Here we summarize our findings following the same order in which the results were presented.

RQ1: ***Which modality of SA (LoS of 3 or 6 months) is more beneficial in the development of L2 speaking performance as measured through CAF?***

Early on in our effort to provide an answer to the question of the impact of SA and LoS, we offered a preliminary answer to RQ1a. When reporting the main effects for *Time* we saw that fluency measures changed significantly from pre- to post test, and moved toward NS baseline rates. GuirIndex, while falling short of statistical significance, also changed notably over time, and approached baseline data values. We attribute these changes to the time learners spent studying abroad in an English-speaking country. We did not see any significant changes in learners' syntactic complexity and accuracy scores from pre- to post-test.

When we initially addressed the role of LoS, we found that the only measures included in the present study in which learner groups performed differently to a significant degree were fluency measures SR and Pruned SR, yet we suspected that these significant results had more to do with divergent initial fluency levels between these two groups than differences related to SA LoS modalities. Thus, we found few changes over time except for speech rate, and no clear differences between learners whose SA experiences varied in length. These were our initial findings when learner groups were analysed together through a series of two-by-two ANOVAs.

We then performed t-test that did not take into account *Time* and *Group* factors together, but rather tested the impact of SA and LoS separately. In doing so we found that SA-3m and SA-6m were seen to progress differently over their respective SA periods. SA-3m registered significant changes in E/ASU, SR and Pruned SR from pre- to post-test, but SA-6m's scores did not change significantly during SA. This apparent lack of change on the part of SA-6m could be due to the fact that SA-6m already performed at a higher level at pre-test than SA-3m who had more 'room to grow' linguistically. Indeed, follow up t-tests performed on learners' pre-test fluency scores indicated that there was a significant difference in level between groups at pre-test.

When we ran the independent-samples t-tests again for post-test scores we found that SA-3m had 'caught up' to SA-6m in that two of three variables that showed significant group differences at pre-test, no longer distinguished between groups. Differences in SR, which continued to be significant at post-test, just barely reached significance, suggesting that SA-3m had made considerable progress in this measure as well.

When testing post-SA gains in the two LoS groups, we found no significant differences between groups. Thus, SA-3m and SA-6m also gained similarly during SA, and could not be distinguished statistically. This finding reinforces the idea that overall, there was a lack of differences between groups differing in LoS during SA.

When including NNS/NS comparisons at several points in these analyses, we saw that a number of CAF variables seemed to indicate a trend toward NS norms in both learner groups (e.g. GuirIndex, E/ASU, SR, Pruned SR). However, only SA-6m's GuirIndex scores progressed to the point at which their performances were no longer statistically distinguishable from those of NSs at post-test when this measure had been significantly different for these two groups at pre-test.

Finally, we examined individual data in order to know the percentage of students who experienced gains toward target-like use of these L2 elements and linguistic forms. We found that in both LoS groups, gainers outnumbered non-gainers on the whole. SA-3m data contained only one variable, CoordC/ASU, in which gainers did not outnumber non-gainers (45.45% gainers). The rest of the syntactic complexity measures indicated that gainers made up over half of the participant sample. And in lexical diversity, accuracy and speech rate, we found that gainers made up more than 60% of the sample. Percentagewise, there were not quite as many gainers in SA-6m as we had found in SA-3m. This was likely due to the fact that the short-stay group exhibited a lower initial level as compared to SA-6m and had more to gain from the experience. When LoS participants were grouped together into a single, NNS group, we found similar proportions of gainers and non-gainers; likewise, lexical diversity, accuracy and fluency were the measures in which a greater percentage of learners improved. That being said, chi-square tests did not find significant differences in the proportions of gainers and non-gainers by their LoS groups.

All in all our analyses of SA and LoS, have shown that SA significantly impacted fluency as measured through SR and Pruned SR measures, and LoS was not seen to be impactful in terms of post-SA outcomes. The lack of impact of LoS was true for both CAF scores over the SA period as well as net gains accrued. Nonetheless, we saw that the majority of SA participants did benefit from their time in the TL country. In some measures (e.g. error ratios), nearly 70% of learners experienced at least some degree of gain as a result of SA, while in syntactic complexity measures there were approximately the same percentage of gainers as non-gainers.

In short, we did not find that any one LoS was more effective than the other in adding to learners L2 proficiency in these CAF measures. Furthermore, the impact of SA itself was quite limited and best seen in measures of speech rate. We did see however, that in terms of simply making progress in the L2, regardless of whether or not that progress is statistical, the majority of learners did indeed benefit from SA to some degree. This was true for both LoS groups and for learners in general.

RQ2: ***Is linguistic development in the L2 as measured through CAF different for learners with different initial levels of proficiency?***

We first dealt with RQ2 by examining the relationship between initial level and post-test outcomes across the CAF variables. *Pearson's* correlations were run for pre- and post-test scores. Results showed positive correlations between pre-test scores and WD/ASU, E/ASU, E/C, SR and Pruned SR at post-test. Thus, students who produced more words per units at pre-test were more likely to continue producing more words per unit at post-test. Similarly, those who made frequent errors before SA were also more likely overall to make errors at post test and high fluency subjects at pre-test remained higher fluency subjects at post-test.

Where correlations between pre- and post-test scores resulted in only five significantly correlated pairs, highly significant, very strong correlations were found between initial CAF scores and gains over the SA period in all 9 CAF variables. What is more, all of these relationships were negative correlations. Therefore, we see that subjects who had relatively low scores at pre-test gained most, while those who already scored on the higher end of the spectrum at pre-test did not experience remarkable gains during SA.

As GRAMSCOR was included in the study as a measure of initial level, we went on to look at correlations between pre-SA scores on SALA grammar and cloze tests and post-test gains finding no significant correlation between the two variables. Therefore we could not relate pre-programme GRAMSCOR with gains obtained during SA.

Next, using median split analysis we grouped learners by whether they exhibited a high or low initial level at pre-test as reflected in CAF scores. In doing so we came up with two groups for each CAF variable to which we assigned high and low-level labels. We also split learner groups into high/low categories based on their Pre-test GRAMSCOR results, in this case, using K-means cluster analysis.

The results of Wilcoxon signed rank tests showed that there was a significant effect of SA for the low initial level group (based on pre-test CAF scores) in 7 CAF variables included here. Changes in WD/ASU and CoordC/ASU in the low-level group did not reach significance. Effect sizes for SA in the low-level group were medium (e.g. in C/ASU, E/C) to large (e.g. in GuirIndex, SR, Pruned SR). Results for the high initial level group, demonstrated little change over the SA period although changes in CoordC/ASU and GuirIndex did reach significance. The changes in CoordC/ASU, experienced by the high-level group represented increased instances of coordination, which is a movement away from NS values. GuirIndex was the only measures in which both high and low-level groups changed significantly over the course of SA. The high-level group significantly increased their lexical diversity, while the low-level group successfully reduced their lexical diversity so as to more closely represent target-like levels in GuirIndex.

When testing for group differences between high and low initial level participants, we found that at pre-test, high and low level groups based on CAF scores differed significantly across all CAF variables. Effect sizes were also quite large with *r* values above .80. Upon arrival from SA, high and low level groups had converged in syntactic and lexical complexity scores, yet still performed differently to a significant degree according to accuracy and fluency measures. The effect size of *Group* was also down

from pre-test, around .50. These results suggest that low-level students were able to catch up to high initial level participants, at least in measures of complexity.

When we tested for between-group differences based on high/low GRAMSCOR, we did not find any significant results, corroborating the lack of significant correlations between initial grammar level and gains that we saw previously through *Pearson's* correlations

To answer our second research question, we did find linguistic development in the L2 as measured through CAF to be different for learners with different initial levels of proficiency. We saw that in general terms, the lower the initial level the greater the change over time and the greater the gains. These findings fit well with previous research into the role of initial level on SA outcomes. In our analysis of initial level, this advantage for low initial level was best seen in correlations between pre-test CAF scores and gains accrued during SA as well as in the comparison of gains where learners are arranged by high/low initial level.

RQ3: ***To what extent do previous periods abroad in the target language country (previous SA and previous abroad experiences in general) affect the outcome of the present SA experience?***

In this section we analysed previous stays abroad and previous, non-study related periods abroad as they impacted the current SA outcomes in terms of changes in scores over time, and in terms of gains. We first looked at the impact of previous SA experiences, then, in a separate set of analyses, we examined the impact of other non-academic stays abroad. We then tested for significant group differences in gains between these two separate groupings of subjects.

Syntactic complexity and accuracy scores were submitted to mixed between-within subjects ANOVAs with *Time* (the current SA) as the within-subjects factor, and Previous SA (yes/no) as the between-subjects factor. We found no significant changes over the SA period in these variables, nor were there any remarkable between-groups differences.

Several attempts at transforming GuirIndex, SR and Pruned SR variables did not result in improved normality for these distributions. Therefore, lexical diversity and fluency scores were submitted to Wilcoxon Signed Rank tests, in order to determine any significant effects of *Time*, Similarly, pre- and post-test scores in GuirIndex, SR and Pruned SR were submitted separately to Mann-Whitney U-tests of independent samples in order to determine whether or not any significant effects for *Group* could be seen through these measures at the different testing times.

Wilcoxon tests indicated that those who had not studied abroad before significantly reduced GuirIndex scores as a result of the current SA but did not demonstrate significant changes in SR and Pruned SR as a result of the current SA. Those who had studied abroad before did not show significant changes over time in GuirIndex, but, repeaters did change significantly over time in SR and Pruned SR. Effects of *Time* for this second group had effects sizes, where *r* values were superior to .50 in both SR and Pruned SR.

When median pre-test lexical diversity and fluency scores for first-timers and repeaters were submitted to Mann-Whitney U-tests we found that between-groups differences in pre-test GuirIndex scores did not reach significance, nor were any significant differences found at post-test. Likewise, differences in SR and Pruned SR scores between first-timers and repeaters did not reach significance at pre-test, and fell just short of significance at post-test.

A similar set of analyses was performed were subjects were arranged by whether or not they had spent periods in English-speaking countries prior to the SA in question. The main effects for *Time* were not found to be statistically significant in syntactic complexity and accuracy measures. The between-subjects effects showed no significant results for these variables although E/C very nearly reached significance.

GuirIndex and fluency measures had been left out of the ANOVA analyses as we could not obtain acceptable normality in these distributions. To see if any significant effects of *Time* had impacted lexical diversity and fluency outcomes, we submitted GuirIndex, SR and pruned SR scores to Wilcoxon Signed Rank tests. To test for significant effects for

Group, at pre- and at post-test, scores in GuirIndex, SR and Pruned SR were submitted to Mann-Whitney U-tests.

Wilcoxon test results indicated that those who had never been abroad to an English-speaking country before significantly reduced GuirIndex scores as a result of the current SA, but did not demonstrate significant changes over the SA period in SR or Pruned SR. Those who had been abroad before did not show any significant changes over time in GuirIndex, nor did they change significantly over time in SR and in Pruned SR. Thus, first-timers differed from repeaters in that they did demonstrate significant drops in GuirIndex as a result of the SA period, while repeaters did not show any statistically significant changes during SA in GuirIndex, SR or Pruned SR.

When median pre-test lexical diversity and fluency scores for first-timers and repeaters were submitted to Mann-Whitney U-tests we found that between-groups differences in pre-test GuirIndex scores did not reach significance nor were any significant differences found at post-test. Differences in SR and Pruned SR scores between first-timers and repeaters did reach significance at pre-test and group differences held true at post-test as well. Effect sizes for *Group* were all very similar, with medium effect sizes in all three variables tested here

Our results suggest that only first-timers progressed in lexical diversity over the SA period, approaching NS levels, but only repeaters progressed to a significant degree in fluency. No significant differences between groups were found in GuirIndex at either testing time, fluency measures, on the other hand, were found to indicate significant group differences in SR and Pruned SR at both pre- and post-test, suggesting that repeaters already performed with higher levels of fluency at pre-test and maintained their advantage over the first-timers over the SA period examined here.

In summing up our look at the impact of prior non-academic abroad experiences on the current SA, we did that first-timers' performances changed significantly in GuirIndex over the current SA period, moving toward NS levels for this task, while repeaters did not. However, between-groups differences in this measure did not reach significance at pre- or post-test. Participants who had been abroad before had higher mean fluency scores at pre -test than did their first-timer peers. Repeaters, performing with a faster SR

than first-timers prior to SA continued to outperform first-timers to a significant degree at post-test. The non-parametric tests we report here confirmed significant group differences across testing times in SR and Pruned SR.

Previous SA and previous abroad experiences in general do not clearly affect the outcome of the present SA experience. Not having gone on a previous period abroad (either SA or other) seemed to enhance students' chances of improving scores in GuirIndex. Previous SA did not seem to have any positive effects on the current SA, and previous non-academic periods abroad seemed to contributed to a superior fluency level at the start of SA, an advantage that repeater students maintained over first-timers even at post-test. Yet, between-groups differences were only seen in fluency score in the case of previous SA, and were not found at all in the case of previous experience. Had we been able to test these factors together, we may have obtained other results

CHAPTER 7

DISCUSSION

The previous chapter presented the results obtained in relation to the three research questions originally presented in Chapter 4. We analysed the language samples produced by 47 English language learners through a role-play task performed just before and upon return from SA periods varying in length by approximately three months. Along with learner data we included 24 native English speakers' productions elicited through the same task, as a baseline reference. In the present chapter we further comment on the results and discuss them with references to previous research efforts discussed in Part I. This chapter is again developed around the three main research questions and following the order of previous chapters.

7.1 Main effects of SA and LoS.

We begin our discussion with the first research question that was concerned with ascertaining whether or not SA and LoS had a significant effect on learners' linguistic ability once back in their home country. More specifically, our first research question read as follows:

> *Which modality of SA (LoS of 3 or 6 months) is more effective in producing gains in L2 speaking performance as measured through CAF?*

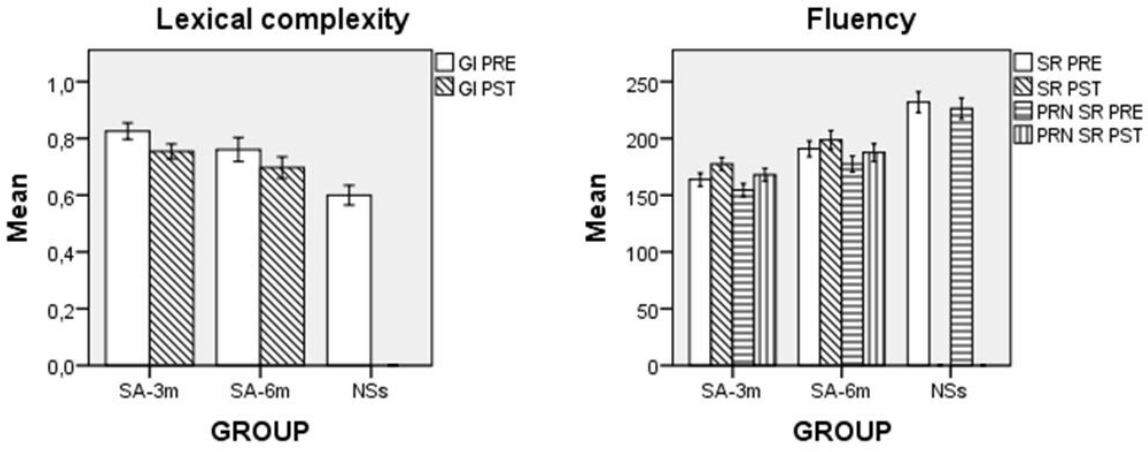
We went about answering this question by taking on a number of different approaches to the data. We began our statistical analyses looking at pre- and post-SA CAF scores and took into account effects of *Time*, that is, the SA experience in general, regardless of LoS, and secondly, effects for *Group*, i.e. the impact of participating in a short-stay programme or a long-stay programme.

7.1.1. Principal changes during SA.

An initial look at mean CAF scores for different LoS groups indicated increases in complexity measures except for WD/ASU and GuirIndex which registered decreases in lexical diversity for both learner groups. We can also observe drops in both error ratios

and increases in Speech rate measures for these groups at post-test. However, with this grouping of subjects, these changes over the SA period were only significant in SR and Pruned SR. Nonsignificant decreases in GuirIndex came close to reaching significance and are included in the illustrative graph of these mean scores (see Figure 7.1) NS values are also included (shown as mean pre-test scores).

Figure 7.1 Mean lexical complexity and fluency scores SA-3m, SA-6m and NSs. Error bars: +/- 1 SE.



In all three of these measures these changes over time represent changes toward NS values for this particular task. Fluency increases over time for both learner groups, but does not reach NS levels. Therefore, we can attribute this increase in fluency toward a more native-like rate (at least in these SR measures) to the current SA. These changes are likely due to an increase in automatisation of L2 forms as a result of being immersed in the L2 environment (Segalowitz & Freed 2004). Changes in GuirIndex, however represent a negative gain in lexical diversity, that is, students are using a more narrow range of vocabulary items in this task upon return from SA. While at first glance these results seem to indicate a backsliding in vocabulary use, when measured against the NS indices we see that NSs produced even less lexically diverse language when performing this task. Therefore, we interpret this drop in GuirIndex as a narrowing of the gap between NNS and NS performances, much in the same way SR and Pruned SR scores moved toward NS values. The effect size in those measures that changed significantly over time, SR and Pruned SR, were seen to be large. Therefore, just as previous studies have found (e.g. Segalowitz & Freed 2004; Collentine & Freed 2004; Freed et al. 2004; Valls-Ferrer 2011; Mora & Valls-Ferrer 2012; Valls-Ferrer and Mora 2014), we can

also consider L2 fluency to be one of the speech dimensions that flourishes in a SA context.

As far as measures of complexity and accuracy are concerned, the most notable point to take in is the quite consistent lack of significant changes over time as an effect of SA. When arranging learners by LoS groups and testing for the effects of SA, none of the four structural complexity measures changed significantly over the observation period (WD/ASU, C/ASU, SUBC/ASU and CoordC/ASU). Nor did the two measures of accuracy (E/ASU and E/C) indicate any significant changes in learner production over the SA period. This lack of change in syntactic complexity and accuracy has been observed often in SA literature starting with the quite early work of Möehle and Raupach (e.g. Möehle 1984; Möehle & Raupach 1983; Raupach 1983, 1984) in which the lack of gains in accuracy and complexity were, at the same time, accompanied by markedly improved fluency. SALA publications that follow a similar design to the one used here have also obtained similar results for these dimensions. Valls-Ferrer (2011) and Mora and Valls-Ferrer (2012) both found little growth in accuracy scores and no significant gains in complexity were detected in either study. At the same time, the authors report that the absence of gains in accuracy and complexity was accompanied by robust gains in fluency. Nonetheless, it should be noted that other SALA publications have reported gains in accuracy after SA (e.g. Pérez-Vidal & Juan-Garau 2011; Juan-Garau 2014), and will be discussed further in the sections that follow.

7.1.2. Did LoS even matter?

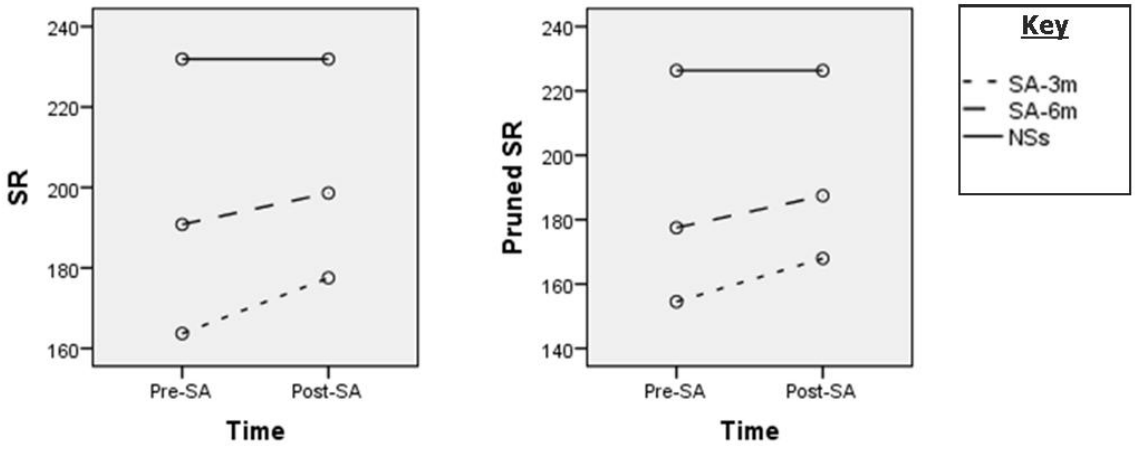
In our look at LoS we found that the only dependent variables in which learner groups performed differently to a significant degree were fluency measures SR and Pruned SR. These differences were found when *Time* and *Group* were assessed together through two-by-two ANOVAs, assuming that initial proficiency level in these groups was very similar at pre-test.

At the beginning of Chapter 6 we presented some preliminary statistical analysis that examined group differences based on the GRAMSCOR variable. In our analysis we found no significant group differences based on pre-test lexico-grammatical knowledge, nor was there any compelling information in profile questionnaires to suggest that learner groups differed in their command of English. If anything, we would expect SA-

3m to demonstrate superior initial proficiency as this group was made up entirely of translation students. Given their choice of study, one might expect SA-3m to demonstrate a superior level of L2 expertise as compared to SA-6m participants who came from a variety of academic majors and were not being trained as language experts. Therefore, from the perspective of statistical analysis these groups were assumed to part from similar linguistic backgrounds and competencies. Nonetheless, upon examining the data more closely, it seems that SA-6m was the stronger of the two groups going into SA. Indeed when *Group* was tested separately from *Time*, t-tests performed on learners' pre-test fluency scores indicated that there was a significant difference in level between groups at pre-test that could easily account for the significant between-groups differences in fluency found in the ANOVA results.

On account of the issue of the divergent initial levels, we calculated CAF gains from pre- to post-test so as to sidestep potentially confounding group differences. Gains are independent from initial level and could pick up on group differences even where participants parted from varying initial levels, differences that would otherwise be lost in the mix. When we tested post-SA gains in the two LoS groups, we found no significant differences between groups showing that SA-3m and SA-6m also gained similarly during SA, so much so that the one LoS group could not be distinguished statistically from the other. This finding reinforces the idea that overall, there was a lack of difference in how the two groups progressed in the L2 during SA at least in the CAF measures we included here.

Figure 7.2 plots of mean fluency scores (SR and Pruned SR).



In order to determine whether these differences in level persisted at post-test we repeated the t-test analyses for post-test scores. Of course, we already know that this was not the case; SA-3m caught up to SA-6m with considerable efficacy. The levelling effect was such that only SR scores between groups remained significant at post-test and with a borderline *p*-value of .046 and a slightly smaller effect size than that found at pre-test.

As we were unable to pick up on any benefit to the additional months spent abroad, i.e. LoS having no statistical effect upon L2 outcomes in terms of CAF scores or gains, our results are in line with some previous LoS studies yet seem to contrast with others. Ife et al. (2000), a study of growth in vocabulary breadth and knowledge during SA, also dealt with different SA durations: one or two semesters. As in the current study, participant groups also initiated SA with different initial levels of proficiency. In the case of Ife et al., students were already identified as having an 'intermediate' or 'advanced' L2 level and this was factored into the analysis. The authors found LoS was a significant contributor to gains across levels. This study was much more sophisticated in its treatment of vocabulary than what we offer here, and although we do not employ the same measures, if the same mechanisms that promoted dramatic growth in vocabulary in Ife et al. are in place during our participants' SA, we might expect that different stay durations would cause a differential impact on learners groups' lexical diversity. Of course, this was not the case as we found no significant between-groups effects for LoS in GuirIndex.

Llanes and Muñoz (2009) elaborated a study of very short stays abroad with LoS of 3 to 4 weeks and looking into gains in specific CAF measures, two of which appear in our study (SR as measured through syllables per minute. and E/C). As with our participants, there was notable variation in proficiency levels at pre-test. And all in all, the sample they drew from was quite varied, with ages ranging from 13-22 years, and some subjects participating in study programmes and others in paid work. Nonetheless, the authors found that when proficiency level was held constant, the participants who stayed for 4 weeks showed a greater decrease in the number of errors over those who stayed only 3 weeks. In contrast, our results did not indicate any group differences in these measures with SA durations differing by nearly three months.

On the other hand, we find that our results are in line with Llanes and Serrano (2011) who report on the lack of impact of an additional month (i.e. a month added to an otherwise two-month stay) in the target language country in the production of advanced level adult learners. This study examined the linguistic domains of fluency, accuracy, syntactic and lexical complexity in oral and written production, finding no added benefit to increasing LoS by a month. Although the unit in the denominator of the speech measures differed from ours (T-units instead of AS-units) the measures they employed were analogous to several of ours, namely WD/TU, C/TU, Guiraud's Index and E/TU. Similarly, no significant effects for LoS were found in their study, just as our results have suggested. They conclude that "'longer' may not necessarily be 'better'… in all contexts, especially as far as participants in an advanced level are concerned" (p.106).

Our results are also in line with a study conducted over the course of a full academic year where data were collected at three points in time and using CAF measures that are similar to the measures we include here. Serrano, Tragant and Llanes (2012) found that English learners studying in the UK during an Erasmus exchange programme had already accrued considerable gains in fluency (syllables per minute) and lexical diversity (Guiraud's Index) in the period of time corresponding to their first semester abroad (September to December). This period of time also roughly corresponds to a midpoint between the two SA durations we have examined (4 versus 3 and 6 months). The authors found that this first semester alone was enough time for significant progress to be made in fluency and lexical diversity, but students were not seen to progress any further during the following semester except in accuracy (Errors per T-unit). We have also found considerable effects for *Time* were measures of lexical diversity[41] and fluency are concerned, but did not find any significant improvement in accuracy or complexity measures even over the longest stays included in our study (i.e. one full semester). We can also add that significant improvement in SR and GuirIndex in our subjects' production was observable well before the semester mark as SA-3m has demonstrated.

[41] The p-value for GuirIndex only neared significance.

Therefore, in light of our results and barring any other unseen dissimilarities between our two sets of subject groups, it may have been possible to detect significant gains in fluency and lexical complexity in Serrano and colleagues' informants' oral production even as early as three-months into SA. Likewise, had our subjects continued their SA past the semester mark perhaps then we would have seen tangible improvement in accuracy and/or complexity just as Serrano et al. (2012) have found. They report that while accuracy gains were detected from Time 2 to Time 3, significant gains in complexity appear only when taking into account the full academic year, that is, from Time 1 to Time 3.

Did LoS even matter? Given these results, we have found that the only measures included here in which learner groups perform differently to a significant degree were fluency measures SR and Pruned SR and this could be attributed to differences in initial level alone. All other measures, while useful in revealing group differences between NNSs and NSs in some cases, did not significantly distinguish between SA-3m and SA-6m performances. If LoS did have an effect over these groups, we were unable to detect it and thus cannot claim any clear advantage of any one SA duration over the other.

7.2. Initial level and its effects

Initial level is notoriously impactful when it comes to SA outcomes (e.g. Brecht et al. 1995; Lapkin et al. 1995; Golonka 2006; Davidson 2010; Valls-Ferrer & Mora 2014). In fact, we have already seen its influence in the first analyses reported in the Results section. Knowing that divergent initials levels were likely at work in subjects' development over the SA period, we dedicated a research question to this factor:

> *Is linguistic development in the L2 as measured through CAF different for learners with different initial levels of proficiency?*

We remember from Chapter 6 that correlations between pre- and post-test scores showed significant relationships to exist between pre-programme scores in WD/ASU, the two error ratios and the two fluency measures, yet no significant relationship was found between pre- and post-test scores in the remaining complexity variables. All of these pairings were positively correlated, simply indicating that those who started off with higher pre-test scores were more likely to have higher post-test scores. Therefore,

when correlating pre- and post-test scores we are merely re-examining the performance gap we saw in previous analyses where students were grouped by LoS. Since this is not the aspect of initial level we are looking to isolate, we went on to perform correlations between pre-test level (measured through CAF scores) and the corresponding gains obtained by post-test. Previous studies have quite consistently pointed to the generalised impact of initial level in that learners with lower initial levels tend to gain more when compared to more advanced L2 speakers (e.g. Brecht et al. 1995; Lapkin et al. 1995; Golonka 2006; Valls-Ferrer & Mora 2014; although. also see the discussion of Rivers 1998 and Davison 2010 in section 1.3.2).

When we correlated pre-programme level with gains we found significant results across the board and in negative correlation, that is, a low score in one variable was associated with a high-score in the other. Thus, these results are in line with previous work that has identified increased gains for participants who went into SA with lower initial proficiency levels.

Davidson (2010) found that pre-programme control of grammar (along with listening ability) was a robust predictor of gains in speaking proficiency. However, when we correlated pre-programme scores on tests assessing grammar knowledge (GRAMSCOR) and CAF gains occurring over the SA period, we found no significant correlations. Although, we take into consideration that, while not reaching significance, these correlations were negative in all variable pairings except for the accuracy measures. This could suggest that pre-test command of grammar and lexis were associated with gains in that lower level students tended to gain more in terms of complexity and accuracy, while better control of grammar prior to SA resulted in a lack of further gains in accuracy. But again, these correlations did not reach significance and any inferences based on these results are purely anecdotal. Similarly, when we divided learners by whether they had high or low GRAMSCOR scores at pre-test, these high/low groups did not perform differently to a significant degree in terms of post-test CAF scores or gains. Our lack of significant results where initial level of lexico-grammatical knowledge is concerned does not coincide with those of Barquin (2012), who did find that these same tests were able to predict differential outcomes over time. However, it should be taken into account that Barquin's study was of L2 writing development during SA, and did not include oral data as in the present study.

Furthermore, her approach to the analysis was slightly different from ours, and could also potentially account for the different results between studies. We conclude that while lexico-grammatical knowledge may be useful in predicting SA outcomes in writing proficiency, we did not find these tests to be particularly helpful in predicting gains in oral proficiency as measured through CAF.

As detailed in Chapter 6, we then went on to group subjects by initial level in each of the 9 CAF variables. Overall, we found that subjects who started SA with lower initial levels tended to experience greater gains than their high-level peers, confirming and further qualifying the highly significant correlations that were seen in previous analyses of initial level. We found that when learners were grouped by high/low initial level, low-level learners accrued greater gains in all 9 CAF variables and also more frequently improved significantly over time than did the high initial level group.

7.3. Previous periods abroad

Llanes and Muñoz (2009) found the duration of SA period prior to the current SA to be a significant predictor of gains (that is, during the current SA). However Brecht et al. (1995) found that previous immersion in the TL country impacted listening skills but not speaking skills in the then current SA. Through linguistic profile and SA conditions questionnaires, we had access to information about students' abroad experiences that took place prior to the current observation period. We first examined the effect of previous SA, grouping participants according to whether or not they had previously studied abroad in an English speaking country. With this arrangement of subjects, we did not find any significant effects for *Time* syntactic complexity and accuracy measures; nor did we find any between-groups differences in learners grouped by previous SA. This same lack of significant results was true for groups arranged by previous experiences abroad in measures of syntactic complexity and accuracy.

As we have already discussed, we were unable to achieve normality in three test variables through transformations. This was true for previous SA (yes/no) and previous experience (yes/no) groupings. Therefore, GuirIndex and SR and Pruned SR scores were submitted to Wilcoxon Signed Rank tests, in order to determine any significant effects of *Time* in these variables. Similarly, pre- and post-test scores in lexical diversity and fluency were submitted separately to Mann-Whitney U-tests in order to determine

whether or not any significant effects for *Group* could be seen through these measures at the different testing times.

Wilcoxon results indicated that those who had not studied abroad before significantly reduced GuirIndex scores as a result of the current SA but did not demonstrate significant changes in SR and Pruned SR as a result of the current SA. Conversely, those who had studied abroad before did not show significant changes over time in GuirIndex, but did change significantly over time in SR and Pruned SR. In the repeaters group the *Time* factor had large effect sizes where *r* values were superior to .50 in both SR and Pruned SR.

When median pre-test lexical diversity and fluency scores for first-timers and repeaters were submitted to Mann-Whitney U-tests we found that between-groups differences in pre-test GuirIndex SR and Pruned SR scores did not reach significance. At post-test, we did not find any significant differences between groups, but differences in SR and Pruned SR scores between first-timers and repeaters fell just short of significance.

When we performed these same tests for students grouped by whether or not they had spent non-study related time abroad we found only first-timers reduced their GuirIndex scores to a significant degree during the current SA and that repeaters showed no significant progress during SA. Significant group differences were not found in GuirIndex, but were seen in fluency measures.

Unfortunately, we could not factor in all these variables together but rather had to test each factor and group separately. Had we been able to test for interactions in these data some of the outcomes we see here may have been better accounted for.

7.4. A note on native speaker comparisons

In the current study we utilized NS data as a references point when examining learner performances. These comparisons have been made mainly through descriptive statistics such as including mean NS scores alongside those of the L2 users. In the analysis of LoS we also included NNS-NS comparisons in the form of statistical analysis (Section 6.2.4.). The use of NSs as the benchmark in language teaching practices and SLA research has gleaned considerable criticism in recent years (e.g. Cook 1999). Here we

acknowledge the potentially negative connotations associated with NS-NNS comparisons, where NS norms are seen as the only goal and the users of the L2 are relegated to a second class status as 'faulty' users of the language. This is a position we reject, as we are more than aware of these subjects' high levels of competence in their own native languages, the vast majority bilingual in Spanish and Catalan, and in their additional languages. However, we also defend the use of NS baseline data in this case, as we have seen that it has proven helpful in determining where changes in learners' CAF scores might represent positive or negative gains toward some higher level of proficiency, in the case, and for lack of a better means of comparison, represented in NS scores. NS performances have provided special insight into learners' outcomes in GuirIndex, where we saw consistent and often significant drops in these scores over time. Knowing NS subjects performed with even lower degrees of lexical diversity on this particular task allowed us to interpret these changes in learner data as important gains occurring over the SA period.

7.5. CAF dimensions

Now that we have discussed the issues surrounding the three principal research questions that shaped the present study, in the following sections we will consider our findings in each specific CAF speech dimension in light of previous research.

7.5.1. On syntactic complexity

The general lack of gains in the syntactic complexity measures we have included here could be related to previous findings by Isabelli (2003) and a study by Serrano, Tragant and Llanes (2012). Isabelli notes that more advanced aspects of syntactic complexity may emerge more slowly for SA participants who start SA at a more advanced level. Already starting off at a high proficiency level leaves less to be gained during SA and it is also likely that the more salient, and thus more readily acquired aspects of the TL syntax have already been integrated into the learners L2 system leaving only the most subtle nuances of grammar still to be worked out. Since these forms are less salient or less frequent, more time is needed to take them up. When taking this into consideration for our learners, although we did see quite a bit of internal variation, students did have at least an upper-intermediate level of English. Even through her instruments and methodology varied considerably from ours, and the L1 and TL did not match those of our study, Isabelli reported that less salient, more abstract aspects of L2 Spanish syntax

began to emerge in learner language between 4 and 9 months into SA. Similarly, Serrano and colleagues found that syntactic complexity gains were significant only after a full academic year in an English speaking country. Given these previous observations, it could be that our subjects simply did not have enough time to further develop syntactic complexity having come into SA with an already considerably high proficiency level.

Isabelli points out the lack of empirical studies that directly address the specific question of whether or not longer stays really are better than shorter stays for acquiring advanced L2 features. This may be due in part to the fact that in the past measures of complexity were thought of as rather straight forward indices of growth (e.g. more subordination=more advanced language). However, today there is more of a consensus that sees complexity as a highly intricate, multidimensional construct that is not necessarily easy to capture in learner language.

> Linguistic complexity measures cannot be validated simply by showing that they increase in the course of acquisition. Developmental timing may give an indication of the difficulty of a grammatical construction or subsystem. (...) whether, or to what extent, structural complexity increases over time needs to be established empirically rather than be taken for granted. (Bulté & Housen 2012: 36-37.)

Those studies that have tackled syntactic complexity development over a SA period generally focus on the emergence of very specific parts of speech or surface forms of grammar (e.g. Guntermann 1995: Learners' correct choice of *ser* or *estar* in Spanish). And do not take into account the advancedness of such forms. The acquisition of underlying, abstract syntactic features that are infrequent in the input likely come more slowly, yet little has been done in terms of research efforts to find out.

When posing the question of 'how long is long enough?' Both Isabelli (2003) and Serrano et al. (2012) suggest that it is only in the second semester abroad that advanced learners begin to register significant gains in the domain of syntactic complexity. Isabelli goes on to recommend the academic year abroad as the most beneficial SA modality for the acquisition of more complex aspects of language.

The delayed emergence of advanced L2 structures is one possible explanation of why we did not see any gains in structural complexity measures. However, we may not have to delve so deeply into issues of advanced learner development in order to provide a plausible explanation of why participants did not improve in these measures. It may be the case that we did not see significant development in syntactic complexity because the task itself did not did not elicit particularly complex structures.

First of all, the role-play was dialogic and highly interactive, often eliciting very short and structurally simple utterances. Michel, Kuiken, & Vedder (2012) in their look at task complexity and interaction based in the Cognition hypothesis (Robinson 2005b) found that the manipulation of the factor +/- *monologic* significantly impacted CAF outcomes. More specifically, they report that dialogic tasks were found to consistently guide L2 performances towards greater accuracy, lexical complexity, and fluency, all areas in which we have achieved significant results at some point in our analysis. Furthermore, in the same paper, the authors report that interactive tasks raised the accuracy and fluency of L2 oral task performance and decreased structural complexity; a factor that potentially affected the outcomes we saw in the current study.

The fact that the role-plays were by nature quite interactive may have also pushed L2 performances toward greater fluency, and away from complexity, an effect of interaction that has been suggested by Robinson (2001). The thought is that as elements of interaction increase (e.g. turn taking, clarification requests, negotiations of meaning, etc.), opportunities to build complex structures decrease. Fluency however, is seen to be enhanced by higher levels of interaction. If these effects were also at work during the role-play task, it could be that these interactions provided a favourable environment for the measure of fluency, while simultaneously impeding the capture of the elements of structural complexity that we have examined here (i.e. general complexity, subordination and coordination).

A second element of the task that may have influenced complexity outcomes was related to a design detail of the role-play. While performing the role-play, participants had a visual support that contained photos of sample ‘decoration schemes’ that they could refer to while performing the task and that could be seen by both participants at

all times. This detail allowed students to easily employ deictic forms in order to refer to specific elements in the photos. With visual support in hand, the frequent use of deixis in these role-plays enabled speakers to directly relate their utterances to the imagined space represented in the photograph without the need to further complexify their language to accomplish their communicative goals. Take for example, the following excerpt from the sample role-play we have included in Appendix B.1. Deictic forms are underlined and in bold.

> *DEC: alright, alright, then we can mix mix them.
> *HOU: put **that** sofa **that** comfortable sofa in **this** rustic living room.
> *DEC: yeah. I think we could put **that** sofa in **this** table and **those** chairs and keep the fire **here**.

In this example, it is clear that the participants are referring to, and likely physically pointing to images on the paper they have in front of them with no reason to employ more complex syntactic structures.

At the same time, it should be noted that Mora and Valls-Ferrer (2012) obtained similar results in CAF domains when analysing an interview task that was far less interactive and more monologic than the role-play. The authors report robust gains in fluency moderate improvement in accuracy and as with our data, a generalised lack of gains in complexity.

All in all, it is likely that some or indeed all of these factors were at work in our failure to detect significant changes in complexity over SA. Perhaps different measures of structural complexity, for example a measure that captures complexity at the phrasal level[42], would have indicated changes in learners' relatively advanced production not registered in the present study. Furthermore, the effects of task complexity go beyond the scope of the current study, but had we compared tasks in which the factors of +/- *Interaction* and +/- *Monologic* varied systematically, we may have been better prepared to isolate and explain the effects these had on complexity outcomes.

7.5.2. On lexical complexity

[42] Barquin (2012) found evidence of syntactic complexity development at the phrasal level in written data provided by SALA informants who went on SA for three months.

In our study, we found consistent decreases in lexical diversity from pre- to post-test. This drop in GuirIndex held true for learners when arranged by LoS, intial level, Previous SA, and Previous experience abroad. We have noted already that upon initial examination, this shift toward a less diverse use of L2 vocabulary could appear to represent a reverse in learner progress toward more target-like language. In general, it would be expected that as learners advance in the L2, their production would include more varied language. However, when we included NS baseline data, we found that decreases in lexical diversity could be interpreted as growth toward native-like vocabulary use for this task. We can know this thanks to the inclusion of NS data at several points in our analysis both in simply comparing mean scores, and through statistical tests. One such test showed that one learner group (SA-6m), having differed significantly from NS in GuirIndex at pre-test, no longer registered significant differences with NS productions at post-test. This indicated that SA-6m and NS scores in GuirIndex and on this task had coalesced during SA to the point that NNS productions no longer varied significantly from those of NSs. If it were not for the NS baseline reference point, the changes in learner behaviour may have been interpreted as an inexplicable backsliding on the part of the learners.

Admittedly, our ability to tap lexical complexity in the current study is quite limited with GuirIndex as the only lexical complexity measure included in the analysis; even so, Guiraud's index appears to faithfully document a movement toward a more advanced use of vocabulary in this particular role-play task. We argue that the negative gains seen in this measure could have been due to learners having acquired a more precise, more appropriate, albeit, more simple use of lexis during SA. Although referring to writing progress at the time, is seems that the observation by Wolfe-Quintero et al. (1998) that "more advanced subjects may tend to use more reduced forms." likely holds true in this case.

Although we are unaware of any other study were significant improvment in lexical diversity is seen though negative gains. Pérez-Vidal and Juan-Garau (2011) report on a group of SALA participants, analogous to our SA-3m group (that is. subjects who participated in a three-month SA), whose gains in lexical diversity as measured through Guiraud's index did not reach significance. Although falling short of corroborating our repeatedly finding negative gains, Pérez-Vidal and Juan-Garau's results for the same

role-play task are likely the product of a similar process that does not yet show significant negative gains in lexical diversity.

7.5.3. On accuracy

In the current study we did not find any significant improvement in accuracy scores in terms of error rate at the unit and clausal level as an effect of *Time* where subjects were grouped by LoS or previous periods abroad. Nor were accuracy scores able to distinguish between learner groups in any of the arrangements we have explored here except where initial level was concerned. These findings do not coincide with those of Pérez-Vidal and Juan-Garau (2011) and Juan-Garau (2014) who did see some significant growth in accuracy over the SA period.

We did find, as we found for all CAF scores, that gains in accuracy were strongly correlated with pre-test level in that, the lower the level at pre-test, the greater the gains during SA. Similarly, we found that when learners were arranged into high/low initial level groups, that lower initial level learners gained more and more consistently than high-level learners in terms of progress over time, in post-test scores and in gains.

In the case of the study by Pérez-Vidal and Juan-Garau the discrepancies between their study's results and those of the current work could be attributed in the different accuracy measures used: errors per word, versus errors per AS-unit and errors per clause. However, in the case of the study by Juan-Garau, the measures, although based on the T-unit rather than the AS-unit, are quite comparable to ours. Further steps would have to be taken in order to determine why such similar research designs would yield such different results. However, it should be noted that Mora and Valls-Ferrer also found only moderate improvement in accuracy after SA but on a different task.

Our accuracy results do not fully coincide with the previous finding of these SALA studies, but do fall in line with studies of a similar demographic whose performances are measured through similar metrics. Studies by Llanes and Serrano (2011) and Serrano, Tragant and Llanes (2012) also reported nonsignificant results for accuracy as far into SA as the three-month and the semester mark respectively. Likewise, we did not register important improvements in accuracy even in the longest LoS modality, a stay of 6 months.

7.5.4. On oral fluency

Oral fluency was significantly impacted by the SA experience regardless of how learners were grouped (by LoS, initial fluency level, previous SA (yes/no) or previous experience abroad). Measures of fluency were also the only measures that were able to distinguish between learner groups in terms of pre- and post-test scores when arranged by LoS and previous experience abroad[43].

As we discussed in the literature review back in Chapter 1, oral fluency is one of the domains of L2 production that most consistently undergoes development as a result of SA. Early on in SA research increased fluency was already considered one of the primary linguistic benefits of a period abroad (Lennon 1990a; Dekeyser 1991; Towell et al 1996). More recent studies continue to report increases in fluency during SA, finding significant improvement in a number of aspects of fluency upon return from a sojourn abroad (Segalowitz and Freed 2004; Trenchs-Parera 2009; Serrano et al.; Valls-Ferrer 2011; Mora and Valls-Ferrer 2012; Valls-Ferrer and Mora 2014) to which we can now add the present work. In the current study, our assessment of fluency is limited to only two measures of SR. However, what we are able to see in our data through SR and Pruned SR fits well with previous findings for fluency development in SA contexts, and is in line with findings reported in other SALA-based studies such as Valls-Ferrer (2011), Mora and Valls-Ferrer (2012), Valls-Ferrer and Mora (2014). Furthermore, our results expand upon previous findings, confirming the impact of the SA periods studied in the SALA project on oral fluency, and through a different, highly interactional task.

7.6. Summary of Chapter 7

In this chapter we have discussed the main findings of the current study, and how they fit with the relevant literature. We commented on the main effects of SA and pointed to GuirIndex, SR and Pruned SR as the variables that we have seen to most consistently change over the SA period. We have also discussed the lack of change in structural complexity that we find throughout the study. We offered some possible explanations

[43] We have already mentioned that this difference in fluency scores was already present at pre-test, but in the case of Previous experience abroad, higher fluency scores at pre-test were only present in the repeaters group, thus suggesting that prev. exp abroad contributed to their high level at pre-test.

related to students' relatively advanced proficiency level and to the nature of the task itself that could contribute to understanding why structural complexity does not seem to undergo any noteworthy changes over time.

We have also discussed our findings related to LoS comparing our results with those of the few studies specific to LoS that came before the present study. With reference to LoS we came to the conclusion that if there was a significant impact of LoS in CAF outcomes, we were unable to observe it in this case and through these measures.

Our results mesh well with the majority of previous findings related to initial level in that we have found that students who scored low in CAF measures at pre-test tended to experience greater gains than those students who already had a superior level prior to going abroad. We discussed our testing of relationships between variables through correlations and in grouping subjects by high/low initial level. Both of these techniques led us to the same conclusion: that during SA, low-level learners gain more and more consistently demonstrate significant changes over time than more proficient learners.

Our look at previous periods abroad showed similar effects for *Time* to those we saw in the LoS analysis in that GuirIndex, SR and Pruned SR were impacted by SA. In this measure, first-timers benefitted more than repeaters, while in SR and Pruned SR it was the repeaters who started out with an advantage over first-timers and maintained it over the course of SA. Analysis of gains did not reveal any further group differences. Given our results, we fell short of being able to substantiate the findings from previous studies in which previous periods abroad were seen to influence current SA outcomes except perhaps where lexical complexity and fluency were concerned. Yet these findings do not discard unseen effects of other factors.

We concluded our discussion of the results by reflecting on the evolution of each CAF variable separately. We discussed our findings in light of previous studies with particular emphasis placed on studies of a similar demographic and during SA periods spent in English speaking countries (Llanes & Serrano 2011; Serrano et al 2012). In the following chapter, our last, we establish some conclusions based on the findings we have discussed at length here. We also consider some limitations to the current study and outline some possibilities for future research.

CHAPTER 8

CONCLUSIONS AND FUTURE RESEARCH

In the Part II chapters we have presented and discussed the present empirical study; first, examining the impact of SA programmes differing in LoS on learner production as seen through measures of CAF. We then examined the role initial level had on how learners progressed in L2 English over the abroad period, and looked into how previous periods in English-speaking countries had potentially influenced the outcomes of the SA experiences in question here.

We found that when analysing SA and LoS together, the impact of SA in general was rather limited in terms of the statistical results. Changes in learner production over time were limited to fluency measures, yet were also seen to some degree in lexical complexity. When seeking an answer to the question of the impact and implications of different LoS options, we did not find any compelling evidence to suggest that LoS had any detectable impact on these groups of learners.

Our results where the role of initial level is concerned, very much mirror well established patterns that have been discussed in the previous literature of SA contexts. We found that in general terms, those students who entered SA with lower speaking proficiency were more inclined to gain during the SA period as compared to those students who came to SA with strong oral skills. One link to initial level that did not give us any significant results was a measure of initial grammatical and lexical knowledge. We were unable to associate high/low initial level in GRAMSCOR with any SA outcomes.

When groups arranged by LoS were compared to NSs' performance of the same role-play task, we found that syntactic complexity measures did not distinguish between NNS and NS groups while lexical complexity, accuracy and fluency did clearly mark group differences based on participants' condition as (non)native speakers of English. During SA we noted a number of important trends toward NS norms especially in

lexical complexity, accuracy and fluency, and even found that those learners who spent six months abroad converged with NSs in terms of GuirIndex scores taken at post-test.

Our analysis of how periods spent abroad prior to the current one influenced the present SA outcomes was able to confirm a significant effect for *Time* as seen in the progress made in GuirIndex, SR and Pruned SR, but only revealed significant between-groups effects for GuirIndex among students who reported (not) having previous SA experiences. In this case, first-time SA participants benefitted from the current SA to a greater degree than did the repeaters group.

Our modestly robust results make it difficult to provide any definite recommendations about how long is long enough in the case of advanced level university learners of English. However, we do side with the idea that the longer the better and although the observation period of our study did not go beyond the semester mark, evidence from previous studies related to LoS point to the year abroad as the best option for high-level learners. Further research is needed in order to be able to design SA programmes that meet participants needs' in terms of SA duration, taking into account that an optimal period spent abroad may be different for different level groups.

8.1 Some limitations in the present study

Our study of SA and LoS presented a number of noteworthy limitations having to do with a number of design element and some details associated with the analysis, as these were less than ideal in some aspects.

First, the design of the study has its advantages and disadvantages. A longitudinal design is useful in reducing the problem of dubious comparability of groups. Since in this approach, subjects represent their own matched pairs across testing time, it is assumed that all other issues of individual variation, apart from the progress learners make while away are held constant. However, this type of design introduces issues related to the potential effect of task repetition across testing times; effects that could be confounded with the effects of the learning context itself and/or other factors were have examined here (e.g. LoS, Previous periods abroad).

Second, while NSs provided baseline data, no learner control group was included in the analysis. NS data was able to provide a target value for each CAF variable, but since we did not have access to longitudinal CAF scores for learners who did not participate in an abroad programme, no SA/AH comparisons can be made at this time. Furthermore, we were unable to gather data at the three-month mark from those students who went on to complete 6 months abroad. Although a call for participation in a remote data collection was issued, only three SA-6m subjects participated in data collection half-way through their SA experience. Not having tested students a testing time that corresponded to the post-test testing time of SA-3m would have limited our ability to attribute any further progress in the long-stay group to the additional time abroad. This rather important deficiency in our design was mitigated by the fact that we did not detect any significant effects for LoS. Therefore, looking further into how students progressed over the additional three months abroad, may have proven to be superfluous anyway in the end. Nonetheless, for future research into the differential effects of LoS, a set of comparable testing times across groups would be highly recommendable.

The last design element that we will discuss here had to do with the lack of control of programme variables and time on task in terms of time spent learning/using the TL during SA. Since SA conditions and how students spent their time abroad were not factored into the analysis of the current study, we have little to say about how these variables potentially influenced post-test outcomes. As we have discussed in literature review, variation in SA conditions, contact opportunities with the TL and other elements of the context are notoriously difficult to control for. Here we have only briefly mentioned SA destinations (host countries) and type of accommodation with the purpose of providing a general description of the SA programmes. Yet, undoubtedly a number of other factors were at work in the relative efficacy of the different SA situations.

Next, we draw the reader's attention to some limitations of the present study that had to do with the analysis we performed. First, we acknowledge the rather small sample size of the SA-6m group where only 14 English learners participated in data collection at both relevant testing times. This number was set against groups of subjects of 33 in the case of SA-3m and 24 in the case of NSs. Had we been able to increase the long-stay sample size, we may have been better equipped to detect significant changes over SA

and/or differences between learner groups. However, when subjects were grouped differently (e.g. by previous SA, yes/no) we achieved an arrangement of subjects in very even sample sizes (N=24, N=23 respectively). Interestingly, we obtained results for the effects of *Time* through this grouping of subject (i.e. significant results in GuirIndex, SR and Pruned SR) that closely reflected the results of the LoS analysis. This causes us to consider that the uneven samples sizes tested in our response to RQ1 (on LoS) did not influence our results to a great degree.

Another shortcoming related to the analysis of CAF was that we incorporated relatively few objective measures in our study. Due to time constraints, the inclusion of additional CAF measures was not possible; however, it would have been interesting to include some complementary indices of proficiency in order to more thoroughly examine the data. First, while we include measures of general complexity (WD/ASU) subordination (C/ASU, SUBC/ASU) and coordination (CoordC/ASU), we were lacking a measure of structural complexity at the phrasal level. Had we included such a measure, we may have been able to detect significant changes and/or gains in learner production as a result of SA, or even LoS. A measure of this kind would have been especially interesting to include in our analysis as previous studies suggest that students at advanced levels may show improvement in L2 complexity at the phrasal level, while other levels of analysis (unit, clausal, sentential) often remain constant after a given treatment, in this case, perhaps due to a ceiling effect. Similarly, dimensions of accuracy and fluency were measures through only two measures each. Furthermore, these pairs of measures tapped very similar aspects of these dimensions. Accuracy was assessed through error rate, and at the unit and clausal level, both tapping learners frequency of error. Fluency was measured through speech rate, standard, and pruned, the only difference being instances of hesitation phenomena (repetitions, formulations, self-corrections) that were removed from the Pruned SR syllable count. Had we included measures that tapped learners' ability to produce accurate speech (e.g. % error-free units) or measures of breakdown fluency, to give some examples, we may have obtained more robust or different results for these speech dimensions.

Finally, in our analysis we did not include any subjective proficiency measures. And those objective measures that were included where obtained by a single rater. In our opinion it would have been particularly interesting to include a global measure of

linguistic performance, and so correlate these with the objective measures. Unfortunately, this improvement on the present analysis will have to be left for further research at a later date, as will the inclusions of multiple raters in the processing of the objective measures.

8.2 Future research

In Chapter 3 of Part I we reviewed three recent shifts in SLA research that have acquired growing acceptance in the SLA community. We first looked at how a team of researchers have made headway in recent years in the teasing apart of the monolith construct of speaking proficiency. Members of the WiSP project have sought to remedy the circular reasoning associated with the study of oral proficiency up until now in which the truth of the observations about proficiency cannot be established independently of the conclusion (De Jong, Steinel & Florijn Hulstijn J. H. 2012a). If these are not separated out, we are left to conclude that certain elements of L2 speech production indicate more advanced performance because examples of 'more advanced' speech performance exhibit these elements. Of course, this circular argument, while potentially valid for routine language assessment, may lead to erroneous conclusions when applied to the construct of proficiency as a whole. As an alternative these authors propose measuring functional adequacy, the relative success of conveying information through speech, separately from knowledge and processing skills, where 'knowledge skills' have to do with the speaker's declarative knowledge about the language and 'processing skills', with the rapidity and accuracy with which they process linguistic information. This novel approach to proficiency awakens an interest in the application of this methodology in specific SLA subfields, not least of which, the study of specific contexts of learning such as SA.

In Chapter 3, we also discussed how certain researchers in the field of SLA have developed an interest in the construct of functional adequacy in the L2; most notably perhaps, Gabrielle Pallotti (see Pallotti 2009). We have already mentioned above that one of the shortcomings of the present study is the absence of a subjective measure of global proficiency with which to relate objective measures. Given the highly naturalistic context associated with SA, learners may vary well acquire language behaviours which add to the functional adequacy of their communications. Therefore, had we the opportunity to include a measure of global proficiency in the current work, functional adequacy would have been the preferred candidate. Yet, this widening of scope will have to wait for future studies.

The third area of innovation discussed in Chapter 3 was the emergence of Dynamic Systems Theory in the field of SLA. It occurs to us that an approach to SLA based in DST that already assumes that context and use of the L2 within that context are highly complex, dynamic and interconnected systems, may prove a useful frame for the study of language acquisition in SA contexts. As DST becomes more consolidated in the study of SLA, it would be interesting to explore how this theory, and the methodologies associated with it, might be applied to the study of SA and perhaps LoS as well.

On a more personal note, the development of the current study has fuelled several research interests that I would like to explore in the future. The first has already been mentioned above, namely exploring how functional adequacy develops over a period of SA. I see the inclusion of such a measure in my research as an attainable, short-term goal, and would very much like to develop a study around this concept. I would also like to test whether or not multiple short stays abroad allow for cumulative effects in learners' speaking proficiency. In the development of this study, it came to my notice that quite a few subjects had spent approximately one month of their summer break abroad in English-speaking countries. A number of subjects had gone on a series of such stays over as many as four consecutive summers. I would be interested in finding out if this practice of spending a few weeks abroad each year would have the same effect as spending an equal, yet uninterrupted amount of time in the TL country, such as a semester-length programme.

Finally, another of my longer-term research interests is in developing and testing effective ways of brining elements of immersion contexts to language students studying in their home countries. While SA has become phenomenally popular in recent decades, a parallel development has come on the scene of language teaching-leaning in the form of instant and constantly available access to communication technologies. Several SA authors have seen technology as a principal culprit in abroad participants' inability to connect with the host culture, thus impeding progress in the L2. Coleman gives us some perspective on just how far SA has come in this sense telling of his own abroad experience in the 1960s when the journey between his home in the UK and his host institution in France lasted over 24 hours, and his principal means of communication with friends and family back home was by letter (Coleman & Chafer 2010). My own SA experience required that computer mediated communication with my family be

accomplish in internet cafes at a half-hourly rate. And my pre-paid phone card allowed me to call the US from public phone booths a couple of times a week. Less than 15 years later, students are perpetually connected through mobile internet connections, free instant messaging and online telephone services. While some would lament the presence of the home culture in the host culture, we perhaps over look the opportunities to turn the tables, brining the TL home to foreign language students. In the case of L2 English teaching-learning there are endless possibilities for students access sources of input and interaction offered through new technologies. It is clear that the extent to which SA participant are 'immersed' in the TL culture has evolved tremendously with the advent of global communication technologies, the next stages of self-immersion at home could have a similar impact on AH language learning.

REFERENCES

Adams, R. (2006). Language learning strategies in the study abroad context. In M. DuFon & E. Churchill (Eds.) *Language learners in study abroad contexts.* 259-293. Clevedon, UK: Multilingual Matters.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign language. *TESOL Quarterly*, 13(2), 219-227.

Altbach, P., & Bassett, R (2004). The brain trade. *Foreign Policy* 44, 30-1.

Anderson , J. (2009). *Cognitive psychology and its implications* (7th ed). New York: Worth.

Anderson, S. (2012). Differential gains in oral proficiency during study abroad: The role of language learning aptitudes. (Doctorial dissertation). Retrieved from (https://repository.library.georgetown.edu/)

Atkinson, D. (2002) Toward a sociocognitive approach to second language acquisition. *Modern Language Journal*, *86*, 525–545.

Association of Universities and Colleges of Canada [AUCC]. (2011). *Trends in Higher Education: Volume 1 – Enrolment*. Retrieved February 16, 2012, from: http://www.aucc.ca/media-room/publications/trends-in-higher-education-volume-1-enrolment.

Avello, P. (2013). L2 phonological development in speech production during study abroad. (Doctorial dissertation). Retrieved from Tesis Doctorals en Xarxa. (http://hdl.handle.net/10803/128624)

Avello, P., & Lara, A. R. (2014). Phonological development in L2 speech production during Study Abroad Programmes differing in length of stay. In C. Pérez-Vidal, (Ed.), *Language Acquisition in Study Abroad and Formal Instruction Contexts.* (pp. 137-166). Amsterdam, John Benjamins.

Bardovi-Harlig, K. (1992). A second look at T-unit analysis: Reconsidering the sentence. *TESOL Quarterly, 26,* 390–5.

Barquin, E. (2012). *Writing Development in a Study Abroad Context*. (Doctorial dissertation). Retrieved from Tesis Doctorals en Xarxa. (http://hdl.handle.net/10803/96196).

Batstone, R. (2002). Contexts of engagement: a discourse perspective on 'intake' and 'pushed output'. *System, 30,*1–14.

Beattie, J. (2008). Evaluating the acquisition of English listening comprehension skills in the framework of the SALA Project. *Proceedings of the V Congreso*

Internacional de Adquisición del Lenguaje (AEAL), Journal Aula Abierta. Instituto de Ciencias de la Educación, Universidad de Oviedo.

Beattie, J. (2014). The 'ins and outs' of a study abroad programme: The SALA exchange programme. In C. Pérez-Vidal, (Ed.), *Language Acquisition in Study Abroad and Formal Instruction Contexts.* (pp. 59–84). Amsterdam, John Benjamins.

Bhatia, T., & Ritchie, W. (2012). *The handbook of bilingualism and multilingualism* (2nd ed.). Malden, MA: Blackwell.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: The case of systematicity. *Language Learning, 33,* 1–17.

Boersma, P., & Weenink, D. (2014). Praat: doing phonetics by computer [Computer program]. Version 5.3.78. Retrieved from http://www.praat.org/

Bradshaw, J. (2012, February 2). University leaders want more Canadians to study abroad. *The Globe and Mail.* Retrieved from http://www.theglobeandmail.com

Brecht, R., Davidson, D., & Ginsburg, R. (1995). Predictors of foreign language gain during study abroad. In *Second language learning in a study abroad context.* B. Freed (ed.). Philadelphia: John Benjamins.

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.* Amsterdam: John Benjamins.

Butler, Y. G. (2012). Bilingualism/multilingualism and second-language acquisition. In T. Bhatia, & W.Ritchie. *The handbook of bilingualism and multilingualism* (2nd ed.) (pp.109-136). Malden, MA: Blackwell.

Byrnes, H., & Sinicrope, C. (2008). Advancedness and the development of relativization in L2German: A curriculum-based longitundinal study. In L. Ortega, & H. Byrnes (Eds.). *The longitudinal study of advanced L2 capacities* (pp. 109-138). New York: Routledge.

Canale, M., & Swain, M. (1980). Theoretical bases of com-municative approaches to second language teaching and testing. *Applied linguistics, 1*(1), 1-47.

Carlson, J. S., Burn, B., Useem, J. & Yachimowicz, D. (1990). *Study Abroad: The Experience of American Undergraduates.* New York: Greenwood Press.

Carroll, J. (1967). Foreign language proficiency levels attained by language majors near graduation from college. *Foreign Language Annals, 1,* 131-51.

Carroll, J., & Sapon, S. (1959). *Modern Language Aptitude Test.* The Psychological Corporation: San Antonio, Texas.

Caspi, T., & Lowie, W. (2013). The dynamics of L2 vocabulary development: A case study of receptive and productive knowledge. *RBLA , Belo Horizonte, 13* (2), 437-462.

Chaudron, C. (2001). Progress in language classroom research: Evidence from *The Modern Language Journal*, 1916–2000. *Modern Language Journal*, 85, 57–76.

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, MA: MIT Press.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences.* Hillsdale, NJ:Erlbaum.

Cohen, A. D., Paige, R. M., Shively, R. L., Emert, H., & Hoff, J. (2005). Maximizing study abroad through language and culture strategies: Research on students, study abroad program professionals, and language instructors. Final Report to the International Research and Studies Program, Office of International Education, DOE. Minneapolis, MN: Center for Advanced Research on Language Acquisition, University of Minnesota.

Coleman, J. A. (1998). Language learning and study abroad: the European perspective. *Frontiers, 4*, 167-203.

Coleman, J. A. (2005). Residence abroad. In J. A. Coleman, & J. Klapper (Eds.) *Effective learning and teaching in modern languages*. London: Routledge.

Coleman, J. A., & Chafer, T. (2010). Study abroad and the Internet: physical and virtual context in an era of expanding telecommunications. *Frontiers* XIX, 151–167.

Collentine, J. (2004). The effects of learning contexts on morphosyntactic and lexical development. *Studies in Second Language Acquisition, 26*(02), 227-248.

Collentine, J. (2009). Study Abroad Research: Findings, Implications and Future Directions. In M. H. Long & C. J. Catherine (Eds.) *The Handbook of Language Teaching* (pp.218-233). New York: Wiley.

Collentine, J., & Freed, B. (2004). Learning context and its effects on second language acquisition . *Studies in Second Language Acquisition*. 26, 153-171.

Cook, V. (1999), Going beyond the native speaker in language teaching. TESOL Quarterly, 33 (2), 185–209

Corder, S. P. (1967) The significance of learners errors. *International Review of Applied Linguistics, 5*, 161–70.

Council of Europe, Modern Languages Division (2001). *Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge University Press.

Crookes (1991). Second language speech production research: a methodologically oriented review. *Studies in Second Language Acquisition. 13* (2), 113-32.

Dalton-Puffer, C. (2011). Content-and-language integrated Learning: from practice to principles? *Annual Review of Applied* Linguistics, 31, 182–204.

Davidson, D. (2010). Study abroad: When, how long and with what results? New data from the Russian front. *Foreign Language Annals, 43* (1), 6-26.

De Bot, K. (1992). A bilingual production model: Levelt's "Speaking" model adapted. *Applied Linguistics, 13*(1), 1-24.

De Bot, K. (2002). Cognitive processing in bilinguals: Language choice and code-switching. In R. B. Kaplan (Ed.). *The Oxford Handbook of Applied Linguistics* (pp. 287–300). Oxford: Oxford University Press.

De Bot, K., Lowie, W., & Verspoor, M. (2005). *Second Language Acquisition, an advanced resource book.* London: Routledge.

De Bot, K., Verspoor, M. & Lowie, W. (2005). Dynamic Systems Theory and applied linguistics: The ultimate so what? *International Journal of Applied Linguistics, 15* (1), 116-118.

De Bot, K., Lowie, W. & Verspoor, M. (2007) A Dynamic Systems Theory approach to second language acquisition. *Bilingualism: Language and Cognition, 10* (1), 7–21.

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

De Jong, J., & Van Ginkel, L. (1992). Dimensions in oral foreign language proficiency. In L. Verhoeven & J. De Jong (Eds.). *The construct of language proficiency: Applications of psychological models to language assessment.* John Benjamins.

De Jong N.H., Steinel, M.P, Florijn, A.F., Schoonen, R., & Hulstijn J. H. (2012a). Facets of speaking proficiency. *Studies in Second Language Acquisition, 34,* 5-34.

De Jong N.H., Steinel, M.P, Florijn, A.F., Schoonen, R., & Hulstijn J. H. (2012b). The effect of task complexity on functional adequacy, fluency and lexical diversity in speaking performance 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.* Amsterdam: John Benjamins.

De Jong, N. H., Groenhout, R., Schoonen, R., & Hulstijn, J.H. (2013). Second language fluency: Speaking style or proficiency? Correcting measures of second language

fluency for first language behavior. *Applied Psycholinguistics*, available on CJO2013. doi:10.1017/S0142716413000210.

De Jong, N. H., & Wempe, T. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. *Behavior Research Methods 41* (2), 385-390.

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

DeKeyser, R. (1990). From learning to acquisition? Monitoring in the classroom and abroad. *Hispania, 73*(1), 238-247.

DeKeyser, R. (1991). Foreign language development during a semester abroad. In B.Freed (Ed.), *Foreign Language Acquisition: Research and the Classroom* (pp. 104-119). Lexington, MA: D. C. Heath.

DeKeyser, R. (2007). Study abroad as foreign language practice. In R. Dekeyser (Ed.). *Practice in a second language: Perspectives from applied linguistics and cognitive psychology*. Cambridge: Cambridge University Press.

DeKeyser, R. (2010). Monitoring progress in Spanish as a second language during a study abroad program. *Foreign Language Annals, 43* (1), 80-92.

Dell, G. (1986) A spreading activation theory of retrieval in sentence production. *Psychology Review, 93* (3), 283-321.

Dewaele, J.M., Housen A., & Wei, L. (2003). *Bilingualism: Beyond basic principles: Festschrift in honour of Hugo Baetens Beardsmore.* Multilingual Matters.

Dewey, D. P. (2004). A comparison of reading development by learners of Japanese in intensive domestic immersion and study abroad contexts. *Studies in Second Language Acquisition, 26*, 303–327.

Dewey, D. P. (2008). Japanese vocabulary acquisition by learners in three contexts. *Frontiers, 15*, 127-148.

Díaz-Campos, M. (2004). Context of learning in the acquisition of Spanish second language phonology. *Studies in Second Language Acquisition, 26* (2), 249-273.

Diessel, H. (2004). *The acquisition of complex sentences*. Cambridge: Cambridge University Press.

DuFon, M., & Churchill, E. (2006). Evolving threads in study abroad research. In M. DuFon & E. Churchill (Eds.). *Language learners in study abroad contexts.* Clevedon, UK: Multilingual Matters.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

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

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

European Commission. (2003). *Promoting Language Learning and Linguistic Diversity: an action plan 2004 – 2006.* Brussels, COM(2003) 449 final. Retrieved from http://www. http://eur-lex.europa.eu

European Commission. (2012). *The ERASMUS Programme – studying in Europe and more*. European Union, 1995-2012. Retrieved from http://ec.europa.eu/education/

European Commission (2013). *Erasmus – Facts, Figures & Trends. The European Union support for student and staff exchanges and university cooperation in 2011-12.* Luxembourg: Publications Office of the European Union

Eurostat (2009). Key Data on Education in Europe 2009. European Commission: Brussels.

Félix-Brasdefer, J. C. (2004). Interlanguage refusals: Linguistic politeness and length of residence in the target community. *Language Learning 54* (4) 587–653.

Field, A. (2009). *Discovering statistics using SPSS* (3rd ed.). London: Sage.

Freed, B. (1990). Language learning in a study abroad context: The effects of interactive and non interactive out-of-class contact on grammatical achievement and oral proficiency. In J. Alatis (ed.), *Linguistics, language teaching and language acquisition: The interdependence of theory, practice, and research* (pp. 459-77), Georgetown University round table on languages and linguistics. Washington, DC: Georgetown University Press.

Freed, B. (1995a). Language learning and study abroad. In B. Freed (Ed.). *Second language learning in a study abroad context*. Philadelphia: John Benjamins

Freed, B. (1995b). What Makes Us Think That Students Who Study Abroad Become Fluent? In B. Freed (Ed.). *Second language learning in a study abroad context.* Philadelphia: John Benjamins

Freed, B. (1998). An overview of issues and research in language learning in a study abroad setting. *Frontiers, 4*, 31–60.

Freed, B. F. (2000). Is fluency, like beauty, in the eyes (and ears) of the beholder? In H. Riggenbach (Ed.), *Perspectives on Fluency* (pp. 243-265). Ann Arbor: The University of Michigan Press.

Freed, B., Segalowitz, N., & Dewey, D. (2004). Context of learning and second language fluency in French: comparing regular classroom, study abroad, and

intensive domestic immersion programs. *Studies in Second Language Acquisition*, *26* (2) 275-301.

Freed, B., So, S., & Lazar, N. (2003). Language learning abroad: how do gains in written fluency compare with gains in oral fluency in French as a second language? *ADFL Bulletin*, *34*(3): 34-40.

Ferguson, C. (1971). Absence of copula and the notion of simplicity: A study of normal speech, foreigner talk and pidgins. In D. Hymes (ed.) *Pidginization and creolization of languages* (117-140). Cambridge: Cambridge University Press.

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

Gass, S., & Mackey, A. (2007). Input, interaction and output in second language acquisition. In B. VanPatten & J. Williams (Eds.), *Theories in second language acquisition: An introduction* (pp. 175-200). Mahwah, NJ: Lawrence Erlbaum.

Gass, S., & Selinker, L. (1994). *Second Language Acquisition: An Introductory Course.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Gass, S., & Varonis, E. (1985). Variation in native speaker speech modification to non-native speakers. *Studies in Second Language Acquisition*, *7*, 37-58.

Generalitat de Catalunya (2010). *Informe de Politica Lingüística.* Retrieved from: http://www20.gencat.cat/

Generalitat de Catalunya (2012). *Informe de Politica Lingüística.* Retrieved from: http://www20.gencat.cat/

Gilabert, R. (2006). The simultaneous manipulation of task complexity along planning time and [+/- here-and-now]: Effects on L2 oral production. In M. P. García-Mayo (Ed.). *Investigating tasks in formal language learning* (pp.44-68). Multilingual Matters.

Golonka, E. (2006). Predictors revised: Linguistic knowledge and metalinguistic awareness in second language gain in Russian. *Modern Language Journal 90*, 496-505.

Gregg, K. (1984). 'Krashen's monitor and Occam's razor', *Applied Linguistics 5*, 79–100.

Griffiths, R. (1991). Pausological research in an L2 context: A rationale, and review of selected studies. *Applied Linguistics 12* (4), 345-364.

Grosjean, F. (2010). *Bilingual: Life and Reality*. Cambridge, MA: Harvard University Press.

Guntermann, G. (1995). The Peace Corps experience. In B. Freed (Ed.). *Second language acquisition in a study-abroad context.* Philadelphia: John Benjamins.

Hakuta, K. (1986). *Mirror of language: The debate on bilingualism*. New York: Basic Books.

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

Harley, T. (2008). *The psychology of language* (2nd. ed.) Hove: Psychology Press.

Harmer, J. (2001). *The Practice of English Language Teaching* (3rd. Edition). Longman.

Hatch, E. (1978). Discourse analysis and second language acquisition. In E. Hatch (Ed.), *Second language acquisition: A book of readings* (pp. 402-435). Rowley, MA: Newbury House.

Højen, A. D. (2003). *Second-Language Speech Perception and Production in Adult Learners before and After Short-Term Immersion.* PhD dissertation, University of Aarhus.

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

Housen, A. Kuiken, F., & Vedder, I. (2012a). Complexity, accuracy and fluency: definitions measures and research. In A. Housen, F. Kuiken & I. Vedder (Eds.) *Dimensions of L2 Performance and proficiency: Complexity, accuracy and fluency in SLA.* Amsterdam: John Benjamins.

Housen, A. Kuiken, F., & Vedder, I. (2012b). Epilogue. In A. Housen, F. Kuiken & I. Vedder (Eds.) *Dimensions of L2 Performance and proficiency: Complexity, accuracy and fluency in SLA.* Amsterdam: John Benjamins.

Hulstijn, J.H., Schoonen, R., de Jong, N.H., Steinel, M.P., & Florijn, A.F. (2012). Linguistic competences of learners of Dutch as a second language at the B1 and B2 levels of speaking proficiency of the Common European Framework of Reference for Languages (CEFR). *Language Testing, 29*, 202-220.

Ife, A., Vives Boix, G., & Meara, P. (2000). The impact of study abroad on the vocabulary development of different proficiency groups. *Spanish Applied Linguistics, 4*(1), 55-84.

Institute of International Education. (2012a). "Host Regions and Destinations of U.S. Study Abroad Students, 2009/10 - 2010/11." *Open Doors Report on International Educational Exchange*. Retrieved from http://www.iie.org/opendoors

Institute of International Education. (2012b). "Top 25 Destinations of U.S. Study Abroad Students, 2009/10 -2010/11." *Open Doors Report on International Educational Exchange*. Retrieved from http://www.iie.org/opendoors

Institute of International Education. (2012c). "Top 25 Places of Origin of International Students, 2010/11-2011/12." *Open Doors Report on International Educational Exchange*. Retrieved from http://www.iie.org/opendoors.

Isabelli, C. (2003). Study abroad for advanced foreign language majors: Optimal duration for developing complex structures. In H. Byrnes & H. Maxim. *Advanced foreign language learning: A challenge to college programs* (AAUSC series), (pp.114-130). Boston: Heinle.

Isabelli, C. (2004). The acquisition of NSP properties in SLA: Some effects of positive evidence in a natural learning setting. *Hispania, 87,* 150-162.

Isabelli, C. (2006). Study abroad social networks, motivations and attitudes: Implications for second language acquisition. In A. DuFon, & E. Churchill (Eds). *Language learners in study abroad contexts* (pp.231-258). Clevedon: Multilingual Matters.

Isabelli, C., & Nishida, C. (2005). Development of the Spanish subjunctive in a nine-month study-abroad Setting. In D. Eddington (Ed.). *Selected Proceedings of the 6th Conference on the Acquisition of Spanish and Portuguese as First and Second Languages* (pp.78-91). Somerville, MA: Cascadilla.

Jackson, J. (2006). Ethnographic pedagogy and evaluation in short-term study abroad. In M. Byram, & A. Feng. *Living and studying abroad* (pp. 134-156). NY: Multilingual Matters.

Jackson, J. (2010). *Intercultural Journeys: From study to residence abroad.* Palgrave/Macmillan.

Juan-Garau (2014). Oral accuracy growth after formal instruction and study abroad: Onset level, contact factors and long-term effects. In C. Pérez-Vidal (Ed.), *Language Acquisition in Study Abroad and Formal Instruction Contexts* (pp. 87–110). Amsterdam, John Benjamins.

Juan-Garau, M., & Pérez-Vidal, C. (2007). The Effect of Context and Contact on Oral Performance in Students Who Go on a Stay Abroad. *Vigo International Journal of Applied Linguistics 4*: 117-134.

Johnstone, B., & Marcellino, W. (2010). Dell Hymes and the ethnography of communication. In R. Wodak, B. Johnstone, & P. Kerswill (Eds). *The Sage Handbook of Sociolinguistics*. London: Sage Publishers.

Juan-Garau, M. & Pérez-Vidal, C. (2007). The effect of context and contact on oral performance in students who go on a stay abroad. *Vigo International Journal of Applied Linguistics*. *4*, 117-135.

Kelo M., Teichler U., & Wachter B. (2006). Toward improved data on student mobility in Europe: Findings and concepts of the Eurodata study. *Journal of Studies in International Education, 10* (3), 194-223.

Krashen, S. (1976).1 Formal and informal linguistic environments in language learning and language acquisition. *TESOL Quarterly, 10*, 157-168.

Krashen, S. (1977). Some issues relating to the monitor model. In H. Brown, C. Yorio, and R. Crymes (Eds.). *On TESOL '77* (pp.144-158). Washington D.C.: TESOL.

Krashen, S. (1985). *The Input Hypothesis: Issues and Implications.* London: Longman.

Kinginger, C. (2009). *Language learning and study abroad: A critical reading of research*. New York: Palgrave Macmillan.

Kormos, J. (2006). *Speech Production and Second Language Acquisition.* Mahwah, NJ: Lawrence Erlbaum Associates.

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

Kuiken, R., & 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.143-169). Amsterdam: John Benjamins.

Kuiken, F., Vedder, I., & Gilabert, R. (2010). Communicative adequacy and linguistic complexity in L2 writing. In I. Bartning, M. Martin, & I. Vedder (Eds). *Second Language Acquisition and Testing in Europe, Eurosla Monographs, 1: EUROSLA.*

Lafford, B. (2004). The effect of the context of learning on the use of communication strategies by learners of Spanish as a second language. *Studies in Second Language Acquisition, 26*, 201-225.

Lantolf, J. P. (1994). Sociocultural theory and second language learning: Introduction to the special issue. *Modern Language Journal, 78*, 418–420.

Lanzendorf, U. (2006). Foreign students and international students. In M. Kelo, U. Teichler, & B. Wächter (Eds.). *Eurodata: Student Mobility in European Higher Education* (pp. 7-55). Bonn: Lemmens Verlags- & Mediengesellschaft mbH.

Lapkin, S., Hart, D., & Swain, M. (1995). A Canadian interprovincial exchange: Evaluating the linguistic impact of a three-month stay in Quebec. In B. Freed

(Ed.). *Second language acquisition in a study-abroad context* (pp.67-94). Philadelphia, PA: John Benjamins.

Larsen-Freeman, D. (1991). Second language acquisition research: staking out the territory. *TESOL Quarterly***,** *25* (2) 315-350.

Larsen-Freeman, D. (2002). Language acquisition and language use from a chaos/complexity theory perspective. In C. Kramsch (Ed.). *Language Acquisition and Language Socialization: Ecological Perspectives.* New York: Continuum.

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., & Long, M. H. (1991). *An Introduction to Second Language Acquisition Research.* New York: Longman.

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

Lennon, P (1990b). The advanced learner at large in the L2 community: Developments in spoken performance. *IRAL, 28(4),* 309–324.

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

Lewis, M. P., Simons, G.F., & Fennig, C.D. (2014). *Ethnologue: Languages of the World, Seventeenth edition.* Dallas, Texas: SIL International. Online version: http://www.ethnologue.com.

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

Llanes, À., & Serrano, R. (2011). Length of stay and study abroad: Language gains in two versus three months abroad. *RESLA, 24,* 95-110

Long, M. (1983). Linguistic and conversational adjustments to non-native speakers. *Studies in Second Language Acquisition, 5,* 177-193.

Long, M. (1981). Input, interaction and second-language acquisition. In H. Winitz (ed.) *Native language and foreign language acquisition. Annals of the New York Academy of Sciences* (Vol. 379, 259-278).

Long, M. (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). San Diego, CA: Academic Press.

Lord, C. (2002). Are subordinate clauses more difficult? In J. Bybee & M. Noonan (Eds.). *Complex sentences in grammar and discourse: Essays in honor of Sandra A. Thompson* (pp. 223-234). Amsterdam: John Benjamins.

Mackey, W. (2012). Bilingualism and multilingualism in North America. In T. Bhatia & W.Ritchie. *The handbook of bilingualism and multilingualism* (2nd Edition) (pp.707-724). Malden, MA: Blackwell.

MacWhinney, B. (2000). *The CHILDES Project: Tools for Analyzing Talk* (3rd Edition). Mahwah, NJ: Lawrence Erlbaum Associates

McLaughlin, B. (1987). *Theories of second language learning.* London: Edward Arnold.

Meara, P. (1994). The year abroad and its effects. *Language Learning Journal*, *10* (1), 32-38.

Michel, M. C., Kuiken, F., & Vedder, I. (2012). Task complexity and interaction: (combined) effects on task-based performance in Dutch as a second language. In L. Roberts, C. Lindqvist, C. Bardel & N. Abrahamsson (Eds.), *Eurosla Yearbook:* Volume 12 (pp. 164–190). Amsterdam: John Benjamins

Miller, L., & Ginsberg, R. (1995). Folklinguistic theories of language learning. In B. F. Freed (Ed), *Second language acquisition in a study abroad context* (pp. 293–316). Philadelphia: John Benjamins.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. New York: Multilingual Matters.

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL Review of Applied Linguistics* 107-108, 17-34.

Mitchell, R., Tracy-Ventura, N., McManus, K., Richard, L. & Romero de Mills, P., with Dewaele, J.M. (2013, April). *The influence of social networks, personality and placement type on language learning during residence abroad: Preliminary findings of the LANGSNAP project.* Paper presentation at Residence Abroad, Social Networks and Second Language Learning, the Centre for Applied Language Research, University of Southampton.

Mora, J. C. (2006). Age effects on oral fluency development. In C. Muñoz (Ed.), *Age and the Rate of Foreign Language Learning* (pp. 65-88). Clevedon: Multilingual Matters.

Mora, J. C. (2008). Learning context effects on the acquisition of a second language phonology. In C. Pérez-Vidal, M. Juan-Garau, & A. Bel (Eds.), *A portrait of the young in the new multilingual Spain* (pp. 241-263). Clevedon: Multilingual Matters.

Mora, J. C. (2014). The role of onset level on L2 perceptual phonological development after formal instruction and study abroad. In C. Pérez-Vidal (Ed.). *Language Acquisition in Study Abroad and Formal Instruction Contexts* (pp.167-194.). Amsterdam: John Benjamins.

Mora, J. C., & Valls-Ferrer, M. (2012). Oral fluency, accuracy and complexity in formal instruction and study abroad learning contexts. *TESOL Quarterly 46*(4), 484-500.

Möehle, D. & Raupach, M.. (1983). *Planen in der Fremdsprach.* Frankfurt: Peter Lang.

Möehle, D. (1984). A comparison of the second language speech of different native speakers. In *Second language productions.*In H. Dechert, D. Möehle, & M. Raupach (Eds.) (pp. 26-49). Tubingen: Gunter Narr.

Norris, J. M., & Ortega, L. (2001). Does type of instruction make a difference? substantive findings from a meta-analytic review. *Language Learning, 51,* 157–213.

Nunan, D. (1992) *Research methods in language learning*. Cambridge: Cambridge University Press.

OECD (2013), *Education at a Glance 2013: OECD Indicators*, OECD Publishing. Retrieved from http://dx.doi.org/10.1787/eag-2013-en

Ortega, L. (2011). Second language acquisition. In J. Simpson (Ed.). *Routledge handbook of applied linguistics.* New York: Routledge.

Pallotti, G. (2009). CAF: Defining, Refining and Differentiation Constructs. *Applied Linguistics 30* (4), 590-601.

Pavlenko, A. (1999). New approaches to concepts in bilingual memory. *Bilingualism: Language and Cognition, 2*, 209–230.

Pavlenko, A. 2002. '"We Have Room for But One Language Here": Language and National Identity in the US at the Turn of the 20th Century', *Multilingua, 21* (2/3), 163–96.

Pérez-Vidal, C. (2011). Language acquisition in three different contexts of learning: Formal instruction, Study Abroad and Semi-immersion (CLIL). In Y. Ruiz de Zarobe, J.M. Sierra, & F. Gallardo del Puerto (Eds.). *Content and Foreign Language Integrated Learning: Contributions to Multilingualism in European Contexts* (pp. 25-35). Bern/Berlin: Peter Lang.

Pérez-Vidal (2014). Study abroad and formal instruction contrasted. In C. Pérez-Vidal, (Ed.), *Language Acquisition in Study Abroad and Formal Instruction Contexts.* (pp. 17–58). Amsterdam, John Benjamins.

Pérez-Vidal, C. (Forthcoming). Languages for all in education: CLIL and ICLHE at the crossroads of multilingualism, mobility and internationalization. In M. Juan-Garau & J. Salazar-Norguera (Eds.). *Content-based learning in multilingual educational environments.* Dordrecht: Springer

Pérez-Vidal, C., Trenchs, M., Juan-Garau, M., & Mora, J. C. (2007, April). *El factor 'Estancia en el país de la lengua meta' en la adquisición de una lengua extranjera (inglés) a corto y medio plazo: Objetivos y metodología del proyecto S.A.L.A.* Paper presented at the XXIV Congreso Internacional de la Asociación Española de Lingüística Aplicada, Madrid (UNED).

Pérez-Vidal, C., & Juan-Garau, M. (2011). The effect of context and input conditions on oral and written development: A study abroad perspective. *International Review of Applied Linguistics in Language Teaching, 49*(2), 157-185.

Pérez-Vidal, C., Juan-Garau, M., Mora, J. C., & Valls-Ferrer, M. (2012). Oral and written development in formal instruction and study abroad: differential effects of learning context. In Muñoz, C. (Ed.) *Intensive Exposure Experiences in Second Language Learning* (pp.213-233). Bristol: Multilingual Matters..

Pica, T., Young, R., & Doughty, C. (1987). The impact of interaction on comprehension. *TESOL Quarterly, 21*, 737-758.

Pineda Herrero, P., Moreno Andrés, M. V., & Belvis Pons, E. (2008). La movilidad de los universitarios en España: estudio sobre la participación en los programas Erasmus y Sicue. *Revista de educación, 346*, 363-399.

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

Polanyi, L. (1995). Language learning and living abroad: Stories from the field. In B. Freed. *Second Language Acquisition in a Study Abroad Context* (pp. 271-291). Philadelphia: John Benjamins.

Porter, P. (1986). How learners talk to each other: Input and interaction in task-centered discussions. In R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 200-222). MA: Newbury House.

Raupach, M. (1983). Analysis and evaluation of communicative strategies. In C. Faerch and G. Kasper (eds.). *Strategies in interlanguage communication* (pp.199-201). London: Longman.

Raupach, M. (1984). Formulae in second language speech production. In H. Dechert, D. Moehle.,& M. Raupach (Eds.). *Second language production* (pp.114-137). Tubignen: Gunter Narr Verlag.

Rees, J., & Klapper, J. (2008). Issues in the quantitative longitudinal measurement of second language progress in the study abroad context. In L. Ortega, & H.

Byrnes. *The longitudinal study of advanced L2 capacities* (pp.89-105). New York: Routledge/Taylor & Francis.

Regan, V. (1995). The acquisition of sociolinguistic native speech norms: Effects of a year abroad on second language learners of French. In B. Freed (Ed.). *Second language acquisition in a study-abroad context* (pp.245-268). Philadelphia, PA: John Benjamins.

Richters, E., & Teichler, U. (2006). Student mobility data: current methodological issues and future prospects. In M. Kelo, U. Teichler & B. Wächter (Eds.). *Eurodata: Student Mobility in European Higher Education* (pp. 80-97). Bonn: Lemmens Verlags & Mediengesellschaft mbH.

Riggenbach, H. (1991).Toward an understanding of fluency: A microanalysis of nonnative speaker conversations. *Discourse Processes, 14* (4) 423-421.

Rivers, W. (1998). Is being there enough? The effect of homestay placements on language gain during study abroad. *Foreign Language Annals, 31*(4), 492-500.

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

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

Robinson, P. (2005a). Attention and memory during SLA. In C. Doughty & M. H. Long (Eds). *The Handbook of Second Language Acquisition*. Blackwell.

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

Ruiz de Zarobe, Y. & Jiménez-Catalán, R. M. (2009). *Content and language integrated learning: evidence from research in Europe*: Multilingual Matters.

Saito, Y. (2003). Investigating the construct validity of the cloze section in the Examination for the Certificate of Proficiency in English. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, 1.

Sasaki, M. (2004). A multiple-data analysis of the 3.5-year development of EFL student writers. *Language Learning, 54*, 525-582.

Sasaki, M. (2007). Effects of study-abroad experiences on EFL writers: A multiple-data analysis. *Modern Language Journal, 91*, 602-620.

Sasaki, M. (2011). Effects of varying lengths of study-abroad experiences on Japanese EFL students' L2 writing ability and motivation: A longitudinal study. *TESOL Quarterly 45*(1), 81-105.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11* (2), 129-158.

Schmidt R. (1993). Awareness and second language acquisition. *Annual Review of Applied Linguistics*, 13, 206-226.

Segalowitz, N., Freed, B., Collentine, J., Lafford, B., Lazar, N., & Díaz-Campos, M. (2004). A comparison of Spanish second language acquisition in two different learning contexts: study abroad and the domestic classroom. *Frontiers, 10*, 1-18.

Segalowitz, N. & Freed, B. (2004). Context, contact, and cognition in oral fluency acquisition: learning Spanish in at home and study abroad contexts. *Studies in Second Language Acquisition, 26*, 173–199.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics 10* (3), 219–231.

Serrano, R., Tragant, E., & Llanes, A. (2012). A longitudinal analysis of the effects of one year abroad. *Canadian Modern Language Review/ La Revue canadienne des langues vivantes, 68* (2), 138-163.

Skehan, P. (1996). Second language acquisition research and task-based instruction. In J. Willis and D. Willis (Eds.). *The challenge and change in language teaching*. Oxford: Heinemann.

Skehan, P. (1998). *A cognitive approach to language learning*. 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.

Skinner, B. F. (1957). *Verbal Behavior*, New York: Appleton-Century-Crofts.

Spolman, 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.

Sunderman, G. & Kroll, J. (2009). When study abroad experience fails to deliver: The internal resources threshold effect. *Applied Psycholinguist, 30* (1), 79 – 99.

Sutton, R. & Rubin, D. (2004). The GLOSSARI Project: Initial findings from a system-wide research initiative on study abroad learning outcomes. *Frontiers, 10* (4), 65-82.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds), *Input in second language acquisition* (pp. 235–253). Rowley, MA: Newbury House.

Swain, M. (1995). Three functions of output in second language learning. In G. Cook & B. Seidlhofer (Eds.). *Principle and practice in applied linguistics: Studies in honour of Henry G. Widdowson* (pp. 125-144). Oxford: Oxford University Press.

Swain, M. (2000). French immersion research in Canada: recent contributions to SLA and applied linguistics. *Annual Review of Applied Linguistics, 20*, 199–212.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.). *Handbook of research in second language teaching and learning* (pp. 471-483). Mahwah, NJ: Lawrence Erlboum Associates.

Tabachnick, B., & Fidell, L. (2006). *Using Multivariate Statistics*, 5th Ed. NY, Pearson.

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

Travakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. in R. Ellis (Ed.). *Planning and Task Performance in a Second Language* (239-273). John Benjamins.

Trenchs-Parera, M. (2009). Effects of formal instruction and a stay abroad on the acquisition of native-like oral fluency. *Canadian Modern Language Review, 65*(3), 365-393.

Valls-Ferrer, M. (2011). *The Development of oral fluency and rhythm during a study abroad period*. (Doctoral dissertation). Retrieved from Tesis Doctorals en Xarxa. (http://hdl.handle.net/10803/52064).

Valls-Ferrer, M., & Mora, J. C. (2014). L2 fluency development in formal instruction and study abroad: The role of initial fluency level and language contact. In C. Pérez-Vidal (Ed.), *Language Acquisition in Study Abroad and Formal Instruction Contexts* (pp.111-136). Amsterdam, John Benjamins.

Van Geert, P. (1994). *Dynamic systems of development: Change between complexity and chaos*. NewYork: Harvester.

VanPatten, B. (1989). Can learners attend to form and content while processing input? *Hispania, 72*, 409-417.

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

Vermeer, A. (2000). Coming to grips with lexical richness in spontaneous speech data. *Language Testing, 17* (1), 65-83.

Verspoor, M., Lowie, W., & De Bot, K. (2008). Input and second language development form a dynamic perspective. In T. Piske & M. Young-Scholten. *Input Matters in SLA* (pp. 62-80). Multilingual Matters.

Wang, Y. (2009). *Anxiety in English language learning: A case study of Taiwanese university students on a study abroad programme.* (Doctorial dissertation). Retrieved from White Rose eThesis online. (http://etheses.whiterose.ac.uk/id/eprint/5396).

Weigrand, E. (2011). Paradigm changes in linguistics: from reductionism to holism. *Language Sciences, 33*(4) 544-549.

White, L. (1989). *Universal grammar and second language acquisition*. Philadelphia, PA: John Benjamins.

White, L. (1991). Adverb placement in second language acquisition: some effects of positive and negative evidence in the classroom. *Second Language Acquisition Reseach, 7*, 133-161.

White, L. (2003). *Second language acquisition and universal grammar*. Cambridge: Cambridge University Press.

Wolfe-Quintero, K., Inagaki, S., & Kim, H.-Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity*. Honolulu, HI: University of Hawaii, Second Language Teaching & Curriculum Center.

Yuan, F., & R. Ellis. (2003). The effects of pre-task planning and on-line planning on fluency, complexity, and accuracy in L2 monologic production. *Applied Linguistics 24* (1), 1-27.

APPENDICES

APPENDIX A.1. Role-play task instructions and visual support

SALA / IND / ROLE PLAY / SU	T 1

ROLE PLAY (A): House owner's role

In this part of the test we would like you two to engage in a conversation around the following topic: **REDECORATING YOUR DINING ROOM**

HOUSE OWNER'S ROLE

You are tired of the decoration in your dining room. You want a big change, but you don't know exactly what you want. The only thing you are sure of is that you do NOT like really modern, minimalist design.

You also have a limited budget so you want to spend as little as possible.

You initiate the conversation by saying:

I'm tired of this old decoration. I want a big change. How can you help me?

SALA / IND / ROLE PLAY / SU	T 1

ROLE PLAY (A): House owner's role

In this part of the test we would like you two to engage in a conversation around the following topic: **REDECORATING YOUR DINING ROOM**

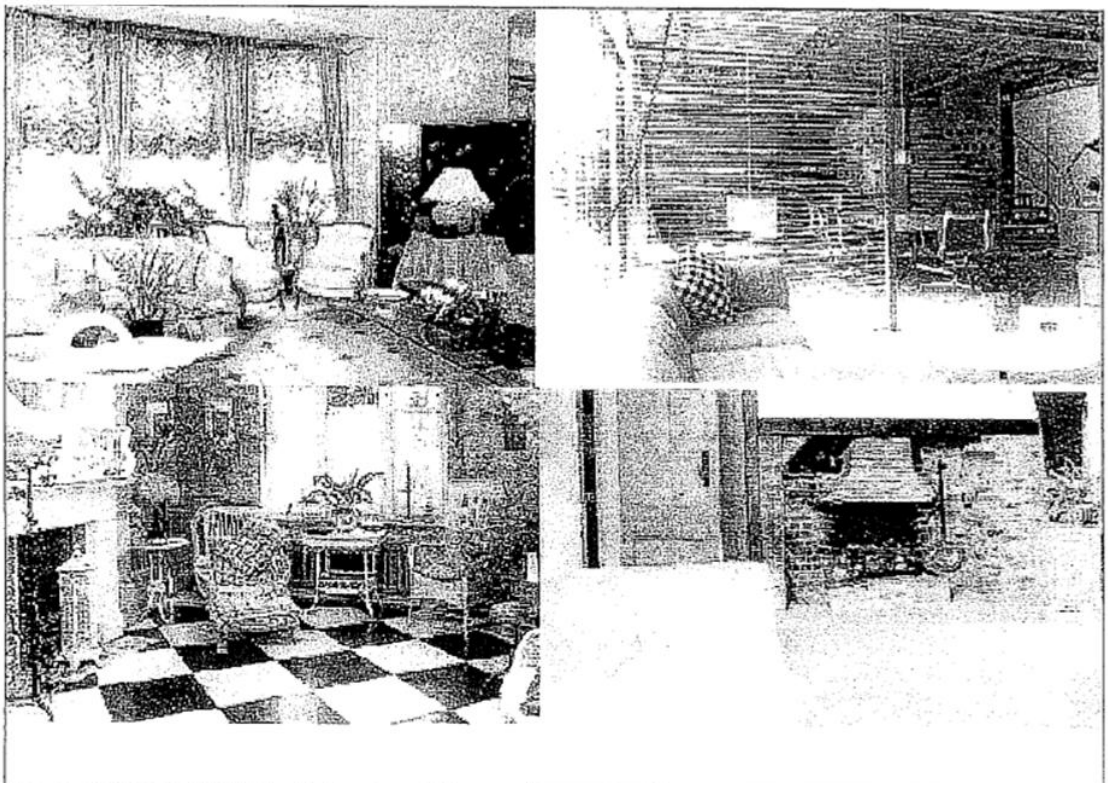
HOUSE OWNER'S ROLE

You are tired of the decoration in your dining room. You want a big change, but you don't know exactly what you want. The only thing you are sure of is that you do NOT like really modern, minimalist design.

You also have a limited budget so you want to spend as little as possible.

You initiate the conversation by saying:

I'm tired of this old decoration. I want a big change. How can you help me?



APPENDIX B.1 Sample transcribed role-play: SA-3m (pre-SA)

*HOU: I'm tired of this old decoration. I want a big change. how can you help me?
*DEC: well I have um different options for you but I would um (1.12) I'd say this this one is definitely the best one and you should buy (0.88) this (0.54) xxx. it's fantastic you know?
*HOU: I don't want such a (0.93) modern decoration. I want (0.56) like (1.10) xxx.
*DEC: but why why not? it's much better. you know?
*HOU: xxx
*DEC: if you buy this one no one will go into your home.
*HOU: because modern things change. and now maybe I like this but then in two years time (0.50) I would hate this second decoration and I would have to change <another time>.
*DEC: you see but uh this is some modern thing that is also classic and you you'll have it for for many years it not that modern if you look at it.
*HOU: I don't have enough money to (0.74) modern <xxx>.
*DEC: <alright> but if you if you if you buy something cheap it it will become ex(pensive) expensive. because if you buy this (0.81) shitty um rustic livingroom it will break. the sofa is so so so bad I mean it will break in two days (0.52). I wouldn't I would never buy it.
*HOU: xxx sofa. yeah it seems comfortable.
*DEC: no it is. it's so uncomfy. <it isn't comfortable at> all.
*HOU: I've come to try it.
*DEC: uh (0.82) <it's
*HOU: well <we could change>
*DEC: why don't you> try this one?
*HOU: uh we could we could change the sofa and the rest of the livingroom (0.63) like this rustic (1.36) livingroom.
*DEC: what do you mean? like you can (0.56) you could uh buy this one but with the rustic uh
*HOU: ah no no no.
*DEC: right. no! you see que it's this this so(fa) it's not just the sofa it's that
*HOU: I don't have enough money.
*DEC: do you think
*HOU: I'm poor.
*DEC: fire is like? you you prefer to have like this heating and everything more comfortable and easier and.
*HOU: I think this this is more comfortable because it's like more familiar more.
*DEC: but it has no no light. it has no light and no tv and
*HOU: I can I can put it light and tv.
*DEC: but it will be so expensive he's wanting to it everything (0.58). and it's not that expensive if you look at the price okay?
*HOU: why don't we mix both?
*DEC: alright alright. then we can mix mix them.
*HOU: put that sofa that comfortable sofa in this rustic <livingroom>.
*DEC: <yeah> I think we could put that sofa in this table and those chairs and keep the fire here.
*HOU: well we'll talk about it.
*DEC: alright.
@End

APPENDIX B.2 Criteria and examples in the application of AS-units to oral data following Foster et al. (2000).

AS-units
The **AS-unit** is "a single speaker utterance consisting of an independent clause. or sub-clausal unit. together with any subordinate clause(s)" (p.365).
Example 1.
A: | so you're a traditional person eh? |
B: | yes I am |

Independent clauses
"An independent clause will be minimally a clause including a finite verb" (p.365)
Example 2.
A: | that's right |
B: | look here |

Independent sub-clausal units
An independent sub-clausal unit is considered to be "one or more phrases which can be elaborated to a full clause by means of the recovery of ellipted elements from the context of the discourse. In this extract from a role-play taken from SA-6m. we can see that the second line produced by speaker A and the utterance of speaker B are counted as AS-units. The utterances in bold contain no finite verb. but their status as AS-units can be established through the recovery of ellipted material understood from the context.

Example 3.
A: | so I would suggest you :: to have {few}(1.11) not big but | you know (0.74)| **|{with personality} few pieces with their own personality so. |**
B: **| what about this open space? |**

Here, as interpreters of the speakers meaning would understand something along the lines of: (*I would suggest you have) few pieces with their own personality. And what (do you think) about this open space?* We count these instances as AS-units but not as clauses.

Subordinate clauses
"A subordinate clause will consist minimally of a finite or non-finite verb element plus at least one other element" (p.366). These additional elements can be a subject. object. complement. or an adverbial element.

Example 4.
A: | but I know :: that I hate {modern the} the modern decoration and this stuff of minimalism and things like this |

In speaker A's turn above we find 1 AS-unit and 2 clauses (1 independent clause. 1 subordinate clause)

Level two criteria
The role-plays were divided into AS-units following the **level two criteria**. described in Foster et al (2000). This level is intended to lend itself to highly interactional data where a considerable portion of the data results in only minimal units such as one-word. minor

utterances and echoed responses. In example 5. elements that were excluded from the AS-unit count according to level two criteria are in bold.

Example 5.
A: **okay**
B: **okay** |and what floor do you want? |
A: **eh**
B: | the black and white one? |
A: **yeah**

Minor elements and exact repetitions
We further specify here the exclusion from the final AS-unit count of a string of one-word. minor elements. Example 6 is an illustration of this. We also exclude exact repetitions of a previous utterance as in the fragment in Example 7.

Example 6.
A: **okay okay well** (0.76) |{um I'm feeling um} I'm feeling more confident about my decision|
B: **uhuh**

Example 7.
A: | uhuh. what is the (1.19) cheapest (0.95) one?
B: **ah the cheapest one uh.**
A: | because I have a limited budget |
B: | okay the cheapest one is {is the the the this} the first one |
A: **uhuh**

Self-corrections and false starts
Self-corrections and false starts were not in the AS-unit count unless they met the criteria to be counted as an AS-unit. False starts and self-corrections are enclosed in brackets as in Example 8.

Example 8.
A: | {**the flowers the the**} all the flowers you want to :: you can have |

Non-syntactic criteria
In runs of speech where the speaker uses elements of coordination or subordination to hold the floor. but where the connectors have no other clear grammatical or cohesive function. a AS-unit boundary was drawn were a pause of 0.5 seconds was accompanied by rising or falling intonation. Measures of pause time were rounded to the nearest hundredth of a second 0.50. Pitch contours automatically generated in PRAAT (Boersma & Weenink 2014) were also consulted whenever necessary. In Example 9 pitch is not marked. but rising/falling intonation was present in the excerpt.

Example 9.
A: |do you have any suggestions :: as to how I could um perhaps personalize this minimalist um decor in my new living room :: um in order to make it more comfortable for me **because**| (pause of 0.62 seconds) um (pause of 1.70 seconds) | **because** I'm not sure :: that this style is exactly :: what I'm looking for. |

Interruptions and scaffolding

We dealt with interruptions applying the same global criteria for determining AS-unit/clausal boundaries on a case by case basis. taking into consideration that a unit may be produced across turns. In Example 10. speaker B is credited with one AS-unit and two clauses over two turns even though speaker A interrupts briefly. attempting to take the floor.

Example 10.

B: |yeah it's not really comfortable :: to have
A: um I think.
B: wheelchairs in the table|

Where **scaffolding** occurred. the speaker who completed the previous turn (through scaffolding) was credited with having produced a complete AS-unit. In Example 11 we see that speaker A completes speaker B's previous turn. providing the object. "another chair". Both A and B were considered to have produced an AS-unit.

Example 11.

A: | but I think :: a glass table wouldn't fit with {the} the chairs |
B: | we can find eh |
A: | **another chair** |

APPENDIX C.1. Grammer and cloze test instructions and examples

Instructions and first two items from SALA 'Grammar I' task.

Instructions:

Finish each of the following sentences in such a way that it is as similar as possible in meaning to the sentence printed before it.

EXAMPLE: Despite Jack's strange clothes. everybody ignored him.
ANSWER: Nobody took .....
ANSWER: *Nobody took* notice of Jack's strange clothes.

1. Please don't use the shower after midnight.

Would you mind
.....

2. The weather was fine at the seaside last Saturday.

We had
.....

Instructions and first four items from SALA 'Cloze' task.

Instructions:

Fill **EACH** of the numbered blanks in the following passage with **ONE** suitable word.

The lady who liked adventure

It was one of those impulse buys that can happen while shopping. Mary Bruce was in London looking for a nice dress ..... (1) she noticed a showroom with a light aircraft for ..... (2) at a terribly reasonable price. Mrs. Bruce went away to ..... (3) on a dress. It did not suit her. The plane ..... (4).

APPENDIX C.2 Complete listing of SALA tests and questionnaires.

Written tests	**Characteristics**
Grammar I*	sentence rephrasing (15 minutes)
Grammar II	gap-fill requiring the conjugation of verb forms (10 minutes)
Cloze*	286-word story in which subjects provide a single appropriate word for each gap (15 minutes).
Listening comprehension	audio recording and pen and paper listening comprehension test with multiple-choice and fill-in-the-blank questions on the content (20 minutes).
Phonological perception	audio recording and pen and paper test of phonological perception presenting groups of three nonsense words pronounced as in English where two items were identical and one contained a different phoneme known to be perceptually opaque for Spanish/Catalan speakers (20 minutes).
Composition	Argumentative essay in response to the prompt: *Someone who moves to a foreign country should always adopt the customs and way of life of his/her new country. rather than holding on to his/her own customs.* (30 minutes).
Oral tests	
Read aloud	Student read a short text out loud: "The North Wind and the Sun" (5 minutes approx.)
Interview	Students answered a series of seven questions about their everyday experiences at university. (7 minutes approx.)
Role-play*	Students performed a role-play in which student A played the role of a homeowner looking to redecorate his/her living/dining room and student B played the role of a professional decorator (7 minutes approx.)
Questionnaires	
Background questionnaire*	recorded demographic information. as well information about informants L1 and L2 use.
Attitude questionnaire	recorded subjects opinions about intercultural contact and language learning
Study abroad conditions questionnaire*	recorded specific information regarding the stay abroad that occurred during the observation period.

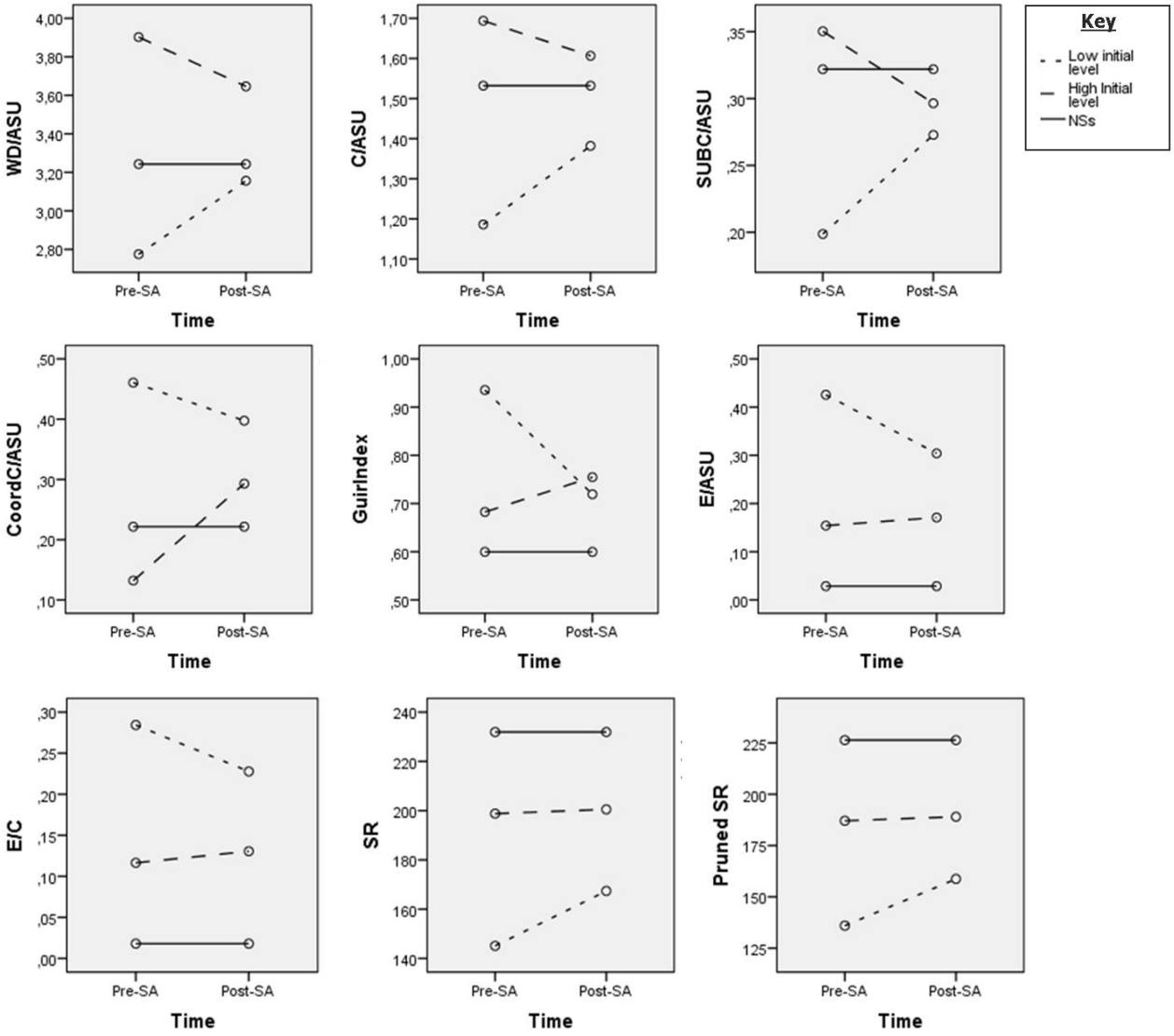
*tests used in the elaboration of the present study

APPENDIX C.3 Diagram of original SALA (three-month SA) design. Taken from Pérez-Vidal & Juan-Garau 2011:170.

Academic years represent one of three cohorts on record. Note: design not applicable to SA-6m.

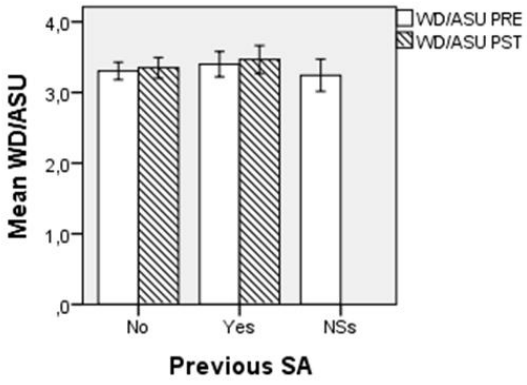
Year	2005/06			2006/07			2007/08		
Term	1	2	3	1	2	3	1	2	3
Treatment	FI (40h)	FI (40h)	—	SA	FI (40h)	FI (40h)	—	—	
	T1		T2		T3			T4	

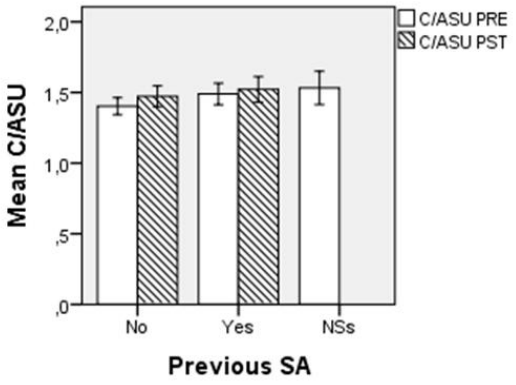
APPENDIX D.1 Line graphs of mean CAF scores with NNS arranged by high and low initial level, NSs.

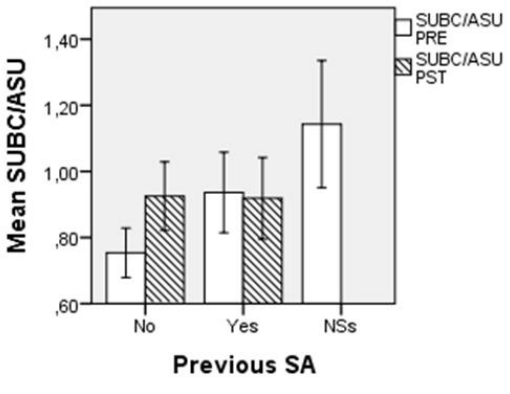


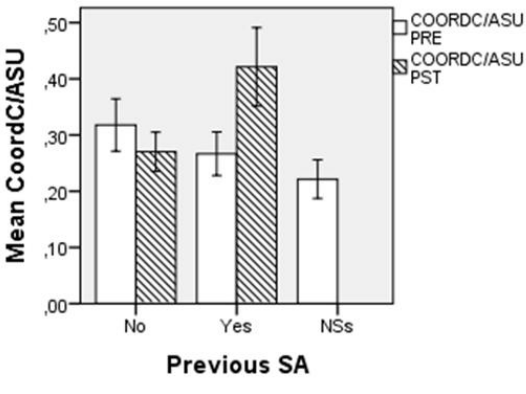
APPENDIX D.2 Bar graphs of mean syntactic complexity and accuracy scores, and median lexical diversity and fluency scores for learner groups Previous SA (yes/no) NS were testing only once (means/medians shown as pre-test scores).

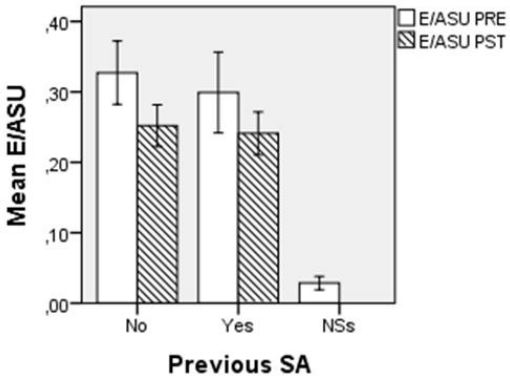
Mean syntactic complexity and accuracy scores. Error bars: +/- 1 SE

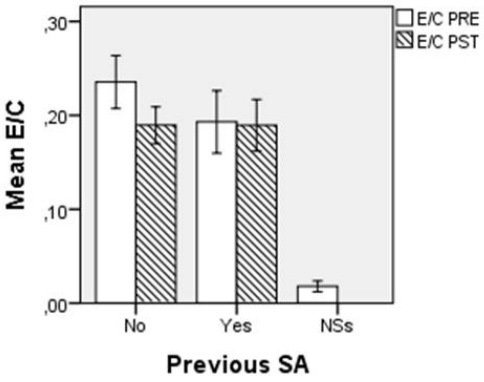




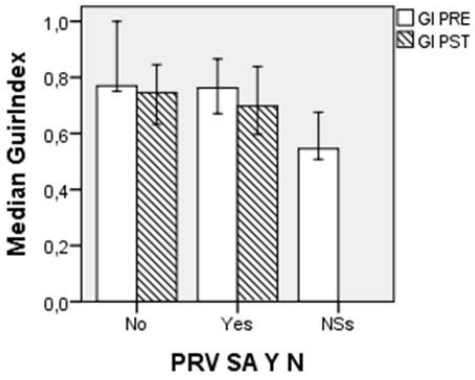


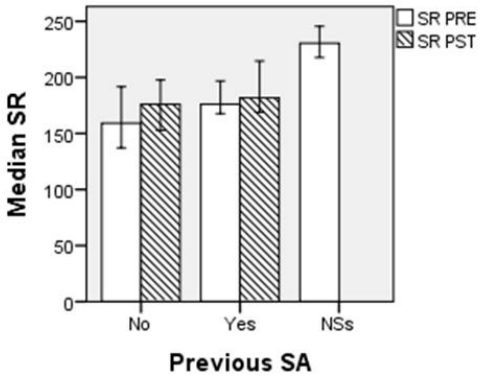


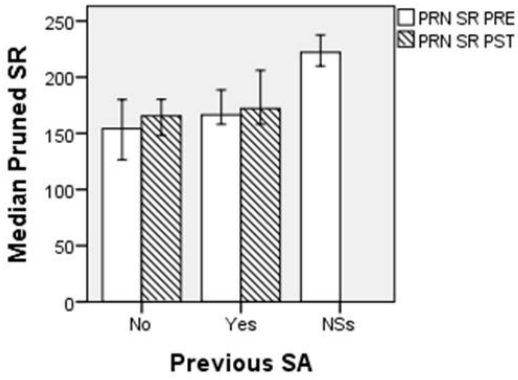




Median GuirIndex, SR and Pruned SR scores. Error bars: 95% CI

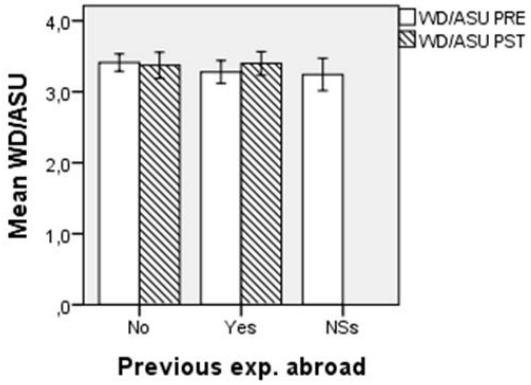


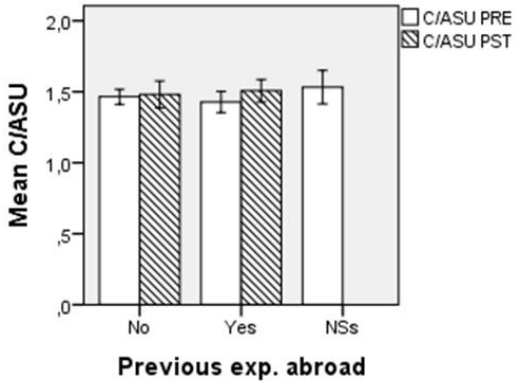


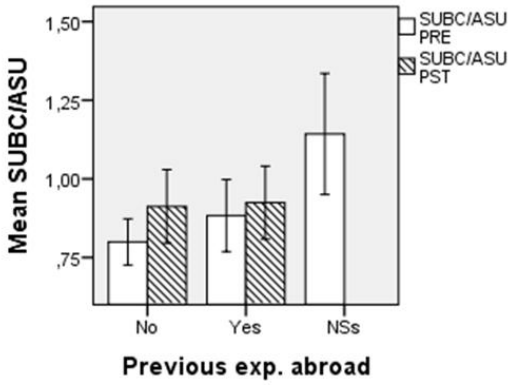


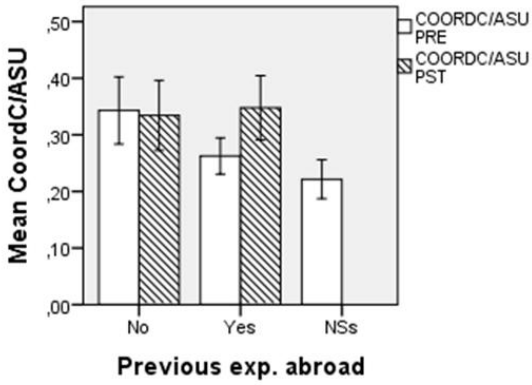
APPENDIX D.3 Bar graphs of mean syntactic complexity and accuracy scores, and median lexical diversity and fluency scores for learner groups Previous experience abroad (yes/no) NS were testing only once (means/medians shown as pre-test scores).

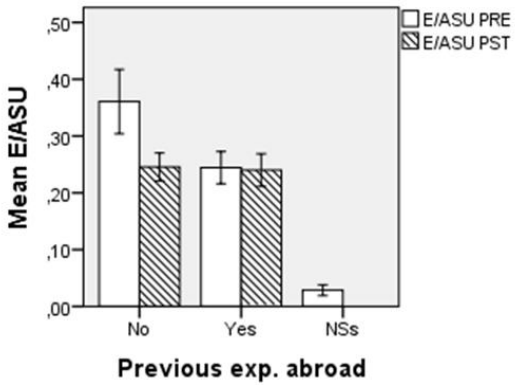
Mean syntactic complexity and accuracy scores. Error bars: +/- 1 SE

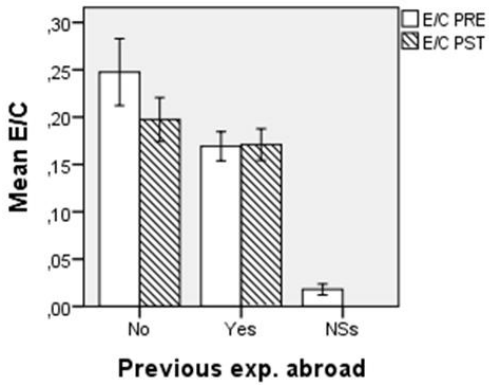












Median GuirIndex, SR and Pruned SR scores. Error bars: 95% CI