THE MEASUREMENT OF THE COMPLEXITY, ACCURACY, AND FLUENCY OF WRITTEN ARABIC

A Dissertation
submitted to the Faculty of the Graduate School of Arts and Sciences of Georgetown University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Arabic and Islamic Studies

By

Michael Raish, M.A.

Washington, D.C.
April 19, 2017
THE MEASUREMENT OF THE COMPLEXITY, ACCURACY, AND FLUENCY OF WRITTEN ARABIC

Michael Raish, M.A.

Thesis Advisor: Karin C. Ryding, Ph.D.

ABSTRACT

This study investigates the multicomponential nature of L2 Arabic writing by adapting a number of direct measures of linguistic complexity, accuracy, and fluency (CAF) to the measurement of written Arabic texts produced by L2 Arabic learners and Arabic Native Speakers (NSs). Previous studies of L2 writing have established that such measures are reliable indicators of learners’ developing interlanguage, however few efforts have thus far applied CAF analyses to spoken or written Arabic. The results of the measures explored here are triangulated with learner responses to several independent estimates of Arabic proficiency, including a novel “short-cut” estimate in the form of an Arabic C-test. Instead of a longitudinal design, in which a learner or group of learners are asked to provide data at different stages of L2 acquisition, the current study entailed the collection Arabic production along a spectrum of writing ability, spanning from learners at the end of their first year through Arabic NSs.

Significant findings of this study include the fact that learners appear to “move” toward more natively-like written production as their overall facility with the Arabic language increases. Furthermore, learners scores on the various selected CAF measures correlate highly with each other, as well as with the selected independent measures. C-test consistently emerges via regression as the most reliable predictor of variability in written CAF among the selected explanatory variables, followed closely by self-estimated overall Arabic ability. Conversely, however, this study establishes that sentence-based measures are shown to be unreliable for the
modeling of learner-produced Arabic texts, as well as the fact that communicative Arabic writing does not necessarily benefit from lemmatization prior to lexical analysis, in spite of features of the Arabic script and morphology that may argue otherwise. Additionally, this analysis indicates that the commonly used independent measure Years of Study is an unreliable predictor of learner CAF among the participants in this study. Ultimately, the measures and analyses presented here represent an initial effort to model L2 Arabic writing according to widely used CAF measures, however the findings presented here nevertheless have important implications for the Teaching Arabic as a Foreign Language (TAFL) field.
This dissertation could not have been written without the continued and ongoing support of many individuals, all of whom are deserving of my sincere thanks and gratitude.

My parents and siblings have always been a constant source of encouragement, and I am deeply grateful for support. I am also grateful to Carolyn, who selflessly assisted in some way in nearly every stage of this dissertation. Logic dictates that I couldn’t have finished it without her.

I sincerely thank the members of my committee, including my adviser Dr. Karin Ryding, who untiringly provided feedback, support, and guidance. Dr. Ryding has long been an appreciated mentor and advocate. Dr. John Norris was likewise a constant source of insight and support—the instruments and study design used here grew out of projects that I began in his classes, and his ongoing help was crucial. Dr. Cristina Sanz was likewise an unwaveringly positive influence during all phases of this study, generously providing direction and encouragement.

I likewise extend my sincere gratitude to all of the individuals who participated in the many stages of this study, including the participants, as well as those good friends who helped recruit participants. I thank them for their contributions and interest in Arabic.

My sincere thanks to all.
## Table of Contents

Chapter 1: Introduction .................................................................................................................. 1

1.1 Background .......................................................................................................................... 1

1.2 Study Overview .................................................................................................................... 3

Chapter 2: Review of Literature .................................................................................................... 7

2.1 Approaches to Measurement and Assessment of L2 Arabic ...................................................... 7

2.2 Alternative Testing Methods .................................................................................................. 12

2.3 Challenges in the Assessment of Arabic Writing ................................................................... 14

2.4 The Complexity, Accuracy, Fluency Framework ..................................................................... 19

2.5 Complexity .......................................................................................................................... 21

2.6 Lexical Complexity .............................................................................................................. 34

2.7 Accuracy ............................................................................................................................. 37

2.8 Fluency .................................................................................................................................. 40

2.9 Genre Effect .......................................................................................................................... 43

Chapter 3: The Development of an Arabic C-Test ....................................................................... 50

3.1 Introduction .......................................................................................................................... 50

3.2 Background .......................................................................................................................... 51

3.3 Developing Arabic C-tests .................................................................................................... 57

3.4 Developing a Pilot C-test ...................................................................................................... 73

3.5 Learner Testing ...................................................................................................................... 80

3.6 Results and Rasch Analysis .................................................................................................. 82

3.7 Discussion and Conclusions .................................................................................................. 86

Chapter 4: Methodology ............................................................................................................... 89
4.1 Research Questions

4.2 Participants and Recruitment

4.3 Instruments and Procedures

4.4 Scoring and Coding

4.5 Syntactic Complexity

4.6 Lexical Complexity

4.7 Accuracy

4.8 Fluency

Chapter 5: Results

5.1 Results Overview

5.2 C-test and Related Independent Variables

5.3 C-test and Unrelated Independent Variables

5.4 C-test Correlation with Independent Variables

5.5 Arabic Writing at Three Levels

5.5 Usefulness of Lemmatization

5.6 Usefulness of Hapax Index

5.7 CAF measures

5.8 Sentence-Based Measures

5.9 Regression Analysis

5.10 NS Writing and L1 Writing Ability

5.11 Genre Effect in Arabic Writing

Chapter 6: Discussion and Implications

6.1 Introduction
6.2 Research Question 1: Arabic CAF Measured at Three Levels ...................... 221
6.3 Research Question 2: Independent Variables that Explain Written Arabic CAF .... 234
6.4 Research Question 3: Genre Effect on Written CAF ...................................... 237
6.5 Limitations ........................................................................................................... 241
6.6 Implications .......................................................................................................... 243

Chapter 7: Conclusion .................................................................................................. 252

References .................................................................................................................... 255

Appendix A: Arabic C-Test .......................................................................................... 279
Appendix B: English Background Questionnaire .......................................................... 283
Appendix C: Arabic Background Questionnaire ............................................................ 288
List of Figures

Figure 1: Arabic as an nth Language .................................................................................. 81
Figure 2: Arabic Learner Educational Status ..................................................................... 92
Figure 3: Languages Spoken or Studied in Addition to Arabic ......................................... 93
Figure 4: Average Time for Completion of Writing Tasks (in minutes) .............................. 94
Figure 5: C-test Score vs YoS .......................................................................................... 150
Figure 6: Fluency ............................................................................................................. 176
Figure 7: Syntactic Complexity ......................................................................................... 179
Figure 8: Lexical Complexity ............................................................................................ 182
Figure 9: Accuracy (EfT/T) ............................................................................................. 185
Figure 10: Words per Sentence (W/S) ............................................................................... 188
Figure 11: Coordination (T/S) ........................................................................................ 191
Figure 12: Fluency Across Genre ..................................................................................... 209
Figure 13: Syntactic Complexity across Genre ................................................................. 211
Figure 14: Accuracy across Genre .................................................................................... 214
Figure 15: Lexical Complexity across Genre .................................................................... 217
List of Tables

Table 1: "dispersing the protesters" ........................................................................................................... 64
Table 2: Modified Arabic Deletion Strategy .............................................................................................. 66
Table 3: “complaint” Variants—All Newspapers .......................................................................................... 70
Table 4: “complaint” variants—Moroccan at-Tajdid .................................................................................. 71
Table 5: Use of جهوي jihawiyy by Region ................................................................................................... 72
Table 6: Candidate Texts ............................................................................................................................. 76
Table 7: 10-Text C-test .................................................................................................................................. 83
Table 8: Key Item Quality Statistics ........................................................................................................... 84
Table 9: Final 5-Text C-test Item Quality Statistics .................................................................................... 85
Table 10: Self-Assessment Correlation with C-test vs. Reported Semesters of Arabic Study .... 86
Table 11: Summary of CAF Measures ......................................................................................................... 107
Table 12: Arabic T-units in Narration ......................................................................................................... 110
Table 13: Simple T-Units ............................................................................................................................. 113
Table 14: Simple T-Units ............................................................................................................................. 114
Table 15: Complex T-Units .......................................................................................................................... 115
Table 16: Most Frequent Types ................................................................................................................... 129
Table 17: NS Verbal Mood Error ............................................................................................................... 135
Table 18: Colloquial Lexical Items .............................................................................................................. 138
Table 19: Colloquial Lexical Items .............................................................................................................. 138
Table 20: Summary of CAF Measures ......................................................................................................... 147
Table 21: Learner C-Test Scores .................................................................................................................. 149
Table 22: Learner C-Test Scores According to Group Membership ............................................................. 152
Table 23: Learner Years of Study .................................................................................................................. 153
Table 24: Learner Self-Estimated Overall Ability ............................................................................................ 154
Table 25: Learner Self-Estimated Writing Ability .............................................................................................. 155
Table 26: Frequency of Interaction with Native Speakers .................................................................................. 156
Table 27: Literacy Practices ............................................................................................................................. 158
Table 28: Self-Estimated L1 Writing Ability ...................................................................................................... 160
Table 29: Correlation Coefficient between C-test score and Independent Measures ........................................ 161
Table 30: NS Persuasion Response .................................................................................................................. 164
Table 31: HP Persuasion Response .................................................................................................................. 165
Table 32: LP Persuasion Response .................................................................................................................. 167
Table 33: Un-Lemmatized and Lemmatized Versions of Guiraud’s Index ......................................................... 169
Table 34: Index of Lexical Hapaxes ................................................................................................................ 172
Table 35: Fluency (W/M) .................................................................................................................................. 174
Table 36: Syntactic Complexity ....................................................................................................................... 178
Table 37: Lexical Complexity ........................................................................................................................... 181
Table 38: Accuracy (EfT/T) .............................................................................................................................. 184
Table 39: Words per Sentence (W/S) ............................................................................................................... 187
Table 40: Coordination (T/S) ........................................................................................................................... 190
Table 41: Correlation Coefficient between C-test score and CAF ................................................................. 195
Table 42: Independent/Explanatory Variable Intercorrelation (n=35) .............................................................. 197
Table 43: Syntactic Complexity Regression Model ........................................................................................ 199
Table 44: Lexical Complexity Correlations .................................................................................................... 200
Table 45: Lexical Complexity Regression Model ............................................................................................ 201
Table 46: Accuracy Correlations ................................................................. 202
Table 47: Accuracy Regression Model ...................................................... 203
Table 48: Fluency Correlations ................................................................. 204
Table 49: Fluency Regression Model ......................................................... 204
Table 50: Correlations between NS Self-Estimate and CAF ....................... 206
Table 51: Fluency (W/M) across Genre ..................................................... 208
Table 52: Syntactic Complexity (W/T) across Genre ................................. 210
Table 53: Accuracy (EfT/T) across Genre ............................................... 213
Table 54: Lexical Complexity (Guiraud’s Index) across Genre ................... 216
Chapter 1: Introduction

1.1 Background

Although commonly described as a “less commonly taught” language, enrollment in university-level Arabic programs in the United States continues to be healthy (Goldberg, Looney, & Lusin, 2015), and indeed Arabic has seen a steadily increasing interest by applied linguistics researchers. Recent efforts have shed important light on issues such as Arabic acquisition during study abroad (Shiri, 2015; Trentman, 2013, Raish, 2015), Arabic-specific considerations regarding vocabulary acquisition (Moser, 2013; Ricks, 2015; Golonka et al., 2015), and aspects of the oral production of Arabic heritage speakers (Albirini, 2013; Albirini, Benhamoun, & Saadah, 2011). To date, however, Arabic remains under-researched in comparison with many other major languages, especially in terms of literacy development and literacy-related issues (Mahfoudhi, Everatt, & Elbeheri, 2011, p. 1012). In other words, much of the knowledge available to frame, structure, and guide the acquisition of other languages is simply unavailable for Arabic; this remains particularly notable in the context of writing.

Issues of Arabic literacy and literacy development are of undeniable importance to Arabic learners, educators, and millions of native speakers, but such issues have been relatively under-investigated in the context of Arabic second/foreign language (L2) acquisition. Although a few studies have approached the issue of Arabic writing among learners at different levels of achievement/proficiency (Shakir & Obeidat, 1992; Khaldieh, 2000), Arabic instructors, program administrators, and learners continue to lack useful information regarding the role of writing in Arabic instruction. What should communicative writing in Arabic “look like,” and how can it best be measured in an educational context? This dissertation attempts, in an albeit limited
fashion, to address these and other questions related to Arabic writing and literacy development among learners.

Recent years have seen developments in the field of Applied Linguistics that hold promise for the conceptualization and measurement of Arabic writing. Particularly, researchers have developed a range of empirical measurements of linguistic complexity, accuracy, and fluency (CAF) that provide direct access to procedural elements of linguistic production. The CAF framework is often employed in relation to Skehan’s (1998) Trade-off Hypothesis, in which learners are assumed to possess finite attentional resources available for devotion to either linguistic complexity, accuracy, or fluency in their L2 production. The CAF triad has also been used in conjunction with Robinson’s (1995, 2001) cognition hypothesis, which presents a framework for conceptualizing the cognitive demands posed by linguistic tasks, and argues for the careful sequencing of pedagogic tasks. Thus far, the bulk of research employing the CAF framework has investigated on oral, rather than written production (Byrnes & Manchon, 2014), in spite of the fact that the modality of writing is understood to play a facilitative role in the development of general L2 proficiency (Williams, 2012). To date, only a small number of studies have analyzed the linguistic production of Arabic learners within the CAF framework in either the oral or written modes.

The current study attempts to depict the varied writing competencies of Arabic learners and Native Speakers (NSs) according to direct measures of written CAF. This effort seeks to add to the nascent but growing body of literature concerned with writing in Arabic in several ways, namely by (a) expanding the use of CAF measurements to the analysis of written Arabic, (b) exploring the CAF measures posited to be most effective at discerning between the writing of L2 Arabic learners at different levels, (c) establishing NS norms for the selected CAF measures, and
(d) introducing a novel “short-cut” proficiency measure (C-test) for estimating the general proficiency level of Arabic learners. In so doing, this study sheds much-needed light on the development of written proficiency among Arabic learners, as well as the production of NSs when posed the same communicative tasks as learners.

1.2 Study Overview

This study relies on the measurement and manipulation of three interacting variables: global Arabic proficiency, estimated via an Arabic C-test and other independent measurements; direct measures of written CAF; and task variability (genre types) established via three written prompts designed to elicit different types of written production. Although these variables have seen limited exploration in the relevant literature with regard to Arabic acquisition, the studies that interrogate them have employed diverse and divergent methodologies, in turn complicating their interpretation. Researchers simply do not know what to expect from either NSs or learners when written production is elicited and analyzed in this way. Therefore, the current study employs a number of specific CAF measures that have been shown to be successful in the modeling of linguistic production in languages other than Arabic. Although these related studies have established that the L2 acquisitional process is generally marked by a predictable increase in fluency, accuracy, and complexity, the exact nature of this process among Arabic learners remains unclear.

The first phase of this study consists of developing the Arabic C-test for use as an independent measure of global proficiency. This multi-step process of test development and validation entailed multiple rounds of consultation and piloting with NSs, followed by pilot testing with learners and selection of the 5 C-test texts that ultimately formed one of the
experimental tasks presented to participants during the study’s second phase. The challenges presented by the Arabic language and script to C-test development are numerous and include regional lexical variation, Arabic derivational and inflectional morphology, the general lack of learner-appropriate authentic texts at the lowest levels of difficulty, etc. However, each of these challenges to Arabic C-test creation can be minimized by following a thorough program of C-test development that includes reliance on native speakers to identify problematic items and ensure the suitability of texts, corpus research to identify alternative acceptable forms, and the modification of original texts to make them more suitable for both the C-test format and Arabic learners, who themselves may have been exposed to a very limited sample of authentic texts and genres at the time of testing. The resulting C-test was subsequently used as an independent, concurrent estimate of learner Arabic proficiency.

The second phase of this study entailed the collection of written responses from 35 Arabic learners and 14 Arabic NSs to a series of written prompts. The selection of writing prompts was motivated by the desire to target a variety of linguistic structures and strategies while at the same time maintaining an ecologically sound experimental design. The current study does not rely on holistic measurement of the quality of participants’ written products per se (cf. Wolfe-Quintero, Inakagi, & Kim., 1998), rather it will investigate the relationship between learners’ scores on a number of direct CAF measures and several global proficiency estimates (e.g., C-test scores, self-estimates, and reported years of study). A subset of the learners who participated in this study provided follow-up interviews to confirm their perceptions of task item difficulty.

All participants were likewise asked to supply information regarding their literacy practices in Arabic. Finally, Arabic learners were asked to respond to the Arabic C-test, resulting
in a numerical score representing overall proficiency being associated with each learner. The specific operationalizations of the various CAF measures applied to learner and NS writing in this study are based in part on the production patterns of NS participants, in addition to available literature, and are therefore illustrated with copious examples, particularly in the discussion of accuracy, lexical complexity, and syntactic complexity, in order to allow for the replication of the selected measures in future investigations. Taken together, these sources of information will help to build a picture of developing Arabic L2 writing proficiency.

The analysis employed in this study relies on a comparison of learner performance according to a number of direct measures of CAF with independent variables such as reported Years of Study, C-test score, and self-assessments. First, the results of the learner-completed C-test were examined, and learners were broken into a High Performance (HP) and a Low Performance (LP) on the basis of C-test score. Following this, the various collected CAF measures are explored in detail, revealing that NSs are indeed more complex, accurate, and fluent writers of Arabic than learners, although the HP and LP groupings are shown to correspond with predicted divergence in performance according to these same measures. This analysis likewise investigates the usefulness of a lexical lemmatization process, a hapax index of lexical complexity, and two sentence-based measures of CAF, determining that none of these are beneficial for CAF-based analysis of the Arabic writing collected in this study. Next, four standard multiple regression models are presented that explain differential CAF performance in relation to independent measures of Arabic proficiency. Regression analysis reveals the significant finding that C-test score is consistently shown to be the best predictor of variation in the selected CAF variables. Finally, this study explores the effect of writing genre on CAF,
finding that writing in the persuasion/argumentation prompt is marked by more syntactically complex, but less accurate writing in all groups.

This study concludes with a discussion of the key limitations and implications associated with the current effort. Notably, the conclusions supported by this study are based on a limited pool of participants. This in turn weakens any statements that can be made on the basis of the correlations and regression models presented below. Additionally, this study investigates written CAF using a single measure per construct. However, the measures discussed below nevertheless prove to be successful at depicting differential rates of written CAF among learners and NSs, and most show statistically significant relationships with the selected independent variables. The results of this study argue strongly for a reconceptualization of the role of writing in the Arabic curriculum—rather than representing merely an avenue through with to reinforce grammatical or lexical knowledge, writing is a key process through which Arabic should be acquired. Arabic learners appear to acquire the structures accessed by measures of written CAF in a way that significantly corresponds with their own increasing self-confidence, as well as overall proficiency. However, the impressive diversity of Arabic NS writing practices noted by this and other studies indicates that Arabic learners would benefit from a “realistic” model of high-proficiency writing as it occurs in varying communicative contexts.
Chapter 2: Review of Literature

2.1 Approaches to Measurement and Assessment of L2 Arabic

The question of what makes a foreign/second language user “proficient” in a given language, as well as how to best measure this proficiency, is both theoretical and practical in nature. This measurement, and the constructs that it is claimed to target (e.g., proficiency, fluency, facility, etc.) is an issue that has been a source of serious debate among researchers in fields as diverse as linguistics, psychology, and neurology. The ability to effectively and accurately measure L2 proficiency is of crucial importance to issues such as language program evaluation, L2 acquisition research, and to high-stakes decisions determining whether an individual has reached a particular proficiency level (Bachman & Clark, 1987, p. 21). The latter might include decisions on whether to grant or withhold credit, whether to hire a given individual, decisions regarding accreditation, salary-related decisions, among others. In the context of Arabic language education and research, practitioners currently have a number of assessment formats at their disposal, whether it be in the form of end-of-unit exams tied to a textbook, standardized proficiency assessments such as the Oral Proficiency Interview (OPI), or commercially available tools such as the Computer Assisted Screening Tool (CAST) (Malone & Montee, 2010). However, many of these formats can be prohibitively expensive in terms of time and money to administer to individual learners. A number of additional approaches to the measurement of proficiency have yet to be expanded to Arabic, or have thus far only been used for research purposes.

Within the context of Arabic instruction, widely used assessment practices and formats have traditionally emerged from prevailing trends in second/foreign language education in general. Until the emergence of the audio-lingual and communicative methods of L2 instruction
in the 1960’s and 1970’s, tests in and of Arabic were primarily translation based, as Arabic instruction itself was largely geared towards training learners to read and translate texts (Rammuny, 1999). Although Arabic instruction gradually moved towards an emphasis on oral skills, course materials and assessment practices remained rooted in literary and written forms of Arabic (Ryding, 2006). The effort to empirically measure and predict L2 proficiency dates back to the era following World War II, in which various entities in the United States government invested considerable resources in identifying and achieving language training outcomes (Sparks et al., 1997). Later, the American Council for the Teaching of Foreign Languages (ACTFL) and the Educational Testing Service were tapped by the Department of Education to expand the earlier work of the Interagency Language Roundtable and Foreign Service Institute to develop the ACFTL Proficiency Guidelines (Lowe & Stansfield, 1988). This effort was undertaken in response to a perceived lack of ability to assess general language proficiency on the part of language professionals employed by the US government (Liskin-Gasparro, 1984). The ACTFL Oral Proficiency Interview (OPI) is a proprietary instrument used by ACTFL to measure and rate language users’ oral proficiency. At their core, the ACTFL guidelines represent a series of “can-do” statements, proceeding from the activities and tasks that it is assumed that language users at each of the ACTFL levels are able to successfully accomplish. ACTFL proficiency guidelines and examples of learner production at each of the articulated levels currently exist for the four core skills of reading, writing, listening, and speaking.

The ACTFL guidelines and the OPI both emerged from the growing acceptance of a communicative model of language proficiency, articulated by Canale and Swain (1980), in which learners are not assessed on the basis of their declarative grammatical knowledge, or the ability to translate foundational texts, for example. Rather, in a communicative model of language use,
learners are ideally assessed on the basis of their ability to use the target L2 in a contextually appropriate manner. I.e., communicative assessment formats such as the OPI take into account “the importance of context beyond the sentence level to the appropriate use of language” (Bachman & Clark, 1987, p. 23). The communicative context of language use encompasses both the larger discourse in which individual utterances or sentences are embedded, as well as the sociolinguistic situation that affects the nature of that discourse. Arabic language professionals have long struggled to articulate an assessment framework for L2 Arabic education that is simultaneously capable of both accommodating the wide variety of registers, codes, levels, etc., that Arabic NSs use on a daily basis, as well as being realistically applicable to traditional classroom settings. For example, Eisele (2006) has proposed that Arabic educators adopt a flexible assessment framework that takes into account learning context, target country/Arabic variety, and learner goals, allowing for the conflation of achievement and proficiency tests in order to inform learning content and the achievement of the same.

The ACTFL proficiency guidelines and the OPI have undeniably been positively influential in the relatively quick spread of proficiency-oriented instruction in the United States, and in the TAFL field in particular. Indeed, the growing influence of the ACTFL guidelines led to the elaboration of guidelines specific to MSA, as well as the elaboration of proficiency-oriented instructional materials for learners of Arabic (Barhoum, 1989). However, the OPI has been critiqued for being time-consuming and costly to administer and rate, as well as for being predicated on a series of guidelines that lack an empirical basis (Liskin-Gasparro, 2003; Norris, 1997). In addition, OPI scores are known to be subject to extra-linguistic factors such as anxiety (Young, 1986; MacIntyre, Noels, & Clément, 1997) and subjectivity among trained raters (Brown, 2005).
Many have likewise noted that, in the case of Arabic, “MSA-only” may be an inappropriate approach to the assessment of oral proficiency (Rammuny, 1999, p. 165). After all, testing for oral proficiency conventionally includes the assessment of functions such as correct interpretation of conversational pragmatics, socially and culturally appropriate wording and content, and knowledge of the subject matter under discussion (Cheng et al., 2009). In Arabic, “native-like” performance of these functions certainly requires some facility with registers other than MSA. However, Belnap and Abuamsha (2015) have noted that, in an Arabic OPI, test takers who do not use the variety of language implicitly or explicitly requested by interviewers may be disadvantaged. Interviewers are known to sometimes insist on MSA, which may contradict a test taker’s individual language background, the relative linguistic and social appropriateness of MSA in a given communicative context, and ACTFL’s own guidelines for assessing Arabic that stress accommodation.

The ACTFL guidelines do not represent the only attempt to grapple with the difficulty of quantifying the performance of language users within the construct of communicative competence (Bachman, 1990; Bachman & Palmer, 1996; Canale & Swain, 1980) or general language proficiency. The Common European Framework of Reference for Languages, in contrast to the ACTFL guidelines, represents more of a policy document expressing goals for instruction and learning. Hulstijn (2011) refers to the constructs of Basic Language Cognition (BLC) and Higher Language Cognition (HLC), with BLC being common to all native speakers and entailing mainly core components of L2 proficiency and “purely linguistic competencies”, and also being limited to the modality of oral language (speaking and listening). HLC is argued to be related to the subject of discourse (“topics addressed in school and colleges”), modality of discourse (speech or writing), and the use of uncommon lexical items or morphosyntactic
structures. Importantly, Hustijn holds that HLC processes can be used to distinguish between native speakers, rather than solely between native speakers and learners.

This delineation of levels of L1 Proficiency actually conforms, to a certain degree, with the way in which Arabic proficiency has often been portrayed in the field of Arabic instruction. For example, Wahba (2006) has argued for the concept of the “educated native speaker” as a theoretical target for Arabic learners, stating that education is the single most important factor influencing Arabic speakers’ “ability” to use different forms of discourse. Importantly, however, the process of elaborating speech levels in Arabic has traditionally relied upon ascribing socioeconomic values to speakers (i.e., “educated”) rather than describing attributes of the discourse itself as in the BLC/HLC system described by Hultsijn. On the other hand, Tracy-Ventura, McManus, Norris, and Ortega (2014) argue for literacy-independent measures of proficiency, claiming that literacy-dependent measures may in fact be tapping into processes governed by HLC, rather than BLC, and may be unreliable measures of the types of BLC assumed to separate NSs from L2 learners. This may be doubly true for Arabic, in which the forms and items typically used in the written and oral modes can be quite divergent.

Currently, the topic of oral proficiency and proficiency testing dominates the discussion of assessment in the TAFL field, however there is no current consensus for how best to assess communicative competence in language learners (Kuiken, Vedder, & Gilabert, 2010, p. 83). Although the ACTFL guidelines and the OPI are widely used in Arabic educational contexts in the United States (Abdalla & Al-Batal, 2011), language program administrators and instructors have noted a lack of tools, knowledge, and resources for effectively utilizing them (Al-Batal, 2007, p. 270). In a survey of Arabic instructors in the United States, Abdalla and Al-Batal (2011)
found that over 40% of Arabic instructors reported having received “no training of any kind in proficiency testing” (p. 15).

2.2 Alternative Testing Methods

Given these challenges, Arabic language educators, researchers, and program administrators charged with assessing the varied Arabic competencies of their students may choose to eschew institutional use of the OPI and instead incorporate one of several alternative, often highly abstract forms of proficiency assessments. Many of these can be both easily scored and easily administered to large numbers of test takers, thus avoiding many of the major costs associated with the OPI. Recent years have seen the development of a number of so-called “short-cut” measures of L2 proficiency, such as the Cloze, C-test, Elicited Imitation Test (EIT), and vocabulary size test, which aim to provide estimates of test takers’ global proficiency and related competencies without resorting to the elicitation of a large amount of oral or written production.

Arabic language professionals needing to quickly (and roughly) estimate a learners’ or group of learners’ facility in Arabic have access to a number of formats developed for research purposes in Arabic or other languages. For example, the EIT format, in which test takers are rated on their ability to repeat increasingly long chunks of the target language currently exits for Chinese (Wu & Ortega, 2013) and French (Tracy-Ventura et al., 2014) and seems to hold promise for the measurement of competency in potentially any variety of spoken Arabic. Cheng et al. (2009) describe the development of the commercially available Versant Arabic Test (VAT), which distinguishes between “facility” and “proficiency” in Arabic. Facility is argued to be limited to the ability to understand and appropriately respond to everyday topics in a spoken
language, while “oral proficiency,” is a broader construct that “includes additional aspects such as correct interpretation of the pragmatics of the conversation, socially and culturally appropriate wording and content and knowledge of the subject matter under discussion” (p. 1). Ricks (2015) presents the development of an automatically scored vocabulary size test for Arabic that allows for the quick estimation of learners’ and native speakers’ Arabic lexical knowledge. The VAT, EIT, and vocabulary size test can be used to separate test takers into spectrums of proficiency (useful in placement-related decisions, for example) even if they do not rest on analysis of the rich and varied production elicited by the OPI or a related oral or written test.

Spolsky (1971) introduced the concept of Reduced Redundancy tests, which function on the principle that by adding “noise” to L2 content (i.e., removing information), test designers can tap in to an individual language user’s ability to manipulate, produce and reproduce the L2. Reduced-redundancy tests include the Cloze format, in which a number of whole words (e.g., every seventh word) are deleted from an intact stretch of discourse and test takers are ask to restore them. Oller (1979) argues that Cloze tests target what he termed a grammar of expectancy, or “g,” referring to a learner’s ability to predict the L2. For Davies (2001b), prediction is “the central element in language performance and … a test which captures that prediction is likely to be more valid than one that does not” (p. 358).

Cloze tests have seen some use in the measurement of literacy-related Arabic competencies (Abu-Rabia & Siegel, 2003; Parry & Child, 1990; Noor, 2007), as well as in isolated studies of Arabic L2 production (e.g., Mansouri, 1995). A similar approach, the C-test, in which learners restore large numbers of half-deleted words in a series of texts, has likewise been explored in the context of L2 Arabic assessment (Khoshaba, 2004). Operating on the basis of the Reduced Redundancy Principle (Babaii & Ansari, 2001), Cloze and C-tests require test
takers to reconstruct information that was removed from the target L2 input by test designers. Reduced redundancy tests are argued to discriminate between learners at different levels, as well as between high achieving learners and NSs on the basis of these restorations. The C-test, while currently largely restricted to the testing of learners of German and English, has been shown to correlate well with other measures of L2 proficiency in a number of studies (Eckes & Grotjahn, 2006; Daller & Phelan, 2006). Arabic, by nature of its character-based script and relatively standardized orthography (in MSA), lends itself to C-test development projects such as that described by Lee-Ellis (2009) for Korean. The processes of developing and validating an Arabic C-test are described in the subsequent chapter.

2.3 Challenges in the Assessment of Arabic Writing

Most of the assessment formats mentioned above have primarily been applied to oral production, or to a general, “latent” construction of L2 proficiency. However, the measurement and assessment of learner writing remains important for researchers, language instructors, program administrators, etc. In addition to representing a key avenue through which learners’ interlanguage is developed (Williams, 2012), communicative writing¹ can be expected to be the main modality through which modern language learners interact with target cultures located overseas. Currently, though, commercial and theoretical options available to Arabic language professionals for the assessment of learner writing are limited. Commonly used in placement-related decisions, the New York University Foreign Language Proficiency Test for Arabic includes a writing component, as do the Arabic Proficiency Test (APT) and the STAMP 4S

¹ “Communicative writing” refers here to writing composed in the context of tasks requiring the exchange of information, ideas, and opinions via the medium of written language, rather than loosely or un-connected sentences composed by learners to demonstrate or refine knowledge of a grammatical or lexical item (Shih, 1999, pg. 21).
(Standards-based Measurement of Proficiency; Clark, 2012) offered by Avant Assessment. ACTFL, via Language Testing International, offers online and offline versions of the Writing Proficiency Test (WPT) for Arabic, scored according to the ACTFL guidelines. Both the STAMP 4S and the WPT are proficiency-aligned instruments that provide holistic, human-scored ratings of test takers’ writing. Beyond these commercially available instruments, however, individuals concerned with measuring the writing competence (communicative or otherwise) of Arabic learners have little to go on in the way of published studies, or even measurement constructs to conceptualize the posited development of Arabic writing ability, which in turn remains a particularly murky subject.

Certainly, producing written Arabic is a taxing activity, especially at the lower levels. Even the mechanics of writing need to be re-learned by most Arabic learners as they adapt to the novel script written in the opposite direction than they are used to. Furthermore, producing written versions of the forms that learners have studied in class is more difficult than receiving them via input, because production requires more knowledge about the form in question. For example, learners may very well recognize the item bint “girl” when they hear it, but correctly producing it in a written (or spoken) passage requires the ability to inflect it for definiteness (al-bint), number (banaat), and often case (e.g., bint-un, bint-an, etc.). The development of L2 writing requires rich and varied sources of input, however the input to which learners have been exposed at the time of data collection is unclear. Even in a curricular context in which different sections of an Arabic course progress through the same material, the amount of time devoted to individual readings, the types of outside texts introduced, and the emphasis placed on writing practice is highly variable.
ACTFL attempts to elucidate the differences between the writing of learners at different levels of the writing proficiency scale in the Arabic-specific notes for the 2012 ACTFL writing proficiency guidelines. These notes point out that that writing is traditionally the least-emphasized of the four skills in the Arabic classroom, and that efforts at writing assessment fall far behind those made at assessing oral proficiency. ACTFL provides some information to Arabic language professionals about the writing performance of Arabic language users at each of the main ACTFL proficiency levels, however this information is either scant, vague, or both. For example, readers are informed that writers at the distinguished level are able to use “MSA and/or vernacular in accordance with contextual requirements of usage” in order to sustain distinguished-level performance across all tasks. The Arabic-specific notes for writers at the various Novice sublevels are more complete, stating that writers at the Novice Mid level, for example, use “short separate sentences with no specific order,” have “serious communication problems beyond formulaic language,” and are “only aware of semantic value of words; not grammatical function” (ACTFL, 2012). However, what are Arabic language professionals to make of these statements when assessing learner writing in or outside of the classroom? And how can researchers hope to quantify them?

Insights from research into languages other than Arabic have the potential to inform the empirical assessment and measurement of written L2 Arabic production. It is known, for example, that lexical measures are highly predictive of holistic ratings of L2 writing (e.g., Crossley et al., 2011; Ferris 1994), and it is also clear that written discourse exhibits a much wider lexical range than spoken discourse (Nation, 2006). Clearly, speaking and writing differ significantly in terms of process and product, although a number of researchers argue that similar measures should be applied to both modalities (e.g., Serrano, Tragant & Llanes, 2012). Studies
of writing proficiency often include holistic measures of learner production (e.g., Sparks et al., 1997). Researchers may also choose to incorporate concurrent, independent measures of L2 proficiency, which may take the form of an OPI, Cloze test, C-test, or a commercially available format.

As for the written production of advanced L2 learners, context clearly plays a role in writing development. Li and Schmitt (2009) studied the writing of an advanced Chinese learner of English studying in an English-speaking university over the course of a year, finding that the learner in question acquired lexical phrases, used them with increasing appropriateness, and became a more confident writer over the course of study. However, even advanced learners remain separated from native speakers in crucial (and measurable) ways. One key area of divergence is lexical knowledge, which might be represented in terms of diversity of forms employed in writing. Graduate-level English learners, for example, have been shown to lack derivative knowledge (i.e., the ability to derive different forms) of academic-level vocabulary items (Schmitt & Zimmermann, 2002). Even among L1 writers, it is clear that academic discourse is not the first natural language of any writer (Tang, 2012). Instead, academic writing is a specialized competence which both native and non-native speakers must acquire (Hyland, 2016), typically through education.

Within the context of Arabic education, the teaching and assessing of writing is a tricky, and potentially ideologically charged practice. Due to the fact that MSA is viewed by many as representing the path of least resistance towards oral proficiency (Daher, 1983, p. 102; Mansouri, 2000, p. 16), many Arabic programs in the United States have responded to the advent of the proficiency movement by articulating a “proficiency through MSA” philosophy (Rammuny, 1990). Thus, it is typically assumed in the case of Arabic that L2 Arabic writing—
communicative or otherwise—will be taught and assessed only within the relatively limited register of MSA. In reality, however, Arabic NSs employ a wide range of codes, registers, levels, scripts, etc., when writing (e.g., Palfreyman & Khalil, 2003; Warshauer et al., 2002). Even the history of print media in the Arab world is more complicated than might be assumed—Fahmy (2011) convincingly establishes the fact that printing newspapers in colloquial Egyptian Arabic is almost as old a practice as printing them in MSA. More generally, it is known that MSA is not the native language of any population of Arabic speakers (Al-Wer, 1997), as it is largely restricted to the context of written discourse and broadcast media (McCarus, 2011) and typically acquired by NSs through the process of formal education (Taha, 2013). Even within MSA, there is a great deal of variation between countries and communities of use (Ibrahim, 2009; Parkinson, 2010).

To date, a number of studies have investigated limited aspects of either L1 or L2 Arabic writing remains limited. Shakir and Obaidat (1991), discussed below, present and early attempt to quantify aspects of written L2 Arabic production, finding a positive effect for a term of university Arabic study. Al-Semari (1993) found that Saudi L1 Arabic writers used similar revision strategies when composing texts in either English or Arabic. Similarly, Noor (2007) found that Saudi participants’ ability to successfully complete a Cloze test in Arabic partially predicted certain elements of their written English compositions. More generally, it is known that the literacy practices of Arabic NSs are often highly diverse. Hallajow (2016) describes the various codes and languages (including colloquial Arabic and English) used by Syrian university students online, and Bianchi (2013) similarly explores the social and communicative role of alternate script choices by L1 Arabic writers in an online environment. If the “educated native speaker” (Wahba, 2006) is taken as a model for L2 Arabic acquisition, the varied communicative
and digital literacy competencies of Arabic NSs should be addressed. These and similar studies argue strongly for a conceptions of L1 and L2 Arabic literacy that are capable of incorporating this diversity.

2.4 The Complexity, Accuracy, Fluency Framework

The abovementioned move towards communication-bound, holistic measurements of proficiency along a scale such as the ACTFL oral proficiency guidelines necessarily raises the question of what these ratings mean in terms of quantifiable elements of learner’s production. The OPI is a particularly common and influential format in the context of Arabic education (Gebril & Taha-Thomure, 2014), however what are researchers to make of the statement that that a Superior speaker is more “fluent” than an Advanced speaker, for example? After all, the ACTFL guidelines and the CEFR both represent a series of “can-do” statements that describe the functions which language learners at various levels of proficiency can be expected to carry out, not statements about learners’ ability to produce a given number of words per minute, or sentences containing a given number of clauses, for example. Alternative assessment formats such as the Cloze, C-test, and EIT offer researchers ability to quantify and produce a distribution of learners’ varied Arabic L2 facilities, however these tests do not rest on the analysis of discourse or communicative production.

In recent decades, applied linguistics researchers seeking to quantify elements of production in learner discourse elicited by tasks such as the OPI have developed a number of discrete, analytic measures that are more fine-grained than holistic ratings. These measures are used to represent the quality sub-discursive elements of learners’ interlanguage in terms of the posited complexity, accuracy, and fluency (CAF) of production. The CAF triad is particularly
influential in the fields of applied linguistics and second language acquisition, in which researchers are often concerned with quantifiable and empirically valid measurements of learners’ interlanguage. To date, evidence exists that the multiple grammatical structures accessed by CAF analyses seem to be acquired by L2 learners in a predictable developmental order (Purpura, 2014), although it has conversely been argued that a given learner’s performance according to CAF measures is highly individualistic and may differ greatly from the group average (Larsen-Freeman (2006)).

Direct measures of CAF, in contrast to the holistic/global ratings offered by an OPI, are able to shed light on the developmental processes that L2 learners undergo. The CAF framework sees particular use in conjunction with Skehan’s (1998) Trade-off Hypothesis, which states that learners engaging in communicative tasks are forced to allocate differing amounts of their mental resources when attending to the complexity, accuracy, and fluency of their production. Furthermore, an increase in resources devoted to any of these three areas is understood to result in a corresponding decrease in measures pertaining to the remaining areas. CAF measures also see use in conjunction with Robinson’s (1995, 2001) Cognition Hypothesis, which provides a framework for conceptualizing elements of task complexity and argues for a careful complexity-oriented sequencing of pedagogic tasks. In general, researchers posit a balance between form (complexity and accuracy) and fluency, with increased attention to one leading to decreased effectiveness in the other (Skehan, 2009). In order to quantify these interacting constructs, researchers frequently rely on direct CAF measures that take the form of “ratios, frequencies, or formulas,” (Norris & Ortega, 2009, p. 1). These measures, when applied to both written and oral L2 production, are likewise thought to represent global estimates of learners’ L2 abilities.
CAF measures are more common in the field of SLA research, while holistic approaches such as the ACTFL OPI are often used to make decisions about individual learners (i.e., granting credit, graduation, etc.) or on the program level. The widespread adoption of the CAF model in applied linguistics research is due in large part to the fact that, per Polat and Kim (2014), it “provides a sophisticated framework for investigating the multicomponential nature of language use and development” (p. 186). Currently, in spite of the undeniable utility of both holistic ratings such as OPI ratings and discrete measures such as CAF measures, the relationship between the two remains somewhat unclear (Kuiken et al., 2010). Finally, the relationship between oral CAF and written CAF is likewise muddy. Although the majority of studies employing CAF measures to date have investigated oral production, a few studies have established that each modality can elicit differential performance from learners. Recent studies have found, for example, that L2 learners may produce more grammatically complex language in the written mode than in the oral (Hakansson & Norrby, 2007), that L2 learners may make more syntactical errors when writing than when speaking (Agren, et al., 2012), and that lexical complexity can be much higher in the written mode. I.e., L2 learners use significantly more diverse vocabulary when writing about a topic as opposed to speaking about it (Agren et al., 2012).

2.5 Complexity

Complexity in L2 production refers to the “size, elaborateness, richness and diversity of the learner’s linguistic L2 system” (Hausen & Kuiken, 2009, p. 464, emphasis in the original). Complexity can be viewed as “the use of sophisticated forms (e.g., past passive modals), complex constructions (e.g., subordination), and various other late-learned production units”
(Purpura, 2013, p. 119). On the syntactic level, complexity might refer to amount of subordination or other clausal measures, or mean length of unit of production (sentence or T-unit), while lexical complexity is often operationalized in the form of ratios such as lexical diversity, sophistication, and uniqueness. It is generally assumed that learners’ linguistic production becomes more complex as they become more proficient in a language (i.e., as their interlanguage develops). However, the measurement of complexity is recognized as the most nebulous leg of the CAF tripod (Hausen & Kuiken, 2009). At its heart, complex L2 production is that which is “elaborate and varied” (Ellis, 2003, p. 340), and the production of complex language is achieved through “expanding or restructured second language knowledge” (Wolfe-Quintero et al., 1998, p. 4).

2.5.1 Syntactic Complexity

Syntactic complexity can be operationalized in a number of different ways, although these measures often take the form of ratios based on a unit of measurement, e.g., clauses, sentences, or T-units. Al-Jabr (2006) identifies three sources of syntactic complexity in Arabic and English texts, namely a) language-specific syntactical features, e.g., whereas English tends to be characterized by a high degree of subordination, Arabic prefers coordination, b) text genre, and c) “the particular author’s (idiosyncratic) lexico-grammatical preferences” (p. 206). In the case of written Arabic, we might add d) the authors underlying level of Arabic proficiency (in the case of learners), and d) the author’s familiarity with the forms, structures, and genres typical of formal/written MSA.

In general, syntactic complexity is understood to increase concurrently with learners’ L2 proficiency. This has been convincingly shown to occur with clausal measures such as clauses
per T-unit, dependent clauses per T-unit, and dependent clauses per total number of clauses (Kuiken & Vedder, 2012, p. 145). Researchers may choose to view production as more syntactically complex if it contains a greater number of clauses per T-unit (Kuiken et al., 2010), or longer T-units in general (Gaies, 1980). However, syntactic complexity in the production of highly advanced L2 learners can sometimes appear to decrease (according to these ratio measures) as advanced learners employ alternative syntactic devices to complexify production (Norris & Ortega, 2009).

Researchers investigating oral data have a number of analytical units at their disposal, including the analysis of speech unit (AS-unit), which Foster et al. (2000) define as “an independent clause, or sub-clausal unit, together with any subordinate clauses associated with either” (p. 365). Likewise, the unit of measure might take the form of the utterance or turn, used to calculate Mean Length of Utterance (MLU) or Mean Number of Sentences per Turn. Other researchers have employed semantically-organized units such as the idea unit (Larsen-Freeman, 2006). The idea unit typically represents message segments that consist of a topic and a comment that are separated by the language user from contiguous units. Oral data also allows for the use of units organized around phonemic quality, such as the tonal-unit or phonemic clause (see Foster et al., 2000). In the case of written data, language users themselves might have the ability to determine this unit, e.g., via punctuation in the case of the sentence, although sentences are argued to be unreliable for the analysis of syntactic complexity for a number of reasons (Norris & Ortega, 2009). Finally, the units of production used in analysis might themselves be characterized by varying amounts of import, or weight, in learner discourse. This has led some researchers to analyze written L2 production, for example, according to relative T-unit weight (Evans et al., 2014).
In the written mode, a preponderance of studies analyze syntactic complexity via the T-unit. Originating with Hunt’s (1965) identification of the “minimal terminable unit,” (p. 37), the T-unit is commonly operationalized as “one main clause with all subordinate clauses attached to it” (p. 20). Researchers have some flexibility in determining the specific operationalization of T-unit used in individual studies. Hunt himself proposed several slightly divergent definitions of the T-unit, and separate definitions of the T-unit have been proposed for oral and written data (see Foster et al., 2000). Often, the T-unit represents the smallest structure that could be considered a grammatical sentence. This unit “contains one independent clause plus any number of other types of clauses, including adverbial, adjective and nominal clauses” (Wolfe-Quintero et al., 1998, p. 85). T-units are useful because they allow researchers to distinguish between simple and compound sentences, thus providing a more fine-grained basis of comparison between writing samples assumed to represent varying levels of proficiency (Schneider & Connor, 1990, p. 415). Foster et al. (2000), in their excellent discussion of units for the analysis of syntactic complexity, point out that T-units may be more suitable for the analysis of written, rather than oral data, due to the lack of additional phenomena such as pauses and intonation changes in writing (p. 363).

Once a stretch of discourse has been broken into units of analysis, such as the T-unit, complexity can be calculated on the basis of ratios incorporating that unit. In the domain of L2 writing, mean length of T-unit, or Words per T-unit (W/T), is the single-most-used measure of complexity (Norris & Ortega, 2009, p. 566; Ortega, 2003; Wolfe-Quintero et al., 1998). Some researchers argue that W/T actually accesses fluency, rather than syntactic complexity, due to the fact that W/T is a length measure and does not take into account how writers achieve length, but rather considers each unit semantically equal (Wolfe-Quintero et al. 1998). According to this
view, length measures such as W/T, and rate measures such as words per minute, both access the construct of fluency. However, Norris and Ortega (2009) argue that the mean length of T-unit measure accesses overall or general syntactic complexity. This type of measurement is held to be dependent on the measurement of “any length-based metric with a potentially multiple-clausal unit of production in the denominator” (pg. 561).

Gales (1980) has argued that analyses based on T-units are not suitable for writers at lower levels of proficiency. However, Ishikawa (1995) notes that the boundary between lower and upper levels of proficiency are not necessarily clear, and that analyses of smaller units such as the clause are abundant and can be carried out in a number of different ways. Ishikawa found W/T useful for accessing differences between learners at lower levels that were obscured by other measures. Ortega (2003) observed that an instructional period of roughly a year was required for substantial changes in the syntactic complexity of L2 writing to manifest.

Additionally, units of measure such as the T-unit can be combined with units that learners themselves define in written production in the form of sentence complexity measures such as number of T-units per sentence (Tsang & Wong, 2000). This measure can be easily calculated in digital written texts.

2.5.2 Measuring Syntactic Complexity in Written Arabic

Syntactic complexity in written and oral Arabic is a rich topic for investigation, as various forms of Arabic discourse can consist of a wide variety of syntactic structures and formulations. To date, several studies have interrogated this construct. For example, Tweissi (1990) investigated surface sentence structure in Arabic, finding that NSs use more pro-forms (i.e., to include otherwise optional subject pronouns) when speaking to NNSs. The author
speculates that NS use of pro-forms when speaking to NNSs was intended to aid non-native comprehension in interactions, in spite of the fact that Arabic is a pro-drop language; that is, subject pronouns are often optional (Farghaly & Shaalan, 2009). Fareh (1988) investigated lexical cohesion (the rate of use of certain cohesive devices per unit of analysis) in written Arabic texts, although this study was organized around the paragraph, rather than the T-unit.

In order to empirically measure aspects of syntactic complexity in Arabic linguistic production, however, a unit of measurement is needed. As mentioned above, the T-unit has been widely used in the analysis of L2 writing, although its use in the domain of L2 Arabic assessment remains extremely limited. Vann (1979) provides examples simple T-units in English texts, contrasting an adult’s utterance, “the man who I saw yesterday runs fast,” as opposed to a child’s imitation, “I saw the man and he runs fast” (p. 6, underline in the original). The former utterance is composed of a single T-unit while the latter contains two, even though both could be considered intact sentences on the basis of punctuation. Bardovi-Harlig (1992) describes the T-unit as one independent clause and its dependent clauses. Thus, the phrase “Bill got up and Mary sat down” consists of two T-units within one sentence. The Arabic equivalent ٠ قائم بيل وجلسات ميري can also be classified as consisting of two T-unit constructions in Arabic contained within a single sentence. While he does not explicitly employ T-units, Al-Jabr (2006) agrees that sentences are particularly unreliable units of analysis in Arabic texts, and proposes an alternative conception of “sentence” beyond the punctuation-delineated units of discourse traditionally defined as such. The alternative “sentence” described by Al-Jabr (2006) in his study “does not refer to the grammatical (nominal/verbal) unit but rather to a language unit that conveys a complete thought. As such, this unit can consist of one or a series of clauses” (p. 207). This proposal of a unit for the analysis of written Arabic appears to resemble both T-unit
and the idea unit, in that it is a semantically-oriented unit, but capable of encompassing a range of varyingly complex clauses.

Studies of written data may easily choose to employ measurements based on the sentence—i.e., a series of main and subordinate clauses necessarily separated by a period—however this unit may be problematic in the analysis of written Arabic data due to the large amount of variability in sentence type (complex, compound, simple, etc.) and length. Arabic sentences in particular are known to be longer than their English equivalents due to both “the absence of strict punctuation rules,” as well as conventions of written Arabic discourse that often result in the coordination of what in English would most likely be several sentences (Farghaly & Shaalan, 2009, p. 9). Thus, a “sentence” may span an entire paragraph and contain multiple T-units/clauses/etc., joined by cohesive devices used to create coordination and subordination.

Often, written data composed by learners is less than clean-cut, and may contain any number of fragments, salutations, “missing” pronouns, etc. These instances require individual researchers to make choices that emerge from their data. For example, take the following sentence composed by a lower-level writer in the current study in response to a description task: بنیات كثيرة، وكل البنايات قريب من غرفتي. bināyāt kathīra, w-kul l-bināyāt qarīb min ġurfatī “Many buildings, and all the buildings [are] close to my room.” In this selection, we find a fragment, “many buildings,” followed by an intact T-unit containing a gender-agreement error. What should be done with the fragment “many buildings”? Ishikawa (1995) does not include fragments or exclamations as T-units, while other researchers do. The researcher may choose to (a) count the fragment as an independent T-unit, (possibly inferring that the author had intended to express “there are many buildings,” (b) ignore the fragment, as it is not able to stand on its own as minimal sentence, or (c) include the fragment with the following T-unit. Compositions
would be considered less complex if analyzed according to the first option (more and shorter T-units) and more complex according to the third (fewer and longer T-units). It is then incumbent upon individual researchers to delineate and justify their choices in light of the nature of their data. The Methodology chapter of this dissertation describes the specific approaches to T-unit identification used in the present study of written Arabic.

A number of studies have employed T-unit analysis for various purposes in Arabic, and several of them offer insights for the current study. Albirini (2013) describes using T-units in the analysis of oral data produced by Egyptian and Palestinian heritage speakers of Arabic, calculating error-free T-units for accuracy and a T-unit complexity ratio for syntactic complexity. Albirini reproduces Young’s (1995) definition of a complex T-unit as being composed of “a matrix plus subordinate clause, two or more phrases in apposition, and fragments of clauses produced by ellipsis” (p. 38). Tweissi (1990) examined the “foreigner talk” (FT) produced by L1 Arabic speakers in recorded phone conversations with both other NSs, as well as L2 Arabic speakers. The Arabic used in interaction with NNSs was shown to be significantly less complex lexically, although Arabic FT was shown to be characterized by longer T-units than production in interaction with other NSs. Tweissi speculated that L1 Arabic speakers who participated in the study tended to provide more pre- and post-verb elaboration when speaking to foreigners (p. 313), resulting in longer T-units in their speech. A second speculation is that, unlike other examples of “foreigner talk”, Arabic FT simply tends to be more complex than ordinary speech, which would render this measure language specific within the context of FT.

In a unique study analyzing the relationship between L1 Arabic writing ability and L2 English writing ability, Noor (2007) reports asking a population of adult Saudi L2 English students to complete a Cloze test in Arabic and English, as well as prepare compositions in both
languages. However, the Arabic compositions were translated into English prior to analysis, and T-unit analysis was presumably only conducted on both texts in English. Both groups of texts (translated Arabic and L2 English) were then rated holistically for quality, as well as according to measurements of syntactic complexity (W/T) and accuracy. The author found that “higher competence in Arabic” (operationalized as participants’ score on the Cloze test) “has a significant positive effect on the overall performance of the students’ writings in English” (p. 421). This study is significant in its extension of the knowledge that L1 writing ability has on L2 writing performance to NSs of Arabic. However, it is limited by the fact that participants’ Arabic compositions were analyzed after being translated to English, limiting the extent to which statements about them can be trusted. Furthermore, the author did not adequately describe the process through which the Arabic Cloze test was created, stating merely that the “Arabic version of the cloze test was used to measure the Ss competence of Arabic. A passage from a reading book for the intermediate students was randomly chosen” (p. 414). Readers are not informed, however, whether the participants were familiar with the text in question, how many words were deleted from the text and at what frequency, whether the Cloze test was piloted before it was administered to participants, etc.

In an earlier longitudinal study of Arabic L2 writing development, Shakir and Obeidat (1991) present an attempt to quantify “maturity” in the writing of L2 Arabic learners. Utilizing a pre-/post-test design anchored to a term of study, the authors collected written compositions written by 18 adult university-level learners of Arabic produced during a standardized proficiency test. The authors calculated

a) number of words per text

b) number of simple sentences per text
c) number of compound, complex, and compound-complex sentences per text

d) number of missing words per test

e) number of cohesive connectives per text

f) the average score obtained on each test (partially reproduced from Shakir & Obeidat, 1991, p. 69)

The researchers found that, after the period of study, participating Arabic learners tended to produce significantly more words per text (i.e., they were writing more fluently), although the number of sentences barely increased (i.e., their writing was more syntactically complex). Sentences produced after the period of study were likewise generally more complex, i.e., displayed more coordination and subordination than sentences produced during the pre-test. The authors argue that Arabic “immature writing tends to be more repetitive, explicit, and concrete, whereas mature writing tends to be less repetitive and less explicit” (p. 76). This study represents a significant contribution to the study of Arabic L2 writing by delineating specific measures capable of quantifying Arabic learner production. Significantly, these measures were also tied to an external and concurrent measure of overall Arabic proficiency (the “AATA test,” presumably the APT discussed above). Concurrent measures of proficiency, while often time-consuming and costly for researchers to administer, ensure that results obtained can be generalized beyond the narrow context in which the data was collected. However, the use of the sentence as a unit of measurement limits the conclusions that can be drawn on the basis of this study.

Subordination and Coordination

A number of studies have tackled syntactic complexity from the standpoint of subordination, which is typically depicted as the joining together of multiple simple sentences by one or more conjunctive particles. Norris and Ortega (2009) describe measures of subordination
as deriving from the counting of all the clauses in a given piece of linguistic production and dividing them by a unit of production, such as the T-unit (p. 558). However, what constitutes a clause has not been clearly defined across disciplines or across languages (Yang, Lu & Weigle, 2015). In Arabic, Mansouri (2000) operationalized subordination as “the existence of a subordinate clause linked to the main clause by means of the equivalent of the English [that]” (pg. 35). Subordination in Arabic occurs through the use of complementizers, relative pronouns or indefinite relative clauses, and other conjunctive particles. Shakir and Obaidat (1991) found that “immature” Arabic written production tends to be less cohesive—writers “array facts in a list-like text,” and those conjunctives that are present are “overused”. The authors found that texts composed before a term of study were less frequent and restricted to ʔaw “or”, wa “and”, and li- or likay “in order to.” Post-test texts, however, contained more conjunctives overall, as well as more types of them. The complementizers ʔanna and ʔan introduce clauses that are used to “explain, justify or elaborate on what has already been said in the main clause” (Mansouri, 2000, p. 35). The nominal complementizer ʔanna affects the cases of the relevant sentence’s subject and predicate, while the verbal ʔan shifts the verbal mood of the verb it modifies.

Coordination, on the other hand, refers to the conjoining of main clause through the use of cohesive devices, such as a conjunction. For Norris and Ortega (2009), coordination is expected to be the best source of information about learners’ syntactic complexification at lower proficiency levels, while subordination is a more useful indicator for learners at intermediate levels (p. 563). Ruiz-Funes (2014) examined coordination in the writing of Spanish learners by operationalizing it as the number of T-units per sentence, in the assumption that coordination will decrease as L2 proficiency increases (Ruiz-Funes, 2014, p. 174; Wolfe-Quintero et al. 1998).
However, the structure of Arabic, and especially formal written Arabic, leads to complications in identifying and separating longer segments of writing into their constituent clauses. Thompson-Panos and Thomas-Ružić (1983) point out that the syntactic structure of Arabic complicates the use of clause-level analyses such as subordination. For example, both the antecedent and relative clause in Arabic can both stand alone as complete sentences (i.e., intact T-units); “neither is subordinate, at least not in the surface structure” (p.617). Thus, the Arabic relative clause is argued to function according to a coordinate, rather than subordinate relationship. Furthermore, and regardless of their classification for the purposes of analysis, many instances of subordination may have to be identified “by hand” by individual researchers, rather than through the use of computer software to search for a given structure. Although Arabic relative pronouns initiate relative clauses and are used to refer to discourse actors who have already been established in the preceding discourse (Mansouri, 2000, p. 36), they only appear in definite relative clauses (see Ryding 2005, p. 322). Indefinite relative clauses therefore need to be visually identified before a decision can be made regarding their relative independence as T-units.

2.5.3 Subordination and Coordination

A number of studies have tackled syntactic complexity from the standpoint of subordination, which is typically depicted as the joining together of multiple simple sentences by one or more conjunctive particles. Norris and Ortega (2009) describe measures of subordination as deriving from the counting of all the clauses in a given piece of linguistic production and dividing them by a unit of production, such as the T-unit (p. 558). However, what constitutes a clause has not been clearly defined across disciplines or across languages (Yang, Lu & Weigle,
In Arabic, Mansouri (2000) operationalized subordination as “the existence of a subordinate clause linked to the main clause by means of the equivalent of the English [that]” (pg. 35). Subordination in Arabic occurs through the use of complementizers, relative pronouns or indefinite relative clauses, and other conjunctive particles. Shakir and Obaidat (1991) found that “immature” Arabic written production tends to be less cohesive—writers “array facts in a list-like text,” and those conjunctives that are present are “overused”. The authors found that texts composed before a term of study were less frequent and restricted to ʔaw “or”, wa “and”, and li- or likay “in order to.” Post-test texts, however, contained more conjunctives overall, as well as more types of them. The complementizers ُنَانِةٌ ʔanna and ُنَآنِةٌ ʔan introduce clauses that are used to “explain, justify or elaborate on what has already been said in the main clause” (Mansouri, 2000, p. 35). The nominal complementizer ʔanna affects the cases of the relevant sentence’s subject and predicate, while the verbal ʔan shifts the verbal mood of the verb it modifies.

Coordination, on the other hand, refers to the conjoining of main clause through the use of cohesive devices, such as a conjunction. For Norris and Ortega (2009), coordination is expected to be the best source of information about learners’ syntactic complexification at lower proficiency levels, while subordination is a more useful indicator for learners at intermediate levels (p. 563). Ruiz-Funes (2014) examined coordination in the writing of Spanish learners by operationalizing it as the number of T-units per sentence, in the assumption that coordination will decrease as L2 proficiency increases (Ruiz-Funes, 2014, p. 174; Wolfe-Quintero et al. 1998).

However, the structure of Arabic, and especially formal written Arabic, leads to complications in identifying and separating longer segments of writing into their constituent
clauses. Thompson-Panos and Thomas-Ružić (1983) point out that the syntactic structure of Arabic complicates the use of clause-level analyses such as subordination. For example, both the antecedent and relative clause in Arabic can both stand alone as complete sentences (i.e., intact T-units); “neither is subordinate, at least not in the surface structure” (p.617). Thus, the Arabic relative clause is argued to function according to a coordinate, rather than subordinate relationship. Furthermore, and regardless of their classification for the purposes of analysis, many instances of subordination may have to be identified “by hand” by individual researchers, rather than through the use of computer software to search for a given structure. Although Arabic relative pronouns initiate relative clauses and are used to refer to discourse actors who have already been established in the preceding discourse (Mansouri, 2000, p. 36), they only appear in definite relative clauses (see Ryding 2005, p. 322). Indefinite relative clauses therefore need to be visually identified before a decision can be made regarding their relative independence as T-units.

2.6 Lexical Complexity

In addition to syntactic complexity, L1 and L2 written production is known to be characterized by variable rates of lexical complexity. As vocabulary/lexical knowledge is central to the act of writing, it is expected that learners who have access do differing amounts of that knowledge should tend to produce writing of varying complexity. Additionally, lexical measures are known to be extremely important in the context of holistic ratings of linguistic performance. Crossley et al. (2011), for example, report a strong link between lexical measures and holistic ratings of L2 writing. In a study that applied lexical measures to English L2 writing samples grouped according to proficiency level, this link was shown to be especially strong at the
beginning and advanced levels. Lexical knowledge has likewise been shown to be statistically related to measures of L2 reading comprehension and reading ability (Schmitt 2010; Milton, 2009; Jeon & Yamashita, 2014).

Within the context of Arabic, it is known that individual speakers can display differing amounts of lexical sophistication depending on the context of interaction. Tweissi (1990), for example, established that the speech of L1 Arabic speakers is significantly less lexically complex when produced in interaction with non-native Arabic speakers. Several researchers have likewise investigated the relationship between lexical knowledge and the Arabic acquisitional process (Khoury, 2008; Khaldieh, 1996). Golonka et al. (2015) found that the acquisition of vocabulary among Arabic learners is highly affected by factors such as learner effort and amount of context, although differences between experimental groups were minimized after a week, noting that repeated exposure facilitates vocabulary acquisition.

Lexical complexity has been operationalized in the form of a number of varied measures argued to target several sub-constructs. Lexical sophistication, for example, can measured by determining what percentage of the words in a given passage do not appear on a list of the most frequent written words in the target language. Laufer and Nation (1995) counted as “sophisticated” those items in participants’ writing samples that did not appear on a list of the 2000 most frequent words in the target language, however the exact cut-off point for “sophisticated” Arabic lexical items when examining learners’ production remains unclear. A possible starting point for researchers investigating lexical sophistication in Arabic is consultation of the most frequent items contained in A Frequency Dictionary of Arabic: core vocabulary for learners (Buckwalter & Parkinson, 2014). This dictionary was compiled via a corpus of Arabic that is almost entirely composed of written texts of various genres, making it
more appropriate for use in conjunction with learners’ written production than similar frequency lists based in whole or in part on oral data. Studies of lexical aspects of Arabic L2 writing may also choose to use measures such as the vocabulary size test developed by Ricks (2015), which could easily be used as an independent measure of a writer’s receptive lexical ability.

A second widely-used measure of lexical complexity, lexical richness or lexical diversity, is typically measured via a type/token ratio, in which the number of types of lexical items in a passage is compared with the total number of words (tokens) in that passage (Purpura, 2013). However, due to the fact that shorter passages have relatively few tokens in general, they tend to repeat types less frequently than longer passages, and a simple type/token ratio may artificially represent lower-level learners’ production as more lexically diverse than that of advanced learners. Therefore, a number of studies employ Guiraud’s index of lexical richness in which an algebraic adjustment for text length is made by dividing the number of types by the square root of the number of tokens (e.g., Kuiken et al., 2010). Guiraud’s index has been shown to successfully compensate for disparate text lengths, an important feature in studies eliciting production from learners at disparate levels of L2 acquisition. Tidball and Daller (2008) found that such measures of lexical richness are related to external variables such as teacher judgement of an essay’s overall quality.

However, the question of which words should be considered as representing different “types” is one that researchers have a great degree of freedom in determining. In English, for example, plural nouns might be considered as the same type as the corresponding singular form. It seems likely, for example, that a language user is calling on a similar area of lexical knowledge when producing both girl and girls, or book and books. Irregular plural forms in English are rarer, such as woman and women. In Arabic, however the majority of nouns are
pluralized according to a complex morphological system referred to as “broken plurals,” and the resulting singular and plural forms are often much farther apart (morphologically and orthographically) than their English, Spanish, German, etc., counterparts. Indeed, the complex process of producing plural nouns in Arabic has been investigated as a feature that separates heritage from native speakers of Arabic (Albirini et al., 2011).

2.7 Accuracy

Accuracy typically refers to the number (and often type) of errors that are produced by a language user in a given stretch of discourse within a limited amount of time. Although often undefined (see Polio, 1997), errors can be understood as deviations from the linguistic systems (phonological, morphological, syntactic, lexical, etc.) that an L1 language user might be expected to produce. This conceptualization of errors is in line with that employed by Ruiz-Funes (2014), following Yuan and Ellis (2003), in which an “error” is considered to be “any digression in syntactical, morphological, and lexical norms, but not in punctuation or capitalization (Ruiz-Funes, 2014, p. 174). Accuracy is likewise understood to be a key area in which even high-achieving L2 learners differ significantly from native speakers. In the oral mode, the learner-produced deviations from target language/NS norms that form the basis of accuracy measures are sometimes operationalized as “disfluencies” (e.g., Trenchs-Perrera, 2009), but in the written mode these are often referred to as “errors.” Accuracy is often operationalized in the CAF framework in terms of “errors” per unit of time or linguistic production, e.g., errors per sentence, errors per 100 words, or errors per T-unit (Kuiken et al., 2010), or it can entail the classification of errors according to type or severity (Abe & Tono, 2005).
There is a wide-ranging strand of inquiry into the nature of accuracy in written L2 production (Polio 1997), however the published studies investigating the written production of Arabic learners, either within the CAF framework or otherwise, remain limited. Early studies of L2 Arabic production such as those by Al-Ani (1972), Hana (1972) and Rammuny (1976) are situated within the error analysis frame and examine topics such as L1 interference and error classification. For example, Hana (1972) used a system of error classification to analyze the errors of Arabic learners reading aloud, representing an early effort at the quantification of Arabic L2 production. Al-Ani’s (1972) study of written L2 Arabic likewise relied on a taxonomy and analysis of learner errors, although this study found that errors in written Arabic are difficult to categorize. Rammuny’s (1976) study attempted to explain written errors in relationship to factors such as teaching pedagogy and language transfer. Although these error analysis studies contributed to knowledge about error production and the quantification of the same among Arabic learners, they did not expand to touch on other aspects of Arabic acquisition such as stages of interlanguage development, developmental sequences, etc. More recently, Brosh (2015) investigated orthographic errors in L2 Arabic writing, identifying nine categories of error that include practices such as “wrong letter choice,” and “omission of taa marbuuta” (pg. 592).

Naqvi et al. (2014) attempted to empirically demonstrate that the reversed scriptural directionality when comparing Arabic to English led to predictable patterns of error occurrence among adult Omani L2 learners of English completing written compositions in English, presumably contributing that the reverse is true, i.e., that English-speaking learners of L2 Arabic may also commit written errors in predictable ways when manipulating a reversed-directionality script. Mansourī’s (1995) study of the relationship between discourse cues and learner grammatical production was based in part on learner production in the form of written essays.
Importantly, advanced learners of Arabic were found to produce far fewer written linguistic errors than lower-level learners, confirming the fact that accuracy is a key measure through which advanced and beginning Arabic learners diverge. Mansouri’s findings argue strongly for a systematic, cumulative and predictable movement of Arabic learners from less complex syntactical structures to more complex ones.

In the context of L1 Arabic, it is known that even NS writers tend to predictably commit “mistakes” while writing, especially on the orthographic/spelling level. This is in keeping with previous studies that have established predictable rates of spelling errors among Arabic NSs. Some researchers have gone so far as to conclude that spoken Arabic and “Literary Arabic,” i.e., MSA, are processed as differently as first and second languages by NSs (Ibrahim & Aharon-Peretz, 2005), leading to divergences from the norms of MSA when NSs engage formal communicative writing. Previous study of spelling in Arabic has found that phonological (i.e., the substitution of one letter for a similar-sounding letter) errors are the most common type of error among Arabic-speaking children (Abu-Rabia & Taha, 2006).

Scott and Tucker (1974) introduced the measure of error-free T-units per total T-units (EFT/T) in their study of L2 English produced by L1 Arabic speakers, seeing improvement in both the complexity and accuracy of participants’ L2 English over a course of study. This measure is now widely used in CAF analyses of written L2 production (Polio & Shea, 2014; Lin et al., 2016; Larsen-Freeman, 2006). Wolfe-Quintero et al. (1998) claim in their meta-analysis of studies of L2 writing that the EFT/T measure is the most effective measure of written accuracy. They note that studies employing learners at different proficiency levels completing timed compositions have collectively found error-free T-units to be the most significant unit of analysis for accuracy.
Although the ratio of EFT/T is a commonly-used measure of written accuracy, researchers have noted a number of issues that complicate the use of this measure across proficiency levels. Evans et al. (2014) points to the difficulty in classifying what types of structures constitute an error, how to identify errors, and what to do when multiple errors occur in the same item. Early studies of written L2 Arabic conducted within the error analysis framework, for example, classified written errors according to a series of categories, e.g., orthographic/phonological, lexical, structural, and stylistic (Rammuny, 1976). Polio and Shea (2014) likewise question whether or not this measure is appropriate for lower-level learners, as beginning learners may make errors in every T-unit or, conversely, produce many more T-units overall than higher-level learners, thus attaining a higher score on this measure. Additional concerns arise when considering whether errors should be ranked according to type or severity. For example, Evans et al. (2014) point out that ratios that do not take into account the relative severity of different types of errors may be unreliable. Ishikawa (1995), for example, considers accuracy in combination with an analysis of discourse errors. Thus, per Polio (1997), the use of error-free T-units is a way to “get at the quantity of errors but not the quality” (pg. 114). Furthermore, as Polio and Shea (2014) note, studies examining differences between groups of learners may find no or little change between groups, as advanced learners may compose progressively longer T-units, but not more error-free T-units.

2.8 Fluency

Fluency refers to the ability of language users to rapidly produce language (Skehan, 1998). This construct can touch on how quickly (or slowly) individuals speak and write, and it likewise interacts with the constructs of accuracy and fluency according to Skehan’s (1998)
Trade-off Hypothesis—the more that L2 language users attend to speaking accurately, for example, the less attentional resources they are posited to have available to speak quickly, and vice versa. Fluency is a highly variable construct, and it intuitively “makes sense” that different individual language users will produce language at different rates. It is known, for example, that even L1 language users tend to vary in their rates of oral fluency (Derwing et al., 2009). In the context of L2 acquisition, even advanced L2 speakers have been shown to produce language more slowly than L1 speakers (Towell et al., 1996), making the construct of fluency potentially predictive of differences between learners and NSs.

Fluency has been analyzed in both the oral and the written modes, often using the simple and easy-to-calculate measure of words per minute (W/M) of production during timed linguistic tasks. This simple temporal measure of production can be further complexified, e.g., through the use of ratios such as syllables per second or minutes on task (Ellis & Barkhuizen, 2005). Ellis and Barkhuizen (2005) discuss the various quantitative fluency measures available to researchers. In general, these measures can be broken down into temporal measures of rate, quantity, and pausing (Ginther et al., 2010). Measures of rate fluency are typically based on the number of words or syllables produced per unit of time, such as by minute or by conversational turn. Ginther et al. (2010) find that speech rate is the most general and inclusive of these measures. Fluency has also been operationalized in terms of breaks in production. For example, a researcher may choose to investigate breakdown fluency by measuring the number or length of pauses in oral discourse (Hausen & Kuiken, 2009; Skehan, 2009). Similarly, repair fluency investigates learner behaviors such as “reformulation, repetition, false starts, and replacements” (Skehan, 2009, p. 513). Blake (2006) presents fluency in terms of temporal variables (e.g., the number of syllables per second or minute on a task), hesitation variables (e.g., number of false
starts, repetitions, reformulations, replacements, or other disfluencies), and the quantity of production (e.g., the response time or the number of syllables in a response).

Fluency and Complexity can be problematic constructs to measure and interpret in the written mode—researchers typically only have access to what students do on the paper (or within a word processor). Several of the rate measures described above, such as articulation rate (i.e., excluding silent pauses but including filled pauses in fluency measurement) are argued to represent the efficiency of the speaker in producing language, although many are not easily transferrable to the measurement of written data. Most measures are likewise unable to take into account any elements of planning or revision that might have a negative impact on measures such as words per minute while at the same time having a positive impact on other CAF measures. For example, Latif (2012) points out that, in tasks that require more planning, learners may appear to be more fluent according to this measure if they eschew planning. According to the Trade-off Hypothesis, cognitive factors “compete” with each other for learner attention as a learner composes a text, and increased attention towards accuracy may have a detrimental effect on linguistic fluency, for example (Skehan & Foster, 1997; Foster & Skehan, 1996). Therefore, it is not clear that written complexity and written fluency will increase together in a linear relationship in tandem with overall proficiency (Wolfe-Quintero et al., 1998).

In the modality of writing, Wolfe-Quintero et al. (1998) view fluency as tied to the number of words or structures employed over time. Some L2 writing researchers have expanded the study of linguistic dysfluencies from the oral mode (breakdowns, pausing, repetition, false starts, etc.) into the written, such as Ellis and Yuan’s (2004) measure of “reformulations”, (i.e., crossed-out and changed words) divided by the total number of words in the sample. Chenoweth and Hayes (2001) found that written fluency tends to substantially increase as learners have more
experience with the target L2. Perez-Vidal and Juan-Garau (2009) determined that the measurement of textual features related to essay length and amount of production over time are the most efficient measures of written fluency.

2.9 Genre Effect

Measures of linguistic complexity, accuracy, and fluency in written and oral production are not only subject to influence by learner-internal factors, such as overall proficiency level or Working Memory Capacity (WMC). Rather, they are likewise subject to learner-external factors such as task type and genre, presence or absence of planning time, and learner anxiety. In a writing task, these factors “could include familiarity with the topic, skill in writing, and communicative purpose” (Laufer & Nation, 1995, p. 308). Furthermore, researchers are able to manipulate task elements in order to affect learner performance on the analyzed CAF measurements. Skehan and Foster (1999) present a model of task complexity in which tasks that are more cognitively familiar to participants require less processing and thus produce more fluent output. Thus, a “personal task” built on information well known to participants presents a reduced cognitive load (p. 98). Although the effect of task type and task complexity on learner oral production has received a great deal of research attention in recent decades, the interaction between task design and written production is beginning to catch up (see Byrnes & Manchon, 2014; Kormos, 2011; Ishikawa, 2007). However, although task design/task complexity clearly plays a role in CAF measurement of learner production, this relationship remains under-researched in the modality of writing (Ruiz-Funes, 2014, p. 164).

Task design or task type has received particular interest by researchers and has been shown to have an influence on direct measures of L2 writing performance. Kaldieh (2000)
argues that written proficiency in L2 writing tasks is task-bound, indicating that distinct tasks targeting different genres or types of discourse have the potential to elicit data that contains different types of structures and discourse strategies. Carell and Conner (1991) have likewise argued that descriptive and persuasive tasks elicit different forms of learner production, and that persuasive tasks are relatively more difficult for L2 writers to complete. Evans et al. (2014) echo previous research (Lee & Anderson, 2007) that establishes that different writing topics can elicit different scores on CAF measures. Foster and Skehan (1996) similarly explore the complex ways in which task design affects the complexity, accuracy and fluency of learner output, finding that CAF measures are indeed subject to differing task design. For example, decision making tasks are “more interactive, and … [require] the capacity to relate a set of moral values to a series of decisions” (p. 98).

Description is understood to be a basic linguistic function that even lower-level L2 speakers and writers can be expected to have some familiarity with. This is reflected in ACTFL’s (2012) guidelines for speaking and writing proficiency, in which descriptive functions rank as among the first that language users acquire. Familiarity with the topic described also influences the performance of description, learners are more able to write effectively about familiar topics than unfamiliar ones (Huang, 2009). This corresponds with Skehan’s (2009) contention that experimental tasks based on familiar information tend to advantage accuracy and fluency over complexity. It might be expected, therefore, that Arabic learners completing a description task, in which they are asked to describe a familiar place, a family member, a pet, etc., may elicit more fluent and accurate, but less complex, production. After all, description stands as among the first activities that L2 learners engage in in the classroom.
Narratives likewise represent one of the most widely taught written genres of linguistic production in foreign language instruction, especially at the beginning level (Kormos, 2011, pg. 148). In the oral mode, narration has been found to emerge before description in some populations of L2 Russian learners (Rifkin, 2010). Among NSs writing in their own language, the narrative function is acquired at a young age and “heavily practiced throughout childhood in listening to and telling stories” (Kellog, 2001, p. 177). Tavakoli (2014) notes that, due to the typical amount of learner familiarity with the task genre, “narratives are considered as ecologically valid tasks that lend themselves well to teaching and research…” (p. 226). Among L2 learners, Skehan (2009) finds that narrative tasks tend to elicit more complex, but less fluent and accurate production, while tasks “based on concrete or familiar information advantage accuracy and fluency” (p. 511). Skehan and Foster (1999) likewise argue that the composition of narratives supported by visual material but that nevertheless requires a degree of organization of information is a less cognitively demanding task than decision making tasks. In general, tasks containing structure, such as the presence of a visual aid in a narrative task, are known to elicit greater accuracy and fluency than complexity in learner output (Skehan, 2009).

Argumentative or persuasive tasks are also widely used in L2 acquisitional and CAF research. The act of taking a position, justifying it, and attempting to convince an audience to take action based on this justification is widely regarded as among the most difficult linguistic functions a speaker/writer can be expected to successfully carry out in an L2. Yang and Sun (2012) argue that this is comparatively more difficult than other written functions such as narration and description. Furthermore, argumentative/persuasive skills are argued to develop according to a predictable sequence, and the linguistic function of argumentation to be language-specific (Golder & Coirier, 1996). This is due in large part to the fact that
persuasion/argumentation requires the use of linguistic structures as well as genre-specific rhetorical devices (Ferris, 1994).

The language-specific strategies required for successfully editorializing across different languages was examined in detail by Neff van-Aerstalaer and Dafouz-Milne (2008), who compared the editorial writing of professional writers and novice learners of English and Spanish. The authors found significant differences in the usage of the analyzed textual and interpersonal markers between both novice and expert writers, as well as between expert writers of the two languages. This is somewhat intuitive, as writers emerging from different literary cultures and composing texts for audiences within those distinct cultures will naturally frame their arguments in different ways. The authors argue that opinion articles represent a particular type of persuasive text

whose final aim is to convince their readership by means of logical (i.e. textual) as well as emotional (i.e. interpersonal) strategies.

Opinion articles are eminently subjective … but at the same time they need to appear to be established on objective bases in order to be accepted… (pg. 88).

The composition of an opinion piece thus necessitates the strategic use of discourse markers in a way that is informed by the norms of this genre of writing that are necessarily language-specific. Ruiz-Funes (2014) employed a task in a study of L2 writing among advanced university-level learners of Spanish in which participants were asked to read an unfamiliar article on a “topic of international significance” and argue against or for the article’s main thesis. This type of task requires participants to “present ideas in a critical, organized manner in order to persuade their readers of their own position” (p. 172). Furthermore, this task is assumed to be more cognitively
distant, according to Skehan and Foster’s (1999) model, in terms of topic, genre and task type, thus it is more demanding in terms of processing (Ruiz-Funes, 2014, p. 172).

L2 learners’ production of persuasive writing is often characterized by stylistic diversions from NS norms (Kellog, 2001). For writers of Arabic, a persuasive/argumentative task may prove to be particularly effective at revealing differences between L2 and NS writers with regards to their lexical and syntactic performance, as well as “on the frequency of cohesive markers that a particular … writing task can be expected to elicit” (Kormos, 2011, 149). Al-Jabr (2006) argues that Arabic editorials are characterized by a frequent occurrence of coordinated clauses and phrases intended to have a pragmatic influence on the reader. Texts within this genre are typified by the strategic use of particular syntactic structures (e.g., coordination) in order to make a rhetorical impact (p. 209). Within L2 studies, this genre of linguistic production is understood to both reflect writers’ ability to display and implement their knowledge of the differences between spoken and written language, as well as the state of their competence in written discourse (Yang & Sun, 2012; pg. 33). Even NSs tend to display this variation when responding to persuasive/argumentative prompts, as the functions required to appropriately compose a response are often not fully developed at all in NSs except for in the most advanced writers (Kellog, 2001).

Given the influence of task design, two learners at comparable proficiency levels may receive quite divergent scores on a given CAF measure due to a number of factors outside their control. Even NSs will likely show some degree of variability in writing, especially in the context of a persuasive/argumentative task. Therefore, researchers aiming to capture a range of expression and to control for the influence of learner-external factors in participant writing production may choose to collect multiple samples of each participant’s written production in
response to prompts within multiple genres. Furthermore, researchers should strive to collect this production in a way that is ecologically sound. Manchon (2014) holds that written tasks should be intended to prompt the activation of various forms of L2 knowledge for the purpose of “communicative engagements in the personal, social, professional or academic sphere” (29). Therefore, written prompts framed as pieces of personal communication (letters, emails, etc.) may be more effective at stimulating this activation. Alanen et al. (2012) concur, stating that writing tasks used for assessment purposes should be designed to have “some measure of authenticity” in terms of the types of production and processes needed to compose them as well as to elicit types of production familiar to participants (p. 188).

A number of studies have examined the relationship between textual genre and L2 Arabic production. Although many of these have examined oral, rather than written data, they are nevertheless relevant to the study of written L2 Arabic. For example, Fakhri’s (1984) study of a high-achieving learner of Moroccan Arabic notably contributed the fact that learners of Arabic compensate for a lack of proficiency by through the use of communicative strategies, and that the use of these strategies is in turn affected by the genre of discourse. Fakhri found that Arabic learners sometimes show the ability to adopt certain NS-like morphosyntactic features such as contextually appropriate pronoun dropping. This is relevant to the current study’s examination of certain syntactic features such as sentence word order that are presumed to be affected by discourse genre. In the written mode, Fadel (2014) employed quantitative and qualitative analysis of NS and NNS expository writing in English and Arabic, specifically examining topic signals and discourse strategies, however Fadel’s analysis employed the identification of sentences, rather than T-units. Fareh (1988) found that Arabic expository discourse is
characterized by conjunctions that serve a thematic role, indicating that their presence may be indicative of a maturing writer.
Chapter 3: The Development of an Arabic C-Test

3.1 Introduction

Given that the OPI can be costly to administer, and that years of study is known to be an imperfect estimate of learners’ Arabic ability, researchers investigating the development of elements of Arabic production would benefit from the inclusion of access to alternative, concurrently collected, independent estimates of global Arabic ability. However, previous researchers have pointed to the fact that the number of high-achieving Arabic learners in the United States is limited (in comparison with other more-commonly taught languages), as well as the fact that most Arabic programs in the United States are relatively small, thus they lack the resources to develop their own in-house instruments (Winke & Aquil, 2013, p. 221). If such instruments are developed, they are typically the work of a small number of faculty members, and their applicability outside of the curriculum from which they emerge is often unclear.

This chapter describes the development of an Arabic C-test for use as a concurrent estimate of global Arabic proficiency to be included in the experimental design described in the subsequent chapter. Arabic presents a number of language-specific challenges to the C-test format, including difficulties posed by the Arabic script, identifying level-appropriate texts, and the presentation of multiple possible correct responses to a given C-test item due to the nature of Arabic morphology and morphosyntax. However, these challenges can be overcome by following a careful process of pilot testing and text revision, as well as by applying a modified C-test deletion strategy specific to Arabic, all of which are outlined below. Following a process of text selection, refinement, and pilot testing with native speakers, 10 candidate C-test texts were administered to 27 learners of Arabic. The results of this administration were analyzed via
Rasch analysis in order to select the five best-working texts, which in turn formed the C-test used during the main data collection phase of this dissertation.

3.2 Background

3.2.1 The C-test as a Test of Reduced Redundancy

Regardless of format, tests of language proficiency are designed to “elicit encompassing evidence of the range of functional language abilities and communicative competencies in a target language” (Wu & Ortega, 2013, p. 680). Different testing formats target different constructs – the OPI, for example, explicitly aims to assess “proficiency,” while the abovementioned Versant Arabic Text (VAT) targets language “facility,” rather than proficiency (Cheng et al., 2009). As implied by the term, “short-cut” proficiency assessment formats aim to provide a rough sketch of test takers’ proficiency in a target language without attempting to collect and assess the full range of linguistic production that might otherwise be required to make meaningful statements about test takers’ proficiency levels. Falling into this category of assessment formats, the C-test is designed to be administered to large numbers of test takers at once, to be easily and quickly scored, and to target a wide range of linguistic structures and difficulty levels. C-tests have been developed for a variety of assessment purposes, including language program placement (Norris, 2006) and for use as concurrent estimates of overall L2 ability in applied linguistics research (cf. Tidball & Treffers-Daller, 2008).

The C-test and the related Cloze test format operate according to the principle of reduced redundancy, introduced by Spolsky (1971). Tests operating according this principle ask test takers to reconstruct information that was removed from a target language text by test designers. While the Cloze format asks test takers to restore whole words that have been removed from a
text, the C-test asks learners to restore sections of individual words that have been half-deleted across several texts. By decreasing the ratio of “signal to noise,” reduced redundancy tests are argued to discriminate between learners at different levels, as well as between high achieving learners and NSs. Oller (1979) describes Cloze tests as targeting what he termed a grammar of expectancy, or “g”., which Davies (2003b) in turn describes as revolving around a learner’s ability to predict the L2. It has been noted, however, that language proficiency cannot be distilled into a single unitary value or entity (Bachman & Clark, 1987), therefore learner scores associated with a C-test or other related format should be carefully considered in conjunction with other variables such as number of years of L2 study, self-assessments, and scores on alternative testing formats.

Similar to the Cloze format, in which several whole words are deleted from a stretch of discourse, the C-test typically consists of several carefully-chosen authentic texts targeting several distinct levels of “difficulty.” In each of these texts, the second half of every other word is deleted, beginning with the second word in the second sentence and continuing for a total of 25 deletions per text (Klein-Braley, 1997). If a word has an odd number of letters, the larger “half” of the word is deleted. Furthermore, words consisting of a single letter (e.g., “I” in English texts) are not included in counting for deletion. The C-test can be seen as a continuation of a particular approach to Cloze test design, in which every fifth word of the source text is entirely deleted (Alderson, 1979). The C-test format typically requires much more knowledge about morphology and items’ word-internal structure, while the Cloze format requires a high degree of higher-order, discourse-level processing. For test designers, the Cloze format poses the additional challenge of often permitting multiple possible correct responses and may take more
time than a C-test to complete because it requires much more reading on the part of participants in order to arrive at a similar number of test items.

Test takers completing a C-test are typically presented with 4-5 constituent texts, which together contain 100-125 dichotomously-scored items typically administered within a span of 30-45 minutes. An example C-test sentence in an English text would appear as

Hope f____ the be____, but pre____ for t____ worst.

Test takers are asked to restore the half-deleted items to the best of their ability. The correctly restored C-test sentence above would appear as

Hope for____ the best__, but prepare for the__ worst.

This sample C-test sentence requires test takers to restore a variety of lexical and morphosyntactic information, including prepositions (for), the definite article (the), the formation of superlative adjectives (best), etc. By targeting every-second word in a text for deletion, the C-test format naturally requires knowledge of the target language lexis and morphology, as well as some elements of praxis and discourse strategies. Klein-Braley (1997) describes how C-tests operate via the reduced redundancy principle, stating that

[i]n order to restore the original text the learners must make use of all the clues available in the remaining portions of the text and in their heads. There may be phonological, morphological, syntactical, collocational, textual and also contextual, pragmatic, logical, situational clues in the text - and no doubt many others. The learners must process the mutilated text interactively, matching the rules they have developed against the text input (p. 53).

Babaii and Ansary (2000) agree that the micro- and macro-level processing targeted by the C-test conforms to the reduced redundancy principle, which “emphasizes that both a global and a local
knowledge are required to supply the missing elements in a distorted linguistic message” (p. 216).

A number of studies have investigated the construct validity of C-tests, or included secondary measures of language proficiency as part of the research design when piloting and implementing C-test-based studies in order to validate claims made about the relationship between a C-test score and general level of language proficiency. Daller and Phelan (2006) successfully compared learner C-test performance to performance on the Test of English for International Communication (TOEIC), finding a significant increase in scores on both measures administered as part of a pre- and post- design in a study abroad program. Eckes and Grotjahn (2006) likewise showed that learner scores on a German C-test and the TestDaF (Test of German as a Foreign Language) correlate highly with each other, and argue that the German C-test used in their study measures the same underlying linguistic capacities as the TestDaF.

Discretely scored tests such as the C-test can provide certain advantages over formats such as the OPI, which have been critiqued for producing scores that can be subjective and vague (Liskin-Gasparro, 2003). Unlike the OPI, the C-test can be easily administered to large numbers of test takers at the same time and quickly scored by a small number of raters. The C-test thus lends itself to use in applied linguistics research, allowing researchers to quickly separate test takers “into cross-sectional groups based on a normal distribution of their language abilities” (Wolfe-Quintero et al., 1998), a key requirement of quantitative cross-sectional studies investigating L2 proficiency. However, it should be noted that each population of examinees is different, and a C-test developed for a specific use in a specific institutional context should not be expected to function equally well when employed for a novel use or in a novel context. Furthermore, as a written measure, the C-test is literacy-bound and does not take into account
learners’ ability in other domains of language use, such as the oral proficiency targeted by the OPI. The current study will employ an original Arabic C-test in a manner similar to that used by Kuiken and Vedder (2012), namely to divide L2 learners into groupings representative of general L2 proficiency for the purposes of an analysis of other aspects of learners’ L2 capacities, namely their written production.

3.2.2 Examples of C-tests in Arabic

C-tests are most frequently used to measure learner proficiency in English (e.g., Daller & Phelan, 2006) and German (e.g., Eckes & Grotjahn, 2006). Although the format has been adapted to languages written in non-Latinate scripts, such as Korean (Lee-Ellis, 2009), published examples of reduced-redundancy tests developed for Arabic remain extremely limited. Several researchers report using the Cloze format in order to assess the oral and written capabilities of both Arabic learners and NSs (Abu-Rabia & Siegel, 2003; Parry & Child, 1990), and the Cloze format has likewise been employed in a few studies of Arabic L2 production (e.g., Mansourí, 1995). However, the existing literature describing the adaptation of the C-test format to Arabic, as opposed to the Cloze, is extremely scant. A single researcher, Matti P. Khoshaba Al-Bazi, has published a few examples of Arabic C-test texts as part of a larger “Integrative Arabic Assessment Test” that includes reading comprehension and Cloze test sections (Khoshaba, 2004). Although Khoshaba’s efforts represent an initial effort into Arabic C-test development, rather than a large-scale test development and validation study, his research nevertheless offers insights for other researchers.

The important test development steps that Khoshaba reports having followed include (a) drawing on experienced Arabic instructors to select level-appropriate texts, (b) pilot testing these
texts with Arabic NSs, and (c) controlling for the amount of time that learners were allotted to complete the C-test. Each of these steps is crucial for ensuring the appropriateness of constituent texts for inclusion of an Arabic C-test, as well as ensuring the reliability of the results. Additionally, Khoshaba describes experimenting with both Left-Hand (LH) and Right-Hand (RH) deletions in Arabic. Due to the reversed directionality of the Arabic script (when compared to roman scripts), LH deletions are word-final, while RH deletions are word-initial. The question of whether to target the beginning or the end of lexical items for deletion touches on more than directionality, however. LH and RH deletions necessarily target different sets of morphosyntactic structures in addition to the core lexical item. Deletions occurring at the beginning of lexical items in Arabic (RH deletions) fall on structures such as the Arabic definite article \( \text{الـ} \ al- \), as well as a number of prefixes including \( \text{بـ} \ bi- \), \( \text{لـ} \ li-/la- \) and \( \text{فـ} \ fa-. \) These prefixes do not represent part of the lexical item, but rather correspond to such English prepositions as “in,” “for,” etc. LH deletion in Arabic tends to target structures such as suffixes that form parts of certain verb conjugations, as well as pronominal suffixes expressing possession. The feminine marker \( \text{ة} \ - \ a/-at \) also occurs exclusively as a suffix. Khoshaba found that both RH and LH deletions elicit learner scores that correlate highly with each other, but that LH deletions (i.e., word-final deletions) are easier for Arabic learners to restore. LH deletions correspond with the word-final deletions employed in most C-tests of English and other languages written in roman scripts. Unfortunately, the examples of actual C-test texts published by Khoshaba are limited in number, and they are also unaccompanied by descriptions of which structures and items were considered subject to deletion by the test developer. A close reading of these example texts suggests that in most instances, all Arabic characters that form the targeted lexical item were considered potentially subject to deletion. The implications and limitations of this approach, as well as an
alternative strategy for creating C-test items that function well with the Arabic script, are discussed below.

3.3 Developing Arabic C-tests

In spite of the early, apparently successful effort to adapt the C-test format to Arabic, the information available to test developers on how best to overcome the challenges posed by Arabic to C-test design remains extremely scant. A number of factors set Arabic apart from better-studied Into-European languages with regard to C-test design. These factors include (a) the fact that written Arabic (MSA) requires a significant amount of education to manipulate; (b) the phonemic, consonantal nature of the Arabic script; (c) the status of Arabic as a Semitic, root-based language; (d) the presence prefixed prepositions and articles; and (e) regional variation in written Arabic. The current section discusses the challenges posed by these characteristics of Arabic, as well as successful strategies used to overcome them.

3.3.1 Education, MSA, and the C-test Format

As a literacy-bound test, Arabic C-tests will likely be limited to texts composed in Modern Standard Arabic (MSA), an elevated register of Arabic that is the primary vehicle of formal written discourse in the Arab world. MSA is understood to be derived from classical literary Arabic. It is typically acquired in schools, used in broadcast media, and is not spoken as a native language by any Arab population (McCarus, 2011). The first time that Arabic-speaking

---

2 It is certainly possible to develop C-tests for Arabic colloquial varieties, however the large amount of variation in the systems typically used to write these varieties, as well as the difficulty in locating suitable colloquial texts, renders the development of colloquial Arabic C-tests a challenging task. However, institutions that teach a variety or varieties of Colloquial Arabic may find the C-test a useful format for placement or achievement tests, assuming orthographic flexibility in test scoring.
children are exposed to written MSA is when they begin the process of literacy acquisition, usually within the context of formal education (Taha, 2013). Therefore, the structures and items that make up MSA remain, in a sense, foreign to NSs of Arabic until they are acquired in school. C-test development projects that rely on NSs during the pilot testing phase may benefit from a consideration of the relative educational levels or careers of NS respondents. Holes (2004) sees education level as a key factor in the deciphering of written Arabic texts:

Arabic writing is almost entirely phonemic. However, the practice of not writing vowels or other orthographic signs with phonological significance (a particular target of educators and script reformers, both Arab and non-Arab) makes it difficult even for educated Arabs to read with complete accuracy unless they have had a thorough grounding in normative MSA (p. 90).

The literacy skills required to successfully complete a C-test, such as reading and spelling, are likewise “affected by the specific linguistic and orthographic features of [Arabic]” (Taha, 2013, p. 726).

Non-native leaners of Arabic likewise typically acquire knowledge of MSA through formal education, thus C-tests that are limited to MSA employ the form of language most familiar to traditional Arabic learners. However, MSA is by definition restricted in its domains of use (Al-Wer, 1997). An MSA-based C-test may therefore reveal less about the Arabic capabilities of non-traditional learners such as heritage learners, who often excel in the oral mode but struggle in the written mode. Additionally, heritage learners are often speakers of varieties of language that are often not the focus of traditional education (i.e., speakers of colloquial Arabic, rather than MSA) (Valdés, 2001). Ultimately, C-tests used to estimate the proficiency of Arabic
learners are likely to be limited to MSA, the register most familiar to learners and which NSs
typically acquire via formal education.

3.3.2 Arabic Script and Orthography

At first glance, the Arabic script appears to lend itself quite well to the development of C-
test texts in Arabic. Like English, German, and other languages written in a Latinate script,
Arabic is written in a script that is composed of discrete letters/characters that correspond quite
well to Arabic’s consonantal and vowel phonemes. Additionally, although the directionality of
the Arabic script is reversed (i.e., right to left) when compared to that of English, C-test
designers are still able to target individual words for word-initial or word-final mutilations. This
study follows the example of Khoshaba (2004) whose published Arabic C-test texts are
composed of word-final mutilations. However, a number of factors inherent to the Arabic script
and morphology complicate a simple adoption of a one-to-one counting of characters in a word
and deleting half of them, as is typically done in C-tests of European languages.

The Arabic script is typically described as an *abjad*, or a consonantal script in which the
indication of long vowels is obligatory, but certain elements such as short vowels, doubled
consonants and markers of grammatical case are omitted (Gruendler, 2011). According to Holes
(2004), “the written Arabic word is essentially a consonantal skeleton: the three qualities of short
vowel, which are just as much phonemes as the consonants that make up the written word, are
not normally marked in writing except in the Koran and children’s reading primers… (p. 89).
Arabic items in which all diacritical marks and short vowels are imparted are orthographically
“transparent,” that is all of the phonological information necessary for identification and
pronunciation is represented. However, the vast majority of items in MSA texts are unvowelled
and thus orthographically “opaque,” and information about the phonological realization of these items must be inferred from knowledge of context and the Arabic morphological system (Saady et al., 2015, p. 126).

The vast majority of written texts in Arabic eschew reproduction of the short vowels, and readers are typically left to decipher the resulting opaque sequences of consonants based on their own knowledge of these systems. Arabic orthography is thus marked by an abundance of both opacity and homography\(^3\) (Taha & Saiegh-Haddad, 2015). For example, the graphical sequence علـْم may be read as either علم عِلْم “science,” علم عَلِِ تم “he knew,” علم عَلِِّا تَما “he taught,” علم علـِه “it was known,” علم علـِه “it was taught” or علم عالـه “flag.” Arabic NSs infer the correct pronunciation and meaning of this three-consonantal-grapheme sequence from context – the diacritic vowel markings are redundant (Al-Busaidi & Al-Saqqaf, 2015), and are thus rarely included in the vast majority of Arabic texts. These features of the Arabic script become problematic for C-test design when combined with the inflectional and derivational morphology of the language. This combination results in a high frequency of items that, due to the lack of short vowels, and the frequency of word formation “patterns,” present test takers with multiple possible correct responses.

3.3.3 Arabic Morphology

As a Semitic, root-based language, Arabic words are typically formed via the interaction of a lexical root and a morphological pattern. This interaction results in a number of classes of

\(^3\) Researchers in other fields of linguistics have cited these factors as difficulties inherent to developing materials for Arabic. For example, researchers developing Natural Language Processing applications cite script directionality, lack of capitalization, variable letter shape (i.e. a “cursive” writing system), and a high degree of homography (Farghaly & Shaalan, 2009).
items, derived from the same root, that are both semantically and graphically related. Combined with the lack of short vowels in most texts, this can result in significant ambiguity, as in the علم­­lm example above. As the C-test relies on the deletion of half of the letters used to represent a given word, this system implies that remaining un-deleted word-halves will often be composed of identical combinations of root- and pattern-letters, thus leading to the possibility of multiple “correct” responses to the same mutilation. The presence of multiple correct responses to a given mutilation is problematic in the scoring of a binary, discretely scored test such as the C-test, in which each item should have only one correct response.

Most Arabic words are composed of a tri-literal root consisting of three consonants mapped onto one of several morphological patterns. Successful restoration of C-test deletions thus requires a functional knowledge of both derivational and inflectional Arabic morphology. The morphological patterns themselves are formed by combining vowels and additional consonants, while the roots often carry much of a lexical item’s semantic meaning. These root-pattern combinations in turn create categories of words that are related to each other semantically (i.e., derived from the same meaning-bearing root) and graphically (i.e., employing the same consonants in the same places according). This process is predictable and logical, and Arabic lexical items are thus characterized by “an orthographic code that focuses on root letters, whose locations are systematically pre-determined by the word patterns” (Velan & Frost, 2011, p. 144).

The inflectional system of Arabic is likewise complex, and is composed of a series of prefixes and suffixes to indicate person, number, gender, tense, and sometimes case (Abu-Rabia & Taha, 2006). Thus, per Alhawary (2009), “a given word can have a large number of related words sharing the same semantic core meaning” (p. 1). Elkateb et al. (2006) note that
most Arabic words are created by applying distinct derivational patterns to some root, relating the two not only in form and meaning but determining their syntactic category as well. New Arabic words can always be coined from an existing root according to the standard derivational patterns. It is also possible to organize sets of Arabic words into distinct semantic fields according to the root from which they are derived (p. 18).

Knowledge of the Arabic root and prosodic pattern system has been shown to play an important role in the processing of input by NSs of Arabic (Boudelaa & Marlsen-Wilson, 2004; Velan & Frost, 2011) and in restoring the phonological information omitted by the Arabic script (Taha & Saiegh-Haddad, 2015). Learners’ developing knowledge of this system undoubtedly plays a role in their ability to correctly restore a C-test text.

A dichotomously scored test such as the C-test should ideally include items that can only be scored as either correct or incorrect, reducing ambiguity and the potential for rater bias, as well as increasing the reliability and internal consistency of the testing instrument. However, the complex interaction of the Arabic script with the Arabic root and pattern system often results in C-test items that present test takers with multiple possible correct responses. The problematic items that emerged from the current project were primarily identified during the NS pilot testing phase, and such items require the test designer to conduct secondary research and text modification in order to avoid the appearance of multiple possible correct answers in the final version of a C-test. For example, the pilot version of this C-test included a mutilation in text posited to represent a high difficulty level that fell on the verbal noun تفريق tafriq “dispersing” in the sentence وأطلقت القوات الغاز المسيل للدموع لتفرق المتظاهرين “[“for dispersing”] the protesters.” However, the original item tafriq was not supplied by a single native speaker participant when completing this item. Most native speakers supplied the alternate
These two items share the same root ف – ر – ق, although they are mapped onto different prosodic patterns (the original being derived from a form II verb and the preferred NS response onto a form V verb). In addition to the shared root, the two patterns themselves (II and V) are related, with V often representing a reflexive version of II (Badawi et al., 2004, p. 60). When subjected to C-test mutilation, the two items appear identical to test takers: tafriiq is rendered as تفـ and tafarruq is rendered identically as تفـ (with both appearing as تفـ). Thus, test takers have the option of two semantically related items that appear identically when mutilated.

Even if a seemingly problematic C-test item presents multiple correct responses, however, it cannot necessarily be assumed that each of the possible responses are equally “acceptable.” In the case of the example, above, a number of NS respondents from a number of dialect regions and professional backgrounds provided grammatically acceptable alternatives to the problematic deletion, but a process of follow-up corpus research revealed that these alternatives form collocations that are apparently not used in practice. The arabiCorpus tool (see Parkinson, 2011) is particularly useful for Arabic C-test designers, allowing users to search a variety of corpora organized by genre and thus determine the relative prevalence of alternative responses provided by NS respondents during the pilot testing phase. In this case, searches of the All Newspapers subcorpus for the two variants tafriq and tafarruq revealed that the original tafriq, rather than the tafarruq popular with NS participants, is a much more common collocation in media Arabic.⁴

⁴ Searches conducted on 2/1/2016
Table 1: "dispersing the protesters"

<table>
<thead>
<tr>
<th>Search string</th>
<th>Gloss</th>
<th>Total instances in All Newspapers subcorpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>تفرق المتظاهرين tafrijq n. (original)</td>
<td>“dispersing the protesters”</td>
<td>167 (0.12 instances / 100,000 words)</td>
</tr>
<tr>
<td>تفرق المتظاهرين tafarruq n.</td>
<td>“splitting up the protesters”</td>
<td>2 (0 instances / 1000,000 words)</td>
</tr>
</tbody>
</table>

The results displayed in Table 1 indicate that the original tafrijq is by far the most common item of the two in newspaper Arabic, and the second, tafarruq, is attested only twice in the corpus. However, in light of the fact that both collocations are attested, and that NS respondents in this study preferred the rarer tafarruq, a decision was made to consider this item as problematic and to shift the deletions in this particular C-test text so that this particular item was not mutilated. In this case, the particle قد qad was inserted at the head of the sentence to increase the number of words in the problematic sentence by one, thus shifting each of the deletions to the following word. Alternately a discourse and aspect marker, qad does not carry meaning of its own, but is often used to indicate the rhetorical structure of a text in written Arabic (Badawi et al., 2004; Sarig, 1995). Thus, by inserting qad at the head of the target sentence, the mutilations in a sentence can be shifted in order to avoid problematic items without significantly altering the meaning of the sentence.
3.3.4 Arabic Orthography and C-test Deletion Strategy

In addition to the inherent ambiguity present in an MSA text due to the absence of vowel markings, the presence of several prefixes argues against the adoption of the standard deletion strategy commonly used in C-tests of languages written in Latinate scripts. In most C-tests, deletion consists of simply counting the constituent letters of a word and deleting half, or “half+1” if the word is composed of an odd-number of letters. Of chief significance for C-test designers is the fact that the Arabic script incorporates a number of prepositions that function as prefixes to the lexical items they modify, as well as the prefixed definite article. Holes (2004) writes that, “[a]lthough by any criteria separate words, wa ‘and’, fa ‘so’, bi ‘in, by, with’ and li ‘to, for’ are written as if they were part of the word that follows them. Functionally similar words that are ‘heavier’ monosyllables or bysllabic, for example ʔaw ‘or’, fii ‘in’, ‘alaa ‘on’, are not so written” (p. 92).

Furthermore, although the standard rules of MSA orthography generally prevent the occurrence of single-character words, individual and geographic variations in writing style often result in the conjunction ٖwa- “and” being rendered as a stand-alone particle (Parkinson, 1999). C-test designers must thus scan individual candidate texts and bring the instances of and in line with the orthographic norms of Arabic, according to which this conjunction is an affix. In some cases, such as texts taken from a Moroccan newspaper (see Parkinson & Ibrahim, 1999), this may entail a large number of modifications to an edited text, in order to ensure that mutilations fall on every other word. It is also possible for prepositional affixes to combine with the definite article to form a four-character prefix. The addition of four additional letters to a lexical item represents a significant increase in the total number of letters available for deletion in a given word in a C-test.
The current C-test responds to the noted features of the Arabic script by adopting a slightly modified deletion method that, like deletion in Latinate scripts, entails the counting of individual characters/letters and deleting the second half or half+1 of letters in a word. However, due to the factors discussed above, a few minor adaptations are required to increase suitability for Arabic:

a) Definite article الـ “the” ignored for deletion

b) Prepositional prefixes ف- “so”, ب- “in, by, with” and ل- “to” ignored for deletion

c) Word-final structures such as pronoun suffixes, the feminine marker, plural markers and elements of verb conjugations are counted as part of the lexical item for deletion.

For example, construction وبالعربية “and in [the] Arabic” is a nine-character string in which five characters are not part of the core lexical item (conjunction = 1, definite article = 2, preposition = 1, feminine marker = 1). If half of the targeted item’s characters are deleted as in a typical C-test, a test taker would be confronted with the mutilation وبالعربية “and in [the]________”. Thus, the entire targeted lexical item would be deleted, leaving the C-test taker no elements of the original lexical item to reconstruct. This is much closer to the nature of Cloze deletions in which entire items are removed. Compare a) the “standard” C-test deletion strategy counting and deleting all letters in a given item equally, and b) the modified deletion strategy used in the current C-test:

Table 2: Modified Arabic Deletion Strategy

<table>
<thead>
<tr>
<th>Original</th>
<th>“standard” deletion</th>
<th>Modified deletion</th>
</tr>
</thead>
<tbody>
<tr>
<td>وبالعربية “and in [the] Arabic”</td>
<td>________</td>
<td>________</td>
</tr>
<tr>
<td>w-bi-l-ساربيأ</td>
<td>w-bi-l-_______</td>
<td>w-bi-l-٥_______</td>
</tr>
</tbody>
</table>

66
The modified strategy used here thus avoids deleting entire content words, but retains deletions on syntactic information (e.g., gender, number, verb inflections, etc.) encoded at the end of the word, bringing the Arabic deletions more in-line with deletions that occur in C-tests of European languages.

3.3.5 Regional Variation in MSA

Regional variation in MSA represents a fourth complicating factor in the development of Arabic C-tests. Although the spoken dialects of Arabic are multitudinous, the written register of MSA is often argued to be relatively standard across regions. Nevertheless, MSA texts composed in disparate geographic regions often exhibit significant variation on the lexical level (Ibrahim, 2011). Depending on the source of Arabic C-test texts, regional variation will likely play a role in the ability of NSs to correctly restore deletions. In order to minimize the confounding effect played by this variation, C-test designers must conduct pilot testing with NSs, as well as follow-up interviews subsequent to test administration, in order to identify the locations of this variation. Following this, test designers may choose to consult corpora in order to inform decisions regarding whether or not the identified regional variants are common enough to warrant retention in subsequent versions of the test, or whether these items should be replaced with variants that appear to be more region-neutral. Finally, this process of C-test creation, administration to NSs, and follow-up interviews has the potential to contribute to the nascent body of knowledge regarding lexical variation in MSA/media Arabic in general (see Parkinson, 2010).

The original batch of candidate C-test texts in the current project included a Moroccan MSA text that participating NSs from other dialect regions were unable to successfully complete
with 90% accuracy. The text in question is a Moroccan news article about electronic piracy that, when converted to a C-test text, presented a mutilation that NS respondents unanimously restored with lexical items not present in the original article. The possible responses to this mutilation are illustrative of the multiple layers of complexity that complicate C-test design in Arabic. In addition to the fact that all identified possible responses to this mutilation begin with the same combination of consonants (thus presenting test takers with an identical mutilated item), the original version of this item in the Moroccan article was not supplied by any NS respondents, all of whom self-identified as NSs of dialects geographically distant from Moroccan.

### 3.3.6 Regional Variation in “complaint”

The mutilation in question originally fell on the item بـشكاية bi-shikāya, in which the prefix بـ bi- is affixed to the lexical item شكاية shikāya “complaint” in the context of a newspaper article describing a complaint to judicial authorities about illegal electronic activity. C-test takers were thus presented with the mutilation بـشـ b-shk. However, during administration of the pilot test to NSs, this mutilation was consistently reproduced as شكوى shakwā “complaint”, or شكوى shakāwā “complaints,” rather than as the original. In addition to those alternatives supplied by NS participants in this study, several additional items might likewise be considered “correct” responses to this mutilation on the basis of their meaning and spelling, including a spelling

---

5 The original article from which this text was adapted was published by the Moroccan al-Masaae on October, 23, 2014, although it has since been removed. A version of the article could still be accessed as of 2/1/2016, however; see Ṣāliḥ (2014).

6 All transliterated pronunciations of this and related items in the current subsection were reproduced from Wehr (1993, p. 565-566)
variant of “complaints,” شكوى shakwā,7 as well as the much-less common items شكوى shakwa “complaint” and شكية shakīya “complaint.” All six possible responses to this mutilation are semantically related to each other, being derived from the same Arabic root ش–ك–ي sh–k–y and are likewise composed of either four or five characters. Thus, due to the fact that only the first two characters of the target item are included before the mutilation, as well as to the fact that the item’s internal vowels are not represented, the mutilation بـشـك b-shk in a text about a legal complaint could be completed with any one of at least six items.

This item was the subject of scrutiny during the NS follow-up interviews in order to determine why NS respondents consistently completed the mutilation with items other than the شكية shikāya used in the original Moroccan newspaper text. Four NS respondents who identified as speaking Lebanese, Egyptian, Saudi and Omani dialects, reported that the original item in the Moroccan text, شكية shikāya, is perceived as strange in the context of a newspaper article about a criminal case. Two participants specifically stated that the original شكية shikāya has a religious or antiquated connotation, rather than the legal connotation invoked by the more-common شكوى shakwā, although both can be glossed as “complaint.”

Subsequent analysis using the arabiCorpus tool (Parkinson, 2011) seems to confirm NS responses during the follow-up interviews regarding the Moroccan text in question. Table 3 presents the results of a series of searches in the arabiCorpus within the All Newspapers subcorpus.8 For ease of interpretability, the item in the original Moroccan text (شكية shikāya) has been bolded:

---

7 The substitution of word-final ﻲ for ﻰ is an extremely common orthographic feature of MSA produced in Egypt. Both variants are pronounced as ّa, however.
8 Searches conducted 12/19/2014
Table 3: “complaint” Variants—All Newspapers

<table>
<thead>
<tr>
<th>Item</th>
<th>English Gloss</th>
<th>Total Occurrences</th>
<th>Instances / 10,000 Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>شكوى shakwā</td>
<td>“complaint; letter of complaint”</td>
<td>8,411</td>
<td>6.21</td>
</tr>
<tr>
<td>شكاوى shakāwā (pl.)</td>
<td>“complaints; letters of complaint”</td>
<td>5,277</td>
<td>3.9</td>
</tr>
<tr>
<td>شكاوي shakāwā (pl.)</td>
<td>“complaints; letters of complaint”</td>
<td>642</td>
<td>0.47</td>
</tr>
<tr>
<td>شكاية shikāya</td>
<td>“complaint; accusation”</td>
<td>100</td>
<td>0.07</td>
</tr>
<tr>
<td>شكوة shakwa</td>
<td>“complaint; grievance”</td>
<td>22</td>
<td>0.02</td>
</tr>
<tr>
<td>شكية shakiya</td>
<td>“complaint; accusation”</td>
<td>10</td>
<td>0.01</td>
</tr>
</tbody>
</table>

Confirming the responses of NS respondents to initial versions of the pilot C-test, the identified possible answers of شكوى shakwā and شكاوى shakāwā are many-times more common in Arabic newspapers than the original Moroccan شكاية shakāya. However, the All Newspapers subcorpus of the arabicCorpus includes source texts published in several regions of the Arab world. A secondary search targeting only Moroccan newspaper articles in the Tajdid 2002 subcorpus, which is composed of texts taken from the Moroccan newspaper al-Tajdid, reveals that the discrepancy between the Moroccan article and the responses of NSs is likely due to regional

---

9 English glosses reproduced from Wehr (1993).
lexical variation in MSA. Within the Moroccan newspaper articles in the Tajdid 2002 subcorpus, the original item in Text 13, شكاية, is actually more common than both of the most common items across the texts that compose the All Newspapers subcorpus, شكوى and شكوى.

Table 4: “complaint” variants—Moroccan at-Tajdid

<table>
<thead>
<tr>
<th>Item</th>
<th>English Gloss</th>
<th>Total Occurrences</th>
<th>Instances / 10,000 Words</th>
</tr>
</thead>
<tbody>
<tr>
<td>شكاية shikāya</td>
<td>“complaint; accusation”</td>
<td>76</td>
<td>2.6</td>
</tr>
<tr>
<td>شكوى shakwā</td>
<td>“complaint; letter of complaint”</td>
<td>48</td>
<td>1.64</td>
</tr>
<tr>
<td>شكوى shakāwā</td>
<td>“complaints; letters of complaint”</td>
<td>48</td>
<td>1.64</td>
</tr>
</tbody>
</table>

Parkinson (2010) demonstrates the existence of regional variation in MSA, and the apparent Moroccan preference within newspaper articles for the شكاية shikāya variant appears to be an artifact of this variation, especially given that fully 76% of instances of شكاية shikāya in the All Newspapers subcorpus occur in the Moroccan al-Tajdid articles.

3.3.7 Regional Variation in Term for “regional”

The Moroccan al-Massae text about electronic piracy was also the site of another instance of possible regional variation identified via follow-up interviews with NS respondents. In the original newspaper article, the feminine adjective جهوية jihawīya “provincial; regional” appeared in the phrase المواقع الجهوية al-muwāqi‘u l-jihawīya “regional [web] sites.” However, several NS respondents were unable to reproduce the original adjective جهوية jihawīya or
expressed the fact that it was contextually inappropriate, stating that it is limited to certain tribal or ethnic connotations. In contrast to شكاية، however, an arabiCorpus search for the adjective جهوي establishes its use in a variety of regional newspapers:

Table 5: Use of جهوي by Region

<table>
<thead>
<tr>
<th>Subcorpus (newspaper)</th>
<th>Country</th>
<th>Occurrences of جهوي</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ghad01</td>
<td>Jordan</td>
<td>393</td>
</tr>
<tr>
<td>Tajdid02</td>
<td>Morocco</td>
<td>379</td>
</tr>
<tr>
<td>Ghad02</td>
<td>Jordan</td>
<td>235</td>
</tr>
<tr>
<td>Hayat96</td>
<td>London/Lebanon</td>
<td>137</td>
</tr>
<tr>
<td>Hayat97</td>
<td>London/Lebanon</td>
<td>132</td>
</tr>
<tr>
<td>Ahram99</td>
<td>Egypt</td>
<td>19</td>
</tr>
<tr>
<td>Masri2010</td>
<td>Egypt</td>
<td>6</td>
</tr>
<tr>
<td>Watan02</td>
<td>Kuwait</td>
<td>4</td>
</tr>
<tr>
<td>Thawra</td>
<td>Syria</td>
<td>4</td>
</tr>
<tr>
<td>Shuruq</td>
<td>Egypt</td>
<td>2</td>
</tr>
</tbody>
</table>

It is notable that newspapers emanating from certain large print media markets such as Egypt appear to make little use of this adjective, in contrast to the papers from Jordan, Morocco and London/Lebanon included in the corpus. Thus, NS C-test respondents who have consumed media primarily from the Egyptian market may be less able than their peers from other geographic regions to restore the جهوية mutilation.

In order to reduce the possibility of regional lexical variation affecting the reliability of C-test results while maintaining the contextual integrity of the texts, problematic items such as
شكاية shikāya and جهوية jihawīya that were consistently reproduced differently by NS respondents were modified or replaced with more-common but semantically similar items. In the case of شكاية “complaint,” this item was changed to نهيمة tuhma “accusation.” The new item نهيمة consists of five characters like the original problematic item; however, it was correctly restored by subsequent NS participants. Likewise, جهوية jihawīya was changed to الالكترونية liktrūnīya “electronic,” thus the problematic phrase المواقع الجهوية al-muwāqiṣu l-jihawīya “regional [web] sites” becomes the much-more common المواقع الالكترونية al-muwāqiṣu l-iliktrūnīya “web sites.”

3.4 Developing a Pilot C-test

After designing a method of deletion that accounts for the idiosyncrasies of the Arabic script, a pilot C-test consisting of 17 candidate texts was administered to 10 NSs of Arabic. Following the analysis of these responses and a process of text modification in response to many of the issues discussed above (i.e., multiple possible correct responses and deviation from orthographic norms in the original texts), 10 texts were selected and administered to 27 learners of Arabic at different levels. Following an analysis of these participants’ responses, the pilot C-test was narrowed to the five best-performing texts using Rasch analysis, and the participants in the current study were administered these five of these texts at the same time that their writing samples were collected.

3.4.1 Text Selection

A C-test is ideally composed of a series of authentic texts (i.e., texts composed by and for NSs) arranged in increasing order of posited difficulty, however the questions of how to assess text difficulty, and where to select texts from, pose test designers with a series of choices that
will affect the final form of the C-test. In general, texts inherently differ in the difficulty they pose to readers. Klein-Braley (1997) argues that “[t]here is some quality inherent in texts which makes some texts absolutely more difficult than others.” (p. 51). This quality is likely a combination of the lexical, morpho-syntactic, and pragmatic structures from which it is composed, as well as the test-taker’s familiarity with the topic of the text. On a practical level, a text which is too difficult for a group of test takers will return extremely low scores among all examinees, regardless of Arabic proficiency level, thus making it impossible to distinguish between them. Similarly, texts which are too “easy” will allow all test takers to score extremely high, making it impossible to distinguish between test takers on the basis of latent proficiency level. C-test designers should therefore select texts that eschew specialized terminology and that do not require specialized knowledge to interpret, in order to decrease the likelihood that any test takers will be advantaged on the basis of that (extra-linguistic) knowledge (Klein-Braley & Raatz, 1984).

In the case of Arabic, C-test designers may choose to consult the ACTFL guidelines for reading proficiency for examples of genres and structures determined by ACTFL to be appropriate for Arabic learners at the various levels of ACTFL proficiency. In addition to the general ACTFL guidelines for reading proficiency, ACTFL has supplied a number of Arabic-specific notations and examples of texts for each level (ACTFL, 2012). For example, the provided texts at the novice level are described as being composed of “basic,” and “common” vocabulary and consist of a series of lists, food menus, schedules and street signs – i.e., genres of written Arabic that are not appropriate for the creation of a C-test. Intermediate-level texts are simple biographical and descriptive texts, and the ACTFL examples increase in length and complexity until the distinguished level. Examples at the distinguished and superior level are
described as “rich with cultural references and complex linguistic structures,” and contain sentences that are “very heavy on cultural and semantic depth,” while example texts at the superior level are consistently described as “[posing] many thoughtful ideas and questions and [containing] cultural references.” It is likely that Arabic C-test designers will find more authentic texts that conform to the higher levels of ACTFL proficiency than they will texts that conform to the lowest levels.

The candidate texts initially selected for this phases of C-test development were drawn from a variety of sources—see Table 6 for a summary of sources and topics/genres. The lowest level of posited text difficulty proved the most difficult for which to identify suitable candidate texts, thus two of the texts in the original pool of 17 were drawn from Arabic L2 course materials. This parallels C-test development projects such as that described by Lee-Ellis (2009) for Korean, in which the lowest-level texts were constructed by drawing on L2 course materials. The original pool of candidate texts likewise included several texts adapted from websites aimed at Arabic-speaking children, however none of these “children’s” texts emerged from Rasch analysis as suitable for inclusion in the final 5-text C-test. Future Arabic C-test designers may find that Wikipedia is a useful source of mid-level texts, as Wikipedia articles tend to be written for the general reader, and articles from genres familiar to Arabic learners (e.g., biographical texts, descriptions of physical places, etc.) abound. However, texts selected from Wikipedia typically need to be significantly altered before they are rendered as C-test texts—this process entails shortening or removal of superfluous sections and the removal of formatting, in addition to bringing the text into line with Arabic orthographic norms.
3.4.2 Native Speaker Administration

A total of 17 texts were converted into candidate C-test texts and administered to native speakers for pilot testing. Of these 17 texts, a total of 14 were also evaluated by a group of Arabic instructors ($n = 5$) and non-Arabic-instructor native speakers ($n = 10$) who in turn rated the texts on a 1-5 scale of perceived difficulty. These texts are arrayed in increasing order of expected difficulty in Table 6. All texts examined by instructors were regarded to pedagogically appropriate for an Arabic course, and the average instructor ratings of “difficulty” generally correlate well with NS ratings. After the texts had been reviewed by Arabic instructors and judged to be appropriate for Arabic learners, they were rendered into 25-mutilation C-tests. Following this, they were administered to a series of NS respondents who self-reported as hailing from the Saudi Arabia ($n = 6$), Oman ($n = 1$), Egypt ($n = 1$), and Lebanon ($n = 2$).

Table 6: Candidate Texts

<table>
<thead>
<tr>
<th>Text</th>
<th>Topic/genre</th>
<th>Source</th>
<th>Rank in final pilot</th>
<th>Instructor rating</th>
<th>NS rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Autobiographical</td>
<td>L2 materials</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>Autobiographical</td>
<td>Children’s website</td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Job ad</td>
<td>Jobs website</td>
<td></td>
<td>1.8</td>
<td>1.75</td>
</tr>
<tr>
<td>4</td>
<td>Biographical</td>
<td>Children’s website</td>
<td>3</td>
<td>1.6</td>
<td>1.63</td>
</tr>
<tr>
<td>5</td>
<td>Biographical</td>
<td>Children’s website</td>
<td></td>
<td>1.8</td>
<td>1.6</td>
</tr>
<tr>
<td>6</td>
<td>Narrative</td>
<td>Children’s website</td>
<td>2</td>
<td>1.8</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Autobiographical</td>
<td>Children’s website</td>
<td>4</td>
<td>1.2</td>
<td>1.3</td>
</tr>
</tbody>
</table>

10 It should be noted that the 5 Arabic instructors who provided ratings for 14 texts consisted of both NS and non-NS instructors. Future research may focus on the differences in perceived text difficulty between Arabic learners and NSs, or between NS and non-NS Arabic instructors.
Table 6. (cont.)

<table>
<thead>
<tr>
<th>Text</th>
<th>Topic/genre</th>
<th>Source</th>
<th>Rank in final pilot</th>
<th>Instructor rating</th>
<th>NS rating</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Biographical</td>
<td>Wikipedia</td>
<td>5</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Entertainment news</td>
<td>Online magazine</td>
<td>2</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Press release</td>
<td>Embassy website</td>
<td>6</td>
<td>2.4</td>
<td>1.9</td>
</tr>
<tr>
<td>11</td>
<td>Historical</td>
<td>University website</td>
<td>7</td>
<td>2.4</td>
<td>2.5</td>
</tr>
<tr>
<td>12</td>
<td>Artistic</td>
<td>Cultural website</td>
<td>8</td>
<td>3.2</td>
<td>2.2</td>
</tr>
<tr>
<td>13</td>
<td>News</td>
<td>News website</td>
<td>3.4</td>
<td>3.4</td>
<td>2</td>
</tr>
<tr>
<td>14</td>
<td>Political speech</td>
<td>University website</td>
<td>3.2</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>Entertainment news</td>
<td>Online magazine</td>
<td>3.6</td>
<td>3.6</td>
<td>2.7</td>
</tr>
<tr>
<td>16</td>
<td>News</td>
<td>News website</td>
<td>9</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>News</td>
<td>News website</td>
<td>10</td>
<td>2.6</td>
<td>1.9</td>
</tr>
</tbody>
</table>

The numerical responses encapsulated in the instructor and NS ratings allow for the calculation of a Pearson correlation coefficient between the two values. Those texts for which both perceived instructor difficulty and NS difficulty ratings were collected (14 texts) show a weak correlation between the two measures (r = .56), indicating a somewhat-parallel perception of difficulty between the two groups of respondents. The only rated text to display a significant divergence between the NS and instructor difficulty ratings was Text 14 ("political speech"), with instructors rating this text as 3.2/5 while native speakers unanimously rated it 1/5, the "easiest" possible rating. This discrepancy is possibly due to the fact that this text contains a number of rhetorical devices common in political oratory that, while familiar to educated native speakers, may be completely novel for learners.
Why did the NS participants in this phase of the study unanimously rate the “political speech” text as the easiest at 1/5, while instructors (and later L2 Arabic learners) found it to be among the most difficult texts in the pool? After being converted into a C-test text and after removing Egypt-specific orthographic features (i.e., the use of word-final ـى instead of ـي, and vice versa), the first two sentences of this text appear as follows:

إنه من دواعي فخري واعتزازي أن أكون مصريا وأنعم بانتمائي لهذه الأرض الطاهرة، أرض السلام والمحبة. ويشاركني هذا الشعـ_________ جمع الشـ_________ المصري العظـ______ الذي قـ________ بثورة عظـ________ في الخـ________ والعشرين م________ يناير 2011 ضـ________ الفساد.

It is my great honor and privilege to be an Egyptian and I am blessed to belong to this virtuous land, a land of peace and love. And I am joined in this feeling by all of the great Egyptian youth who undertook a great revolution on the twenty fifth of January 2011 against corruption.

In general, the items on which the deletions in this text fall are relatively common (in, feeling, youth, great, undertook, great, fifth, from, against). However, the rhetorical devices used to compose the text likely render it quite difficult for learners. For example, the above excerpt contains the statement “the twenty-fifth of January,” which taps into the complex numerical system of formal Arabic. This text is assumed to follow a schema for speeches in Arabic which, while familiar to native speakers (hence the unanimous “1/5” NS ratings), are naturally unfamiliar to the typical university-level Arabic learner. One NS respondent, for example, described this text as “super easy – common text/propaganda”. In contrast, Arabic instructors evaluated this text as relatively difficult, and Arabic learners in the pilot testing phase scored quite low on it.\textsuperscript{11} This text was ultimately excluded from the 10-text C-test administered to

\textsuperscript{11} The identification of certain genres of written production in which NS and Arabic instructor/learner perceptions of “difficulty” sharply diverge raises major pedagogical questions – future studies may choose to
learners due to the fact that many NS participants were not able to correctly restore this text’s deletions with at least 90% accuracy. When this text is removed from the calculation of correlation, the agreement between the two sets of ratings climbs to $r = .84$, indicating a strong correlation between instructors’ and NSs’ impressions of text difficulty.

3.4.3 Text Revision and Selection of 10 Candidate Texts

Given that texts composed by and for NSs of Arabic are naturally subject to a great deal of variation, as well as a great deal of structural complexity that often frustrates attempts to directly convert authentic texts into C-test texts, designers of Arabic C-tests are naturally required to make a series of choices that result from necessary modifications to source texts. In the case of the current C-test, items that NS respondents identified as problematic or confusing were confirmed via secondary research to be the result of regional variation in MSA, and were subsequently replaced with region-neutral synonyms. Furthermore, multiple administrations of the pilot C-test to native speakers revealed a large number of items in most candidate texts that presented test takers with multiple possible correct responses to individual mutilations. In these cases, deletions were usually shifted via the non-meaning-bearing (but C-test-deletable) particle *qad* at the head of the sentence, in order to shift subsequent deletions over by one. Given the frequency of this particle in MSA,\(^\text{12}\) it is assumed to not gratuitously change the meaning of the sentence it precedes. The problems posed by the affixed Arabic definite article, prepositions and “and” conjunction are easily solved by ignoring these letters for deletion, ensuring that every mutilation in the C-test retains at least one character of the original “core” lexical item.

\(^\text{12}\) 248.22 instances of *qad* per 1000,000 words in *All Newspapers* subcorpus of *arabiCorpus*. Searched as “adv”, search conducted on February 24, 2016.
The process of test administration and follow-up interviews with NS respondents resulted in the discovery of mutilations in each of the candidate texts that were either consistently answered incorrectly by NS respondents, consistently left blank, or consistently answered with alternative correct responses. In order to further investigate instances in which NS respondents uniformly choose to complete deletions with items that were not present in the original text, follow-up interviews were conducted with four NS respondents from four disparate dialect regions (Egyptian, Lebanese, Saudi, and Omani). During these follow up interviews, NS respondents were asked to explain their responses to a number of items that had typically proven problematic for other respondents or that they had answered in a way that diverged from the source text. Following the identification of problematic items and the analysis of follow-up interviews, every candidate C-test text was modified (to varying degrees) in order to avoid items presenting (a) multiple correct responses and (b) regionalisms.

3.5 Learner Testing

3.5.1 Participants

A total of 27 undergraduate (n = 15) and graduate-level (n = 12) adult learners of Arabic were recruited for participation in this study. The learners were drawn from a variety of levels in order to increase the likelihood that a wide range of proficiency levels would be represented in the responses to the C-test. A majority (n = 17) of these participants reported having lived or studied abroad in Arabic speaking countries. Within this group of Arabic learners, one participant each reported being a native speaker of Italian, Spanish, and Czech, while the remaining 24 reported being native speakers of English. In keeping with general trends regarding Arabic language study in the United States, only a single participant in this phase of the study
reported that Arabic was the only language they had studied in addition to English. All remaining 26 participants reported speaking, studying, or otherwise having learned other languages in addition to Arabic.

![Figure 1: Arabic as an nth Language](image)

Importantly, this study was conducted at an institution that offers advanced and graduate-level coursework in Arabic, and a number of high-achieving learners participated in this phase of data collection.

3.5.2 Self-Assessment Questionnaire

In order to correlate C-test performance with other measurements of the proficiency construct, Arabic learner participants were asked to complete a short self-assessment and language background questionnaire. Self-assessments are widely used concurrent estimates of language proficiency in SLA studies (Shameem, 1998). Participants were asked to rate themselves on a 1-5 scale in each of the four “main skills,” as well as provide a 1-5 assessment of their overall Arabic ability. In addition to these self-estimates, participants were asked to
report on their Arabic learning history, including any study abroad experience, as well as report their educational level (undergraduate, graduate, etc.)

3.5.3 Procedure

A total of 27 Arabic learners were presented with paper versions of the 10 selected candidate texts and given a maximum of 60 minutes to complete all 10 texts to the best of their ability. Participants were instructed to attempt all 10 of the texts but were not restricted as to how much time they spent on an individual text. At the time of test administration, participants were likewise asked to self-assess their Arabic proficiency level in each of the “four skills,” as well as provide an overall assessment of their Arabic proficiency. None of the participants were previously familiar with the C-test format, thus each of them was provided a short introduction and explanation to the format before test administration began. In addition to an explanation of the mechanics of the format, all C-test materials were provided after a cover page that included an example sentence in English taken from the second-level Arabic textbook used in the institution in which data was collected.

3.6 Results and Rasch Analysis

Participant scores on the C-test were analyzed according to the Rasch model, which emerges from Item Response Theory (IRT) (c.f. Henning, Hudson & Turner, 1985). Rasch analysis is predicated on the goal of conceptualizing “the expected performance of individuals on a test item or task as a function of their ability and the difficulty of the item” (Lynch & McNamara, 1998, p. 161). Rasch analysis allows the separation of learners into different levels of “ability,” and test C-test texts (treated as individual items, each of which is scored out of 25
possible correct responses) into different levels of “difficulty,” based on participants’ responses to each of the 10 texts. Analysis of participants’ responses to the 10-test C-test indicate that this instrument was indeed successful at separating learners into different levels of ability, and test items into different levels of difficulty. The current effort entailed several rounds of analysis that compared different combinations of the 10 texts and learners in order to identify the 5 best-functioning texts to form the final version of the 5-text Arabic C-test. As indicated in Table 8, the Rasch analysis, conducted via the Winsteps statistical software, was able to explain 85.91% of variance in learner performance with high reliability on 10-text C-test.

**Table 7: 10-Text C-test**

<table>
<thead>
<tr>
<th>Variance explained by Rasch model</th>
<th>Separation</th>
<th>Strata</th>
<th>Reliability (alpha)</th>
</tr>
</thead>
<tbody>
<tr>
<td>85.91%</td>
<td>4.55</td>
<td>6.41</td>
<td>.95</td>
</tr>
</tbody>
</table>

The initial analysis included a single participant who represented a severe outlier in patterns of performance. This individual did not attempt several of the 10 texts, thus pushing the infit and outfit values of the analysis to above the upper limit of 2.0. After this individual was removed from the analysis, the statistical performance of the C-test items was improved. The analysis was re-run including 26 participants, and one text was determined to be poorly functioning. As shown in Table 9, Text 2 (T2) demonstrates infit and outfit values above 1.20, as well as low point-biserial (Rpbi), relatively low discrimination and high standard error (SE) values. Thus, text T2 was determined to be poorly functioning, and was removed from subsequent analyses.
Following the elimination of T2 and a single poorly performing participant from the analysis, the variance explained by the Rasch model improved to 86.99%. Following this, the analysis focused on identifying items that a) represent relatively equal levels of difficulty, b) contribute to covering the present range of examinee ability, and c) present high point biserial and discrimination values. This process resulted in the removal of items T4, T5 and T8, while items T1 and T10 were retained in order to maintain a spread of item difficulties, even though their measurement qualities were evaluated to be slightly marginal.

Following this process, the final 5-text C-test is composed of T1, T5, T6, T7, and T9, in increasing order of difficulty. Based on their scores on the pilot text, four examinees fall above the difficulty level targeted by T9, the most difficult text, and one falls below the level targeted.
by T1, the easiest text. Following the selection of the five best-performing texts, the variance explained by the Rasch model for the identified 5-text C-test increases to 92.62% (separation = 4.45; strata = 6.26; reliability (alpha) = .95). The item quality statistics presented in Table 10 indicate that the selected texts target distinct levels of examinee ability.

**Table 9: Final 5-Text C-test Item Quality Statistics**

<table>
<thead>
<tr>
<th>Text</th>
<th>Rpbi</th>
<th>Discrimination</th>
<th>InfitT8</th>
<th>Outfit</th>
<th>SE</th>
<th>Measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>T1</td>
<td>.75</td>
<td>1.00</td>
<td>1.04</td>
<td>1.19</td>
<td>.11</td>
<td>-1.59</td>
</tr>
<tr>
<td>T5</td>
<td>.90</td>
<td>1.05</td>
<td>.96</td>
<td>.95</td>
<td>.09</td>
<td>-.43</td>
</tr>
<tr>
<td>T6</td>
<td>.87</td>
<td>1.21</td>
<td>.83</td>
<td>.77</td>
<td>.09</td>
<td>.42</td>
</tr>
<tr>
<td>T7</td>
<td>.90</td>
<td>1.14</td>
<td>.69</td>
<td>.71</td>
<td>.08</td>
<td>.23</td>
</tr>
<tr>
<td>T9</td>
<td>.87</td>
<td>1.27</td>
<td>.63</td>
<td>.66</td>
<td>.10</td>
<td>1.04</td>
</tr>
</tbody>
</table>

The final version of the 5-text C-test was compared with examinees’ responses to the requested self-evaluation of Arabic ability. Spearman’s rho was calculated for each variable, revealing a moderately strong correlation \( r_s = .63 \) between self-assessment of overall Arabic ability and C-test score. The results, presented in Table 11, indicate that the self-assessment of reading has the highest correlation with C-test score, and listening the lowest, not surprising given that the fact that C-test performance involves reading in the first instance.
Table 10: Self-Assessment Correlation with C-test vs. Reported Semesters of Arabic Study

<table>
<thead>
<tr>
<th>Self-Assessment</th>
<th>Arabic C-Test Score (5-Text)</th>
<th>Semesters of Study</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td>.76</td>
<td>.68</td>
</tr>
<tr>
<td>Writing</td>
<td>.66</td>
<td>.58</td>
</tr>
<tr>
<td>Listening</td>
<td>.41</td>
<td>.51</td>
</tr>
<tr>
<td>Speaking</td>
<td>.64</td>
<td>.74</td>
</tr>
<tr>
<td>Overall</td>
<td>.63</td>
<td>.72</td>
</tr>
</tbody>
</table>

Finally, after the final 5-text C-test was identified via Rasch analysis, participants’ scores on those 5 texts were compared with the number of years that they reported having studied Arabic, revealing a moderately strong correlation ($r_s = .66$) between the two variables.

3.7 Discussion and Conclusions

The selected 5 texts can be said on the basis of Rasch analysis to target a range of both examinee ability and text difficulty. Furthermore, the item difficulty as determined by Rasch analysis closely mirrors the posited difficulty of each item prior to testing, as represented in the numbering of the texts. In the final version of the text, however, T7 and T5 fall very close to each other on the scale of item difficulty, thus they are assumed to target a very similar level of examinee ability. Future refinement of this C-test might thus benefit from the removal of one of these items in order to add a new text targeting the evident gap between T1 and T5, or even falling below T1. Given the presence of a number of high-achieving learners in the current participant pool, it is possible that future applications of this C-test to other populations of learners may not see such a broad range of examinee ability. Thus, it seems likely that future
efforts to identify candidate texts for inclusion in additional versions of this test will benefit from
new texts at the lower, rather than higher end of the difficulty scale in order to create more
nuance in discrimination between future populations of learners. Finally, it is perhaps
unsurprising that the area of language performance – elicited via a self assessment – that shows
the highest correlation with performance on the 5-text C-test is reading, given that the C-test is a
literacy bound format that largely does not target productive skills.

The positive correlations between C-test score and participants’ self-assessments and
reported number of years of Arabic study ($r_s = .63$ and $r_s = .72$, respectively) indicates that the C-
test instrument may be tapping into some aspect of latent Arabic proficiency accessed by these
measures. This is in keeping with previous studies that have shown that self-assessments of
language ability correlate well with other areas of linguistic performance (Flege, Mackay, &
Piske, 2002). However, it should not be surprising that number of years of Arabic study does not
correlate exactly with other measurements of Arabic proficiency, as it has been known for some
time that years of study is a consistently unreliable indicator of Arabic ability. In fact, this was
one of the principal findings of the early efforts to systematically measure Arabic ability among
large populations of university-level learners in the 1970’s described by Rammuny (1975; 1983).
However, given that the C-test format is intended to be a short-cut estimate of proficiency, rather
than an alternative to much more revealing formats such as the OPI, this positive correlation
argues well for use of the C-test format in contexts in which a quick, cost-effective measure
would be appropriate.

Although this effort drew on the participation of a relatively low number of Arabic
learners ($n = 26$) after the removal of one problematic participant, the moderate strong
correlations between C-test score and overall self-estimate of Arabic ability, as well as self-
reported semesters of study, together support this measure’s inclusion in the experimental design outlined in the subsequent chapter.
Chapter 4: Methodology

4.1 Research Questions

The current study is motivated by a desire to understand how a number of variables contribute to the development of written complexity, accuracy, and fluency (CAF) in both NSs and learners of Arabic. This dissertation specifically investigates the following questions:

1. What are the main differences between Arabic NSs and Arabic learners at different levels with respect to direct measures of written CAF?
2. What are the most important independent variables (e.g., C-test score, years of study, self-assessment) in determining elements of a given learner’s written CAF?
3. What effect do various written genres (e.g., persuasion, description, narration) have on the selected CAF measures, and do these effects manifest in the writing of both NSs and learners?

4.2 Participants and Recruitment

This study entailed the collection of written responses to three prompts provided by 14 Arabic NSs and 35 L2 Arabic learners drawn from a range of proficiency levels. The pool of NS participants originally numbered 16 (12 female; 4 male), although one male and one female participant were subsequently eliminated from analysis due to a failure to complete all experimental tasks. NS participants were recruited through the personal networks of the investigator. Of the final 14 NS participants, 11 were speakers of Saudi Arabic, two of Egyptian, and one of Lebanese. As noted in the Accuracy section of the current chapter, the inclusion of NSs from different dialect regions likely resulted in the introduction of regional variation to the pool of NS responses, however it is not expected that this variation extends beyond the orthographic level (e.g., region-specific practices for writing the hamza letter). All NS
participants reported having reached or completed university-level education at the time of data collection, including those who had or were currently studying at the BA level \( n = 10 \), the MA level \( n = 3 \), and the PhD \( n = 1 \) level. The educational fields pursued by this group of Arabic NSs are varied, although a majority \( n = 8 \) reported studying medicine as their academic major. One each of the remaining 6 participants report studying education, linguistics, teaching Arabic as a foreign language, law, English, and Islamic Studies.

With one exception, all NS participants report having attended secondary school in Arabic, rather than a foreign language, indicating that this pool of NSs can be expected to be familiar with the type of formal writing that occurs as part of a typical Arabic-language educational trajectory. However, in a reflection of the incredibly diverse array of educational systems and formats offered in the Arab world, only two of the final 14 NS participants reported that the language of instruction in their university education was Arabic, and one participant reported attending university in a mixture of Arabic and English. The remaining 11 participants (8 Saudi speakers, 2 Egyptian speakers, 1 Lebanese speaker) report English as the language of instruction in their university level education. Accordingly, it is expected that the lack of Arabic as a language of instruction in the tertiary education of these 11 participants may be reflected in their responses to the literacy practices survey. Thirteen of the 14 NS participants report having studied English as an L2, two studied an L3, and one studied an L4. When asked to rate themselves against their peers as an Arabic writer on a 1-5 Likert scale, the average of responses from this group of participants was 3.57/5 \( (SD = 0.76) \). This is substantially lower than the self-assessment of L1 writing ability given by Arabic learners, presented below.

The Arabic learner group consisted of 35 participants who were recruited both from their university level Arabic classes, as well as through the researcher’s personal networks. All
participants were enrolled in an academic program at a large American university at the time of data collection. Most participants were recruited via personal visits by the investigator to Arabic classrooms during the second, fourth, and sixth semesters of instruction. Recruitment also targeted high achieving undergraduate and graduate Arabic learners who were completing or had completed upper level coursework in the language.

These participants were drawn from all levels of Arabic acquisition, including individuals who had completed approximately one year of Arabic instruction at the time of data collection, and extending to an upper level of 10 years of Arabic study at the time of data collection, indicating that the higher end of written production and C-test scores collected as part of this study may be close to the higher end of what can be expected from similar studies of Arabic acquisition in the United States. Figure 2 depicts the distribution of educational levels in this population of participants. Importantly, 13 of 35 participants are graduate students, although these graduate students range from intermediate to advanced Arabic learners (based on both self-reported years of study as well as C-test score). The largest overall group of Arabic learner participants are 1st-year students who had completed a single year of Arabic instruction at the time of participation ($n = 11$):
This population of participants was evenly split between female \( (n = 18) \) and male \( (n = 17) \) participants. Most participants reported their academic major as being in a field related to international studies, international relations, politics, Arab studies, etc. Seven participants pursuing Arabic or Arabic literature as a major, and two participants reported their major as linguistics. This population of Arabic learners is reflective of the diverse student body at the institution in which data was collected, presenting a number of countries of birth. The most common country of birth in this group of learners is the United States \( (n = 27) \), followed by the UK/England \( (n = 3) \), China \( (n = 2) \), and one each for Singapore, Ukraine, and Luxembourg. Within this population of learners, Arabic was only infrequently a participant’s first L2. Figure 3 depicts the fact that the vast majority of participants report speaking or studying a third language in addition to Arabic and English \( (n = 16) \) or a fourth language \( (n = 12) \), with some participants studying or speaking even more.
4.3 Instruments and Procedures

This section describes the experimental tasks administered to NS and Arabic learner participants in the current study, as well as the method by which they were used to elicit the written and biographical data analyzed in the current study. Generally, the instruments consisted of three experimental prompts, a C-test (see the previous chapter for a description of this instrument), a self-assessment/language background questionnaire, and a series of post-task follow-up interviews administered to a subset of Arabic learner participants. Following data collection, handwritten responses were digitized via transcription, and all participant responses were compiled into a corpus of Arabic learner and NS writing described and analyzed in the subsequent chapter.

4.3.1 Administration of Experimental Tasks

Data was primarily collected from Arabic learner participants on the campus of an American university that is characterized by a relatively large undergraduate- and graduate-level Arabic program. Arabic learners arranged to complete the experimental prompts at a time of
their convenience, and their participation was supervised by the researcher. In order to decrease the possibility that participant fatigue could influence the results obtained via the experimental tasks Arabic learner participants were administered the five-text C-test, three written prompts, and a questionnaire, in random order. Additionally, the constituent texts of the C-test were randomized for each participant. Participants were afforded breaks as needed or requested. Time for completion of the C-test was limited to 30 minutes, and 20 minutes maximum were allotted for the completion of each of the individual written tasks, resulting in a total maximum participation time of ca. 90 minutes, excluding the subsequent time needed to complete the self-assessment and questionnaire (ca. 5 minutes). However, as the written responses described below are framed as pieces of communication, participants were “allowed” to take less than 20 minutes to complete. Figure 4 presents the average time for completion of the individual writing tasks by each group (NSs and learners) of participants, revealing that learners took longer on average to respond to each prompt than NSs:

![Figure 4: Average Time for Completion of Writing Tasks (in minutes)](image)

NSs were recruited through the researcher’s social networks, with the aim of ensuring that this group of participants is representative (to the extent possible) of literate Arabs educated in Arab
universities. Therefore, with one exception, all NS participants in this study submitted their responses electronically from their home countries. NSs were not administered a C-test. The decision for electronic data collection, as well as for choice of modality (i.e., typing), is justified on the basis of the fact that NSs are anticipated to be adept at typing, as well as the fact that typing is the preferred modality for the types of communicative tasks administered, i.e., two tasks are posed in the forms of “letters” or “message”. The decision to allow typing of responses, as well as electronic submission, was made in order to provide an element of authenticity to the types of production and processes needed to compose them, per Alanen et al. (2012, p. 188).

Furthermore, NS participants were provided with explicit instructions, both directly and imbedded in the experimental tasks, to record their start and stop times for each written task, allowing time on task to be measured and controlled. As part of the same documents containing the experimental tasks, NSs were asked to answer the same language background questions as the L2 Arabic learner participants, as well as to provide a self estimate of their ability as L1 Arabic writers.

4.3.2 Experimental Prompts

Arabic learners and NSs were administered three written prompts, to which they were asked to respond within a maximum time of 20 minutes. These prompts were provided to learners in English, and to NSs in Arabic. This was done in order to minimize the possibility that the language in which a prompt was composed (and a participant’s familiarity with it) would influence the written production elicited therein. The instructions given prior to the C-test completed by Arabic learners are described in the previous chapter. At the head of each page on
which a prompt was presented, participants were provided with the following instructions in either English or Arabic, depending on their L1:

أمامك 20 دقيقة للاجابة على الموضوع التالي. قم بالإجابة على السؤال بشكل تفصيلي وأكتب كل ما تتمكن من كتابته.

You will be given 20 minutes to respond to the following prompt. Please try to answer as thoroughly as possible and write as much as you can.

The motivation for the selection of the three tasks used in the current study is described in the review of literature presented in the first chapter of this dissertation. Briefly, it is known that measures of linguistic complexity, accuracy, and fluency in written and oral production are subject to the influence of task genre. For instance, tasks that are more familiar to learners are held to be less cognitively demanding, and thus elicit more fluent output (Skehan & Foster, 1999). Therefore, the prompts described below target three distinct genres of written production, namely description, narration, and persuasion. The first two of these are common task genres in the L2 Arabic classroom, while the third is understood to be a distinct and difficult type of written activity, even for NSs. The choice of three distinct task types will allow for the collection of diverse written production from each participant, as well as for the investigation of the relationship between written task genre and measures of written CAF. Manchon (2014) holds that written tasks should be intended to prompt the activation of various forms of L2 knowledge for the purpose of “communicative engagements in the personal, social, professional or academic sphere” (29). Therefore, the written prompts used in this study are framed, to the extent possible, as pieces of personal communication (letters, emails, etc.).
Finally, the topic of communicative writing in Arabic necessarily touches on the contentious issue of the *variety* of language to be used, both in the prompts and on the part of the participants. While authentic communicative discourse such as emails to friends can be expected to be composed in colloquial Arabic or a mixed variety by NSs, a newspaper editorial or letter to the editor will most likely be composed in relatively formal MSA. However, due to the “reverse privileging” (Ryding, 2006) that characterizes Arabic instruction in the United States, Arabic learners cannot generally be expected to have received instruction in producing writing that spans multiple registers. Unfortunately, L2 Arabic students are typically not required to produce written products in any register other than formal Arabic/MSA. Although a number of researchers have called for the integration of colloquial Arabic into Arabic L2 curriculums (e.g., Younes, 2006), few argue for explicit instruction in writing colloquial Arabic or mixed varieties. The issue of formality/informality in Arabic written discourse is highly contentious and represents a potentially rich avenue of research for future studies. As Fahmy (2011) argues, any act of writing in colloquial Arabic is inherently a controversial, political act. Thus, although Arabic NSs likely produce a large amount of writing composed in informal registers of Arabic (that may correspond with spoken dialects) in the form of personal emails, text messages, and writing composed to be shared via social media platforms (Hallajaw, 2016), it cannot be expected that Arabic learners will have received instruction allowing them to do the same. Therefore, the Arabic versions of these prompts were composed in MSA, and the English versions and instructions for them made no mention of the type of Arabic that participants were expected to use to respond to them.
4.3.3 Descriptive Task

As a basic linguistic function typically practiced extensively in all stages of L2 acquisition, it is fully expected that all participants in the current study will be familiar with this written genre. In order to minimize the elicitation of linguistic functions typical of the narrative and persuasion functions targeted in the other prompts, the descriptive prompt asked participants to describe their university. This choice was made to ensure that the topic of description was familiar to all participants, as well as to access functions that are frequently practiced in the Arabic classroom. In order to maintain the framing of these prompts—to the extent possible—as acts of communication, this prompt was framed as a “message” to a friend who is considering enrolling:

A friend from high school wants to enroll in a university to continue their higher education, and they would like some information about the campus where you study/studied. Write them a message in which you describe your campus in as much detail as possible.

This prompt was framed as a “message” in English due to the fact that the request for an “email” or a “letter” might prime participants to associate their response with a particular modality (i.e., handwriting or typing). Furthermore, the Arabic رسالة "message” could be interpreted as either “email” or “letter.”
4.3.4 Narrative Task

The narrative task uses a series of pictures that depict actions familiar to lower-level Arabic learners to elicit a narrative in either the past or present tense. The 12 cartoon panels that form the basis of this prompt occur on pgs. 193-94 of the first edition of *Al-Kitaab Fii Ta'allum Al-'Arabiyya: A Textbook For Beginning Arabic* (Part I) (Brustad et al., 1995) and ps. 180-81 of the second edition (Brustad et al., 2004). Although this cartoon is taken from a commonly used Arabic textbook series, a substantial familiarity effect is not expected among the group of Arabic learners who participated in this study. This is due largely to the fact that this cartoon is typically covered in the first semester of the first year of Arabic instruction at the institution in which data was collected, thus no study participants are likely to have been exposed to this cartoon during the semester of data collection. Additionally, the style and pace of Arabic instruction can vary wildly between sections, and it is not guaranteed that Arabic students will have been exposed to this cartoon through their instruction. The cartoon panels roughly depict a day in the life of a male protagonist, following the protagonist during the process of waking up, showering, going to work, returning home, and going to sleep. This picture series was chosen because it depicts a linear temporal progression of events in the daily life of the protagonist, however the prompt itself encourages creativity by asking participants to “tell the story”:

*The following cartoon depicts a normal day in the protagonist’s life. Tell the story of what this person does during their day. Try to tell the story as comprehensively as possible*
4.3.5 Persuasive Task

The persuasive task is fully expected to have been the most “difficult” for all participants to complete. Due in part to the fact that persuasion/argumentation requires many types of language-specific forms, structures, and devices to successfully complete, this genre of linguistic production is argued to reflect both writers’ “ability to signal the differences between the oral and written registers, and their acquisitional state of the written discourse competence in [the] L2” (Yang & Sun, 2012; pg. 33). In the persuasive prompt used in the current study, participants were asked to compose a “letter to the editor of the university newspaper” in which they support or oppose a proposed ban on smoking on their campus. This prompt in particular was the subject of refinement in consultation with several Arabic NSs in order to ensure that both the terminology and concepts related to letters to the editor, student unions, university newspapers, and smoking, could be successfully translated into Arabic for an Arabic-speaking audience.

Your university is studying the issue of banning smoking on campus, and you have been asked to express your opinion on this matter as an active member of the student union.
Draft a letter to the editor of the university newspaper in which you express your desire to support or oppose this idea, and detail your justification for doing so.
4.3.6 Post-Task Follow-Up Interviews

In order to investigate whether the experimental tasks were functioning as planned during the data collection period, a subset of Arabic learner participants were administered an open-ended follow-up interview following their completion of the experimental tasks. These interviews were conducted primarily in order to confirm the posited hierarchy of task difficulty (i.e., that the persuasion task would be the most difficult for learners to complete), as well as to investigate the possibility that the persuasion task in particular may have been too difficult for beginning-level learners to complete due to its targeting of specialized vocabulary.

The persuasive task was expected to be the most “difficult” for Arabic learners to complete, and this expectation was confirmed by the follow-up interviews, in which this task was frequently described by polled Arabic learners as the most difficult of the three. However, the source of this difficulty was more-often-than-not attributed to the lack of forms and structures necessary to form arguments, rather than a lack of health-related vocabulary in particular. A lower-level participant who had completed one year of Arabic study at the time of data collection held that “all three were definitely doable even being a first year student.” In the opinion of this beginning participant, the descriptive task accessed functions rehearsed in the classroom, as Arabic learners have “definitely done schedule things in class a lot,” as well as describing the university. This participant perceived the persuasion prompt as the most difficult, due to a combination of a lack of specialized vocabulary (“I forgot the word for ‘article,’” and “vocab related to student government”), as well as the vocabulary of argumentation, stating that “for me when I was making an argument I said min al-lāzīm.”13 A second lower-level participant responded that it is “harder to argue a point,” and that “not having the passive voice is huge,” for

13 “It is necessary,” as opposed to alternative conditional constructions.
framing an argument. In order to circumvent this shortcoming, this participant reported having “to use a lot of double negatives. We don’t know ‘should,’ and those kinds of things.” This participant held that “more important for forming an argument is being about to say should or would. I guess I can say ‘it is best.’”

When comparing the persuasive task to the descriptive task, a lower-level participant stated that they were able to “keep going on about the campus for much longer,” but that the descriptive task would have likewise been more difficult if it had included a persuasive/argumentative element, stating that “if it was [to] convince your high school friend to enroll in college, I couldn’t do that.” In comparing the descriptive vs the persuasive task, lower-level learners tended to respond with variations of the learner-provided statements that “it was harder to ‘argue a point’” The descriptive task was generally perceived to be difficult by some lower-level learners, as one responded that “describing the campus, my Arabic isn’t good enough…you get these general things, like, oh, I like my classes. I can do it generally but you run out of steam.” This respondent agreed that “description was easier than persuasion, but I just don’t quite have the vocab to say I couldn’t care less about people smoking.” However, the perception of the persuasive prompt as the most difficult of the three was not quite unanimous, as an MA-level learner reported difficulty recalling the vocabulary needed to respond to the descriptive task, stating that “this uses useful vocabulary—I can’t remember any.”

It was anticipated that the lack of task-specific lexical knowledge may have rendered the persuasive prompt particularly difficult for Arabic learners to complete. In order to investigate this possibility, the persuasive-task-specific item تدخين tadkhîn “smoking” was the source of particular inquiry. The item tadkhîn appears quite frequently in the responses analyzed in this study. As described in the section on Lexical Complexity below, this item is the eighth-most
frequent item in the corpus of NS and learner written responses to the experimental tasks. This absolute frequency stands in stark contrast to the relative infrequency of this item in general media Arabic texts. *tadhīn* appears 2.89 times per 100,000 words in the *All Newspapers* subcorpus of *arabiCorpus*, as compared to the appearance of *jāmiʿa*, another frequent noun in this study and fifth-most-common overall, which appears 53.92 times per 1000,000 words.\(^\text{14}\) *tadhīn* was used in the persuasion responses of 27 Arabic learners (out of 35 total learner participants).

4.3.7 Self-Assessment and Language Background Questionnaire

The relationship between Arabic literacy practices, self-assessments of writing ability, and performance on experimental writing tasks forms one of the directions of inquiry of this dissertation, and as such all participants were asked to complete a self-assessment and language background questionnaire concurrently with their responses to the written prompts. These questionnaires were administered in participants’ native languages, i.e., Arabic learners were asked to complete a paper version of this questionnaire in English at the time of data collection, and Arabic NSs were asked to respond to an electronic version of the questionnaire in Arabic before or after they completed the written prompts. The English version of this questionnaire collected basic biographical information relevant to the current study, such as self-reported years of Arabic study, educational status, study abroad experience, major, etc. The Arabic version of this questionnaire, included as an appendix to this chapter, likewise determined educational major and student status, as well as language of instruction in high school and university and main dialect of Arabic spoken by the NS participants. Additionally, all participants were asked to

\(^{14}\) Searches conducted on 12/5/2016
provide a self-estimate of writing ability in their native language, and learners were asked to likewise report their self-estimates of their reading, writing, speaking, listening, and overall Arabic ability. Self-ratings and other forms of self-assessments have been extensively used experimental studies of language acquisition as an additional and task-independent measure of proficiency (Shameem, 1998). These measures have likewise been shown to correlate with linguistic performance according to empirical measures of L2 production (Flege, Mackay, & Piske, 2002).

Arabic learner participants were likewise asked to provide the number of years for which they have studied Arabic. This measure is numerical, and for all participants occurs within a possible range of 1-10 (the upper-limit of years of study in this population of participants was 10 years). As the data analyzed in the current study was collected near the end of the spring semester of an academic year, participants alternatively used whole or half numbers when asked to provide their years of study. For example, a first-year student might alternatively respond with .5 years or 1 years when asked towards the end of the second semester of Arabic study how long they had been studying Arabic. In these cases, the researcher made a decision based on a given participant’s current Arabic class and participants’ responses to a separate question investigating how long they had studied Arabic in elementary, high school or university (assessed via the same questionnaire) in order to arrive at a value for years of study. In general, participants at the end of their second semester were coded as having studied Arabic for one year, and so on.

In the second section of the questionnaire, all participants were asked to report information about their Arabic (L1 or L2) literacy habits. This questionnaire was based in part on the Language Contact Profile, created by Freed, Dewey, Segalowitz, and Halter (2004) for the reporting of L2 contact activities in a study abroad environment, although it was modified to
focus primarily on literacy practices. This section of the questionnaire asked all participants to report how frequently they engage in activities such as a) writing formal vs informal correspondence in Arabic, b) producing academic writing in Arabic, c) consuming or producing Arabic social media, d) reading Arabic novels, etc. It is hoped that the NS responses to this section will shed light on the relationship between NS literacy practices, educational background, and writing performance, as well as provide information of potential use to Arabic language educators as they seek to design curricula aimed at guiding Arabic learners toward a NS-oriented model of Arabic use.

Arabic learners were also asked about their frequency of interaction with Arabic NSs, as a way of providing an additional numerical variable posited to be related to their overall Arabic ability or level of written CAF. Learners were asked, “On average, how often did you communicate with native or fluent speakers of Arabic in Arabic in the last year?” The possible responses included never, a few times a year, monthly, weekly, and daily. These possible responses were respectively accorded a score of 1, 2, 3, 4, and 5 during the coding process.

Finally, all participants were asked to provide a self-assessment of their L1 writing ability as part of the language background questionnaire. Arabic learners, for example, were asked “How do you rate your writing ability IN ENGLISH,” with the clarifying question “i.e., when compared to your peers, how good of a writer do you feel you are in English.” Arabic NSs were posed the Arabic version of the same question. Participants were asked to rate themselves on a scale of 1-5, with descriptors of “low/poor” provided for 1, and “superior/excellent” provided for 5. On average, the Arabic learners in this population of participants view themselves as highly advanced English writers (average rating = 4.69; SD = 0.47).
4.3.8 C-test

As described in the previous chapter, the C-test used in this study is intended as a “short-cut” proficiency measure, designed to be easily completed within 30 minutes and easily scored by researchers. The C-test used here consisted of five texts, each of which contains 25 sub-items in which words originally present in the source text were half-deleted. This resulted in a numerical score out of 125 for each Arabic learner participant. For coding purposes, each constituent text was treated as a “super item”, for which each learner received a score out of 25. Participants were presented the individual C-test texts in random order and allotted 30 minutes for completion of all texts.

4.4 Scoring and Coding

The current study employs a number of discrete measures of written complexity, accuracy, and fluency that have been elaborated in studies of languages other than Arabic. A secondary goal of this dissertation is to explore the utility of several of these measures for application to written Arabic texts, as well as the practical challenges associated with this effort. Although future efforts aimed at investigating the nature of written Arabic may take into consideration additional factors related to the relative quality of written discourse, such as text organization features (Uysal, 2008) and related groups of textual features (Friginal & Weigle, 2014), the current study will not rate texts holistically beyond classifying them as produced by NSs or Arabic learners. Participants’ texts will instead be measured according to a series of objective measures described below. This study will employ ratio-based measures as opposed to simple frequency counts, as they have been argued to be more valid representations of learners’ developing interlanguage (Wolfe-Quintero et al., 1998). Each of the following measures will be
calculated for participants’ production on each individual task, as well as participants’ written production in response to all tasks as a whole.

**Table 11: Summary of CAF Measures**

<table>
<thead>
<tr>
<th>Construct</th>
<th>Type</th>
<th>Measure</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic Complexity</td>
<td>Overall Complexity</td>
<td>Mean Length of T-Unit</td>
<td>W/T</td>
</tr>
<tr>
<td></td>
<td>Coordination</td>
<td>T-units per Sentence</td>
<td>T/S</td>
</tr>
<tr>
<td></td>
<td>Length of Production</td>
<td>Mean length of sentence</td>
<td>W/S</td>
</tr>
<tr>
<td>Lexical Complexity</td>
<td>Lexical Richness</td>
<td>Guiraud’s index</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hapax Index</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>Accuracy</td>
<td>Error-free T-units per T-unit</td>
<td>EfT/T</td>
</tr>
<tr>
<td>Fluency</td>
<td>Rate of Production</td>
<td>Words per minute</td>
<td>W/M</td>
</tr>
</tbody>
</table>

Handwritten learner responses, as well as the responses from a single NS who chose to respond by hand, were subsequently transcribed in Microsoft Word and assigned tags for the pseudonym associated with each participant, as well as for the individual task. After each participant’s responses had been transcribed, they were further subdivided into T-units, with each T-unit forming a new line within the response, and T-units containing errors were tagged. This process allowed for the calculation of several measures which were then entered into an Excel file containing the measures associated with each participant’s responses. The measures calculated in Word include total words, total sentences, total T-units, and number of error-free T-units per response. The CAF measures used in this study were calculated using these scores in Microsoft Excel.
The transcription process represented the first stage in which learner responses (and those of one NS) were altered from their original format. During the transcription process, any short vowels that a participant may have included in the paper response were not rendered in the electronic version. The decision to omit any and all short vowels was made due to the fact that the software used to conduct lexical analyses would recognize two strings of characters as separate types (i.e., separate words), even if the only difference between them was a short vowel. Additionally, all personally identifiable information included by participants in their responses was removed during the transcription phase. For example, many participants explicitly named themselves, their acquaintances, or the institutions where they study as part of their responses to one or more of the experimental prompts. This information was removed, although, in order to maintain accurate counts of words per minute and T-unit length, these identifiable items were replaced by the Arabic word اسم اسم “name.” The replacing of identifiable names of individuals, institutions, buildings, etc., with the Arabic اسم اسم also had the effect of compensating for what would otherwise be elevated measures of lexical diversity in the responses of participants whose production was characterized by long lists of buildings and streets, e.g., in response to the prompt in which they were asked to describe their campus.

4.5 Syntactic Complexity

Syntactic complexity is operationalized in the current study according to three measures, namely Mean Length of T-unit, the ratio of T-units per sentence, a commonly-used measure of syntactic coordination (e.g., in Ruiz-Funes, 2014), and mean length of sentence. However, given the posited unreliable nature of the sentence as a unit of measurement in Arabic texts (described in chapter 2 of this dissertation), it is possible that subsequent analyses will find measures
employing this unit, including coordination, to be unreliable. Sentences are relatively easy to identify on the basis of participant-supplied punctuation marks—periods, exclamation marks, and question marks were counted in this study as terminating a sentence. If a passage contained none of these marks, it was assumed to consist of a single sentence. T-units, however, are much more difficult to identify, and their use requires a series of decisions on the part of the individual researcher working with written Arabic.

4.5.1 Identifying Arabic T-units

As discussed in the preceding review of relevant literature, the current study follows the example of Hunt (1965), in which the T-unit represents the minimal unit into which a stretch of discourse can be broken without leaving any fragments. An examination of the examples provided by Vann (1979) indicates that this strategy is applicable to Arabic texts as well. Vann contrasts an adult’s utterance, “the man who I saw yesterday runs fast,” as opposed to a child’s imitation, “I saw the man and he runs fast” (p. 6, underline in the original). The former utterance is composed of a single T-unit while the latter contains two, even though both could be considered intact sentences on the basis of punctuation. When glossed in Arabic, the adult’s الرجال الذي رأيته أمس يجري سريعا  consists of a main clause and a definite relative clause, the latter of which cannot stand on its own as an intact sentence (الرجل الذي رأيته أمس يجري سريعا is a fragment), thus this utterance consists of a single T-unit. The child’s رأيت الرجل ويجري سريعا can be broken into two self-contained, intact Arabic sentences (رآيت الرجل ويجري سريعا).

By way of example, the following transcribed passage has been arrayed according to T-units. It was composed by a lower-level Arabic learner in response to the written narration task.
The original author’s periods have been retained to indicate sentences, as have orthographic and lexical errors present in the original text:

**Table 12: Arabic T-units in Narration**

1. it became 7 o’clock
2. [and] then he drank coffee with his wife.
3. after that he listened to the radio
4. [and] he read the news.
5. he left his home
6. and used the bus to travel to his occupation.
7. he worked in a bank
8. [and] hated his jobs.
9. he was happy when he left the bank because he was able to play [basket]ball with his friends.
10. he finally returned home
11. and ate with his wife.
12. at eleven o’clock he slept

This composition displayed in Table 16 consists of:

a) 7 sentences (determined via author punctuation)
b) 13 T-units
c) 16 clauses (a subject and a predicate)
Researchers working with Arabic are required to make a number of decisions when attempting to determine where T-unit boundaries occur. For example, should the sentence نزل بيته أخيرا وأكل مع زوجته. nazala bayti-hi ?axīr-an w-?akala ma? sawjati-hi could be glossed as “He finally [returned] home and ate with his wife,” or “he finally [returned] home and he ate with his wife.” Variation in surface structure due to the pro-drop nature of Arabic (Farghaly & Shaalan, 2009) has the potential to drastically increase or decrease the apparent complexity of a text based on how T-units are arrived at. Kim and Elder (2005) noted a similar situation for Japanese and Korean, which “tend to have a null-subject in the surface structure of a clause” (p. 363), as does Arabic. However, according to the definition of T-unit as the “minimal terminable unit (Hunt, 1965, p. 36), the above sentence can be justifiably classified as consisting of two Arabic T-units, due to the fact that each constituent clause can be viewed as an intact sentence.

It is expected that Arabic T-units will tend to be shorter than T-units in English or other European languages. Ryding (2014) notes the concatenative nature of Arabic results in a situation in which an entire predication can be contained within a single word (i.e., yushāhidūna-hā “they watch her”). This statement, if properly punctuated, would be a grammatical sentence, thus researchers working with Arabic are required to determine whether it would constitute a T-unit. Researchers working with written Arabic may be frequently confronted with such sentences. In the current study, a participant describing a university campus wrote ً أستطيع أن أمشي إلى المترو وأذهب إلى أي مكان في (المدينة) ً Astaṭī ِ an ْamshī ِ ilā l-metro w- ِ adhhab ِ ilā ayy makān fi [l-madīna] “I can walk to the metro and go to any place in [the city]”. In English, the statement, “I can walk to the metro and go anywhere in the city” would be considered one T-unit, while “I can walk to the metro and I go anywhere in the city” would be...
considered two T-units. In Arabic, however, in the absence of short vowels indicating verbal mood (i.e., ʔadhhaba or ʔadhhabu), it is possible in this case to either option. Again, however, according to the view of T-units as the minimal terminable unit, i.e., the smallest unit containing a grammatical sentence without leaving behind any unaccounted-for fragments, this utterance can be justifiably viewed as containing two T-units.

Several studies of L1 and L2 Arabic production (both written and oral) have employed the T-unit as a unit of analysis for rate measures, however the authors typically do not go into detail regarding the specific type of T-units used for analysis or discuss how decisions about what constitute T-units were reached. In the absence of a substantial body of work treating the analysis of syntactic complexity in written L2 Arabic texts, this study will attempt to delineate a functional system for T-unit identification, arrived at through examples taken from NS and NNS texts. As noted above, researchers have some degree of freedom in terms of delineating the boundaries that separate T-units.

4.5.2 Subdividing Simple and Complex Arabic Sentences into T-units

Written Arabic is often composed of T-units that vary widely in length, and the average length of T-units in a writing sample is likewise expected to reflect some element of learner’s developing Arabic interlanguage. At lower levels of written proficiency, the question of where one T-unit ends and the next begins is relatively simple. As sentences increase in complexity and as increasingly long clauses are placed into relation with each other via coordination and subordination, however, the process of subdividing sentences into constituent T-units becomes more difficult and requires specific decisions on the part of the researcher. Within the context of this study, every effort was made to base these decisions on trends that emerged from the
responses of the NS participants, and occasionally to confirm these decisions by consulting the 

*arabiCorpus* tool.

In their simplest form, T-units in Arabic can be overlap entirely with learner-punctuated sentences, as in example 1 below, which is taken from the production of an Arabic learner:

**Table 13: Simple T-Units**

<table>
<thead>
<tr>
<th>العربية</th>
<th>“In the morning, [he] gets up at five o’clock for the day. He goes to the bathroom.</th>
</tr>
</thead>
<tbody>
<tr>
<td>في الصباح، اسم يصحو في الساعة الخامسة لليوم. يذهب إلى الحمام.</td>
<td></td>
</tr>
</tbody>
</table>

Example 1 consists of two punctuated sentences, each of which contains a main clause and consists of an intact T-unit. There is no way to further subdivide this sample into smaller T-units without leaving behind unattached sentence fragments. It is expected that, as learners increase in overall or written Arabic proficiency, the complexity of their T-units will increase. However, as Arabic sentences begin to increase in complexity, and owing especially to Arabic’s status as a pro-drop language, the task of identifying T-units within punctuated sentences begins to be more complicated, as in example 2, below:
Example 2 above consists of a single sentence that contains three clauses and an ambiguous number of T-units. In English, this sentence might be glossed as “he drinks coffee with his [wife], reads the newspaper, and listens to the radio,” or it might equally be glossed as “he drinks coffee with his [wife] and he reads the paper and he listens to the radio.” The former gloss would typically be considered a single, complex T-unit, while the latter would consist of three simple T-units. Given that this issue arises quite frequently in written Arabic data due to Arabic’s status as a pro-drop language, the decision was made in this study to take the strictest possible view of the “minimal terminable unit” coined by Hunt (1965), and to count as T-units the smallest possible segment that could be considered a functional sentence without leaving any fragments. Thus, the first example above was classified as consisting of two T-units, while the second was classified as containing three.

T-units become increasingly difficult to tease apart as the proficiency level of an individual writer increases, however, and they also tend to be substantially more complicated.

---

15 This sentence could arguably be coded as containing an error, specifically because the author described the female character depicted in the narrated picture series as a زوج zawj “husband.” However, due to the fact that “husband” was spelled correctly and affixed with a possessive pronoun appropriate for the story’s male protagonist, this was not coded as an error.
when produced within certain genres of Arabic writing. The genre of argumentation or persuasion in written Arabic is particularly complex, given that arguments are typically made and framed within a series of rhetorical devices, connecting devices, lexical repetition, clausal subordination, and coordination. The following passage consists of a single sentence produced by a NS participant in response to a persuasive prompt regarding a proposed university-wide ban on smoking:

*Table 15: Complex T-Units*

"This is due to many causes, for the damage of smoking is not limited to the smoker for whom smoking is a personal issue for him alone; rather the damage of smoking in public places in particular extends to include his surroundings including the environment and life, the air becomes laden and one is oppressed by it, especially those who suffer from asthma, for whom clean air is considered an issue of life or death!"
This sentence illustrates that fact that Arabic sentences can range from short to extremely long, and contain multiple clauses and complex T-units within them. However, T-unit boundaries can nevertheless be identified by searching for the smallest units into which this passage can be broken without leaving fragments. This passage can be subdivided into four individual T-units of varying length, each of which can stand on their own as an intact sentence:

1. وَذَلِكَ يُعود لأسباب عديدة
   “This is due to many causes

2. فَإِنَّ ضرر التدخين لا يقتصر على المدخن الذي كان التدخين مسألة اختيارية بالنسبة له فحسب
   for the damage of smoking is not limited to the smoker for whom smoking is a personal issue for him alone

3. بل ضرر المدخن في الأماكن العامة بالذات يتسع ليشمل محيطه من بيئة وحياة
   rather the damage of smoking in public places in particular extends to include his surroundings including the environment and life
the air becomes laden and one is oppressed by it, especially those who suffer from asthma, for whom clean air is considered an issue of life or death!”

A number of items relevant to T-unit identification in Arabic appear in this sentence. First, it is clear that multiple T-units may appear within the context of a single Arabic sentence, and that these T-units may simply be coordinated with each other, rather than subordinated to each other. Secondly, this sentence demonstrates the unpredictable length of author-punctuated sentences in written Arabic, implicitly arguing for a sub-sentential unit of analysis such as the T-unit. Simply put, Arabic sentences are often as long as the paragraphs in which they appear, and rather than being a sign of an inexperienced under-educated writer, sentence length may instead be determined by other factors such as genre type or familiarity with formal written Arabic.

Finally, this sentence includes a number of particles that can be considered to initiate T-units in certain instances. For instance, NSs in this study used the particles إذ idh, بل bal and إن inna to initiate new sentences, indicating that they can occur at T-unit boundaries (unless إن is linked to the verb قال qāla “to say”). Searches in the arabicCorpus for these particles confirmed that they do indeed often initiate new punctuated sentences, indicating that they may be used to initiate T-units. Therefore, according to the strict definition of the T-unit as the

---

16 Diacritic marks were not transcribed as part of this study, thus it is unclear from orthography alone whether or not an individual string of إن corresponds to the particle إذ or the particle إن inna. In these cases, a judgement was made on the basis of the surrounding context to determine which particle was used.
shortest possible grammatical sentence into which a stretch of discourse can be divided, these
and related particles were considered to initiate new T-units in the current study. Likewise,
interjections such as ُيا إلهيُ “oh God” were coded as independent T-units. Salutations,
such as at the beginning and end of the letters that many participants chose to frame their
responses to the description and persuasion prompts, were counted as independent T-units, even
if they did not represent complete sentences or predications. Finally, several participants chose to
title their responses – titles were likewise coded as intact T-units. These instances formed a small
percentage of the overall T-units produced with the context of this study, however. The
preposition ُفـ “so” was likewise recognized as possibly initiating a new T-unit, e.g. when
coordinating two independent clauses, unless it occurs as part of another construction such as ُمـع
أن ُماـكا ُوـانـنا “however,” or ُإذا ُثـدـحـ “if”.

4.6 Lexical Complexity

This study operationalizes lexical complexity according to two separate-but-related
ratios, namely number of types vs number of tokens (Guiraud’s index), and number of hapaxes
per number of types. A hapax is an item that occurs only once in a given context, i.e., a type that
occurs only once for each participant, or for each participant’s individual submissions. In order
to explore effective methods for the analysis of lexical complexity in written Arabic, two
versions of each of these ratios were prepared, i.e., using a lemmatized and an un-lemmatized
version of each participants’ responses. Both the lemmatized and un-lemmatized versions of
these measures will be considered in the initial statistical analysis of this data, in order to
determine whether the lengthy lemmatization process was successful in adding nuance to the
selected lexical measures. In both cases, all responses were compiled in a corpus containing each
of the three written responses produced by each individual participant and analyzed using the
Lexico 5 software package. Lexico 5 provides several sources of numeric information about each
individual response tagged in this corpus, including numbers of types, tokens, hapaxes, as well as
the most frequent item per response and the number of times that it occurs. Additionally, these
values are likewise provided for each participant, treating each of their three submissions
together. These values were subsequently imported into Microsoft Excel, where they were
converted into the type/token and hapax/type ratios.

This study employs Guiraud’s index of lexical richness, which compares the number of
types of items used by a participant to the total number of items, resulting in a numeric
type/token ratio. Guiraud’s index relies on the additional calculation of dividing the number of
types by the square root of the number of tokens (e.g., Kuiken et al., 2010), providing an
algebraic adjustment for text length that renders this index more suitable for L2 texts which may
be shorter on average. For example, the L2 Arabic composition reproduced in Table 16, above,
consists of 55 words (tokens), and approximately 45 types (i.e., words that are not repetitions of
previously-used words). This passage thus has a relatively high type/token ratio of 0.82, and
effectively demonstrates why such ratios have been critiqued as subject to the influence of
passage length.

The Lexico 5 software also calculated the number of hapaxes, or items that occur only
once in a given context, for both each individual submission as well as for each individual
participant. Mellor (2008) found that the proportion of hapaxes in a given text can be used to
differentiate between learner and NS written production. The analysis described in the
subsequent chapter will conduct lexical analyses for individual tasks, as well as for each
participant’s combined production across all three tasks, using both number of types and number of hapaxes as a unit of analysis.

4.6.1 Semi-Lemmatization of Input Data for Calculation of Lexical Measures

In order to determine whether or not lemmatization of written Arabic is able to reveal lexical nuance that might be obscured in the analysis of un-lemmatized texts, the written Arabic collected in this study was subjected to a process of partial lemmatization. As noted in the review of literature above, the calculation of lexical measures for written Arabic texts raises the perennial question of what should be considered to represent a “word”? At their heart, measures of lexical complexity are typically understood to be accessing how diverse, sophisticated, complex, unique, etc., the types of lexical items are that a writer deploys in response to a prompt. Unlike English and other commonly taught European languages, Arabic is a concatenative language in which an entire grammatical sentence may be contained within a single word (see Ryding, 2014). Additionally, Arabic boasts a number of single-character affixes in the form of conjunctions, prepositions, and articles that, when affixed to a lexical item, result new strings recognized as novel items by all software that can be used for the analysis of corpora of written Arabic, including the Lexico 5 software used in the current study. Absent a process of lemmatization or semi-lemmatization as employed in the current study, these dual issues of Arabic’s concatenative nature and multiple affixes combine to greatly increase the apparent lexical diversity in a given Arabic text. However, any process of lexical lemmatization applied to an Arabic text, regardless of how its particulars, will necessarily touch on a number of issues. Researchers in the field of Natural Language Processing have long recognized this dilemma, noting that converting an Arabic text into a series of lexical lemmas interacts with the
unvowelled nature of Arabic texts, leading to graphical conflation of some words; and the complex nature of Arabic word forms, which include single-character conjunctions and prepositions (Dichy et al., 2001).

Morphosyntactic information in Arabic can be represented in the form of affixes, rather than in independent lexical items as is often the case in English or Spanish. The English sentence “we will wait for them” consists of five tokens and five types according to measures of English lexical complexity, while the Arabic سننتظ هم sa-nantażir-u-hum is a single token and type according to available lexical analysis software. Furthermore, the Arabic ستنتظرهم sa-tantażir-u-hum “she will wait for them,” سأنتظرها sa-ʔantażir-u-hā “I will wait for her,” سينتظرني sa-yantażir-u-nī “he will wait for me,” etc., would all be recognized as unique and independent lexical types by software capable of analyzing Arabic corpora, while the English equivalents would consist of largely overlapping sets of repeating types. An analysis of the types and tokens present in the English version of this hypothetical corpus would indicate that the English author employed a high ratio of tokens to overall types, i.e., less lexically diverse language, while an analysis of the Arabic version would indicate an extremely high rate of diversity, with a much lower ratio of tokens to types. Multiplied across all instances of verbal inflection contained in a given passage or a given corpus, the concatenative nature of Arabic will result in a vastly larger number of unique lexical types when compared to an English corpus containing translations of the same constituent texts. A key feature of this type of word formation is that, when Arabic texts are rendered into a corpus, they present a high number of hapaxes, or words that occur only once in the corpus (Appah, 2013, p. 135)

Additionally, the Arabic definite article and conjunction “and” are written as affixes as well, extending sites of apparent lexical diversification to nouns and adjectives, in addition to
verbal inflection. For example, the Arabic strings سورة “a car,” السيرة as-sayyara “the car”, والسيارة, الـسيرة, وـالسيرة, والسيرة wa-s-sayyara “and the car”, بالسيارة bi-s-sayyara “in the car,” وبالسيرة, وبالسيارة wa-bi-s-sayyara “and in the car,” etc., will all be recognized as independent lexical items by a given lexical analysis software, while their English equivalents would recognize that the item “car” had been used in each case. Retaining each of the items described above in the input data would in effect “artificially” increase the amount of lexical diversity present in a given text. This is a key question, however: should the use of a token with a preposition constitute a new type, or should the Arabic token be stripped of the preposition and counted as the original type? The first option will increase the apparent lexical complexity of a response, the second will lower it.

Researchers working with Arabic are left with three broad options to deal with this ambiguity inherent in written Arabic texts, each of which offer trade-offs in terms of time investment on the part of the researcher and the amount of lexical diversity revealed in the final analysis. Researchers may choose to undertake a process of a) full lemmatization, in which each verb, adjective, preposition, etc., is reduced to a single lemma b) partial lemmatization, in which certain items are reduced to a single lemma, but others are retained, or c) no lemmatization, in which the original text is submitted for lexical analysis “as-is”, retaining all inflections, prepositions, articles, etc..

Full lemmatization, such as that which was undertaken to compose the arabicCorpus tool (Parkinson, 2011) and the associated Frequency Dictionary of Arabic (Buckwalter & Parkinson, 2014), entails condensing all verbs (past and present) and verbal nouns into a single lemma, as well as combining singular and plural forms of nouns into a lemma. Although this process would facilitate analyses that may be fruitful in future investigations of written Arabic (such as comparing total number of lexical roots used in relation to Arabic proficiency level, for
example), it ultimately obscures a great deal of semantic and morphosyntactic information present in a given text. In a fully-lemmatized Arabic text, writers who had used the forms تكتب taktub “she writes,” أكتب ʔ aktub “I write,” كتبنا katabnā “we wrote,” كتابة kitāba “writing,” etc., would be recognized as having used four tokens of the same type. This process of reduction may make particular sense in Arabic, which in its most formal form employs 13 distinct pronouns with distinct inflections, as well as mandatory inflection for verbal mood. Full lemmatization would likewise obscure a writer’s choice of a verbal noun vs an inflected verb, for example, or the use of alternate plural forms such as شجرات shajarāt vs أشجار ʔashjār “trees,” which would be collapsed into a single lemma.

The current project implemented a process of semi-lemmatization, in which lexical items were stripped of the definite article, the “and” conjunction, object and possession suffixes, and prepositional affixes. Inflected verbs were reduced to a single lemma (third person, masculine, singular), but verbal nouns were retained as they appeared in participants’ responses. Singular adjectives were reduced to a single gender (i.e., feminine inflected adjectives were reduced to masculine). However, all plural nouns and (human plural) adjectives were retained as they appeared. The goal of this process was to reveal, to the extent possible, the diversity of lexical forms deployed by participants, while not calculating measures of lexical diversity based on input in which every inflected verb or possessed noun, for example, has the potential to represent a novel form, thus “artificially” magnifying the diversity present. Due to the fact that it requires the removal of the definite article, object pronouns, and verbal inflection (all verbs rendered as first-person singular indicative), the semi-lemmatization process implemented in the current study necessarily reduces the morphosyntactic nuance present in the text. By way of example, the Arabic learner-produced T-unit يحبان بعضهما البعض yuḥibb-ān baṣḍ-a-humā l-baṣḍ “they (m.
Additionally, researchers working with Arabic texts that were produced by Arabic NSs must inevitably confront the fact that modern users of written Arabic are likely participants in the incredibly diverse literacy environment described by Hallajow (2016) and Bianchi (2013), in which literacy practices and the orthographic, lexical, and syntactic forms used to carry them out are both highly individualized and highly varied. Therefore, differences in spelling, punctuation, and even differences in whether to inflect a given item according to the rules of MSA or according to those of a participant’s dominant colloquial variety, can lead to a massive statistical diversity of computer-recognized forms. This is equally true when working with Arabic texts produced by learners, in which deviations from standard spelling conventions are abundant. Therefore, a number of decisions are inevitably made during the lemmatization process, including whether or not incorrectly or alternatively spelled words should be “corrected,” ignored in the analysis, or left in their “incorrect” or alternative form. For example, a learner in the current study wrote about a character in the narration prompt that 

"[name] got up before the sun appeared." This T-unit was coded as containing an error\(^\text{17}\), and the spelling error was identified and corrected during the lemmatization process for the purpose of the calculation of lexical measures. If such items are left as-is during the calculation of lexical measures, participants who produce items that contain spelling errors will in-effect be “rewarded” for those errors by being recognized as having produced a rare form.

---

\(^{17}\) In fact this T-unit arguably contains two errors; the second being a possible lexical error in the form of using ظاهير الشمس instead of the standard ظهور الشمس for “sunrise.” Due to the operationalization of accuracy in this study as error-free T-units per T-unit, however, it is not necessary to conduct the secondary research required in order to empirically establish that ظاهير الشمس is an “unacceptable” item.
This lemmatization process clearly requires the researcher to decide and justify the amount of morphosyntactic information they are comfortable removing from a given text in order to reveal information about lexical diversity. Future investigations of lexical complexity in Arabic writing may choose to investigate alternative lemmatization strategies, such as via leaving verbal inflections intact. Additionally, given a large enough written sample from individual participants, it may not even be necessary to fully or partially lemmatize texts for the purpose of lexical analyses. It is important to note the individual decisions made during the lemmatization process described here (i.e., in order to define what constitutes a distinct “word”) are in part arbitrary and up to the researcher to justify. The choices described here were made in the absence of published studies providing an exploration of this issue in the context of the measurement of L2 Arabic production.

4.6.2 Order of Operations for Semi-Lemmatization

After data had been collected, all participant responses were collected into a single word file and labeled with the relevant participant pseudonym and task type. The following process was followed within this document.

1. Remove conjunction و wāw “and” by using the search function to identify a wāw preceded by a space “و” and replacing it with a space. This was not done in instances that this letter was the first letter of a lexical item, e.g., in the common item ولاية wilāya “state.”

2. Remove word-initial prefixes including prepositions and particles by searching for the prefix preceded by a space and replacing them with a space. This step was undertaken individually rather than collectively.
3. Replace certain lexical items with “nonsense” alternatives in order to avoid altering them in the following step in which the word-initial definite article was automatically removed. These items included 
لاِلِـّا “except for”, الله al-lāh “God,” and التحاق ḥāq “joining s.th.”
4. Remove the definite article by searching for “ال” preceded by a space and replacing it with a space.
5. Remove the feminine endings from adjectives by searching for “ة” and replacing it with a space in each case that it is not part of a feminine singular or plural noun (i.e., in the case of طلبة talaba “students”). This process is undertaken individual, rather than collectively, in order to retain the feminine marker on feminine nouns.
6. Reduce inflected verbs to a single inflection, in this case third person, masculine, active voice, present tense, singular. This stage of the process entails searching for variants of verbs and replacing them with the appropriate verbal lemma. For example, searching within the corpus for instances of استيقظ istayqa “he woke up” and replacing them with يستيقظ yastayqi “he wakes up”. 
7. Remove possessive and case endings from singular and plural forms of nouns (which have by now been rendered indefinite). For example, search for all instances of بيته bayt-u-hu “his house” and replace them with بيّت bayt “house”.

In this case, after the initial round of lemmatization was complete, the Word file containing all participant responses and categorized according to author and individual task was
saved as a .txt file and opened in the *Lexico 5* corpus analysis program. This allowed for the separation of the entire corpus into individual types, i.e., types of items that appear in the responses arrayed beginning with those types that occur most frequently in the text. Following this, the researcher visually scanned the list of tokens that appear only once in the corpus in order to identify items that were overlooked during the process described above. For example, items at the beginning of a new line that include a word-initial *wa-* or *al-* would have been overlooked by the find-and-replace function that was instructed to eliminate instances of “و” or “الـ” that were preceded by a space. This stage of the process also identified a number of spelling errors in which a type was spelled incorrectly. In instances in which the researcher could make a reasonable assumption about the intended type, these spelling anomalies were “corrected”, thus reducing the overall number of infrequent types present in the corpus.

This process likewise entails the “correcting” of spelling errors and spelling variants present in the original typed and transcribed responses, in the even that these can be confidently interpreted. For example, a learner participant produced the string *في رأي* “in my opinion,” producing the same letters but in a slightly different configuration than the standard *رأي* – this spelling was “corrected,” thus the participant was credited with using the common type *رأي* “opinion,” rather than the novel *راءي*.

In general, the changes identified during the second stage of lemmatization consisted of overlooked affixes, adjectival feminine markers, object pronouns, inflection of the direct object (accusative case) suffix on masculine indefinite nouns, and remaining inflected verbs. However, every effort was made to respect authors’ choice of regional and register-related spelling variants. For example, one participant produced *دافئ* “warm,” while another provided the alternative spelling reflective of colloquial Arabic morphology *دافي* “warm”. Other items that
were identified during this process included spelling variants that had been previously overlooked but that the Lexico 5 program regarded as unique—for example the corpus contained at this point 20 instances of أخبار "news" and one instance of أخبار "news." As discussed in the discussion of accuracy in this chapter, the presence or absence of the hamza letter was not in and of itself regarded as sufficient for constituting an error, however in order for the Lexico 5 software to treat all instances of “news” as the same type, the hamza-less variant of “news” (and other similar cases) were “corrected”. Clearly, the lemmatization process used in this study resulted in the flattening of a great deal of syntactic and semantic diversity present in the original compositions. For example, all single-character affixed prepositions were simply eliminated, meaning that regardless of how often a participant had used بـ “in” or لـ “for”, that diversity would not be reflected in their scores on lexical measures. Similarly, instances of <i>bi-haadhaa</i> and <i>li-haadhaa</i> were changed to <i>haadhaa</i>. The choices described here do not by any means reflect the only correct way to lemmatize Arabic texts for the purpose of calculating lexical measures—rather they reflect a series of informed choices intended to reveal in as clear a way possible the diversity of lexical types present in the corpus.

4.6.3 Compilation of a Writing Corpus for Lexical Analysis

After the first stage of lemmatization conducted via the find-replace function in Word, Lexico 5 identified 2473 types before this process. Most of the types that were identified and “corrected” in the second stage of lemmatization occurred once or twice. Following this process, the corpus was reduced to 2320 types. The lemmatization of participant responses described above resulted in a text file containing all lemmatized responses that formed the basis of a corpus constructed by the Lexico 5 software package. A check of the corpus at this stage revealed that
the ten most-common items are composed mainly of prepositions, although the influence of the individual tasks and choices made during the transcription and lemmatization process are also evident. As shown in Table 20, prepositions and related particles are by far the most common lexical items used by learners and NSs in this study, however the high frequency of the items جامِعَة jāmiʿa “university”, as well as تدخين tadkhîn “smoking” are undoubtedly the result of the individual experimental tasks, in which participants were asked to both describe their university as well as support or oppose a smoking ban on campus. Additionally, the decision to remove personally identifying information, such as the names of individual participants, specific universities, and names of participants’ friends to whom the experimental task framed as a piece of personal correspondence was addressed, resulted in the high frequency of the item اسم ʔism “name”.

**Table 16: Most Frequent Types**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Type</th>
<th>Transliteration</th>
<th>Gloss</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>في</td>
<td>fī</td>
<td>in</td>
<td>745</td>
</tr>
<tr>
<td>2</td>
<td>أن</td>
<td>ʔan/ʔanna</td>
<td>to/that</td>
<td>441</td>
</tr>
<tr>
<td>3</td>
<td>من</td>
<td>min</td>
<td>from</td>
<td>439</td>
</tr>
<tr>
<td>4</td>
<td>اسم</td>
<td>ʔism</td>
<td>name</td>
<td>368</td>
</tr>
<tr>
<td>5</td>
<td>جامِعَة</td>
<td>jāmiʿa</td>
<td>university</td>
<td>342</td>
</tr>
<tr>
<td>6</td>
<td>إلى</td>
<td>ʔilā</td>
<td>to/toward</td>
<td>294</td>
</tr>
<tr>
<td>7</td>
<td>مع</td>
<td>maṣa</td>
<td>with</td>
<td>188</td>
</tr>
<tr>
<td>8</td>
<td>تدخين</td>
<td>tadkhîn</td>
<td>smoking</td>
<td>179</td>
</tr>
<tr>
<td>9</td>
<td>لا</td>
<td>lā</td>
<td>no/does not</td>
<td>163</td>
</tr>
<tr>
<td>10</td>
<td>على</td>
<td>ʕalā</td>
<td>on</td>
<td>162</td>
</tr>
</tbody>
</table>
4.6.4 Un-Lemmatized Writing Corpus

A second version of the corpus analyzed via Lexico 5 was prepared in which minimal lemmatization was conducted. In this version of the corpus, a simple find/replace function was carried out when all participant responses were compiled in Word, eliminating all word-initial instances of the definite article (i.e., searching for and eliminating "ال" including a preceding space) and word-initial “and” conjunctions (i.e., eliminating all instances of "و" including a preceding space). This version of the corpus retained all NS and learner production as it appeared after transcription, including region-related spelling variants, spelling errors, as well as definite articles, conjunctions, affixes, etc. Participant responses were simply tagged for a participant’s pseudonym and the individual task to which they were responding. This process is necessarily rough, and certainly missed many instances of the definite article and targeted conjunction, however the intent of this operation was to explore what researchers working with written Arabic might consider the “easiest” way of dealing with orthographic complexity in Arabic. This type of un-lemmatized corpus is much easier to compile, and may in fact reveal useful information about the various lexical items used in response to the current study, especially if the texts produced by all participants are sufficiently long. The lemmatized and un-lemmatized corpora will be compared in the subsequent chapter, in order to determine whether or not the detailed lemmatization process described above is necessary, or indeed useful in revealing elements of lexical complexity in written Arabic that would otherwise be obscured.
4.7 Accuracy

This study employs the measurement of error-free T-units per T-unit, introduced by Scott and Tucker (1974) in their study of L2 English produced by L1 Arabic speakers. This measure is now widely used in CAF analyses of written L2 production. Wolfe-Quintero et al. (1998) note in their meta-analysis of studies of L2 writing that the EFT/T measure is among the most effective measures of written accuracy. They note that studies employing learners at different proficiency levels completing timed compositions have collectively found error-free T-units to be the most significant unit of analysis for accuracy. In the context of L1 Arabic, it is known that even NS writers tend to predictably commit “mistakes” while writing, especially on the orthographic/spelling level. This is in keeping with previous studies that have established predictable rates of spelling errors among Arabic NSs.

However, the question of what to consider an error in the context of written Arabic is particularly fraught. Previous studies, such as Al-Ani’s (1972), have attempted to categorize errors in written L2 Arabic production according to type, such as errors in orthography, lexical choice, gender agreement, etc. Brosh (2015) takes a similar approach to the study of orthographic errors in learner-produced Arabic, treating as an error “any deviation from the correct spelling” (p. 591) and providing a taxonomy consisting of nine categories of orthographic error. However, this rigid interpretation fails to accommodate the varied nature of NS spelling in Arabic, and would consequently invalidate or render “incorrect” a great deal of the NS-produced writing collected pursuant to this study. Therefore, given the diverse and multifaceted nature of NS written production in general, a decision was made to avoid the thorny issue of error classification altogether by merely considering whether or not a given T-unit contained an error at all, rather than to classify error according to type. Furthermore, certain written practices
proved to be extremely common among both learners and NSs who participated in this study, and were thus not considered to be errors. These include the use of “colloquial” lexical items, as well as variation in the writing of the letter hamza, both of which are discussed below.

Instances of NS and learner production that were coded as an error involved departures from the morphosyntactic (rather than lexical) norms of MSA, as well as spelling errors judged not to be reflective of colloquial spelling. Examples of items considered to be errors are presented below, and included spelling errors, gender/number disagreement, and departures from the MSA case system that would normally be evident in unvowelled texts, such as failure to represent the mansūb (accusative) case on an indefinite, masculine, singular noun. Errors in gender and number agreement appear to be more common in the current study in the writing of Arabic learners, rather than NSs. For example, a high-achieving Arabic learner described a male character the narration task by writing ثم تخرج من البيت thumma takhruj min al-bayt “then [s]he leaves the house,” substituting the female singular inflection of this verb for the male. This was coded as an error, in spite of the fact that this T-unit could be argued to be grammatically correct in isolation, due to the fact that the inflected verb does not agree in gender with the already-named subject described in the rest of the submission. Ultimately, errors identified in the context of this study typically took the form of agreement errors, lexical errors (i.e., missing or incorrect words), and orthographic errors. In general, T-units containing these errors were much-more-common among Arabic learners than among NSs, although NSs likewise produced errors.

The following section presents classes of items that were either coded as errors or were “accepted” that emerged from the written production of NS production in this study.

---
18 Due to the unvowelled nature of Arabic texts, this T-unit may also be read as thumma takharraja min al-bayt “then he graduated from the house.” This reading would likewise result in this T-unit being classified as containing an error, although the error in this case would be lexical, rather than inflectional.
4.7.1 Instances of NS production coded as errors

In spite of the fact that colloquial items, items determined to be regional spelling variants, and variation in the use of the *hamza* letter, were not coded as errors, the NSs who participated in this study nevertheless produced a large percentage of T-units that contained errors. The identification of the examples presented here allowed for the classification of T-units as either error-free, or as containing an error. In certain cases, follow-up research was conducted to confirm that problematic instances of NS production elicited in the current study were indeed “errors,” or deviations from the norms of written MSA.

4.7.2 Missing, Alternative, or Superfluous letters

NS participants often supplied superfluous letters or items, or alternatively omitted necessary letters from a given lexical item, resulting in a T-unit that was coded as containing an error. For example, a NS participant included a superfluous preposition in the descriptive task, describing their university campus as في بالقرب من المنزل *fi bi-l-qurb-i min al-manzi* “in in proximity to the home.” A second NS wrote يدرسوننا، presumably accidently including a typo in the intended *yudarrisīn-a-nā* “they teach us.” In other cases, a NS participant omitted a letter from a lexical item, resulting in a misspelling. A NS produced the following T-unit during the narration task: يحرص هذا الشخص عل تناول العشاء مع زوجته *yāḥris hādhā šāhshāḥal tanāwwul il-ẓashā? maš zawjat-i-hī* “this person is intent on eating dinner with his wife,” in which the preposition على *ẓalā* “on” is provided absent its final letter.

The decision of which forms to consider “acceptable,” and which forms to consider “unacceptable,” in the context of a study such as this inevitably confronts the individual
researcher with the unpalatable task of weighing in on a highly contentious and ideologically charged debate regarding the “correct” form of written Arabic. In the current effort, every effort was made to take a descriptive view of accuracy, including basing judgments of acceptability on the performance of NS participants as well as usage practices determined by secondary research. This is often a complicated process, however; for instance, one NS participant consistently spelled “university campus” as هرم الجامعة haram al-jamiʿa, rather than the standard حرم الجامعة ḥaram al-jamiʿa. At first glance, this appears to be a misspelling of a common collocation, substituting one very similar letter for another. However, a Google search for the string “حرم الجامعة” nevertheless returns over 5,000 results,19 indicating that this spelling is extant, if not widely used (compared to the over 400,000 instances of “حرم الجامعة”). However, within the curated subcorpora that form the body of the arabiCorpus, most of which are published newspapers, this spelling variant is entirely absent. This can be compared to 181 instances of the standard “حرم الجامعة” in the All Newspapers subcorpus. Therefore, this spelling variant was coded as an error, in spite of the fact that it is clearly used in at least some rare cases by NS writers. It remains unclear at what point a spelling variant becomes common enough to warrant consideration as an “acceptable” form, but in the absence of a consensus on this issue, inclusion in arabiCorpus was taken to be an important indicator of acceptability within the context of the current study.

4.7.3 Case and Verbal Mood, and Related Inflection Errors

Deviations from the complex MSA system of verbal mood inflection and noun/adjective case inflection by both NSs and Arabic learners were also counted as errors. Examples of verbal

19 Searches conducted on 11/30/2016
mood errors in NS production produced in the context of the current study typically take the form of the failure to account for the jussive mood when inflecting verbs that include a root-final vowel. For example, a NS participant produced the T-unit depicted in example 5 in response to the persuasive prompt, in which the participant included a negative particle لَمُ lam that requires the verb it negates to be realized in the imperfect jussive mood. However, the inflected verb produced by this participant was inflected according to either the indicative or subjunctive moods (it has been bolded below):

Table 17: NS Verbal Mood Error

<table>
<thead>
<tr>
<th>Table 17: NS Verbal Mood Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>حيث أن فرض القوانين في الجامعة والعمل على تطبيقها بشكل فعال لن يجدي اذا لم يتفاهم مع قوانين الدولة والمجتمع</td>
</tr>
</tbody>
</table>

haythu ظanna furuṣ-a l-qānūn fī l-jāmiʿa w-l-؟امال ظال تابثiq-i-hâ bi-shakal faṣiḥ lan yajdâ

ٍّدَهَlam tatamâshâ maṣʿ-hu gawâānîn ad-dawla w-l-mujâmaṣʿ ka-kul khârij al-ḥaram al-jâmiʿî

“Whereas imposing laws in the university and working to apply them in an effective way will not work if **they are not in harmony** with the laws of the state and society as a whole outside of the university’s campus”

According to the rules of MSA verbal mood inflection, form VI verbs such as the verb تماشى tamâshâ “to conform to, be in harmony with” that are composed of a triliteral root of which the final root letter is realized as an ٍّلîf maksûra should be realized with a word-final fatḥa in the jussive mood. As in similar cases described in this chapter, in which the possibility exists that the NS production collected in response to the experimental prompts may represent “acceptable” variation from orthographic norms, the arabicCorpus tool was used to confirm the existence of apparent variants in the genre of media Arabic. The classification of the inflection لم تمتشى lam
tatamāshā as an error was confirmed by a series of searches in the “All Newspapers” subcorpus of the the arabicorpus tool, which found no occurrences of the string لم تتماشى or its masculine equivalent لم يتماشى. The arabicorpus does contain two instances of the correctly inflected string لم تتماش و لم يتماشى, and three of the correct string’s masculine equivalent, لم يتماشى, however. This type of error, i.e., errors in the inflection of the jussive mood in verbs with word-final vowel, occurred elsewhere in the corpus of learner and NS responses. For example, a second NS produced the T-unit لكن يوم زوجته لم ينتهي بعد lākinna yawm zawjat-i-hi lam yantahī ba’d “but his wife’s day had still not ended,” in which the problematic لم ينتهي was provided instead of the correct لم ينته lam yantahi. The “All Newspapers” subcorpus of the arabicorpus contains 25 instances of the problematic لم ينتهي, as opposed to 732 of the “correct” لم ينته, thus the former was coded as an error in this study.

Errors in the inflection of the case system that governs the realization of nouns and adjectives in MSA likewise occasionally occurred in the production of NS participants, as in the failure of a NS to inflect an indefinite masculine noun in the accusative case, writing في كل مبنى ستجد محل لبيع القهوة fī kul mabnā sa-tajidd maḥall li-bīṣ-i l-qahwa “in every building you will find a store selling coffee,” whereas according to grammatical rules of MSA, the lexical item in question should have been rendered محلا mahall-an as an indefinite masculine singular noun in the accusative case.

---

20 Searches conducted on 11/30/2016
4.7.4 Instances of NS Written Variation Not Coded as Errors

Just as researchers working with written Arabic are required to justify their choices of what constitutes an “error,” even in the writing of NSs, written Arabic inevitably presents several types of variation that can convincingly be argued to be “acceptable” deviations from written norms, such as they are. Within the context of the current study, two main areas of variation were not regarded as representing errors, namely the “correct” (i.e., correctly inflected) use of colloquial lexical items, as well as common spelling variants.

4.7.5 Inclusion of Colloquial Lexical Items

A number of NS and Arabic learner participants in this study produced lexical items that can be considered to belong to the register of “colloquial” Arabic. The act of writing colloquial Arabic, whether in response to an experimental prompt, in the context of a newspaper article, or in any other supposedly “formal” context is inherently a controversial act (see Fahmy, 2011 for a discussion of this issue). However, absent a consensus regarding this ideologically-fraught issue, it is not necessarily the individual researcher’s responsibility to actively endorse or subvert the lines that allegedly separate the various registers of Arabic writing. Therefore, given that NSs and learners used colloquial items frequently in this study, such items were not coded as errors if they were appropriately inflected.

For example, a NS participant produced the following two T-units in example 6 in response to the descriptive prompt, employing in the process several interesting items that were not coded as errors:
"…regarding my university, I see it as very pretty, and you know that I love trees and nature”

This two-T-unit utterance contains a colloquial item *shayfat-hu* “I see it as” in the main clause of the first T-unit. This item is drawn from the colloquial register of Arabic and is a common item found in many regional dialects, including Egyptian, Gulf and Levantine. This item is likewise correctly inflected for a female speaker (i.e., *shāyfa* vs the male *shāyf*). Given that this item and other colloquial lexical items discussed in this section are correctly inflected and semantically appropriate, such instances were not coded as errors. Additionally, this participant used what can be described as MSA elsewhere, including the verb inflected according to the standards of MSA for the second person feminine singular *taṣrifīn-a* “you (f.s.) know.”

Interestingly, the co-occurrence of these two items (the colloquial active participle *shāyfa* and the MSA verb *taṣrifīn-a*) could be considered an instance of the diglossic code switching described by Bassiouney (2006). This behavior occurred both in the writing of NSs, as well as that of high-achieving Arabic learners, e.g., in the learner-produced T-unit depicted in Table 23:

**Table 19: Colloquial Lexical Items**

"therefore, any subject you want to study, there’s work and opportunity here”
According to the taxonomy of colloquial vs MSA structures laid out by Bassiouney (2006), this T-unit likewise represents an instance of diglossic code switching – while the first clause includes the MSA relative pronoun الذي al-ladhī, the second clause includes the (presumably Levantine) item hōn “here”, as well as the colloquial item ⲳⲧⲧⲧī “there is”. Rather than count this instance of code switching as an error, this T-unit was counted as an error-free T-unit. Whether the combination of items used in this T-unit represent a deviation from the stylistic norms of academic writing remains an open question, however.

Additional colloquial items produced by Arabic learners that were deemed “acceptable” include the item هون hōn “here”, rather than the MSA item هنا hunā, as well as the inflected verb يستنتني yastanā “he waits.” This item appeared in the context of the following learner-produced T-unit:

yastanā ʕalā mawqaʕ l-ʔawtūbīs ʕalā ʔakhad jānib min bayti-himā fī minṭaqaṭ ʔism

“He waits at the bus stop to take [the bus] beside their home in the ‘name’ area”

This T-unit is marked by several items that can be seen to contain a number of elements typical of colloquial Arabic, including the verb yastanā/yastanī,21 rendering the letter ذ dh as it appears in MSA as ذ d as it appears in several colloquial dialects, and possibly the use of ʕalā as a substitute for ʕilā, thus illustrating the complexity inherent in the task of classifying errors in

21 There is speculation as to the source and “correct” form of this verb in MSA, however a search in the “All Newspapers” subcorpus of arabicCorpus (conducted 3/21/2017) returned only two instances of yastanī used as a verb a newspaper article. This occurs in Shabāyk (2010), an editorial clearly written in Egyptian colloquial Arabic. A second instance of this verb occurs in the comment section of an article in the Syrian Al-Thawra newspaper, and co-occurs with several Levantine colloquial forms.
written Arabic. This T-unit was ultimately coded as containing an error on the basis of the string على أخد جانب من بيتهما ʕ alā ʔakhad jānib min bayti-himā. Several additional participants substituted a colloquial spelling variant of a given item, e.g., in the responses of several NS participants who referred to سجائر sijāyyir “cigarettes,” rather than using the standard spelling سجائر sijāʔir. Most instances of colloquial Arabic that appeared in the production of Arabic learners and NSs likewise occurred on the lexical, rather than syntactic level, with the notable exception of the learner-produced negation construction ما عندهم mā ʕand(a/i)-hum “they do not have.”

4.7.6 Common Spelling Variants

Two types of orthographic variation were not classified as errors in the context of this study. These instances of variation were “accepted” largely because they are so common in the writing of NSs, both within the responses collected in this study and in Arabic writing produced in a diverse range of contexts. The most frequent types of spelling variation consisted in variability in writing the letter hamza, as well as substituting the word-final letter ـه –ah for the feminine marker ـة a(t). In addition to these two categories discussed below, certain spelling practices commonly used in Egypt were not counted as mistakes, even if they might be viewed as such by NSs from a different region of the Arab world. The Egyptian spelling style that manifested most frequently in this data was the (mostly) Egyptian practice of writing a word-final ـي i without the accompanying dots (i.e., as ى).

4.7.7 Omission/Placement of hamza

Variable placement of the hamza letter was likewise not coded as an error in the data that forms the basis of this dissertation. Simply put, the rules that allegedly govern hamza placement
in written MSA are adhered to intermittently in practice. The *hamza typically* represents a glottal stop and can appear in written Arabic either “aloof,” i.e., written alone (اء) or written on one of three seats corresponding with concurrent vowel sounds, i.e., it can be written as ِ، ى، or َ. The rules that govern the appearance (i.e., on a seat or aloof) of the *hamza* are fixed and determined in large part by both the *hamza*’s position in the word, as well as adjacent vowel sounds. Additionally, certain classes of verbal nouns (e.g., verbal nouns of form IV verbs) are *hamza-initial*, however other classes of verbal nouns (e.g., verbal nouns of form VIII verbs) are typically written with the alternative form *hamzat al-waṣl*, i.e., the *hamza* may be pronounced, but its shape is orthographically absent. Ryding (2005, pp. 16-20) provides a detailed examination of the various rules that govern how the *hamza* is realized in modern written Arabic.

However, an examination of *hamza* use in the responses of both Arabic learners and NSs indicates that, at least in certain contexts, the appearance of this letter is highly variable.²² Both Arabic NSs and learners evinced a great deal of diversity in terms of including or omitting this letter. Furthermore, certain instances of variable *hamza* placement are clearly regional in nature. For example, an Egyptian NS participant in this study produced spelling variants that subsequent research established as associated with MSA produced in Egypt. These include مسؤول *masʔul* “responsible,” instead of the “standard” مسؤول *masʔul*. A search on the *arabiCorpus* website²³ confirms that the spelling variant مسؤول *masʔul* occurs almost exclusively in Egyptian subcorpora of the *All Newspapers* corpus, providing cursory evidence for the fact that it is indeed an “acceptable” spelling variant in Egyptian print media. In addition to regional variation, NSs and learners also exhibited significant variation regarding the presence or absence of the *hamza* in lexical items in

---

²² In fact, variability in *hamza* use can be easily found in a wide range of contemporary Arabic newspapers. Instances of variation in the realization of this letter include both omission of the *hamza*, as well as inclusion in contents where it “should not” appear, such as part of a form VIII verbal noun, for example.

²³ Search conducted on 9/29/2016.
which this letter “should” appear. For example, a NS participant in this study produced the item يودي rather than the standard يؤدي “leads to”. Due to the sheer frequency of “missing” hamzas, no such instance was counted as an error, regardless of whether or not any other participant in this study produced a given hamza-related spelling variant.

Finally, participants in this study showed a great deal of innovation with regards to the types of seats on which they placed word-medial hamza’s. These letters are ideally governed by a hierarchy of adjacent vowel sounds (as described by Ryding, 2005), however these rules were adhered-to intermittently in the current study. For instance, two NS participants produced the prepositional phrase مع أصدقائه maʕaʔaṣdiqāʔi-hi “with his friends,” in which the hamza appears on the “correct” yāʔ seat. However, one NS participant produced the identical phrase, although with an alternative realization of the hamza: maʕaʔaṣdiqāʔi-hi24 “with his friends.” Variability of hamza seat was likewise evident in an item produced by a separate NS who wrote قارئ rather than the standard قارئ “reader.” As this alternative spelling variant occurred in the context of an attempt to place a hamza in an Arabic word, it was not coded as an error.

4.7.8 Omission of dots/substitution of ـه ah for ـة a(t) in word-final feminine marker

A number of both Arabic NSs and learners employed a spelling strategy in their responses that, while it can be considered a deviation from the norms of written MSA, is actually extremely common in written Arabic, especially non-formal or non-formalized writing contexts. This strategy was clearly evident in the data collected as part of this study, and consists of the occasional or systematic omission of the dots from the word-final feminine marker ـة a(t), known as the تاء مربوطة تاء التأنيث تاء التأنيث “bound tāʔ” or tāʔ marbūṭa “bound tāʔ” or تاء التأنيث تاء التأنيث “feminizing tāʔ”. The tāʔ

24 In this case, the pronunciation of this item indicated by the realization of the hamza conflicts with the pronunciation dictated by the case rules of formal Arabic.
marbūṭa is a spelling variant of the letter ـت which occurs in the word-final position on nouns and adjectives and is used to indicate feminine gender. In practice, it is distinguished from a word-final ـه only by the addition of two dots over the letter. Ryding notes that in most situations in which Arabic words are pronounced in pause form (i.e., pronounced without vocalizing case markers), “the pronunciation of the taa’ marbuuTa becomes haa’. Because the final /h/ is hard to year, it sounds as though the word is pronounced only with a final /al/, the fatHa that precedes the taa’ marbuuTa” (p. 23). However, if the tā? marbūṭa occurs on a word that is not the final word in an ʔidāfa structure, it is pronounced.

Variability in hamza placement (i.e., omitting it in spite of orthographic rules that seem to compel its use) may be related to common practices in Arabic handwriting in which the word-final dots on the feminine marker ـة a(t) are often omitted. Regardless, the fact remains that this practice—omitting the dots from the tā? marbūṭa—is a common feature of both handwritten and typed Arabic. These dots are often eschewed in typed Arabic in spite of the fact that there are separate keys for the hā? and tā? marbūṭa on most common Arabic keyboards. This practice is salient enough that one or more notes on its use have been made on the style guide of the Egyptian Colloquial (maṣrī) version of Wikipedia instructing authors on when to use the feminine marker ـة a(t) and when to use the dot-less ـه ah. Under the Letters section of the style guide, the site includes a note on the feminine marker which instructs writers that the choice of ـة a(t) or ـه ah is determined by pronunciation; that is, if a feminine noun is possessed in an ʔidāfa structure, the feminine marker ـة a(t) should be used, while if the feminine noun does not occur within such a structure, the letter ـه ah should be used. The guide contains the following bullet point regarding use of the feminine marker:25

25 As of November, 2016.
• Feminizing tah: Feminizing tah is written according to the pronunciation[,] either /a(t)/ or /ah/, so you write <I’m going to the library [مكتبة maktabah]> or you write <I’m going to the library of the college [مكتبة الكلية maktabat al-kuliyah]>. (siyāsāt, n.d.)

It is by no means clear whether or not those NS or Arabic learner participants who responded to written prompts within the context of the current study substituted ah for a(t) in ways which conform to the suggestion of the maṣrī version of Wikipedia. It is not even clear, for that matter, whether or not individual authors who contribute to the maṣrī Wikipedia follow this guideline. Rather, this note should be taken as evidence that variability in hamza placement is a common and marked practice. Future research could investigate whether or not this “switch” is conditioned, either in the manner suggested by the maṣrī style guide or otherwise.

Regardless, instances of this substitution present in responses submitted as part of the current study were not coded as errors, due to the fact that this practice is extremely common in handwritten and typed texts produced within a variety of written contexts. For instance, a NS participant in the current study, in describing their university, referred to قسم المباني الإداريه qism al-mabānī l-ʔidāriyya “the section of administrative buildings,” substituting –ah for –a(t) as the feminine adjectival suffix. However, it should be noted that the opposite was not considered to be the case, i.e., that the substitution of a ـ a(t) for aـ ـha/hi was coded as an error. For instance, an Arabic learner produced the following T-unit: َلا احبة لأنه ليس قريب من بناية "اسم" Lā ṭuḥibbu-(hu) li(ʔ)anna-hu laysa qarib min bināyat “ism” “I don’t like it because it is not close to ‘name’ building.” The reason for this choice, i.e., that substituting aـ for aـ is not considered
an error, but that the reverse is coded as an error, is that the first substitution is extremely common, including those NSs who participated in this study. The reverse substitution only occurred in the texts of Arabic learners, and in a sporadic fashion.

4.8 Fluency

Fluency refers to the rapidity with which learners produce language within a specific temporal period (Skehan, 1998). Even NS language users tend to be highly variable in their rates of measured oral fluency (Derwing et al., 2009), and it is fully expected that this variability will abound in the rates at which Arabic NSs compose written texts. Although fluency measures can take many forms in the oral mode, including measures of speech rate, quantity, and pausing (Ginther, 2010), the current study explores the role the interaction between written fluency and other related constructs of written CAF. Ultimately, researchers working with written data typically only have access to the final written product, rather than a record of the product’s composition, meaning that certain measures of oral fluency do not easily transfer to the written mode. Therefore, the current study employs the easy-to-calculate measure of words per minute. This value was calculated for each participant by dividing the total number of words in each submission by each participants’ (reported or observed) time on task for each writing prompt.
Chapter 5: Results

5.1 Results Overview

As described in the Methodology chapter above, all learner and NS participants in this study completed three timed written compositions in response to prompts that aimed to elicit written production oriented the linguistic functions of description, narration, or persuasion. Each participant’s production across all three tasks was coded and analyzed according to the direct measures of written CAF described previously. The results presented in this chapter, in the form of descriptive statistics, correlation analyses, and regression analyses, indicate that several measures of written Arabic CAF have statistically significant relationships with independent estimates of learner proficiency. Additionally, the C-test was shown to be one of the strongest collected measures of Arabic proficiency collected from learners in this study.

This chapter begins by exploring the results of the C-test and its relationship to other independent measures of Arabic proficiency, including self-estimates of proficiency and years of study. Learner C-test scores will be used to break learners into two groups, and subsequent analyses include comparisons of group means to determine whether or not the C-test-derived groupings are statistically significant. This chapter will then move on to present results of CAF measures among these groups. Next, a series of standard regression analyses will be presented in order to determine which independent variables appear to be contributing most to the selected CAF measures. Finally, learner and NS responses will be broken down by genre, and group means will be compared in order to determine whether the genre of written prompt has a measurable effect on the complexity, accuracy, or fluency of a particular group.

The current study compared a number of independent variables, such as C-test score, years of study, and self-estimates of Arabic ability, with the CAF measures displayed in the
Table 24 below. These measures attempt to access the constructs of syntactic complexity, lexical complexity, accuracy, and fluency within the writing of Arabic learners and NSs, and are referred to throughout this chapter according to the indicated codes:

**Table 20: Summary of CAF Measures**

<table>
<thead>
<tr>
<th>Construct</th>
<th>Type</th>
<th>Measure</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic</td>
<td>Overall Complexity</td>
<td>Mean Length of T-Unit</td>
<td>W/T</td>
</tr>
<tr>
<td></td>
<td>Coordination</td>
<td>T-units per Sentence</td>
<td>T/S</td>
</tr>
<tr>
<td></td>
<td>Length of Production</td>
<td>Mean length of sentence</td>
<td>W/S</td>
</tr>
<tr>
<td>Lexical</td>
<td>Lexical Richness</td>
<td>Guiraud’s index</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hapax Index</td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td>Accuracy</td>
<td>Error-free T-units per T-unit</td>
<td>EfT/T</td>
</tr>
<tr>
<td>Fluency</td>
<td>Rate of Production</td>
<td>Words per minute</td>
<td>W/M</td>
</tr>
</tbody>
</table>

The coding of individual samples was conducted in MS Word, the calculation of the CAF measures (typically ratios) for each participant was conducted in MS Excel, and the descriptive statistics reported below were calculated in SPSS. With the exception of $p$ values, or unless otherwise noted, all statistics presented in this chapter have been rounded to two decimal places. Confidence intervals (95%) have been included as part of the descriptive statistics presented below in order to facilitate between-group comparisons—ideally, groups that are significantly different from each other will not present group means that fall within the CIs of another group. Given the relatively small number of participants, power analyses such as Cohen’s $d$ may be unreliable for comparing group means, however $d$ will still be provided for a number of group
comparisons. In general, the various statistics and tests presented below will be explained in the
text, in order to facilitate their interpretation by Arabic language professionals who may be
unfamiliar with their use in the context of an empirical study of linguistic CAF.

The findings presented here will be discussed the subsequent chapter, however at this point these findings establish that (a) written CAF among this group of participants is
significantly and positively related to several independent estimates of learner proficiency, (b)
NSs are indeed more complex, accurate, and fluent writers than even advanced learners, although they exhibit a great deal of within-group variation, (c) lemmatization does not appear to be
necessary for the analysis of lexical complexity in written Arabic, (d) sentence-based measures
appear to be wholly unreliable for the empirical analysis of learner-produced written Arabic, and
(e) persuasive writing in Arabic is measurably more syntactically complex, but less accurate,
than descriptive or narrative writing.

5.2 C-test and Related Independent Variables

The 5-text C-test administered to learners represents an important axis of analysis in this
dissertation, and is likewise shown via the regression analysis presented below to be the
independent variable most consistently related to written CAF. The 35 learners who participated
in the main data collection of this dissertation are a different population than the participants who
participated in the development of this instrument that was described previously. In addition to
the C-test, each Arabic learner is likewise associated with a number of additional independent
variables, several of which represent estimations of Arabic proficiency or writing ability.
Specifically, learners supplied self-estimations of overall Arabic ability and Arabic writing
ability, supplied the number of years they had studied Arabic, estimated the frequency with
which they interact in Arabic with native speakers, and described their frequency of participation in a number of literacy-related practices. Several of these measures are shown here to have a statistically significant relationship with each other (and with CAF), and a few are not, such as reported literacy practices. This section will present the descriptive statistics associated with these measures, as well as explore the results of the C-test, which was used to organize the analysis of CAF measures presented later.

All 35 Arabic learners completed the 5-text C-test described in a previous chapter. As participants were drawn from a range of Arabic ability levels, the resulting C-test scores are quite diverse. Nevertheless, they show a strong positive correlation with several other independent measures of Arabic proficiency, such as self-reported years of study, self-estimates of Arabic ability, and self-reported frequency of interaction (in Arabic) with Arabic NSs. Additionally, as the C-test is expected to create wide distributions among test-takers, the observed abundance of within-group variation is not unexpected, especially in such a small sample size of 35 test-takers. The mean score for Arabic learners on the C-test is 65.51, falling very close to the middle of the 125-point scale of this instrument.

**Table 21: Learner C-Test Scores**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statistic</th>
<th>Score (n=35)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-Test Score</td>
<td>Mean</td>
<td>61.51</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>19.04</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>34</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>105</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>71</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>32</td>
</tr>
</tbody>
</table>
Based on the proportionally large standard deviation value of 19.04 points, it can be seen that this group of C-test scores are dispersed quite diffusely around the mean score of 61.51. Additionally, the observed range of scores is quite wide, ranging from the minimum score of 34 to the maximum score of 105. All C-test scores are displayed below, plotted for ease of interpretation against the alternative independent variable Years of Study (YoS). From the graph displayed in Figure 5, it can be observed that there is a general positive trend resulting from the correlation of these two variables—as learners in this group were associated reported more years of Arabic study, they also tended to score higher on the C-test instrument. Each point in this graph corresponds to a single participant’s associated C-test and Years of Study values:

![Figure 5: C-test Score vs YoS](image)

A Pearson correlation coefficient was computed for these two variables, producing an $r$ value of .70 ($p < .001$). The closer a correlation coefficient ($r$) approaches +1 or -1, the stronger the relationship between the two variables can be said to be. Alternatively, a coefficient of 0 would indicate no relationship. Typically, a coefficient of 1 would indicate an exact linear correlation between two variables, a coefficient of .70 would indicate a strong positive correlation, and a
coefficient above .50 would indicate a moderate positive correlation. The C-test/YoS correlation was likewise used to produce an R\(^2\) value of 0.48. The R\(^2\) value is represents the shared variance between the two measures, or in other words, the extent to which the points on the graph fall on the regression line. In general, the closer the R\(^2\) value is to +1/-1, the more two given variables can be said to overlap (Larson-Hall, 2015). Thus, the amount of variance shared by C-test score and YoS is 48%. This is not a perfectly linear relationship, as this group of participants exhibited a wide range of both possible C-test scores and possible number of years studied. This can be seen in the graph in Figure 5, for example in the fact that C-test scores for learners who had completed one year of study ranged from 34 to 57, and among learners who had completed three years of study C-test scores ranged from 43 to 91. However, given that the C-test is expected to create wide distributions, and that YoS is known to have a nebulous relationship with other empirical measurements of Arabic proficiency (Rammuny, 1983), an R\(^2\) value of 0.48 for these two measures is encouraging.

For the purpose of visualizing the scores associated with the variables investigated in this dissertation, it, it is advantageous to break learners into two groups on the basis of their C-test scores. These groupings will facilitate comparisons used in subsequent analyses. To this end, a cut-off point of 70 was identified with respect to C-test score, allowing for the identification of a Low Performance (LP) and a High Performance (HP) group. This point was chosen (a) because there is a natural 6-point gap in scores between the highest LP score (64) and the lowest HP score (70), and (b) because this cut-off score falls almost exactly in the middle of the measured range of learner scores (69.5). This cut-off point leaves the LP group with 21 participants, and the HP group with 14. The following table presents the descriptive statistics of LP and HP C-test scores.
Table 22: Learner C-Test Scores According to Group Membership

<table>
<thead>
<tr>
<th>Variable</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>C-test score</td>
<td>48.19</td>
<td>81.50</td>
</tr>
<tr>
<td></td>
<td>9.23</td>
<td>9.99</td>
</tr>
<tr>
<td>Range</td>
<td>30</td>
<td>35</td>
</tr>
<tr>
<td>Minimum</td>
<td>34</td>
<td>70</td>
</tr>
<tr>
<td>Maximum</td>
<td>64</td>
<td>105</td>
</tr>
</tbody>
</table>

It can be immediately observed that these two groups are much more tightly clustered than when all learners were considered together, as the SD for each group has been cut in half. Additionally, the HP and LP groups present very similar ranges (35 and 30, respectively). Finally, these groupings—determined solely on the basis of C-test score—correspond to significant differences in group means when used to explore the various independent proficiency-related variables collected via the current study.

5.2.1 Years of Study (YoS)

Arabic learners who participated in this study also responded to a number of questions regarding other aspects of their Arabic abilities and practices via a language background questionnaire. Pursuant to this questionnaire, learner participants were asked to record the number of years they had studied Arabic, or to provide their best estimate. LP learners reported studying Arabic for an average of 2.33 years at the time of data collection, and HP learners for an average of 5 years. Table 28 presents the mean and median YoS values for the HP and LP groups, as well as the 95% CI, i.e., the range of values in which 95% of the measured parameter can be expected to occur. These two groupings can be said to be significantly different on the
basis of a comparison of group YoS CIs, in which the LP mean does not fall within the HP CI [3.70, 6.30], and vice versa:

**Table 23: Learner Years of Study**

<table>
<thead>
<tr>
<th>Variable</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of Study</td>
<td>Mean</td>
<td>2.33</td>
</tr>
<tr>
<td></td>
<td>95% CI-lower</td>
<td>1.57</td>
</tr>
<tr>
<td></td>
<td>95% CI-upper</td>
<td>3.10</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>1.68</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>6</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>7</td>
</tr>
</tbody>
</table>

The significance of difference between these group means is confirmed via an independent-samples t-test. A t-test is associated with a *p* value, which is typically regarded as statistically significant if it is lower than .050. In this case, a t-test reveals that the differences in HP and LP YoS means is statistically significant; t(33) = 4.01, *p* < .001. Finally, t-tests are often reported in conjunction with Cohen’s *d*, an indication of how many standard deviations group means differ by. (Plonsky (2015) cautions that when there is a great deal of variance between group scores, i.e., large values of standard deviation, there is a higher probability of overlap between group scores. This in turn results in a decreased likelihood of statistical significance being attached to the results of a t-test, and their corresponding *d*-values will be smaller (pg. 39). Therefore, it can be said that grouping learners on the basis of C-test score has resulted in two groups that present a statistically significant difference regarding average number of reported years of Arabic study.
5.2.2 Self-estimates of Arabic ability

Arabic learners were also asked to provide self-estimates of their Arabic ability in the four main skill areas (reading, writing, speaking, and listening), as well as an “overall” estimate, on a 5-point Likert scale. As described previously, learners were asked to provide a rating for each area ranging from 1 “very low/beginner” to 5 “native/native-like”. The estimates of overall ability and writing ability were determined to be the most relevant to the current study of Arabic writing and are presented below. The group means for these two estimates (overall ability and writing ability) are quite similar to each other within both the LP and HP groups. For example, the LP group collectively reported a self-estimate of 2.24 for both overall ability and writing ability, while the HP group reported an overall ability of 3.29 and writing ability of 3.07. LP and HP group means were determined to be significantly different for both overall and writing estimates on the basis of a comparison of group CIs. The LP overall estimate (M = 2.24) occurs outside the HP overall CI [2.76, 3.81] and vice versa.

Table 24: Learner Self-Estimated Overall Ability

<table>
<thead>
<tr>
<th>Variable</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Estimate</td>
<td>Mean</td>
<td>2.24</td>
</tr>
<tr>
<td>(overall)</td>
<td>95% CI-lower</td>
<td>1.81</td>
</tr>
<tr>
<td></td>
<td>95% CI-upper</td>
<td>2.67</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0.94</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>4</td>
</tr>
</tbody>
</table>
Table 25: Learner Self-Estimated Writing Ability

<table>
<thead>
<tr>
<th>Variable</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self-Estimate (writing)</td>
<td>Mean</td>
<td>2.24</td>
</tr>
<tr>
<td></td>
<td>95% CI-lower</td>
<td>1.89</td>
</tr>
<tr>
<td></td>
<td>95% CI-upper</td>
<td>2.59</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>2</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>0.77</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>4</td>
</tr>
</tbody>
</table>

The significance of the difference between both groups on these measures is confirmed via an independent-samples t-test. In this case, the differences between LP and HP self-estimates of overall Arabic ability are statistically significant (t = 3.28, p = .003, d = 1.13), and the difference between LP and HP Arabic writing ability self-estimates are likewise significant (t = 3.00, p = .006, d = 1.05). The associated d values indicate a strong effect size, although this power estimate is notably less effective when used in conjunction with small sample sizes.

5.2.3 Frequency of Interaction with Native Speakers

This study also collected information on the self-reported frequency of Arabic learners’ contact with NSs of Arabic on a 1-5 scale, with numbers on the scale representing frequencies ranging from “never” to “daily”. This measure did not differentiate between learners’ contact with instructors and NSs outside of class; rather, it merely asked learners to provide an estimate of how frequently they use Arabic to interact with NSs or “fluent” Arabic speakers. Although the
group means are relatively close to each other, this measure reveals that HP learners collectively report more frequent Arabic interactions with NSs (M = 4.07) than LP learners (M = 2.95):

**Table 26: Frequency of Interaction with Native Speakers**

<table>
<thead>
<tr>
<th>Variable</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Native Speaker</td>
<td>Mean</td>
<td>2.95</td>
</tr>
<tr>
<td></td>
<td>95% CI-lower</td>
<td>2.37</td>
</tr>
<tr>
<td></td>
<td>95% CI-upper</td>
<td>3.54</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>3</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>1.28</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>4</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>5</td>
</tr>
</tbody>
</table>

The HP group is more tightly clustered than the LP group, displaying a narrower range and smaller SD value, and the HP group likewise falls higher on the scale of possible frequency of NS interactions. The difference between these groups can be said to be significant on the basis of a comparison of group CIs, as well as an independent-samples t-test; t (33) = 2.79, p = .009, d = .93).

The HP and LP groupings thus correspond to statistically significant differences in group means according to the independent variables represented by C-test score, YoS, self-estimates, and reported frequency of interaction with NSs in Arabic. The group comparisons admittedly rely on a relatively small number of participants drawn from a wide range of proficiency levels, however these results nevertheless point to the fact that the C-test was an effective instrument for separating learners into two groups that correspond in a statistically significant way with other
independent measures of Arabic proficiency. These groupings will therefore be used below in the presentation of group statistics according to the various examined CAF measures. As will be discussed in the subsequent section, these C-test-determined groups likewise correspond to significant differences in CAF performance.

5.3 C-test and Unrelated Independent Variables

This study also collected a number of independent variables from all learners and NSs that do not appear to be statistically related to group membership. These variables were collected from all participants, and include the composite values of Informal Literacy practices (IL), Formal Literacy practices (FL), and self-estimated L1 writing ability.

5.3.1 Literacy Practices

In order to investigate the posited relationship among learners and NSs between written CAF literacy practices, this study asked all participants to estimate the frequency with which they engaged in a number of reading- and writing-oriented activities. In order to facilitate analysis, the reported frequencies of (a) composing formal emails, (b) composing “academic” writing, (c) reading the news, and (d) reading novels were combined as FL, while the reported frequencies of (a) reading social media, (b) writing in social media, (c) writing personal emails, e.g., “to a friend,” and (d) writing SMS/text messages were combined into IL. As demonstrated below, neither FL nor IL scores correspond in significant way with group membership.

The LP group collectively reported engaging in both FL and IL literacy practices more frequently than HP learners. The LP IL mean was likewise higher than that of the NS group, although the NS group displayed the highest mean FL score. However, the differences between
the LP and HP means for these measures are not statistically significant. For each of these three measures, both the LP and HP means fall within the other group’s CI, and vice versa.

*Table 27: Literacy Practices*

<table>
<thead>
<tr>
<th>Variable</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
<th>NS (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Informal Literacy (IL)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2.69</td>
<td>2.23</td>
<td>2.09</td>
</tr>
<tr>
<td>95% CI-lower</td>
<td>2.18</td>
<td>1.61</td>
<td>1.68</td>
</tr>
<tr>
<td>95% CI-upper</td>
<td>3.20</td>
<td>2.85</td>
<td>2.49</td>
</tr>
<tr>
<td>Median</td>
<td>2.75</td>
<td>2</td>
<td>2.13</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>1.13</td>
<td>1.08</td>
<td>0.70</td>
</tr>
<tr>
<td>Range</td>
<td>3.75</td>
<td>3.75</td>
<td>2.25</td>
</tr>
<tr>
<td>Minimum</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>4.75</td>
<td>4.75</td>
<td>3.25</td>
</tr>
<tr>
<td><strong>Formal Literacy (FL)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mean</td>
<td>2.24</td>
<td>2.13</td>
<td>2.95</td>
</tr>
<tr>
<td>95% CI-lower</td>
<td>1.99</td>
<td>1.79</td>
<td>2.35</td>
</tr>
<tr>
<td>95% CI-upper</td>
<td>2.49</td>
<td>2.45</td>
<td>3.54</td>
</tr>
<tr>
<td>Median</td>
<td>2</td>
<td>2</td>
<td>3.13</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>0.56</td>
<td>0.57</td>
<td>1.03</td>
</tr>
<tr>
<td>Range</td>
<td>1.75</td>
<td>2</td>
<td>3.50</td>
</tr>
<tr>
<td>Minimum</td>
<td>1.5</td>
<td>1.25</td>
<td>1</td>
</tr>
<tr>
<td>Maximum</td>
<td>3.25</td>
<td>3.25</td>
<td>4.5</td>
</tr>
</tbody>
</table>

In addition to the fact that the HP and LP groups present similar scores according to the IL and FL composite variables, these two groups likewise engage in IL and FL at similar rates to the NS group. The only significant differences between groups (on the basis of CI comparison) is between both learner groups and the NS group. The NS FL CI of [2.35, 3.54] is higher than
either the HP or LP means according to this variable, although as noted above the HP/LP distinction here is non-significant. Therefore, it can be said that there appear to be differences in composite literacy practice scores between the learner participants in this study and NSs, but there is no difference between LP and HP learners when groups are determined on the basis of C-test score.

When the individual polled literacy practices that were used to form the FL and IL composite scores were compared across the HP and NS groups, only one individual practice proved to be statistically significant—writing personal emails. For the FL and IL variables, a one-way ANOVA does not show a statistical difference between groups \( (p > .05) \). Therefore, self-reported frequency of engagement in the selected literacy practices does not appear to be statistically related to membership in the LP, HP, or NS groups, and these variables will be excluded from subsequent analyses.

5.3.2 L1 Writing Ability

All participants in this study were asked to rate themselves against their peers with respect to their L1 writing ability. Arabic learners were asked to rate their ability to write English, and Arabic NSs were asked to rate their ability to write Arabic. Both groups of learners ultimately view themselves as much more skilled learners than the participants in the NS group, as the LP \( (M = 4.70) \) and HP \( (M = 4.64) \) means are significantly higher than the NS \( (M = 3.57) \). In this case, a one-way ANOVA showed that group means according to this variable are statistically significant \( (F_{2,46} = 19.20, \ p < .01) \). Comparisons using Tukey’s contrasts found a statistical difference between the NS and LP groups \( (\text{mean difference} = 1.14, \ CI = .67, 1.62, \ p < .001, \ d = 1.90) \) and between the NS and HP groups \( (\text{mean difference} = 1.07, \ CI = .55, 1.60, \ p < .001, \ d = 1.90) \).
.001, $d = 1.66$). There was no statistical difference between the LP and HP groups. No Arabic learner rated themselves lower than 4 on the 5-point scale when asked to estimate their L1 writing ability when compared to their peers, thus the range for both the LP and HP groups is 1 in both cases. The minimum NS score on this measure was 2, however, resulting in a range of 3. These results reflect the fact that learners in the HP and LP groups consistently rated themselves as excellent English writers, while NS participants collectively had a significantly lower estimation of their own Arabic writing ability.

**Table 28: Self-Estimated L1 Writing Ability**

<table>
<thead>
<tr>
<th>Variable</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
<th>NS (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Writing Ability (self-estimate) Mean</td>
<td>4.71</td>
<td>4.64</td>
<td>3.57</td>
</tr>
<tr>
<td>95% CI-lower</td>
<td>4.50</td>
<td>4.36</td>
<td>3.13</td>
</tr>
<tr>
<td>95% CI-upper</td>
<td>4.92</td>
<td>4.93</td>
<td>4.01</td>
</tr>
<tr>
<td>Median</td>
<td>5</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>0.46</td>
<td>0.49</td>
<td>0.77</td>
</tr>
<tr>
<td>Range</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>Minimum</td>
<td>4</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>Maximum</td>
<td>5</td>
<td>5</td>
<td>5</td>
</tr>
</tbody>
</table>

The relatively large range of self-estimated L1 writing ability scores in the NS group indicates that this variable may be useful in subsequent analyses, for example as a proxy for the learner-supplied self-estimated overall and writing ability discussed above.
5.4 C-test Correlation with Independent Variables

In addition to the various comparisons of group means presented above, sets of Pearson’s correlation coefficients were likewise calculated comparing C-test score with other collected numerical variables by treating all Arabic learners as a single group. To this end, correlation was used to explore the relationship between Arabic learner’s C-test scores and the independent variables introduced above:

Table 29: Correlation Coefficient between C-test score and Independent Measures

<table>
<thead>
<tr>
<th>Independent Variable</th>
<th>Self-Estimates</th>
<th>Literacy Practices</th>
</tr>
</thead>
<tbody>
<tr>
<td>Years of Study</td>
<td>.70&lt;sup&gt;b&lt;/sup&gt;</td>
<td>Reading .61&lt;sup&gt;b&lt;/sup&gt; IL -.11</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Writing .58&lt;sup&gt;b&lt;/sup&gt; FL .05</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Listening .64&lt;sup&gt;b&lt;/sup&gt; NS/Year .47&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speaking .61&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Overall .57&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td></td>
<td>L1 Writing -.02</td>
</tr>
</tbody>
</table>

*Note: Pearson correlation. <sup>a</sup><i>p</i> < 0.05, <sup>b</sup><i>p</i> < 0.01

These correlation coefficients reflect the group analyses presented above, namely that certain independent variables correspond in a significant way with the groupings determined on the basis of the C-test (e.g., self-estimates, YoS, and NS contact), while others appear to have no relationship to C-test score (e.g., literacy practices and L1 writing ability). The correlation between C-test score and various self-estimates of Arabic proficiency is moderately positive; for example, self-estimated overall Arabic proficiency has a moderately strong correlation with C-test score (<i>r</i> = .57, <i>n</i> = 35, <i>p</i> > .01). The correlation coefficient between C-test score and writing self-estimate (<i>r</i> = .58, <i>p</i> > .01) is nearly identical. The other collected self-estimates likewise
show a positive correlation with C-test score, including $r = .61$ for reading, $r = .61$ for speaking, and $r = .64$ for listening; all correlations statistically significant ($p > .01$).

The identified correlation between C-test score and self-reported contact with Arabic NSs is weak, but statistically significant, reflecting the significant difference between LP and HP means described above for this measure. There was no identifiable correlation between C-test score and estimated L1 writing ability, or reported IL and FL scores. When examined individually, no single literacy practice (e.g., reading Arabic news or writing impersonal emails) was revealed to correlate positively or negatively with C-test score among all learners, either. Therefore, it can be said that correlation analysis has largely confirmed the effectiveness of the variables described above in distinguishing between groups that were determined on the basis of C-test score. Those variables for which LP and HP group means were significantly different correlate positively with C-test score, with the strongest relationship existing between C-test score and YoS ($r = .70$) and the weakest between C-test score and NS interaction frequency ($r = .47$). Conversely, those variables determined via group analyses to be ineffective at distinguishing between groups lack a correlational relationship with C-test score.

5.5 Arabic Writing at Three Levels

Before presenting the descriptive statistics associated with the selected CAF measures used in this study, it will be instructive to examine examples of Arabic writing produced by members of the NS, HP, and LP groups. This will facilitate the conceptualization of what the numerical values presented below “look like” within the context of meaning-carrying samples of Arabic writing. As described in the Methodology section, NSs were presented written prompts in Arabic, and learners were presented the same prompts in English. As can be observed in the
progression of the CAF measure values through the NS to the HP to the LP samples, the measured variables tend to decrease as proficiency level decreases. However, each group displays within-group variation with respect to the selected variables, and this variation is often substantial.

The passage below presents the first 101 words\(^{26}\) taken from a representative example of a NS response to the persuasion prompt, in which participants were asked to support or oppose a campus-wide ban on smoking in the form of a letter to the editor of their student newspaper:

\(^{26}\) The English translation of this passage is 148 words in length.
The university is currently undertaking the organization of a campaign to eradicate smoking. I see this as a great effort that deserves to be lauded; whereas society sees smoking as a societal scourge that has to be eliminated by diverse means. But I see at the same time that the fruits of that campaign will not be fully derived without the participation of the state and society as a whole, whereas imposing laws in the university, and working to apply them in an effective manner, will not be useful if it does not conform to the laws of the state and society as a whole outside of the university’s campus. I think that among the most influential means with regard to smoking, in this respect, is the media; youth are especially affected by media content and are the ones who form the biggest percentage of those who appear...

<table>
<thead>
<tr>
<th>Construct</th>
<th>Lexical richness</th>
<th>Syntactic complexity</th>
<th>Syntactic complexity</th>
<th>Syntactic complexity</th>
<th>Accuracy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure</td>
<td>Guiraud’s W/S</td>
<td>W/T</td>
<td>T/S</td>
<td>EfT/T</td>
<td>W/M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>9.90</td>
<td>34.80</td>
<td>15.82</td>
<td>2.20</td>
<td>0.91</td>
<td>8.70</td>
</tr>
</tbody>
</table>

Within this passage, relatively complex lexical items have been used in relatively complex sentences/T-units, which in turn were produced with a high degree of accuracy and fluency. This short sample is nevertheless sufficient to illustrate the fact that words/sentence may be an unreliable measure of syntactic complexity in written Arabic, given that sentence length tends to vary so widely both among and between participants. In this sample, the W/S value (34.8) is
more than double the W/T value (15.82). Although the W/S value is quite close to the average W/S value for all NS writing in the persuasive category (M = 40.46), it is far below the maximum value (98.00) and well above the minimum (16.40). This sample’s W/T value of 15.82, however, is quite close to the average of 13.96 W/T for NS writing in the persuasive category, while the observed maximum (21.88) and minimum (9.00) of this measure in response to the persuasive prompt are separated by a much narrower range. This comparison points to the relative unreliability of sentence-based measures, even within the NS subgroup.

The NS sample above can be compared to the 101-word sample presented in Table 35, in which an HP Arabic learner responds to the same prompt:

**Table 31: HP Persuasion Response**

To the reader of this message. Peace be upon you. I hope that this message finds you in good health and fortune. A week ago I heard of the decision of the university to end smoking in the area of the university. I was happy for this news and I thank you for your good decision. Forbidding smoking usually the person begins smoking because something makes him sad or annoys him. Meanwhile he is able to ask for help for his problem or treat himself he begins smoking and doesn’t improve anything in his life. Rather he endangers his health and the health of the people who are listening to smoking. Instead of spending his time smoking this person has to become what annoys him in his life and ask his family and friends for help in ending smoking. Therefore, the university’s decision to forbid smoking will force the students to search for what helps them and improves their health.
Table 31. (cont.)

<table>
<thead>
<tr>
<th>Construct Measure</th>
<th>Lexical Complexity</th>
<th>Syntactic Complexity</th>
<th>Syntactic Complexity</th>
<th>Syntactic Complexity</th>
<th>Accuracy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guiraud’s W/S</td>
<td>7.56</td>
<td>11.20</td>
<td>8.42</td>
<td>1.33</td>
<td>0.58</td>
<td>9.18</td>
</tr>
</tbody>
</table>

This HP participant wrote less accurately and complexly than the NS participant according to all measures in response to the persuasive prompt. Guiraud’s index reveals a measurably lower rate of lexical richness, the HP participant’s T-units in this passage are shorter by several words, and the HP sentences are likewise much shorter. Furthermore, the HP participant has included fewer T-units per sentence (1.33 T/S compared to the NSs 2.20 T/S) However, this participant created written language in this prompt at a slightly faster rate than the NS participant above. This is reflective of the observed overlap in measured fluency rates between the HP and NS group, discussed below.

Finally, the persuasive response of an LP participant represents the lower end of what can be expected from Arabic CAF in response to the persuasive prompt within this group of participants. This LP response is 77 words in length:
Table 32: LP Persuasion Response

مرحبًا مريد الجريدة في جامعة "اسم".

اسمي ليلى وأنا طالبة في الجامعة. أنا من الشرق الأوسط ولكن سفرت إلى أمريكا لأنني أريد أن أصبح دكتورة ولذلك سأساعد الناس في العلم العربي. في جامعة "اسم"، أشترك في مجموعة الطلاب واتقرر أن نساعد المدرسة. يمكننا أن نساعد المدرسة وكل الطلاب إذا كل الناس لا يدخنون. في صفي الطب، درسنا ما يحدث في هذه المدينة ولكن لا نحتجهم في منطقة المدرسة.

إذا ستمعم إلى رأي في هذا الرسالة، شكرا.

ليلى

Hello, director of the newspaper in the university of “name”.

My name is Layla and I am a student in the university. I am from the Middle East but I traveled to America because I want to become a doctor and therefor I will help people in the Arab World. At “name” university, I participate in the student group and we want to help the school. We are able to help the school and all of the students if everyone doesn’t smoke. In my medicine class, we studied what happens in this city but we don’t need them in the area of the school.

If you listened to my opinion in this letter, thank you.

Layla

<table>
<thead>
<tr>
<th>Construct</th>
<th>Lexical Complexity</th>
<th>Syntactic Complexity</th>
<th>Syntactic Complexity</th>
<th>Syntactic Complexity</th>
<th>Accuracy</th>
<th>Fluency</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure</td>
<td>Guiraud’s W/S</td>
<td>W/T</td>
<td>T/S</td>
<td>Eft/T</td>
<td>W/M</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5.81</td>
<td>11</td>
<td>5.92</td>
<td>1.86</td>
<td>0.62</td>
<td>5.92</td>
</tr>
</tbody>
</table>

This sample is shorter overall than the HP and NS samples, it was composed at a slower rate, and it exhibits lower rates of syntactic and lexical complexity. The lexical richness value

27 “Layla” is a participant-provided pseudonym and is not this participant’s actual name.
supplied by Guiraud’s index for this sample is 5.81, which can be compared to the values of 7.56 for the HP sample and 9.90 for the NS sample.

Holistically, these three samples are quite different—they display major differences in organization, lexis, structure, and accuracy. The authors take divergent stances regarding the proposed ban of smoking on campus, as well as embed those stances within different forms of experiential evidence. The LP writer, for example, adopts the position of a fictional Middle Eastern student, while the HP and NS writers appeal to personal experience. Additionally, the LP and HP samples contain elements of what could be considered a “good” letter, namely a salutation, a signature, an introduction, etc. These samples also illustrate that there is often some overlap between the groupings used in this study—the HP writer produced language in response to this prompt at a quicker rate than the NS writer, for example, while the LP writer was more accurate in this prompt than the HP writer. After learner responses to all three prompts are considered together, however, the differences between groups become much more clear, as will be demonstrated in the following presentation of CAF measures.

5.5 Usefulness of Lemmatization

Before moving on to a presentation of the analyzed CAF measures, it is necessary to determine which versions of the measures for lexical complexity to present. As discussed in the previous chapter, the written Arabic collected for this study was subjected to a process of semi-lemmatization in order to facilitate lexical analyses. This process entailed collapsing various lexical types into the same lemma. Written Arabic includes several articles, particles, and prepositions that are written as affixes, i.e., phrases such as “the book,” “in the book,” and “a book” “and the book,” etc. are all single, intact strings of characters and are recognized by
lexical analysis software as separate lexical types. The lemmatization process described in the previous chapter aimed to collapse as many of these types as possible into single lemmas, in order to determine whether or not such a process is helpful or necessary in shedding statistical light on the nature of lexical complexity in written Arabic. The results presented here indicate that, although lemmatization had a statistically significant effect on Guiraud’s index scores by way of uniformly lowering group means, both the lemmatized and un-lemmatized versions of this measure have nearly identical relationships to alternative measures, and indeed can be said to duplicate each other.

The descriptive statistics of the lemmatized (L) and un-lemmatized (UL) Guiraud’s index values are displayed below for the LP, HP, and NS groups:

**Table 33: Un-Lemmatized and Lemmatized Versions of Guiraud’s Index**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statistic</th>
<th>LP – UL</th>
<th>LP – L</th>
<th>HP – UL</th>
<th>HP – L</th>
<th>NS-UL</th>
<th>NS – L</th>
</tr>
</thead>
<tbody>
<tr>
<td>Guiraud’s</td>
<td>Mean</td>
<td>9.29</td>
<td>8.24</td>
<td>11.45</td>
<td>10.15</td>
<td>13.71</td>
<td>12.01</td>
</tr>
<tr>
<td>Index</td>
<td>95% CI-upper</td>
<td>8.72</td>
<td>7.75</td>
<td>10.62</td>
<td>9.39</td>
<td>12.47</td>
<td>11.09</td>
</tr>
<tr>
<td></td>
<td>95% CI-lower</td>
<td>9.88</td>
<td>8.72</td>
<td>12.29</td>
<td>10.90</td>
<td>14.96</td>
<td>12.93</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>9.28</td>
<td>8.27</td>
<td>11.22</td>
<td>10.15</td>
<td>14.01</td>
<td>12.08</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>1.27</td>
<td>1.06</td>
<td>1.44</td>
<td>1.31</td>
<td>2.15</td>
<td>1.59</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>6.23</td>
<td>6.02</td>
<td>9.24</td>
<td>7.82</td>
<td>10.42</td>
<td>9.84</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>6.22</td>
<td>4.56</td>
<td>4.73</td>
<td>4.80</td>
<td>7.89</td>
<td>5.28</td>
</tr>
<tr>
<td></td>
<td>Interquartile</td>
<td>.93</td>
<td>1.01</td>
<td>2.33</td>
<td>1.86</td>
<td>3.12</td>
<td>2.81</td>
</tr>
</tbody>
</table>

The lemmatization process had the expected result of lowering mean scores and reducing standard deviation scores in each of the three identified groups. This is expected, given that the
lemmatization process described in the previous chapter entailed the collapsing of several types of words into a single type (i.e., verbs conjugated for alternative pronouns, removal of affixed prepositions, etc.). In other words, scores according to Guiraud’s index of lexical richness are slightly lower on average if a text has been lemmatized, and the three groups examined here were more closely clumped after lemmatization. Likewise, lemmatization tended to decrease the range of recorded values for each group, i.e., decreasing the distance between a group’s minimum and maximum values, with the exception of the HP group that saw an extremely slight increase in range from 4.73 to 4.80. An examination of the Interquartile Range (IQR) values before and after lemmatization likewise reveals that the lemmatization process did not have a major effect on the reported values. The IQR reflects the distance between the first and third quartiles in a dataset, i.e., the range in which the “middle” half of scores occur, giving an indication of how dispersed a group is. The IQR of lexical richness slightly increased in the LP group after lemmatization, but slightly decreased in the HP and NS groups; however, each both versions of each group’s IQR values remain quite similar, indicating that lemmatization did not radically change the internal composition of any group.

Lemmatization thus lowered the mean scores of all three groups with respect to Guiraud’s index. However, all three groupings remain statistically significant on the basis of a comparison of group CIs, whether the lemmatized means or the un-lemmatized means are compared. That is, no L or UL group mean falls within one of its counterparts’ CIs. Lemmatization tends to decrease the standard deviation associated with this measure, resulting in groups that are slightly less variant. This is expected, given that many types were combined during this process.
5.5.1 Significance of Difference Resulting from Lemmatization

Although lemmatization lowered group means and SD values, the lemmatized and un-lemmatized versions of this measure are near-exact correlates of each other. When treating all learner and NS participants as a single group, the lemmatized Guiraud’s index shows a near-exact correlation with its un-lemmatized counterpart \((r = .98, n = 49, p < .001)\). When learners alone are isolated, the lemmatized/un-lemmatized correlation coefficient of Guiraud’s index is \(.97 (n = 35, p < .001)\). These values are similar for the NS group \((r = .98, n = 14, p < .01)\).

Finally, both versions of learner scores according to Guiraud’s index were correlated with C-test score, resulting in \(r = .69\) for the lemmatized version, and \(r = .70\) for the un-lemmatized version \((n = 35, p < .01)\). A paired-samples t-test was conducted in order to determine whether or not the differences in learner means for the lemmatized and un-lemmatized lexical richness measure were statistically significant. In this t-test, both versions of the scores for all learners were compared. This test indicates that there is a significant difference between both means; \(t (34) = 16.17, p < .001, d = 0.73\). Overall, un-lemmatized Guiraud’s index scores were 1.16 points higher than lemmatized scores \((95\% \text{ CI } [1.01, 1.30])\).

The lemmatization process thus had a statistically significant effect in lowering group means on the lexical richness measure examined here. However, given the fact that the lemmatized and un-lemmatized values of the lexical complexity measures used here are correlated so closely with both each other and with the additional measures described above, it does not appear that lemmatization was successful in strengthening statistical analyses on the basis of ratios employing lexical types. Ultimately, it appears that the effect of lemmatization was very slight. It does not appear that this process resulted in a significant increase in statistical
robustness. Therefore, the lexical measures reported throughout in this chapter will represent the un-lemmatized, rather than the lemmatized values for these measures.

5.6 Usefulness of Hapax Index

After determining that lemmatization is not a prerequisite for statistical analysis of lexical complexity of written Arabic (at least according to the measures used in the current study), an additional question remains regarding the usefulness of the hapax measure described in the previous chapter. This measure used lexical hapaxes (i.e., types that occurred only once in a given context) instead of lexical types in the calculation of Guiraud’s index, arrived at by dividing the total number of types by the square root of the number of tokens. Given the high frequency of hapaxes in written Arabic data due to a number of factors discussed in the previous chapter, it was hypothesized that differences between group means according to a hapax index would prove to be statistically significant. However, the hapax index was shown to in effect represent a reproduction of Guiraud’s index.

**Table 34: Index of Lexical Hapaxes**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statistic</th>
<th>LP</th>
<th>HP</th>
<th>NS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hapax Index</td>
<td>Mean</td>
<td>6.84</td>
<td>9.07</td>
<td>11.05</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>1.21</td>
<td>1.46</td>
<td>1.83</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>5.39</td>
<td>4.99</td>
<td>6.78</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>.97</td>
<td>2.64</td>
<td>2.37</td>
</tr>
</tbody>
</table>

The LP, HP, and NS groupings are significantly different from each other on the basis of a comparison of CIs, however when the hapax index is correlated with Guiraud’s index, it becomes clear that the new index is merely a numerically lower version of the first.
learner and NS participants, there is a correlation of $r = .98$ ($n = 49$, $p < .01$) between Guiraud’s index and the hapax index, and the relationship is the same when only the production of Arabic learners is considered ($r = .97$, $n = 35$, $p < .01$). A paired-samples t-test was conducted to compare learner means on the Guiraud’s and hapax indexes, revealing that average scores on these two measures are indeed statistically different; $t (34) = 32.93$, $p < .001$, $d = 1.43$. However, given the extremely high correlation between these measures, it appears that these two measures duplicate each other even if their group means are significantly different. It is therefore not evident that the hapax measure was helpful within the context of the current study in elucidating any information about the nature of lexical complexity in the written Arabic samples analyzed in the current study. This index will be omitted from further analysis, and the discussion of lexical complexity in this study will thus rely on the un-lemmatized version of Guiraud’s index.

5.7 CAF measures

Thus far, this chapter has presented the results of learner responses to the C-test, divided learners into two groups on the basis of this score, and justified this division on the basis of the statistical relationship between C-test score and several independent estimates of Arabic ability, and presented representative examples of Arabic writing at the three identified levels (LP, HP, and NS). Conversely, the results presented above argue against the need for lemmatization in the analysis of written Arabic lexical complexity, as well as established that the selected lexical hapax measure does not add any clarity over the more-common Guiraud’s index when analyzing the written Arabic collected in this study. The current section will present the descriptive statistics for the remaining dependent variables representing written CAF. The initial results of this study reveal that NSs of Arabic are indeed more complex, accurate, and fluent writers than
their high-performance and low-performance Arabic learner counterparts. However, differences within groups indicate that all three groupings of participants (LP, HP, and NS) include writers with diverse CAF-related abilities. Furthermore, certain measures appear to produce more distinct groupings than others, indicating that they may be more effective at distinguishing between Arabic writers. Sentence-based measures in particular are shown to be highly unreliable in distinguishing between groups.

5.7.1 Fluency

This study relied on the fluency ratio words per minute as a measure of rate of written production. The statistics presented in Table 39 below represent the average of each participant’s total words produced across the persuasion, narration, and description prompts, divided by the total time on task for each of the three prompts.

*Table 35: Fluency (W/M)*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statistic</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
<th>NS (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency</td>
<td>Mean</td>
<td>5.36</td>
<td>7.32</td>
<td>12.45</td>
</tr>
<tr>
<td>(w/m)</td>
<td>95% CI-upper</td>
<td>4.72</td>
<td>5.99</td>
<td>10.41</td>
</tr>
<tr>
<td></td>
<td>95% CI-lower</td>
<td>5.99</td>
<td>8.64</td>
<td>14.48</td>
</tr>
<tr>
<td>Median</td>
<td></td>
<td>4.79</td>
<td>6.44</td>
<td>12.51</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td></td>
<td>1.39</td>
<td>2.29</td>
<td>3.53</td>
</tr>
<tr>
<td>Minimum</td>
<td></td>
<td>3.59</td>
<td>4.15</td>
<td>8.00</td>
</tr>
<tr>
<td>Maximum</td>
<td></td>
<td>7.61</td>
<td>11.42</td>
<td>21.53</td>
</tr>
<tr>
<td>Range</td>
<td></td>
<td>4.03</td>
<td>7.27</td>
<td>13.53</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td></td>
<td>2.71</td>
<td>4.08</td>
<td>4.94</td>
</tr>
</tbody>
</table>
Each of the three identified groups performed differently according to the selected fluency measure of words per minute (W/M) with LP (M = 5.36) writing more slowly than HP (M = 7.32), who in turn wrote more slowly than NS (M = 12.45) participants. This progression from less to more fluent across these three groups makes intuitive sense, as the NS group wrote more than twice as fast as the LP group, on average. Additionally, although the NSs in this study produced written language at the fastest rate, this group also presents a wider distribution of scores than the other two groups, with a standard deviation of 3.53, compared to 2.29 and 1.39 for the HP and LP groups, respectively. This indicates that, as fluency increases across groups, the distributions of the groups themselves tend to widen as well. This is likewise reflected in the fact that IQR increases from LP (2.71) to HP (4.08) to NS (4.94), i.e., the range within which the “middle” half of scores is contained increases across groups, indicating that the LP group is grouped most closely around the mean, while the NS group is the most dispersed. This is reflected in the box plot in Figure 6:
Visually, it is evident that, as group membership changes, average fluency rate changes as well, with a general increase in words written per minute from LP to HP to NS. However, it is likewise evident that the NS group is more dispersed than the other groups, exhibiting more within-group variability than the tightly-grouped LP group, for example. The plot in Figure 6 likewise indicates that the NS group includes a single outlier who wrote much more quickly (when averaged across all three prompts) than the other individuals within that group, which in turn increases the NS standard deviation.

Importantly, no group’s W/M mean is located within the 95% CI range of the other groups, providing an indication of statistical significance to these groupings with respect to writing fluency. Additionally, the width of the CI increases for each group as well, with the LP CI [4.72, 5.99] being narrower than HP [5.99, 8.64], which is in turn narrower than the NS CI
[10.41, 14.48]. This further indicates that the range of possible fluency rates increases as participants increase in fluency, and indicates more diverse fluency rates at the upper end of the proficiency scale, and less diverse fluency rates at the lower end of the scale. The CI comparison partially corresponds with the results of a one-way ANOVA conducted to determine the effect of group membership on differences in W/M means, which found group to be a significant factor ($F_{2,45} = 41.26, p < .001$). This ANOVA excludes the single NS outlier depicted in the plot in Figure 6 above. Post-hoc analysis via Tukey’s contrasts shows that the difference between HP and LP fluency means is statistically significant (mean difference = 1.96, CI = .29, 3.63, $p = .018$). Additionally, the NS-LP comparison (mean difference = 6.39, CI = 4.68, 8.10, $p < .001$) and the NS-HP comparison (mean difference = 4.43, CI = 2.56, 6.30, $p < .001$) are statistically significant. There are large effect sizes for the HP-LP comparison ($d = 1.09$), the NS-LP comparison ($d = 3.41$) and the NS-HP comparison ($d = 1.87$). Therefore, these groupings remain significantly different on the basis of comparing group CIs and as evidenced by post-hoc analysis. Thus, within this population of Arabic learners and NSs, the three groupings presented above thus can be said to correspond to significantly different rates of fluency in written production.

5.7.2 Syntactic Complexity

As with the fluency measure, there is a clear increase in measured syntactic complexity according to the selected Mean Length of T-unit (W/T) measure that corresponds with a movement from the LP (M = 6.07) to HP (M = 7.87) to NS (M = 10.09) groups. Thus, it can be said on the basis of these group means that NSs in this study produced T-units that were longer
on average than those produced by their HP and LP counterparts, averaged across all three written tasks.

*Table 36: Syntactic Complexity*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statistic</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
<th>NS (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic</td>
<td>Mean</td>
<td>6.07</td>
<td>7.87</td>
<td>10.09</td>
</tr>
<tr>
<td>Complexity (W/T)</td>
<td>95% CI-upper</td>
<td>5.67</td>
<td>7.19</td>
<td>8.85</td>
</tr>
<tr>
<td></td>
<td>95% CI-lower</td>
<td>6.48</td>
<td>8.55</td>
<td>11.34</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>6.00</td>
<td>8.00</td>
<td>10.48</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>.89</td>
<td>1.17</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>4.36</td>
<td>6.13</td>
<td>7.38</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>7.97</td>
<td>9.75</td>
<td>14.48</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>3.61</td>
<td>3.62</td>
<td>7.10</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td></td>
<td>1.41</td>
<td>2.12</td>
<td>3.50</td>
</tr>
</tbody>
</table>

Again, as with the fluency measure, the group SD values increase along with group mean, indicating that, as Arabic proficiency increases (operationalized as group membership), the average dispersal around group mean according to the W/T measure likewise increases. In other words, the three groups become less tightly grouped as they produce longer T-units. This is reflected in the IQR values for each of these three groups. Although the total range between maximum and minimum W/T scores is nearly identical in the LP and HP groups, the IQR values of the three groups show a clear progression, as the range of the “middle” halves of each group’s respective scores increases in tandem with overall W/T. This indicates a wider dispersal of scores as fluency increases within this population of participants. The NS IQR (3.50) is more than twice as wide as the LP IQR (1.41). This is evident in the box plot displayed in Figure 7 which presents the dispersal of W/T scores for these groups:
Group variance is greatest within the NS group, reflecting the fact that NS lexical complexity scores are more diverse than those of both groups of learners. It is likewise notable that the maximum value of W/T in the NS group (14.48) is much higher than both the NS mean (10.09), as well as the maximum values in the HP (9.75) and LP (7.97) groups.

However, there is some overlap between these three groups. This is evident in a comparison of the minimum and maximum scores for each group and an examination of the above box plot displayed in Figure 7, revealing that the IQRs (the blue-shaded portions of the plots) of the HP and NS groups in particular share a substantial portion of the y-axis. Ultimately, though, the three groupings remain statistically significant on the basis of a comparison of means.
and CIs. No group’s mean falls within an adjacent group’s CI. This is confirmed via a one-way ANOVA which found a significant effect for group membership in mean W/T scores ($F_{2,46} = 33.40, p < .001$). All group comparisons were found via post-hoc analysis (Tukey’s) to be significant, including HP-LP (mean difference = 1.80, CI = .61, 3.00, $p = .002$, $d = 1.78$); NS-HP (mean difference = 2.22, CI = .92, 3.53, $p < .001$, $d = 1.29$); and NS-LP (mean difference = 4.03, CI = 2.83, 5.22, $p < .001$, $d = 2.65$). Thus far, it can be said that these three groupings correspond to differential performance on the W/T measure of syntactic complexity when participant performance is combined across all three tasks.

5.7.3 Lexical Complexity

Lexical complexity was operationalized in this study as Guiraud’s index of lexical richness. A separate version of this index was calculated using lexical hapaxes, rather than lexical types, however as detailed above this additional measure was not determined to add any additional clarity to the nature of lexical complexity in the samples of written Arabic at hand, thus it has been excluded from the current analysis. The scores presented below represent the numerical value resulting from the division of the total number of types in a participant’s writing samples by the square root of the total number of tokens. Therefore, it is not a representation of a concrete ratio, such as words per minute, or percentage of accurate T-units, rather it is a numerical value for which higher values represent more lexical richness, and lower values represent less lexical richness. As in the other CAF measures discussed here, lexical richness produced by this population of participants increases with group membership. The writing of the LP group is the least lexically rich/complex (M = 9.29), followed by the writing of the HP group (M = 11.45), and then the NS group (M = 13.71).
Table 37: Lexical Complexity

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statistic</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
<th>NS (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Richness</td>
<td>Mean</td>
<td>9.29</td>
<td>11.46</td>
<td>13.71</td>
</tr>
<tr>
<td></td>
<td>95% CI-upper</td>
<td>8.72</td>
<td>10.62</td>
<td>12.47</td>
</tr>
<tr>
<td></td>
<td>(Guiraud’s Index)</td>
<td>9.88</td>
<td>12.29</td>
<td>14.96</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>9.28</td>
<td>11.22</td>
<td>14.01</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>1.27</td>
<td>1.44</td>
<td>2.15</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>6.23</td>
<td>9.24</td>
<td>10.42</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>12.44</td>
<td>13.97</td>
<td>18.31</td>
</tr>
<tr>
<td>Range</td>
<td></td>
<td>6.22</td>
<td>4.73</td>
<td>7.89</td>
</tr>
<tr>
<td>Interquartile Range</td>
<td></td>
<td>.93</td>
<td>2.33</td>
<td>3.12</td>
</tr>
</tbody>
</table>

In contrast to the other CAF measures examined here, however, the LP group appears to exhibit a great deal more within-group variation than the HP group. For example, the LP range is wider than the HP range, although still narrower than the NS range. As evident in the associated box plot displayed in Figure 8, the LP group likewise presents a number of outliers, in contrast to the outlier-less HP and NS groups.
The IQR scores for each group (which necessarily exclude outliers as they represent only the range of the “middle” half of scores) increase in tandem with group membership, proceeding from the very narrow LP (0.93) to the much wider HP (2.33) and NS (3.12) interquartile ranges. The differences between all three groups on the basis of a comparison of CIs are clear, however—no group’s mean falls within another’s CI. This is confirmed via a one-way ANOVA, which finds a significant effect for group membership. It should be noted that this ANOVA excludes the 5 outlying LP participants indicated in the plot in Figure 8, thus \( n = 16 \) for the LP group (\( F_{2,41} = 31.80, p < .001 \)). All between-group comparisons are statistically significant (Tukey’s post-hoc), including HP-LP (mean difference = 2.08, CI = .76, 3.41, \( p = .001, d = 2.00 \)); NS-HP (mean difference = 2.26, CI = .89, 3.63, \( p = .001, d = 1.25 \)); and NS-LP (mean difference = 4.34, CI = 3.02, 5.67, \( p < .001, d = 2.89 \)). These comparisons are also associated

Figure 8: Lexical Complexity
with large effect sizes. Although the ANOVA reveals that these groupings are significantly different from each other even after the LP outliers are removed, the original presence of 5 LP outliers indicates that comparisons with this group according to the selected measure of lexical richness may not be reliable.

One further aspect of these results should be noted, namely that within the NS group, the chosen measure of lexical complexity shows a moderate positive correlation with written fluency (W/M) \( (r = .62, p = .019) \). However, the NS group contains one outlier with respect to the fluency measure, and after this individual is removed, the correlation becomes non-significant. The LP group likewise shows a positive correlation between lexical richness and both syntactic complexity and fluency, however this group contains 5 outliers. After these individuals are removed, neither of these correlations are significant.

5.7.4 Accuracy

As with the fluency and syntactic complexity measures examined above, Arabic writers that participated in this study produced written language at differing rates of accuracy that correspond with group membership. As described in the previous chapter, this study operationalizes accuracy as the ratio of error-free T-units to total T-units, resulting in a numerical value in which a score of 1 is equal to a 100% T-unit accuracy rate. According to the taxonomy of errors described in the previous chapter, the NS group was the most accurate when written performance was averaged across tasks, producing accurate T-units at an average rate of .93 accurate T-units per T-unit. This was followed by the HP group (.72) and the LP group (.56).
Table 38: Accuracy (EfT/T)

<table>
<thead>
<tr>
<th>Measure (EfT/T)</th>
<th>Statistic</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
<th>NS (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>.56</td>
<td>.72</td>
<td>.93</td>
</tr>
<tr>
<td></td>
<td>95% CI-upper</td>
<td>.49</td>
<td>.63</td>
<td>.88</td>
</tr>
<tr>
<td></td>
<td>95% CI-lower</td>
<td>.62</td>
<td>.82</td>
<td>.98</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>.58</td>
<td>.74</td>
<td>.96</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>.13</td>
<td>.16</td>
<td>.09</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>.25</td>
<td>.27</td>
<td>.73</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>.81</td>
<td>.92</td>
<td>1.00</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>.56</td>
<td>.65</td>
<td>.27</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>.18</td>
<td>.16</td>
<td>.11</td>
</tr>
</tbody>
</table>

In contrast to the measures examined above, these three groups tend to become less dispersed according to the EfT/T measure (i.e., more tightly grouped) as they progress upwards on a posited scale of Arabic proficiency. In other words, more proficient writers in this study produced fewer errors, and exhibited lower rates of within-group variation according to this measure. While in the case of fluency and syntactic complexity, NSs were the most dispersed group and LP learners the least dispersed, and in the case of accuracy this situation is reversed. This is evident in a comparison of IQR scores, which decrease as group membership changes, moving from LP (IQR = .18) to HP (.16) to NS (.11). This indicates a gradual tightening of the “middle” halves of each group, as the first and third quartiles move closer together, evident in the box plot presented in Figure 9.
All three groups present a single outlier participant who as significantly less accurate than the group mean, although the NS outlier is significantly closer to the group mean than the LP or HP outliers (both of whom produced T-units at around 30% accuracy). There is some degree of overlap between all three groups (i.e., the LP maximum exceeds the NS minimum), however there is no overlap of the blue-shaded interquartile ranges of each group depicted on the plot in Figure 9. The box plot, combined with a comparison of group IQR scores, makes it evident that the NS group is the most tightly clustered, and the NS group is likewise clustered at the top of the graph in Figure 9, with the group mean approaching 100%. This is both expected and encouraging, given that NSs should be expected to be extremely accurate writers.

As in the fluency and syntactic complexity scores, no group’s mean falls within an adjacent group’s CIs, indicating that the groupings are significant. This significance is confirmed
via a one-way ANOVA. This ANOVA excludes the three outliers (one per group) indicated in
the plot in Figure 9, and finds a significant effect for group membership ($F_{2,43} = 50.38, p < .001$).
All between-group comparisons were found to be statistically significant (Tukey’s post-hoc),
including HP-LP (mean difference = .18, CI = .09, .27, $p < .001$, $d = 1.60$); NS-HP (mean
difference = .18, CI = .09, .28, $p < .001$, $d = 2.10$); and NS-LP (mean difference = .37, CI = .28,
.45, $p < .001$, $d = 3.48$). The very strong effect sizes associated with these comparisons likely
reflect the small numbers of participants in each group, which were each restricted by one
participant after outliers were excluded. Nevertheless, it can be stated on the basis of these
results that the groupings determined via C-test score or NS status have a statistically significant
relationship to written accuracy, operationalized as the ratio of error-free T-units per total T-units
averaged across all three experimental tasks.

5.8 Sentence-Based Measures

This study included two measures of syntactic complexity that employ punctuated
sentences as a unit of analysis. As discussed in the previous chapter, sentences have been argued
to be an unreliable unit of analysis for use in empirical CAF measures (Norris & Ortega, 2009),
leading many researchers to adopt the alternative T-unit when working with written data.
Sentences may be even more problematic with respect to written Arabic, given the great
variability evident in sentence length and punctuation usage in authentic Arabic texts. However,
given that sentences in written texts can be considered an “organically” occurring unit, in that the
authors themselves provide the unit via punctuation, that sentence length is relatively easy to
calculate, and that sentences have not yet been explored as a unit of analysis in empirical CAF
studies of written Arabic produced by learners, the decision was made to include sentence measures in the current analysis.

5.8.1 Mean Length of Sentence (W/S)

As in the alternative CAF measures described above, the Mean Length of Sentence (W/S) measure shows a clear increase across the three groups, as LP (M = 10.50) sentences are shorter than HP (M = 18.62) sentences, which in turn are shorter than (M = 31.97) sentences. Indeed, the NS mean sentence length is more than three times as long as the LP length. However, this neat progression of group means belies an observable unreliability inherent in this measure, namely that these gradually increasing group means are associated with divergent degrees of variance. Group variance ranges from an LP SD of 2.29 to the NS and HP SDs of 20.28 and 21.58, respectively. This indicates that, although the LP group may be more tightly bunched around its mean for this measure, the LP and HP groups are much more dispersed—by an order of magnitude—around their respective means. This is not the case for the four CAF measures investigated above.

Table 39: Words per Sentence (W/S)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statistic</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
<th>NS (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean Length of Sentence (W/S)</td>
<td>Mean</td>
<td>10.50</td>
<td>18.62</td>
<td>31.97</td>
</tr>
<tr>
<td></td>
<td>95% CI-upper</td>
<td>9.46</td>
<td>6.16</td>
<td>20.26</td>
</tr>
<tr>
<td></td>
<td>95% CI-lower</td>
<td>11.55</td>
<td>31.09</td>
<td>43.68</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>10.45</td>
<td>13.22</td>
<td>29.09</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>2.29</td>
<td>21.58</td>
<td>20.28</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>7.70</td>
<td>9.83</td>
<td>13.71</td>
</tr>
</tbody>
</table>
Table 39. (cont.)

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statistic</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
<th>NS (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>W/S</td>
<td>Maximum</td>
<td>17.64</td>
<td>93.33</td>
<td>86.33</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>9.94</td>
<td>83.51</td>
<td>72.62</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>3.29</td>
<td>3.66</td>
<td>25.62</td>
</tr>
</tbody>
</table>

Although the IQR values of the LP and HP groups are quite similar at 3.29 and 3.66, respectively, the NS IQR is more than 7 times wider than these (25.62). This reflects the presence of one or more outliers in the HP group whose presence increased HP. The NS group, on the other hand, evidences much more within-group variation than the learner groups with respect to this measure, resulting in a much larger IQR of 25.62.

![Figure 10: Words per Sentence (W/S)](image)

Furthermore, a comparison of group significance via CIs indicates that the two learner groups are not significantly different, given that the LP mean (10.50 W/S) falls within the HP CI [6.16, 18.89].
A one-way ANOVA confirms this, finding that group membership is a statistically significant factor ($F_{2.46} = 7.74, p = .001$), although post-hoc analysis (Tukey’s) finds that both the HP-LP and NS-HP comparisons (i.e., comparisons between “adjacent” groups) are non-significant. The only significant comparison is NS-LP (mean difference $= 21.47$, CI $= 8.25, 34.68$, $p = .001$, $d = 1.67$). Finally, it was determined that, within the LP group, there is a moderate positive correlation between mean sentence length (W/S) and mean length of T-unit (W/T) ($r = .54$, $p = .012$). However, this correlation was only present for the LP group and was not present for the HP and NS groups. Based on the results presented in Table 43 and Figure 10 above, it appears that the W/S measure may have been successful at depicting a “normal” range of variation within the NS group—the NS distribution resembles those associated with alternative CAF measure scores presented above—but the resulting LP and HP distributions do not appear to be useful for the basis of group comparisons. Given that LP and HP group scores according to this measure are not significantly different, the identified LP W/S-W/T correlation can likely be ignored.

### 5.8.2 Coordination (T/S)

Finally, this study calculated a measure of syntactic complexity in the form of T-units per sentence, in the hope that this ratio would be useful in the discrimination between simple and complex sentences, much in the way that words per T-unit (W/T) is argued to help differentiate between simple and compound T-units. Based on the group means presented below, it can be said that the three identified groups deployed differing numbers of T-units in the average sentence written in response to the three experimental prompts, with LP writers composing the least complex sentences, ($M = 1.74$), followed by HP writers ($M = 2.24$), followed by NS writers.
This mirrors the progression across groups seen in the other examined CAF measures, reflecting a generalized increase in the average number of T-units used per sentence among these groups of participants.

**Table 40: Coordination (T/S)**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Statistic</th>
<th>LP (n=21)</th>
<th>HP (n=14)</th>
<th>NS (n=14)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Coordination (T/S)</td>
<td>Mean</td>
<td>1.74</td>
<td>2.24</td>
<td>3.23</td>
</tr>
<tr>
<td></td>
<td>95% CI-upper</td>
<td>1.58</td>
<td>1.00</td>
<td>2.08</td>
</tr>
<tr>
<td></td>
<td>95% CI-lower</td>
<td>1.90</td>
<td>3.48</td>
<td>4.39</td>
</tr>
<tr>
<td></td>
<td>Median</td>
<td>1.64</td>
<td>1.63</td>
<td>2.18</td>
</tr>
<tr>
<td></td>
<td>Std. Deviation</td>
<td>.36</td>
<td>2.14</td>
<td>1.99</td>
</tr>
<tr>
<td></td>
<td>Minimum</td>
<td>1.40</td>
<td>1.52</td>
<td>1.43</td>
</tr>
<tr>
<td></td>
<td>Maximum</td>
<td>2.81</td>
<td>9.67</td>
<td>8.00</td>
</tr>
<tr>
<td></td>
<td>Range</td>
<td>1.41</td>
<td>8.14</td>
<td>6.57</td>
</tr>
<tr>
<td></td>
<td>Interquartile Range</td>
<td>.41</td>
<td>.32</td>
<td>2.82</td>
</tr>
</tbody>
</table>

However, an examination of group variance in the form of SD values and the ranges associated with each group reveals that this measure is less than reliable. This is evident from the fact that the HP SD value (2.14) is nearly six times greater than the LP SD (.36), and is likewise larger than that of the NS group (SD = 1.99). Furthermore, the IQR scores for each group indicate that, while the LP and HP groups are clustered very tightly around their means (.41 and .32, respectively), the NS group is spread across a large swath of the y-axis (IQR = 2.82), indicating a much higher degree of group variability.
Figure 11: Coordination (T/S)

Each of these three group means falls within an adjacent group’s CI, indicating a lack of significance for the differences between them. This is largely confirmed via a one-way ANOVA, which does find a significant effect for group membership in group T/S scores ($F_{2,46} = 3.78, p = .03$), although as above in the W/S variable, the only significant comparison is NS-LP (mean difference = .99, CI = .17, 2.81, $p = .023$, $d = 1.16$).

The coordination (T/s) measure yielded a strong positive correlation with the first sentence measure discussed above, mean length of sentence (W/S) in all groups, including LP ($r = .67, p = .001$), HP ($r = .99, p < .001$), and NS ($r = .95, p < .001$). The HP and NS coordination coefficients in particular are so strong as to be nearly perfect, which may be difficult to explain at first glance. However, as noted above, sentence-based measures are expected to be unreliable within the context of written Arabic, and in fact the HP group contains one outlier with respect to
both the T/S (coordination) and W/S (coordination) measures. This individual (participant 18 in the W/S and T/S plots in Figures 10 and 11, above) produced the most T-units per sentence of any participant in the LP, HP, and NS groups, and likewise produced the most average words per sentence (93 W/S). This individual, who had completed three years of Arabic study at the time of participation, composed three responses to the three experimental prompts that collectively consisted of 280 words and 3 sentences, i.e., each prompt was expressed within a single punctuated sentence. In fact, this participant’s sentences were on average much longer than even the closest NS participant (86.33 W/S). When this participant is removed from the calculation of correlation, the correlation coefficient between these two variables within the HP group falls to \( r = .45 \) and is no longer statistically significant \( (p = .123, n = 13) \).

Based on an examination of the group distributions associated with both the W/S and T/S measures of syntactic complexity, it does not appear that either of these measures is statistically related to the groupings of learners identified on the basis of C-test scores. In both cases, post-hoc analysis (Tukey’s) failed to find a significant difference between the LP and HP groupings of learners, and indeed between HP learners and NSs. Only the NS-LP comparisons were found to be significant. Although the NS group distributions according to these sentence-based measures appear more “normal” based on a comparison with group distributions rendered by the alternative CAF measures discussed above, neither the mean sentence length (W/S) nor the coordination (T/S) measures appears to have a statistical relationship to overall Arabic proficiency (operationalized as C-test score). As a final check, in order to determine whether or not the two sentence measures might be related to other independent variables in a way that is not revealed by groups determined on the basis of C-test score, Pearson correlation coefficients were calculated comparing both the learner W/S and T/S measures to all independent variables.
In this calculation, all Arabic learners were treated as a single group. The only statistically significant correlation to emerge from this comparison was the weak correlation between coordination (T/S) and learner self-estimated writing ability ($r = .33$, $p = .050$). However, given the weakness of this relationship, as well as the fact that the associated $p$ value falls exactly on the cusp of the traditionally accepted level of statistical significance, this relationship can likely be discarded.

This analysis has shown that each of the four non-sentence-based CAF measures (W/M, W/T, Guiraud’s index, and EfT/T) corresponds in a statistically significant way to the HP and LP groupings of learners determined on the basis of C-test score. This statement is made on the basis of group CI comparisons as well as one-way ANOVAs examining the effect of group membership, followed by post-hoc analysis of between-group comparisons. The only group comparison shown to be non-significant via post-hoc analysis the HP-LP comparison for fluency (W/M), although these groupings remain significantly different on the basis of CI comparison. Each of the four CAF measures is also successful at distinguishing between the HP group and the NS group. Conversely, however, the two sentence-based measures discussed here were not shown to be related in a significant way to group membership among the learners who participated in this study and will therefore be excluded from the regression analyses presented below. Sentence-based measures may be useful when analyzing NS writing. These implications will be explored in the subsequent chapter.

5.9 Regression Analysis

This section presents the results of four standard regression analyses targeting the relationship between four CAF measures and four independent variables representing estimates
of Arabic proficiency. In general, multiple regression shows the extent to which changes in the selected independent variables can predict changes in the dependent variable (in this case, one of the four selected CAF measures), as well as the amount of the relationship contributed uniquely by each independent variable. A number of studies have explored combining CAF measures into a composite variable for the purpose of conducting regression, e.g., the composite fluency index used by Mora & Valls-Ferrer (2012), which was calculated by combining three fluency measures. However, the various constructs of syntactic complexity, lexical complexity, accuracy, and fluency investigated in the current study are diverse, divergent, and anticipated to “trade-off” with each other in complex ways. For these reasons, these four constructs were targeted by four separate calculations of regression.

In order to select the specific CAF measures for use in regression models, the results of the group comparisons above were compared with each variable’s correlation with C-test score, the independent measure used to determine group membership. As displayed in Table 45 below, C-test score shows a significant, positive correlation with each of the CAF measures discussed above, with the exception of the two sentence-based measures that proved ill-suited for the analysis of learner writing:
Each of the four CAF measures that corresponded to statistically significant differences in group means above are likewise associated with a significant positive correlation with C-test score, and these values reflect the fact that one outlying participant was removed from the calculation of the C-test-W/M coefficient. Both sentence-based measures, on the other hand, do not have a correlational relationship with C-test score. On this basis, the four CAF measures selected for regression are words per minute (fluency), mean length of T-unit (syntactic complexity), Guiraud’s index (lexical richness), and error-free T-units per T-unit (accuracy), representing four distinct CAF areas. As noted above the lexical hapax index was dropped from the analysis due to its apparent duplication of Guiraud’s index. Likewise, both sentence-based measures discussed above were excluded due to their unreliability and lack of correlation with other measures. The selected CAF measures, in contrast, tend to show a positive and significant relationship with both each other, as well as the selected independent variables.

The independent (explanatory) variables included in each multiple regression presented below are years of study (YoS), self-estimated overall Arabic proficiency, C-test score, and self-
reported frequency of contact with NSs. Ideally, the independent (i.e., explanatory) variables used in regression will have some logical basis for their inclusion, and in this case they were selected because they were shown to have a significant correlation with other variables, as well as significance when used to compare between the HP and LP groups. It was not known from previous research what role any of the explanatory variables will play in determining learner score on the selected CAF measures in the context of multiple regression.

Admittedly, the explanatory power of this analysis will be limited by the small sample size of 35 learners. However, this analysis nevertheless has the potential to reveal which independent variables contribute the most to the selected CAF scores within this population of learners. There are differing suggestions as to the minimum number of participants required to reliably conduct multiple regression analysis (Larson-Hall, 2015, pg. 184; Jeon, 2015, pg. 137). For example, Tabachnick and Fidell (2012) recommend a participant number of $50 + 8k$, in which $k$ refers to the number of independent or explanatory variables to be investigated. An alternative rule of thumb for the social sciences is 15 participants per explanatory variable (Stevens, 2009). Given the small sample size of Arabic learners here, regression analysis will be limited to explanatory (independent) variables shown above to be statistically related to the various collected CAF measures. As Larson-Hall (2015) points out, the higher the expected $R^2$ value (i.e., effect size) of a regression analysis, the fewer participants likely required (pg. 185). Therefore, it is expected that number of 35 Arabic learners who participated in this study will be sufficient to calculate multiple regression with four predictor variables, with expected $R^2$ values above .25 (Cohen et al., 2003).

Table 46 below presents the bivariate correlation coefficients between the various independent proficiency measures, as well as the correlations between these variables and the
syntactic complexity measure (W/T), the first target of multiple regression. Larson-Hall (2015) notes that correlations above $r = .70$ between explanatory (independent) variables may complicate regression analysis, and a few correlations involving Years of Study reach that level, including with C-test score ($r = .70$) and overall self-estimate ($r = .78$). However, it should be expected that, as an independent measure of Arabic ability, YoS is expected to be highly correlated with other measures of the same construct. All Variable Inflation Factor (VIF) values for the independent variables were determined via the regressions reported below to be under 5 (with the highest being YoS = 3.34), indicating that none of the included variables are sufficiently collinear to harm the model, thus none were excluded on this basis (Larson-Hall, 2015, p. 194).

Table 42: Independent/Explanatory Variable Intercorrelation (n=35)

<table>
<thead>
<tr>
<th></th>
<th>C-test</th>
<th>Years of Study</th>
<th>Self-Estimate</th>
<th>NS Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Syntactic Complexity (W/T)</td>
<td>.57</td>
<td>.39</td>
<td>.54</td>
<td>.18</td>
</tr>
<tr>
<td></td>
<td>.000</td>
<td>.021</td>
<td>.000</td>
<td>.062</td>
</tr>
<tr>
<td>C-test</td>
<td></td>
<td>.70</td>
<td>.57</td>
<td>.47</td>
</tr>
<tr>
<td></td>
<td></td>
<td>.000</td>
<td>.000</td>
<td>.005</td>
</tr>
<tr>
<td>Years of Study</td>
<td>.70</td>
<td></td>
<td>.78</td>
<td>.43</td>
</tr>
<tr>
<td></td>
<td>.000</td>
<td></td>
<td>.000</td>
<td>.010</td>
</tr>
<tr>
<td>Self-Estimate (overall)</td>
<td>.57</td>
<td>.78</td>
<td></td>
<td>.46</td>
</tr>
<tr>
<td></td>
<td>.000</td>
<td>.000</td>
<td></td>
<td>.005</td>
</tr>
<tr>
<td>NS Contact</td>
<td>.47</td>
<td>.43</td>
<td>.46</td>
<td></td>
</tr>
<tr>
<td></td>
<td>.005</td>
<td>.010</td>
<td>.005</td>
<td></td>
</tr>
</tbody>
</table>
As can be observed via Table 46, syntactic complexity (W/T) within this population of learners shows a positive correlation with all selected explanatory variables with the exception of NS interaction frequency. Only this correlation is non-significant ($r = .18, p = .289$). All intercorrelations between explanatory variables are positive and statistically significant. For the sake of brevity, subsequent regression reports will only include the first line of this table, in which the dependent variable is correlated with the four independent variables (n = 35 in all cases).

5.9.1 Syntactic Complexity

In the multiple regression model presented here, the four selected explanatory variables contribute to explain 48% of the variance in W/T score ($R^2 = .48$). Table 47 presents the results of the multiple regression for this combination of dependent and independent variables. The presented values include the Constant (coefficient for the y-intercept of the regression equation), the unstandardized coefficients for each independent variable (used to write the regression equation), the standardized coefficients (versions of the former that are easier to interpret and compare to one another, based on standardized versions of the input values), the t values and significance scores of each variable’s contribution to the model, and the squared semipartial correlation ($sr^2$) value for each variable. The $sr^2$ value indicates how much the $R^2$ value will decrease if the variable in question is removed from the equation; i.e., a higher $sr^2$ value indicates a larger effect.
### Table 43: Syntactic Complexity Regression Model

<table>
<thead>
<tr>
<th>Measure</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>p value</th>
<th>Squared semipartial correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>3.61</td>
<td></td>
<td>4.87</td>
<td>.000</td>
<td></td>
</tr>
<tr>
<td>Self-estimate (overall)</td>
<td>.80</td>
<td>.63</td>
<td>2.92</td>
<td>.007</td>
<td>.38</td>
</tr>
<tr>
<td>Years of Study</td>
<td>-.25</td>
<td>-.44</td>
<td>-1.81</td>
<td>.080</td>
<td>-.24</td>
</tr>
<tr>
<td>NS interaction</td>
<td>-.21</td>
<td>-.20</td>
<td>-1.31</td>
<td>.201</td>
<td>-.17</td>
</tr>
<tr>
<td>C-test score</td>
<td>.04</td>
<td>.61</td>
<td>3.19</td>
<td>.003</td>
<td>.42</td>
</tr>
</tbody>
</table>

By way of example, this multiple regression model predicts that a learner’s W/T score could be predicted by the following equation (see Larson-Hall, 2015, p. 193):

\[
y = 3.61 + (.8)\text{self-estimate} + (-.25)\text{years of study} + (-.21)\text{NS contact} + (.04)\text{C-test score}
\]

In other words, within this population of 35 learners, W/T scores have a positive relationship with C-test score and self-estimated overall proficiency scores, and a negative relationship with YoS and reported frequency of NS contact. In general, it should be assumed that as overall proficiency increases, the average length of a participant’s T-units will likewise increase. The negative value associated with YoS in this model contradicts this expectation, although it should be noted that YoS has a weaker correlation with W/T than either C-test score or self-estimate.

Furthermore, the results of the t-tests associated with each variable’s contribution to the model indicates a lack of significance for the YoS and NS contact variables (\(p = .080\) and \(p = .201\), respectively). The standardized beta values of .61 and .63 indicate that a majority of the variance in syntactic complexity scores can be explained by C-test score and overall self-estimate score,
respectively. This is likewise reflected in the squared semipartial correlation value for these variables (C-test $r^2 = .42$; self-estimate $r^2 = .38$). Both of these relationships are statistically significant. This analysis was likewise not determined to include any outliers on the basis of standardized residuals scores—all points are between $–3.0$ and $3.0$, reflecting the fact that no points are more than three standard deviations away from the mean. On the basis of this regression model, it can be said that C-test score and learner self-estimate of overall proficiency are strong predictors of written syntactic complexity (W/T), while YoS and NS interaction frequency are not.

5.9.2 Lexical Complexity

Lexical complexity (operationalized in this study as Guiraud’s index of lexical richness) is likewise positively correlated with each of the four selected explanatory variables. Unlike the case of syntactic complexity (W/T) above, the weak positive correlation ($r = .35$) between lexical complexity and reported NS contact frequency is statistically significant ($p = .039$), making lexical complexity the only construct for which this correlation is significant.

Table 44: Lexical Complexity Correlations

<table>
<thead>
<tr>
<th></th>
<th>C-test</th>
<th>Years of Study</th>
<th>Self-Estimate</th>
<th>NS Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical Complexity (Guiraud’s index)</td>
<td>.000</td>
<td>.001</td>
<td>.000</td>
<td>.039</td>
</tr>
<tr>
<td></td>
<td>.70</td>
<td>.55</td>
<td>.63</td>
<td>.35</td>
</tr>
</tbody>
</table>

The multiple regression model explaining variance in Guiraud’s index scores in the context of the selected explanatory variables reveals an $R^2$ value of $.58$, indicating that these variables combine to explain $58\%$ of the variance in this lexical complexity measure.
Table 45: Lexical Complexity Regression Model

<table>
<thead>
<tr>
<th>Measure</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>p value</th>
<th>Squared semipartial correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>5.64</td>
<td>6.65</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-estimate (overall)</td>
<td>.75</td>
<td>.47</td>
<td>2.39</td>
<td>.023</td>
<td>.28</td>
</tr>
<tr>
<td>Years of Study</td>
<td>-.15</td>
<td>-.19</td>
<td>-.92</td>
<td>.365</td>
<td>-.11</td>
</tr>
<tr>
<td>NS interaction</td>
<td>-.07</td>
<td>-.06</td>
<td>-.41</td>
<td>.688</td>
<td>-.05</td>
</tr>
<tr>
<td>C-test score</td>
<td>.05</td>
<td>.59</td>
<td>3.49</td>
<td>.002</td>
<td>.41</td>
</tr>
</tbody>
</table>

As in the syntactic complexity model above, the contributions of C-test score and self-estimated overall Arabic ability to lexical richness score among this population of learners are positive and significant. C-test score is the most predictive explanatory variable of lexical richness ($\beta = .59$, $sr^2 = .41$), followed by overall self-estimate ($\beta = .47$, $sr^2 = .28$). The contributions of YoS and reported NS contact frequency are both negative, although they are both likewise non-significant ($p > .05$ in both cases), in spite of the significance of the correlations of these variables with lexical complexity. This analysis was not determined to contain any outlying data points more than three standard deviations away from the mean. C-test score plays a similar role in the determination of lexical richness to its role in the determination in syntactic complexity score above, presenting very similar $\beta$ and $sr^2$ values in both cases. However, in spite of the strong positive correlation between C-test score and self-reported YoS ($r = .70$), the YoS value was not determined via this regression model to be a significant factor in determining learner linguistic complexity.
5.9.3 Accuracy

Unlike the lexical and syntactic complexity measures investigated above, the accuracy measure used in this study proved to be much more variable when compared to the selected independent measures of Arabic proficiency. In this case, the ratio of error-free T-units per T-unit (EfT/T) shows a significant positive correlation with only one explanatory variable, C-test score ($r = .50, p = .001$). The variables YoS, NS contact frequency, and overall self-estimate, on the other hand, show weak positive or negative correlations, and none reach the level of statistical significance ($p > .05$ in all cases). The statistical relationship between this accuracy measure and C-test score may be reflected in the fact that the HP and LP groups (identified solely on the basis of C-test score) presented significantly different group means according to the accuracy measure, as discussed in conjunction with the descriptive statistics for this measure above. This indicates that if the HP and LP groups been determined on the basis of reported YoS, this accuracy measure may not have proven to be a useful distinguisher between the two.

Table 46: Accuracy Correlations

<table>
<thead>
<tr>
<th>C-test</th>
<th>Years of Study</th>
<th>Self-Estimate</th>
<th>NS Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy (EfT/T)</td>
<td>.50</td>
<td>.20</td>
<td>-.02</td>
</tr>
<tr>
<td>.002</td>
<td>.248</td>
<td>.916</td>
<td>.144</td>
</tr>
</tbody>
</table>

The standard regression model explaining variance in the accuracy variable EfT/T presents an $R^2$ value of .41, indicating that the selected explanatory variables account for 41% of the observed variance in accuracy. This is lower than the two complexity-oriented regression models presented above.
Table 47: Accuracy Regression Model

<table>
<thead>
<tr>
<th>Measure</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>p value</th>
<th>Squared semipartial correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>.40</td>
<td>4.09</td>
<td>.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Self-estimate (overall)</td>
<td>-.09</td>
<td>-.55</td>
<td>-2.37</td>
<td>.025</td>
<td>-.33</td>
</tr>
<tr>
<td>Years of Study</td>
<td>.01</td>
<td>.09</td>
<td>.34</td>
<td>.735</td>
<td>.05</td>
</tr>
<tr>
<td>NS interaction</td>
<td>.02</td>
<td>.15</td>
<td>.91</td>
<td>.372</td>
<td>.13</td>
</tr>
<tr>
<td>C-test score</td>
<td>.01</td>
<td>.68</td>
<td>3.37</td>
<td>.002</td>
<td>.47</td>
</tr>
</tbody>
</table>

In this model, YoS and self-reported NS contact frequency have positive relationships with the selected accuracy measure, although these contributions remain non-significant ($p > .05$). In contrast with the complexity-based models above, self-estimated Arabic ability now has a negative relationship with the targeted construct ($\beta = -.55$, $sr^2 = -.33$), indicating that, within this population of learners, higher rates of observed accuracy across all three tasks are actually associated with lower self-reported estimates of overall Arabic ability. C-test score, however, remains the most influential explanatory factor in predicting learner accuracy score within this sample ($\beta = .68$, $sr^2 = .47$). This analysis was not determined to include any outliers on the basis of standardized residuals values.

5.9.4 Fluency

Unlike accuracy, the fluency measure used in this study shows a positive correlation with each selected independent/explanatory variable. The strongest of these is self-estimated Arabic
ability \((r = .52, p < .002)\), followed by C-test score \((r = .49)\), YoS \((r = .39)\), and NS contact \((r = .32)\). However, the significance of the NS interaction correlation is non-significant \((p = .067)\).

One outlying participant was removed from the calculation of these coefficients, therefore \(n = 34\) in all cases. The consistent lack of significance (with the exception of lexical complexity) stands in contrast to the fact that the LP and HP groupings on the basis of C-test score were shown to be significantly different according to this measure above.

*Table 48: Fluency Correlations*

<table>
<thead>
<tr>
<th>Measure</th>
<th>C-test</th>
<th>Years of Study</th>
<th>Self-Estimate</th>
<th>NS Contact</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fluency (w/m)</td>
<td>.49</td>
<td>.39</td>
<td>.52</td>
<td>.32</td>
</tr>
<tr>
<td></td>
<td>.006</td>
<td>.022</td>
<td>.002</td>
<td>.067</td>
</tr>
</tbody>
</table>

The regression model associated with the interaction between fluency and the selected explanatory variables returns an \(R^2\) value of .33, rendering it the weakest regression model of the four investigated CAF constructs, although it contains one fewer participant than the other models as it excludes a single outlying Arabic learner.

*Table 49: Fluency Regression Model*

<table>
<thead>
<tr>
<th>Measure</th>
<th>Unstandardized Coefficients</th>
<th>Standardized Coefficients</th>
<th>t</th>
<th>p value</th>
<th>Squared semipartial correlation sr^2</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Constant)</td>
<td>2.23</td>
<td></td>
<td>1.93</td>
<td>.064</td>
<td></td>
</tr>
<tr>
<td>Self-estimate</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>(overall)</td>
<td>.03</td>
<td>.317</td>
<td>1.49</td>
<td>.148</td>
<td>.23</td>
</tr>
<tr>
<td>Years of Study</td>
<td>-.18</td>
<td>-.227</td>
<td>-.84</td>
<td>.405</td>
<td>-.13</td>
</tr>
</tbody>
</table>
In the model of standard multiple regression that explains fluency performance in relation to scores on the four selected independent variables, these variables account for only 33% of fluency variance. As in the regression models discussed previously, self-estimated overall Arabic ability and C-test score are the most important factors in determining the dependent variable (W/M), with the C-test being the largest determiner ($\beta = .51$, $r^2 = .32$). The C-test is also the only selected independent variable that presents a significant contribution to this regression model ($p = .046$).

Therefore, it can be said that, within this population of Arabic learners, written fluency across three tasks is the least susceptible construct to explanation in the context of the selected independent variables. This determination is made on the basis of the relative weakness of the fluency regression model, as well as the fact that none of the selected independent variables was determined to contribute to the fluency model in a significant way. Accuracy is somewhat more explainable via linear regression, with 41% of variation in the EfT/T measure being attributed to the selected independent variables. Conversely, the two complexity measures (W/T and Guiraud’s index) proved to be the most susceptible to explanation with regard to the four independent variables, with 48% of variation in W/T (syntactic complexity) and 58% of variation in Guiraud’s index (lexical complexity) scores being attributed to them via the standard regression models. Not all independent variables contribute equally, however—while C-test and self-estimated overall ability prove to be consistently significant in the W/T, Guiraud’s index, and EfT/T models, YoS and NS interaction frequency were shown to be non-significant.
contributors in every model. C-test score was the only universally significant predictor of written CAF according to the selected

5.10 NS Writing and L1 Writing Ability

A separate correlation analysis investigates the relationship between NS self-estimated L1 (Arabic) writing ability and NS CAF. As noted above, the mean NS self-estimate of Arabic writing ability was 3.57 on a 5-point scale (SD = .77). The NS group displayed a wide range (4) on this measure, in contrast to the respective ranges of .46 and .49 for the LP and HP groups. The relative diversity in scores on this measure in turn indicates that it can be used as a proxy for the self-estimates of writing and overall ability that formed a basis for the analysis of learner writing.

To this end, Pearson correlation coefficients between NS L1 writing ability estimates and the four CAF measures explored above were calculated.

**Table 50: Correlations between NS Self-Estimate and CAF**

<table>
<thead>
<tr>
<th></th>
<th>Fluency (W/M)</th>
<th>Accuracy (EfT/T)</th>
<th>Syntactic Complexity (W/T)</th>
<th>Lexical Complexity (Guiraud’s)</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1 Writing Ability</td>
<td>.52</td>
<td>.22</td>
<td>.31</td>
<td>.80</td>
</tr>
<tr>
<td><em>p</em> value</td>
<td>.061</td>
<td>.465</td>
<td>.285</td>
<td>.001</td>
</tr>
<tr>
<td><em>n</em></td>
<td>13</td>
<td>13</td>
<td>14</td>
<td>14</td>
</tr>
</tbody>
</table>

NS L1 writing self-estimate values show a positive, strong, and significant correlation with lexical complexity, in the form of Guiraud’s index (_r_ = .80). The positive correlation with fluency approaches, but falls short of statistical significance. The calculation of these coefficients excluded one outlying participant each in the fluency and accuracy coefficients.
5.11 Genre Effect in Arabic Writing

The previous sections treated learner and NS production averaged across all three written tasks. However, an ancillary question posed by this study concerns the effect that written genre has on the selected empirical measures of written CAF. This study likewise aims to explore the extent to which anticipated differences in measured CAF due to genre are found in the writing of either learners or NSs. This issue was explored in detail the literature review chapter, therefore it is sufficient to merely remind the reader here that that factors such as the communicative purpose of writing have been determined to affect measured CAF (Laufer & Nation, 1995). The question examined here is whether or not distinct genres of written production elicit production characterized by differences in the selected CAF measures, i.e., whether genre changes can provide evidence of a “trade-off” effect (Skehan, 1998) in the writing of learners or NSs. In many of the constructs investigated below, the persuasive task appears to elicit substantially different production than either the narrative or descriptive written tasks. This effect is evident in the production all groups, i.e., persuasion tends to elicit the least accurate and fluent, but most syntactically complex writing.

5.11.1 Fluency

In all three groups of participants (LP, HP, NS), the persuasive task elicited the least fluent production. That is, participants in this group produced written language at the slowest rate of words per minute when responding to this task. The differences between tasks across all groups are slight, however. For example, the LP persuasive mean fluency rate (4.52) is only 1.24 words per minute slower than maximum LP fluency rate, achieved in the narrative task (5.76).
The NS persuasive fluency mean (12.68) is only .45 words per minute slower than the fastest NS mean, achieved in the descriptive task (13.13). The HP and LP groups are likewise tightly grouped across tasks, showing highly similar SD values in each task genre. The NS group, on the other hand, shows a wide dispersal around the mean in description (SD = 7.43) and a narrow dispersal in persuasion (SD = 2.32).

Table 51: Fluency (W/M) across Genre

<table>
<thead>
<tr>
<th></th>
<th>Descriptive Task</th>
<th>Narrative Task</th>
<th>Persuasive Task</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>NS (n = 14)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>13.13</td>
<td>12.79</td>
<td>12.68</td>
</tr>
<tr>
<td>(SD)</td>
<td>(7.43)</td>
<td>(4.96)</td>
<td>(2.32)</td>
</tr>
<tr>
<td><strong>HP (n = 14)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>7.43</td>
<td>8.00</td>
<td>6.61</td>
</tr>
<tr>
<td>(SD)</td>
<td>(2.76)</td>
<td>(2.60)</td>
<td>(2.19)</td>
</tr>
<tr>
<td><strong>LP (n = 21)</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>mean</td>
<td>5.68</td>
<td>5.76</td>
<td>4.52</td>
</tr>
<tr>
<td>(SD)</td>
<td>(1.54)</td>
<td>(1.56)</td>
<td>(1.65)</td>
</tr>
</tbody>
</table>

A repeated measures ANOVA was performed with the Greenhouse-Geisser adjustment to examine the interaction between task genre and group means, finding that this interaction was non-significant. The graph presented in Figure 12 displays the differential W/M scores of each group according to each written genre, demonstrating the fact that each group produced language more slowly in the persuasion condition, but that all three groups tended to produce language at much the same rate across all three genres. Furthermore, LP and HP means in all three genres are
quite close to each other, while NS means fall higher on the scale of W/M:

![Fluency across Genre](image)

*Figure 12: Fluency Across Genre*

Taken together, these results reveal that the persuasive task elicited the least fluent production in all groups, although the differences across tasks are slight. Alternatively, it can be said that the persuasion condition had the strongest negative effect on fluency rate in the LP group, and that this effect remained present, but attenuated, in the higher-proficiency HP and NS groups.
5.11.2 Syntactic Complexity

The persuasive prompt elicited the most syntactically complex writing of any prompt; i.e., persuasion elicited longer T-units in all three groups than either the descriptive or narrative tasks. This difference was particularly pronounced in the NS group, who collectively produced persuasive T-units (M = 13.96) that are nearly 5 full words longer than descriptive T-units (M = 9.18), the second-longest in each group. The HP group is characterized by a similar distribution of mean scores, with persuasion eliciting the longest T-units (9.53), followed by description (8.30), although in this case the two means are much closer together. Finally, the LP group also produced the longest T-units in the persuasive task (7.19), likewise followed by the descriptive task (5.87).

Table 52: Syntactic Complexity (W/T) across Genre

<table>
<thead>
<tr>
<th></th>
<th>Descriptive Task</th>
<th>Narrative Task</th>
<th>Persuasive Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS (n = 14)</td>
<td>mean</td>
<td>Narrative Task</td>
<td>Persuasive Task</td>
</tr>
<tr>
<td></td>
<td>(SD)</td>
<td>(SD)</td>
<td>(SD)</td>
</tr>
<tr>
<td>NS (n = 14)</td>
<td>9.18 (2.86)</td>
<td>8.47 (2.33)</td>
<td>13.96 (3.39)</td>
</tr>
<tr>
<td>HP (n = 14)</td>
<td>8.30 (2.40)</td>
<td>6.94 (1.41)</td>
<td>9.53 (1.85)</td>
</tr>
<tr>
<td>LP (n = 21)</td>
<td>5.87 (1.10)</td>
<td>5.78 (0.94)</td>
<td>7.19 (1.51)</td>
</tr>
</tbody>
</table>

A repeated measures ANOVA was performed to examine the interaction between task genre and group means, finding that this interaction was significant using the Greenhouse-Geisser correction ($F_{3.21,73.86} = 7.30, p < .001$, partial eta-squared = .24, power = .99), explaining 24% of variance in the W/T measure. The graph displayed in Figure 13 presents the differential W/T
scores of each group according to each written genre, demonstrating the fact that each group produced more complex language in the persuasive condition, although this increase was most marked among NSs.

**Figure 13: Syntactic Complexity across Genre**

It is noteworthy that even the LP group produced the longest T-units in response to the persuasion prompt. For example, the first-year learner who scored the lowest out of all learners on the C-test produced this complex T-unit in response to the persuasion prompt:
الناس الذين يدخنون سيموتون عندما أصغر من الناس الذين ليس يدخنون

an-nās al-ladhīna yudakhkhinūn sa-yamūtūn ʿanda-mā ʔaṣğar min in-nās al-ladhīna laysa yudakhkhinūn

the people who smoke will die when younger than the people who don’t smoke

Although this T-unit contains at least one error (negation, in bold), it can unambiguously count as complex, containing two definite relative clauses. This complex LP T-unit can be compared, for example, to an extremely long NS T-unit produced in the persuasive condition:

مع كامل الاحترام والتقدير لجميع الزملاء من المدخنين ولكن أنا أؤيد وبشدة منع التدخين في الحرم الجامعي لمكانة الجامعة في المجتمع وأنها مصدر للتعليم والرقي واحترام الاخرين بعدم ادانتهم والسلوك الحضاري بالابتعاد عن جميع ما يزعج زملاء الدراسة أو العمل ويسبب لهم الآدئ


With all respect and appreciation to all of my smoking colleagues, but I emphatically support the ban of smoking on campus, for the status of the university in society, as it is a source of education, advancement, and respect for others while not harming them, and of civilized behavior refraining from anything that disturbs colleagues in study or work and causes them harm.
This NS-produced T-unit is 43 words in length, falling well above the NS average T-unit length in the persuasive condition (13.96), and approaches the upper limit of T-unit complexity that was elicited from either learners or NSs in this study. It consists of a number of clauses of varying lengths, and exhibits a high degree of subordination.

5.11.3 Accuracy

The selected accuracy measure, in the form of a ratio of error-free T-units per T-unit, likewise shows influence of task genre. In all three examined groups, the persuasion task elicited the least accurate written production. This is true among the LP group, which collectively wrote at a rate of 43% accuracy in the persuasive prompt, among the HP group who wrote with 63% accuracy, and the NS group who wrote with 88% accuracy. In many cases, the difference between persuasion accuracy and accuracy rates in the other tasks is not great, with a 6-point difference in the NS group between persuasion and narration, and 4-point difference in the HP group between persuasion and description. The largest such gap is the 16-point difference between LP persuasion accuracy and description accuracy.

Table 53: Accuracy (Eft/T) across Genre

<table>
<thead>
<tr>
<th></th>
<th>Descriptive Task</th>
<th>Narrative Task</th>
<th>Persuasive Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS (n = 14)</td>
<td>mean 0.96</td>
<td>0.93</td>
<td>0.88</td>
</tr>
<tr>
<td></td>
<td>(SD) (0.06)</td>
<td>(0.11)</td>
<td>(0.15)</td>
</tr>
<tr>
<td>HP (n = 14)</td>
<td>0.67</td>
<td>0.82</td>
<td>0.63</td>
</tr>
<tr>
<td></td>
<td>(0.22)</td>
<td>(0.16)</td>
<td>(0.18)</td>
</tr>
</tbody>
</table>
Table 53. (cont.)

<table>
<thead>
<tr>
<th></th>
<th>Descriptive Task</th>
<th>Narrative Task</th>
<th>Persuasive Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>LP ((n = 21))</td>
<td>0.59</td>
<td>0.61</td>
<td>0.43</td>
</tr>
<tr>
<td></td>
<td>(0.17)</td>
<td>(0.19)</td>
<td>(0.28)</td>
</tr>
</tbody>
</table>

It should be noted that in all cases, both learner groups display greater SD values than the NS group, indicating a wider dispersal around the mean. Only the NS persuasion SD value of 0.16 is comparable to learner SD values in any task. This can be compared to the relatively small NS SD value in the descriptive task, for example (0.06).

![Accuracy across Genre](image)

*Figure 14: Accuracy across Genre*

A repeated measures ANOVA was conducted to investigate the interaction between genre and group mean EfT/T score, finding that this interaction was non-significant. However, it can be
observed via the graph in Figure 14 that the persuasive prompt elicited the least accurate language from each group of participants, and that narration elicited the most accurate language from Arabic learners.

The following samples each represent the introductory sentences (after salutations and greetings) of individuals from the LP and HP groups in response to the descriptive prompt, which asked participants to respond to a letter from a friend requesting information about their university:

(HP)

طلاب كثيرة في الجامعة دائما

ṭulāb kathīra fī-l-jāmiʿa dāʾim-an

many students in [are] always in the university

(LP)

في جامعة اسم بنايات كثيرة قريبة من منتصف الجامعة

fī jāmiʿat “ism” bināyāt kathīra qarība min muntaṣifat al-jāmiʿa

in “name” university there are many tall buildings close to the middle of the university

The problematic items in the LP and HP sentences above have are in bold—the HP error consists of an inappropriate particle used after the verb قال qāla “he said,”28 and the LP error morphological, resulting from the feminization of the masculine noun منتصف muntaṣaf “middle.”

---

28 As described previously, instances of colloquial Arabic were not considered to be “errors” in the context of this study, hence the very common construction قال qāla “he said,” was not flagged as problematic. However, this T-unit was coded as containing an error due to the problematic particle used in conjunction with this verb.
5.11.4 Lexical Complexity

As mentioned above, the values associated with Guiraud’s index of lexical richness are not associated with a discrete unit or ratio, such as words per minute or words per T-unit—rather, these values are a representation of the relationship between the numbers of types and tokens included in each participant’s submissions. In this case, the higher the value of the index, the more complexity can be said to characterize a given group. Table 58 presents group means according to this index for the three written tasks, revealing that HP, LP, and NS learners tend to be characterized by distinct rates of lexical complexity.

Table 54: Lexical Complexity (Guiraud’s Index) across Genre

<table>
<thead>
<tr>
<th></th>
<th>Descriptive Task</th>
<th>Narrative Task</th>
<th>Persuasive Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>NS (n = 14)</td>
<td>mean 9.07 (SD 1.82)</td>
<td>7.75 (SD 1.70)</td>
<td>8.95 (SD 1.19)</td>
</tr>
<tr>
<td>HP (n = 14)</td>
<td>mean 7.74 (SD 1.32)</td>
<td>7.46 (SD 0.69)</td>
<td>6.84 (SD 0.82)</td>
</tr>
<tr>
<td>LP (n = 21)</td>
<td>mean 6.35 (SD 1.04)</td>
<td>6.65 (SD 0.98)</td>
<td>5.33 (SD 1.00)</td>
</tr>
</tbody>
</table>

A repeated measures ANOVA was performed with a Greenhouse-Geisser correction, determining that the interaction between lexical complexity variation across genre and group membership is statistically significant (F_{36,91.17} = 12.76, p < .001, partial eta-squared = .36, power = 1.00), in other words explaining 36% of variance in the Guiraud’s index measure. Figure 15 below displays the differential Guiraud’s index scores of each group according to each
written genre, indicating, that while NSs showed a marked decrease in lexical complexity in the narration measure, LP writers saw a notable increase.

![Lexical Complexity across Genre](image)

**Figure 15: Lexical Complexity across Genre**

As will be discussed in the subsequent chapter, the fact that the NS group exhibited the least lexical richness in response to the narrative task may be due to the structure of the task itself—as a picture-supported narration exercise, the number of different types required to complete it is measurably less in this group than the number deployed in the persuasive or descriptive tasks. This is evident in the opening sentences of an NS, HP and LP participant in response to the
narration prompt. Each of these participants had a similar lexical richness score on this prompt (each falls between 6.99 and 7.06):

(kul yawm Ŧinda th-thâmina ṣabâh-an yastayqiz “ism” ʕalâ ʂawt il-munabbih)

Each day at 8 in the morning “name” wakes up to the sound of the alarm clock

(HP)

(istâqaẓa “ism” qabla ẓāhir ish-shams)

“Name” woke up before the appearance of the sun

(LP)

(fî-ṣ-ṣabâḥ, “ism” yaṣḥû fi-s-sâqa 1-khâmisa li-l-yawm)

In the morning, “name” wakes up at 5 for the day

Even in the extremely limited context of these introductory sentences drawn from a single participant in each group, it can be seen that all three are employing similar lexical items to respond to the same picture-supported story. Clearly, each sentence is characterized by differing rates of syntactic complexity and accuracy (e.g., note the spelling and lexical errors in the HP example), however the thematic organization is quite similar.

Writing genre can thus be said to interact with measurable CAF in complex ways within the context of the current study. NSs are observed to produce the least accurate, but the most complex language in the persuasive condition, and this tendency is reflected in writers in the LP and HP groups as well, although the effect is slighter. Lexical complexity scores likewise appear
to reflect task type, at least within the writing of NSs, as the picture-supported description task elicited the least lexically complex writing by this group. Finally, measured written fluency rates did not significantly change across tasks for the NS group—all three task fluency means were very similar for this condition. However, both the LP and HP groups showed a decrease in fluency when responding to the persuasive task, and this decrease reached the level of statistical significance for the LP group.
Chapter 6: Discussion and Implications

6.1 Introduction

In a general sense, the findings presented here combine to establish that there are measurable and statistically significant differences between the writing of Arabic NSs and Arabic learners at different levels of general ability. While this may not be particularly surprising in the context of the study of second language acquisition, this effort nevertheless makes a number of specific and significant contributions to the measurement of Arabic writing. These findings are admittedly limited in scope, and the chief limitation stems from the relatively small sample size. Group sizes range from a maximum of $n = 21$ in the LP group to $n = 14$ in the HP and NS groups. This small size in turn limits the strength of conclusions that can be made on the basis of analyses that involve correlations, especially when outliers may be present. An additional limitation concerns the nature of the employed CAF measures themselves—although each measure has been widely used in comparable studies of oral and written L2 production, the constructs they target are multifaceted and incredibly complex. Future explorations of written Arabic would undoubtedly benefit from a more fine-grained analysis of written L2 Arabic CAF that compares the results of multiple related measures within the same construct.

Even given these and other limitations discussed below, however, the results of the current study seem to corroborate the general consensus that the acquisition of individual structures measured by CAF analysis generally “follows a fixed developmental sequence” (Purpura 2014, p. 120). Certainly, the LP group scored measurably and significantly lower than the HP group in every construct, and the groupings were determined to be statistically significant for every construct except fluency. However, the groupings themselves encompassed a large amount of within-group variation, both according to the selected measures as well as to
estimated proficiency level (according to C-test, YoS, and self-estimates). This fact may serve as a jumping-off point for a more in-depth exploration of written CAF within a single proficiency level. After all, the L2 developmental process is highly individualized, and related studies have found that CAF measures are able to capture some elements of this idiosyncratic development (Larsen-Freeman, 2006).

6.2 Research Question 1: Arabic CAF Measured at Three Levels

The first research question motivating this dissertation is concerned with the nature of differences between Arabic writing when composed by authors at different levels of proficiency. These differences manifested themselves in complex ways within the context of the current study, however the general statement can be made Arabic writers tend to increase in written CAF as they increase in overall Arabic ability.

6.2.1 Lexical Complexity

This study operationalized lexical complexity according to the widely used Guiraud’s index of lexical richness. Lexical measures in general are known to be extremely useful in the conceptualization of L2 writing, as they are known to be highly predictive of holistic ratings of L2 writing (Ferris 1994). Crossley et al. (2011), for example, found a strong link between lexical measures and holistic ratings of L2 writing samples at the beginning and advanced proficiency levels. In the current study, Arabic learners’ score on this measure show a positive and significant correlation with several independent estimates of Arabic proficiency, including C-test score and self-estimate of overall ability. However, current knowledge regarding the development of Arabic vocabulary, and how this process interacts with the broader development
of Arabic interlanguage, remains incomplete. This lack of knowledge stands in contrast with the primacy of vocabulary in the Arabic acquisitional process (Al-Batal, 2006).

Although the acquisition of lexical knowledge by Arabic learners has been investigated by a number of researchers (e.g., Khaldieh, 1996; Moser 2013), the findings of the current study indicate that learner level may complicate the measurement of this knowledge. The Guiraud’s index measure revealed 5 outliers (operationalized as being located two SD values away from the mean) in the LP group. This indicates that, within this population, this measure may not be effective at revealing the lexical complexity of lower-level learners’ Arabic writing. Indeed, after outliers were removed, previously present positive correlations between LP lexical complexity and fluency disappear.

On the whole, NS writers in this study deployed language that was significantly more complex than their LP and HP learner counterparts. This is in keeping with the general expectation that NS writers will deploy more varied and more complex lexical forms than their Arabic learning counterparts. Skehan and Foster (2007) note in a meta-analysis of CAF studies comparing the oral production of NSs and learners that “clearly, and unsurprisingly, native speakers…are able to draw upon a wider range of less frequent lexis while speaking, reflecting their richer, more accessible, and better organized lexicons” (pg. 15). Lexical elements of NS production are likewise known to vary between and within individual NSs, however. Tweissi (1990), for example, demonstrated that individual Arabic NS speakers demonstrate variability in the lexical sophistication of their production depending on the context of interaction. Writing genre was likewise shown in this study to elicit differential rates of NS lexical complexity, with a narrative prompt supported by pictures eliciting the least lexically complex NS writing. Within
the limited context of the current study, NS writers are indeed observed to produce more lexically rich (i.e., complex) language than both groups of Arabic learners.

The difference between learner- and NS-produced lexical complexity is not always stark, however—the wide range of NS Guiraud’s index scores overlaps with a large expanse of the HP range. This indicates that many of the advanced writers in the current study approached “native-like” performance in this construct when their production is averaged across all three tasks, even if the NS-HP group comparison was shown to be statistically significant. Previous research has found that advanced-level language users often remain measurably separated from NS production, such as in derivative lexical knowledge (Schmitt & Zimmermann, 2002).

The relationship between lexical knowledge and other aspects of Arabic production has been the focus of a few studies, including that reported by Ricks (2015), who investigated the relationship between lexical knowledge and independent variables related to overall Arabic proficiency. Noor (2007) likewise found a relationship between L1 Arabic writers’ overall Arabic competency (operationalized as score on an Arabic cloze test) and their L2 English compositions. The current study extends knowledge about this relationship with the contribution that lexical complexity appears, within the current limited sample population of Arabic learners and NSs, to be the construct most strongly related to selected independent estimates of overall Arabic ability. Among learners, the regression model predicting lexical complexity is associated with the relatively high $R^2$ value of .58, i.e., indicating that the selected independent/explanatory variables account of 58% of variation in learner lexical complexity scores. This is the strongest model of the four presented in the Results chapter, and likely reflects a strong relationship between the construct accessed by Guiraud’s index and other important elements that govern Arabic production.
Lexical complexity also showed the strongest relationship among NSs to the collected independent variable, resulting in an extremely strong, positive relationship between NS Guiraud’s index scores and NS self-estimates of Arabic writing ability. That is, the variation among NSs according to this self-estimate was sufficient to produce a correlation coefficient of .80 ($p = .001$). This strong correlation stands in stark contrast to the coefficients associated with the other constructs, none of which are statistically significant (although fluency approaches significance). The fact that lexical complexity is alone in the investigated constructs in being significantly positively correlated with NS self-estimated L1 writing ability merits further investigation. Further studies would likely benefit from a complexification of this construct by employing alternative measures, such as of lexical sophistication (e.g., by comparing production with a frequency list of Arabic lexical items) or lexical uniqueness (e.g., by comparing each participant’s produced lexical types with those provided by their peers).

The results of this study hold an additional important implication for researchers working with written Arabic texts in the finding that neither the lexical lemmatization process nor the hapax-token ratio described in the Methodology chapter appear to have added any significant clarity to the depiction of lexical complexity in the writing of learners. Learners’ and NSs’ scores according to the hapax measure, which is based on the number of types that appear only once in participants’ writing, show an extremely high correlation with collected Guiraud’s index scores, which is based on the total number of types used ($r = .98$ across all participants). The strength of this correlation indicates that these two measures are effectively accessing the same construct. It is possible that additional measures based on the number of hapaxes produced by a given participant may be able to shed light on lexical aspects of Arabic production, although this does not appear to be the case in this analysis. Given the ease with which Arabic writing samples can
be measured for number of lexical hapaxes in the Lexico 5 software used in the current study, measures based on this unit merit further exploration.

The results of this study likewise contribute the relevant finding that the process of lexical lemmatization described in the Methodology chapter is apparently unnecessary for the successful depiction of lexical complexity in written Arabic. Admittedly, the lemmatization process had the expected result of lowering mean group scores and the variance associated with each group, and the two versions of learners’ Guiraud’s index scores were shown to be significantly different via a paired-samples t-test. However, the lemmatized and un-lemmatized versions of this measure show an extremely high, significant correlation with each other, both when all participants’ (r = .98) and when only learners’ (r = .97) scores are compared. As in the case of the hapax/Guiraud’s index comparison above, the lemmatization process therefore does not appear to have added a great deal of clarity in the depiction of lexical complexity in the written Arabic collected in this study. When considered in light of the fact that this process is necessarily time consuming—until software capable of automatically completing this process become publically available, it must be carried out manually—it is not evident that written data needs to be prepared in this way prior to the calculation of lexical measures. Future exploration of this issue is warranted, however, given that lexical complexity emerged in the current study as the construct most closely among learners to independent estimates of ability, and among NSs as the most strongly correlated with estimates of L1 writing ability.

6.2.2 Syntactic Complexity

The current study approaches the analysis of syntactic complexity in written Arabic through the medium of the T-unit and the related words per T-unit measure, and can be added to
Tweissi (1990) and Alбирини (2013), both of which employ the T-unit in the analysis of orally produced Arabic. The general findings of this study corroborate the broadly understood notion that syntactic complexity tends to increase concurrently with learners’ L2 proficiency, as has been established, for example, with measures such as clauses per T-unit (Kuiken & Vedder, 2012). A key contribution of the current study is that the T-unit is appropriate for the analysis of written Arabic, and is associated with the finding that W/T increases concurrently with several independent estimates of Arabic proficiency among Arabic learners. Although it has been argued that this measure may not fully represent the increased complexity of highly advanced L2 learners, who in turn may complexify their production via alternative syntactic devices (Norris & Ortega, 2009), the current study finds that Arabic writers generally tend to produce longer T-units as they become more proficient and self-confident writers. NS writers are shown to produce the longest T-units of all, although they too are subject to a great deal of within-group variation.

The finding that syntactic complexity increases in tandem with overall proficiency, self-evaluated Arabic ability, and years of Arabic study, coincides with the findings presented by Shakir and Obeidat (1991), who found in their study of 18 adult university-level L2 Arabic writers that syntactic complexity (in the form of average sentence length and ratio of complex to simple sentences) increased after a period of study. The current study’s use of T-units can be compared to Shakir and Obeidat’s classification of simple and complex sentences, although the current study did not find average sentence length to be significantly related to independent measures of Arabic proficiency. Although Gales (1980) has argued that T-unit-based analysis may not be suitable for writers at the lower levels of proficiency, the LP group in the current study showed a normal distribution according to T-unit length that is significantly lower than the HP and NS scores. All participants in this study had studied Arabic for at least 1.5 semesters, in
keeping with Ortega’s (2003) observation that L2 writers begin to manifest substantial changes in syntactic complexity after a year of instruction.

Arabic NS participants in the current study did not provide self-estimates of their Arabic writing ability that correlated significantly with their average W/T rates. Among learners, however, the relationship between W/T and the selected independent variables is much more clear. The multiple regression model associated with W/T presents an \( R^2 \) value of .48, the second-largest of the four constructs. Additionally, the LP-HP comparison according to W/T rates is statistically significant, indicating its relationship with overall proficiency (operationalized as C-test score). Taken together, it is clear that this measure is accessing a facet of learners’ Arabic production that is closely related to overall Arabic ability, even if this relationship may not be present among the NSs who participated in this study.

Although the T-unit has been shown here to be much more appropriate for the analysis of written Arabic than the sentence, it too can be complexified, and such analyses may shed more light on the individualized developmental process of linguistic complexity among Arabic learners. For example, future efforts in this vein may benefit from analyzing T-units by weight, rather than length alone, allowing for more a more nuanced depiction of advanced Arabic writing (cf. Evans et al., 2004). Future efforts may likely benefit from the introduction of clausal analyses to the analysis of written Arabic syntactic complexity. Although it is generally much easier to divide essays into T-units than into clauses (Polio & Shea, 2014), both units are regarded as capable of discriminating between writers at different levels of proficiency (Norris & Ortega, 2009; Ortega, 2003). Additionally, alternative approaches to task design may reveal between-group similarities with respect to the deployment of written complexity in the future.
Skehan and Foster (2007), for example, found that the addition of planning time to task design led both learners and NSs to produce more complex language while completing the task.

6.2.3 Accuracy

The results of this study reflect the fact that accuracy is a complex construct, and that the specific operationalization of “error” in a given study design has a substantial effect on resulting accuracy rates when data is coded. The current study attempted to take an emic approach to the operationalization of accuracy, with common practices of NS participants being considered as “valid,” rather than representing an “error.” This is evident in the decision to accept colloquial lexical items and regional spelling variants, especially in the placement of the hamza letter. An alternative approach to accuracy would likely yield wildly different results, however. For example, variations in the placement of the hamza letter could easily and justifiably be considered errors, as they represent departures from the codified morphological and orthographic norms that categorize written MSA. This would also include spelling variants observed in this study, such as the NS-produced مسئول masʔūl “responsible,” for example. Such an approach would greatly decrease rates of NS accuracy (and learner accuracy as well). However, certain elements of NS production present in the current study were nevertheless counted as errors, such as deviation from the norms of MSA verbal inflection (as in the abovementioned string لم تنماشي lam tatamāshā, or in the doubled preposition في بالقرب في bi-l-qurb). These can often be explained as “typos” or related orthographic departures from spelling norms. Future studies may find highly divergent rates of NS accuracy within similar written tasks by applying more restrictive or more permissive conceptions of accuracy in Arabic, however.
It is not inherently surprising that NS writers would commit “errors,” especially in the context of written Arabic prompts composed in MSA, a register with which NSs may have divergent levels of familiarity in producing. The NS writers who participated in this study did indeed produce “errors” according to the current operationalization of accuracy, even if their production is highly accurate overall. Spelling errors in general tend to be common among Arabic NS writers, and have likewise been shown to be extremely common among NS children writers (Abu-Rabia & Taha, 2006). Although the findings of this study should not be considered corroboration of the controversial claim by some researchers that “Literary Arabic” may be in effect a second language for Arabic NSs (Ibrahim & Aharon-Peretz, 2005), the results presented here do reflect a reality in which NS competencies in written Arabic are diverse and varied.

Nevertheless, the operationalization of accuracy in the current study was successful at producing group means that correspond to significant differences between the LP, HP, and NS groupings. In fact, the accuracy measure may represent the only investigated construct in which NS participants collectively showed very little overlap with the HP group of participants. Mean NS accuracy (operationalized as the ratio of error-free T-units to total T-units) was 21 points higher than HP accuracy, and this was the only measure in which there was no overlap between NS and HP IQRs. The HP group was likewise clearly distinct from the LP group, and these two groups likewise show no IQR overlap. Importantly, the NS writers included in this dissertation prove (with one outlying exception) to be extremely accurate writers overall—all NS production is clustered close to the scale’s maximum value of 100% accuracy, and the NS group exhibits a reduced amount of variance on this measure when compared to the LP and HP groups. This indicates that, as Arabic writers increase in overall Arabic ability, the range of their possible accuracy rates becomes increasingly restricted.
The clarity of the distinction between groups according to this measure corroborates Mansouri’s (1995) finding that accuracy is a key measure through which advanced and beginning Arabic learners diverge. The current study adds NSs to the mix, finding that even HP writers tend to be much less accurate than NSs, indicating that accuracy measurements may be good indicators of difference between the writing of NSs and high performing learners. However, Brosh (2015) found that second- and third-year learners of Arabic tended to commit orthographic/spelling errors at similar rates, even though third-year learners tended to be more fluent (i.e., composed longer compositions) overall.

Future studies in this vein may choose to manipulate task-internal elements of the experimental design, for example by alternating the amount of planning time given to L2 Arabic writers. On-task planning time has been shown to promote accuracy at the expense of fluency and complexity among L2 language users. Adams et al. (2014) speculate that this may be due to learners’ use of this time to develop and complexify written content, leading to a more-cognitively demanding task overall, which in turn may elicit reduced accuracy rates (p. 153).

The decision of which forms should be considered “acceptable,” and which forms should be considered “unacceptable,” in a study such as this confronts the individual researcher with the uneasy task of weighing in on a highly contentious and ideologically charged issue. Perhaps the simplest strategy would be to take a strict “MSA only” approach, invalidating all lexical forms that are not found in a standard MSA dictionary, and counting as “errors” all orthographical deviations from these norms. This approach would inevitably render a high percentage of NS writing as containing errors, however, and it would likewise invalidate the incredible diversity of forms that characterizes NS writing in Arabic. Alternatively, researchers may choose an entirely descriptive approach, in which all NS production is assumed to be accurate merely by nature of
having been produced by a NS writer. This approach, while it may have much to offer in the way of informing a practical, empirically based conception of what learner accuracy should aim to resemble, would complicate efforts to model the writing of NSs in various task conditions.

6.2.4 Fluency

Among Arabic learners, fluency shows a positive correlation with each independent variable selected for regression analysis (although NS interaction frequency is non-significant). However, these correlations belie the fact that the multiple regression model targeting fluency is the weakest of the presented models, revealing an $R^2$ value of .33. Additionally, of the selected independent variables, only the C-test was revealed via regression to contribute significantly to variation within this measure, indicating that the relationship between fluency and YoS, self-estimate, and NS interaction frequency are tenuous within this population of learners.

Regardless, the comparison of group fluency means nevertheless holds interesting indications for Arabic practitioners, including the elementary conceit that writers tend to write more quickly as they become more proficient in Arabic. Additionally, variance with respect to fluency increases in tandem with overall fluency rates. This is in keeping with the general finding that written fluency tends to substantially increase as learners have more experience with the target L1 (Chenoweth & Hayes, 2001). In a methodologically similar study of Arabic spelling strategies, Brosch (2015) found that third-year Arabic learners tended wrote on average 40 more words per hour than second-year learners. The NS group in the current study consequently demonstrates the highest overall fluency rates, as well as the highest rate of variance (SD) according to this measure.
The elevated NS fluency variance (when compared to the LP and HP groups) corresponds with the knowledge that L1 users vary in their rates of oral fluency (cf. Derwing et al., 2009), extending this finding the written mode. NS Arabic writers who participated in this study varied widely in terms of the number of written words that they produced in their time on task, ranging from a minimum of 8 words per minute to a maximum of over 22 words per minute. This is not unsurprising given the variation associated with NS fluency rates in related studies. It is known, for example, that NSs in the oral mode produce disfluencies such as breakdowns and pauses much in the way that learners do (Skehan & Foster, 2008). Although such disfluencies may occur at different rates and locations than those produced by native speakers (cf. Davies, 2003a), they would still contribute to variability within a NS group with respect to fluency measures, such as that seen here. The high degree of variability within the NS group with respect to written fluency may have likewise prevented the fluency-L1 writing ability correlation from reaching the level of statistical significance.

Additionally, the apparent high degree of overlap between the HP and LP groups is not necessarily an unexpected finding. Related studies have found that, when compared to L2 learners, NS fluency rates can be a tricky target of inquiry in an experimental setting. Isurin (2012), for example, found that Russian NSs were measured to produce fewer words on average than L2 Russian learners and English-Russian bilingual speakers in the oral mode. Certainly, modality may have played a role in the variation collected in this study, as typing speed undoubtedly interacted with task-internal factors such as writing genre to affect written fluency. Perhaps the most remarkable aspect of the fluency-related findings presented in this study is the fact that this construct appears to be the least related (within this population of participants) to independent measures of Arabic proficiency.
6.2.5 Unreliability of Sentence Measures

The current study includes two measures based in part on the use of learner-punctuated sentences, and neither of these was revealed to be a reliable indicator of complexity in Arabic learner writing collected as part of this study. Although a number of studies of written and oral CAF in other languages have employed sentences, such as in the widely used measure of T-units per sentence as an operationalization of syntactic coordination, for example, this measure was not revealed here to be useful in the depiction of learner writing.

This general unreliability of Arabic sentences corresponds with existing knowledge about the interaction between Arabic NS punctuation practices and syntactic devices for coordination and subordination. It has been noted, for example, that Arabic independent clauses are often linked to each other via discourse markers and connecting particles (Green & Manning, 2010, pg. 395; Ryding, 2005), rather than punctuation such as the periods, question marks, and exclamation points used to differentiate sentences in the current study. This results in Arabic sentences that are often much longer than their English counterparts, and indeed this appears to have been internalized by the single HP participant noted above who composed responses to all three prompts in the form of single sentences, rendering this participant’s W/S and T/S higher than any other learner or NS participant.

Additionally, as noted in the discussion of accuracy in this dissertation, Arabic orthographic practices are often highly variable, and this includes punctuation practices. Although this may not often be the source of explicit instruction in the typical Arabic classroom, Arabic NS punctuation is highly contextualized, and is subject to the influence of factors such as text formality, especially in online contexts (Abbasi & Chen, 2005). This in turn leads to a
situation in which Arabic punctuation practices are highly idiosyncratic (cf. Azni & Al-Thaniyyan, 2012), which in turn renders analyses based on punctuation-determined units such as the sentence highly unreliable.

This unreliability has led some researchers to problematize the use of sentences in the analysis of written Arabic. Al-Jabr’s (2006) study of rhetorical practices in Arabic writing, for example, responded to these challenges by proposing an alternative formulation of the “sentence” as it appears in Arabic texts that closely resembles the definition of T-unit used in the current study (p. 207). However, empirical investigations of L2 Arabic texts via the application of common CAF measures remain extremely scant. The findings presented here confirm the observations of previous researchers regarding the unpredictable length of Arabic sentences, and argue for more fine-grained units such as the T-unit in order to successfully depict elements of Arabic learners’ developing interlanguage.

6.3 Research Question 2: Independent Variables that Explain Written Arabic CAF

This study’s second research question interrogated the relationship between written CAF and several independent measures of Arabic proficiency among learners, including various self-assessments, a C-test score, self-reported years of study, and literacy practices. Although previous research into the predictive power of the C-test format in relation to aspects of L2 Arabic writing is largely nonexistent, Noor (2007) found that a Cloze test was partially predictive of certain elements of the L2 English writing compositions of Saudi L1 Arabic speakers. The current study extends this knowledge to Arabic learners, finding that the C-test and self-estimated overall Arabic ability consistently emerged as the most significantly related to
written learner CAF. This conclusion is made on the basis of various bivariate correlation coefficients and the various regression models presented above.

The independent measure that proved to be the most-unambiguously successful at determining variation in learner written CAF is the C-test. Learner C-test scores fall on a wide scale (out of 125) and the learners in this study demonstrated a wide range of possible scores. The C-test showed a high degree of correlation with other independent measures (e.g., $r = .70$ with YoS), and a high degree of correlation with learner scores on the four investigated CAF variables, with the highest such correlation being lexical complexity ($r = .70$) and no correlation falling below $r = .50$. In the context of multiple regression analyses, the C-test consistently emerged as the most predictive independent/explanatory variable with regard to the four investigated CAF constructs, with the exception of fluency, in which no independent variable’s contribution was deemed to be statistically significant. Indeed, the C-test $r^2$ values in the various regression models are consistently the highest of any independent variable, except for in the context of the regression model targeting fluency, in which no variable’s contribution to the model was revealed to be statistically significant. On this basis, it can be said that the C-test is by far the most predictive independent variable in determining Arabic learner CAF scores.

The self-estimate of overall Arabic ability emerged in the various regression models as the second-strongest explanatory variable of CAF. This is in line with Ricks (2015), who found that a limited self-assessment questionnaire provided data that correlated strongly with measures of Arabic vocabulary knowledge. More broadly, self-assessments or self-estimates have been shown to correlate well with other measures of language ability, providing a measure of concurrent validity (Shameem, 1998). The results of this study argue for the inclusion of such questionnaires in experimental design, as the battery included in the current study required
minimal effort to administer, yet nevertheless proved to be much more reliable than YoS in predicting learner CAF. Additionally, this measure also provided useful information about lexical qualities of NS writing, providing another indication of its relationship to the processes understood to govern linguistic production.

An important finding of this study is that YoS is quite problematic if used as a substitute for other independent measures or estimates of overall proficiency level in an experimental setting. Although this measure is associated in the current study with significant positive correlations with C-test score \( r = .70 \), overall self-estimate \( r = .57 \), and reported NS interaction frequency \( r = .43 \), it consistently emerged as a non-significant contributor to the various regression models targeting learner CAF. This is in keeping with the general knowledge that number of years of Arabic study is an unreliable indicator of overall Arabic proficiency or predictor of Arabic test performance (Rammouney 1975, 1983). Although easy to solicit from participants, and easy to code in numerical form, this measurement does not take into account any number of factors which undoubtedly influence what a language learner can actually do in an L2. Instructional curriculums vary from institution to institution, for example, and may draw on a wide variety of divergent materials. Even within the same Arabic program, individual course sections may vary widely with respect to instruction methods and amount and type of homework assigned. More broadly, Arabic courses may differ in terms of intensity and number of contact hours provided to students. Given this measurement’s ease of collection and centrality to claims of efficacy made on the basis of participation in individual Arabic programs, it makes sense to include in background questionnaires, however it should be included in addition to, rather than in place of, additional measures such as self-estimates.
Finally, the battery of questions investigating learner and NS literary practices was unsuccessful in uncovering between-group differences that corresponded either to NS status or membership in the LP or HP groups. In other words, participants in this study did not supply answers to these questions that are statistically related to their writing ability. One possible cause of this lack of significance accorded to literacy practices is that the investigated practices themselves—such as reading and contributing to social media, writing formal and informal emails, etc.—are relatively ancillary to an individual’s ability to write in Arabic. It is also possible that the questionnaire used in this study investigated the wrong literary practices. Regardless, future research may reveal to what extent an Arabic learner’s interaction with written Arabic influences their ability to compose complex, accurate, and fluent writing, even if this relationship did not emerge from the current study.

6.4 Research Question 3: Genre Effect on Written CAF

This study’s final primary research question investigated the relationship between written task genre and variable performance on CAF. When each group’s performance was broken down according to task, each condition indeed elicited divergent rates of CAF performance in each group, although these interactions were not always revealed by repeated measures ANOVA to be statistically significant. In general, these results indicate that learner and NS performance in the persuasive condition may provide evidence of the type of “trade-off” predicted by Skehan (1998), as this condition is associated with much more syntactically complex, but slightly less accurate production in this group. In general, Robinson’s (1995, 2001) cognition hypothesis predicts that increasing task difficulty will result in a simultaneous increase in both accuracy and complexity, and that complexity and accuracy should correlate, whereas Skehan (1998) predicts
that constructs such as complexity and accuracy will “trade off” with each other, and that devoting increased attentional resources to one will lead to decreased performance in the other (Skehan & Foster, 2007).

The differential rates of written CAF elicited by the various prompts included in this study are in line with a preponderance of studies that have found that writing tasks targeting different genres tend to elicit different linguistic structures (Kaldieh, 2000; Carell & Conner, 1991). In the context of the current study, the persuasive task proved to elicit the most syntactically complex (W/T) writing among all groups, while simultaneously eliciting the least accurate (EfT/T) writing among all groups. In some cases, the difference between the persuasive task and the remaining two tasks according to these measures is striking, such as in the measured decrease in LP writing from 59% and 61% error-free T-units in the descriptive and narrative tasks, respectively, to 43% accurate T-units in the persuasive task. Although this effect is likewise found in the HP and NS groups, it is somewhat attenuated. This corresponds with Skehan & Foster’s (2001) posited tension between accuracy and complexity, in which both constructs are observed to “trade off” with each other, rather than increase in tandem. A similar tension was observed by Ruiz-Funes (2014), who found that, between two writing tasks, the more complex task elicited longer T-units and higher rates of subordination but lower rates of accuracy, coordination, and fluency among adult advanced L2 writers. The current study extends this finding, establishing that that a complex writing task elicits more complex, but less accurate writing even among NS writers.

Ruiz-Funes (2014) speculates that the genre of argumentative essay prompts writers to respond with “a more formal, academic register characterized by longer and more varied sentence structures,” the composition of which in turn reduces the amount and extent of
attentional resources available to writers in attending to the accuracy of their writing (p.183). This certainly appears to be the case in the context of the current study, as NS responses to the persuasion prompt are marked by a notable frequency of discourse markers and devices drawn from the formal register of MSA. The results of the current study likewise indicate that persuasive tasks may be helpful in efforts to distinguish between the writing of learners and NSs, given that L2 learners’ production of persuasive writing is often characterized by stylistic diversions from NS norms (Kellog, 2001). For writers of Arabic, a persuasive/argumentative task may prove to be particularly effective at revealing differences between L2 and NS writers with regards to their lexical and syntactic performance, as well as “on the frequency of cohesive markers that a particular … writing task can be expected to elicit” (Kormos, 2011, 149). Al-Jabr (2006) argues that Arabic editorials are characterized by a frequent occurrence of coordinated clauses and phrases intended to have a pragmatic influence on the reader. Texts within this genre are typified by the strategic use of particular syntactic structures (e.g., coordination) in order to make a rhetorical impact (p. 209). Future studies of Arabic writing may therefore benefit from an attempt to classify or subjectively rate such writing samples, in order to associate differential CAF performance according to genre with empirical indications of register.

The current analysis assumes that the persuasion task represented the most challenging/complex writing task for learners, and this assumption is made on the basis both of the aforementioned follow-up interviews, in which selected participants overwhelmingly rated this task as more difficult, as well as the results of the CAF measures which reveal that this task elicited the most syntactically complex writing among all groups. The persuasion task in the current study is associated with each group’s least accurate production, as well as the least fluent production of LP and HP learners. HP persuasive writing occurs in this task .82 W/M less rapidly
than description (the next-most-fluent task), and LP persuasive writing occurs in this task 1.16 W/M slower than description. This tendency is largely absent from the observed writing of NSs, however, contributing to the failure of a repeated measures ANOVA to find statistical significance in the interaction between task type and group fluency means. The decreased learner fluency and accuracy rates elicited by the persuasive task among Arabic-learning writers seem to corroborate Ruiz-Funes’s (2014) finding that a less complex writing task oriented around a topic more familiar to learners elicited more fluent and accurate, but less complex writing. Taken together, this indicates that complex writing tasks (in the form of an argumentative essay) are associated with less-fluent writing among both HP and LP learners, but that NSs do not appear to be subject to the same fluency constraints.

The results of this study provide an additional indication of the differential effect of task type on learner and NS CAF performance in the relatively restricted NS lexical complexity elicited by the picture-supported narrative task. For example, the narrative prompt was associated with the lowest rate of lexical richness among NS writers (7.75) and the highest rate of lexical richness (Guiraud’s index) by LP writers (6.65), with HP writers falling in the middle (7.46). The narrative condition is thus associated with the lowest rate of NS lexical richness and the highest rate of LP lexical richness. It seems likely that the series of cartoon panels used to support this prompt had the effect of restricting the range of lexical items employed by NS writers by tying this group to a common narrative skeleton. Conversely, the picture panels seem to have had the simultaneous effect of increasing the lexical diversity of LP writing by prompting LP writers to deploy familiar lexical forms in narrating a relatively quotidian scene, leading to more lexically complex production.
The findings of this study do not fully conform, however, with the findings presented by Ruiz-Funes (2014) that HP and LP learners behaved differently in a more-complex and less-complex task condition. Ruiz-Funes found that the more-complex task elicited more syntactically complex, more accurate, and more fluent writing among HP learners, while the less-complex task elicited more syntactically complex and more fluent writing among LP learners. In the current study, however, the HP and LP groups tended to mirror each other’s performance, with the HP group achieving qualitatively higher rates of written CAF than the LP group, but with both groups seeing largely the same task-specific increases and decreases cross writing tasks.

Given the differential rates of written CAF elicited by the various tasks used in the current effort, future studies of Arabic L2 production would likely benefit from a complexification of the broad task types used in the current effort. Various versions of a narrative task, for example, may elicit differential rates of linguistic CAF, in keeping with Tavakoli’s (2014) finding that complex storylines can elicit more syntactic complexity in writing. Such efforts may help to extend Skehan’s (2009) finding that narrative tasks tend to elicit more complex, but less fluent and accurate production. This tendency was not observed in the NS or learner writers in the current study, although the generalized nature of the three writing tasks at hand may have obscured these elements of participants’ production.

6.5 Limitations

The current study is associated with a number of limitations that affect the extent to which the findings presented here can be expected to apply to similar efforts in this vein. Chief among these limitations is the relatively small number of participants (35 learners and 14 NSs) on who’s written production the conclusions discussed here are based. The restricted size of this
participant pool in turn leads to statistical effects such as the exaggeration of effect sizes, such as Cohen’s $d$, and possibly inflating the size of the several strong, positive correlations identified in the previous chapter. Every effort has been made to remove outlying participants from the calculation of regression and correlation coefficients, however this effort has the simultaneous effect of further lowering the number of participating participants.

A second limitation of the current study is the fact that participants were drawn from a wide range of anticipated ability levels. This is reflected in the wide range of YoS scores (range $= 9$ years), as well as the wide range of C-test scores (71 points on a scale of 125 points). Admittedly, the inclusion of such a wide distribution of participants allowed for the capturing of learner-produced writing that falls quite high on the various scales of CAF performance discussed here, and in turn likely contributed to the fact that NS scores on the various measures often appear to be surprisingly close to HP scores on the same measures. However, this wide range of learner abilities likewise likely had the effect of obscuring within-group elements of learners’ writing. If all 35 learner participants had been first- or second-year Arabic learners, for example, this study may have been able to contribute more meaningful information about the varied performance of learners at that level, including how aspects of LP CAF vary from year to year.

Finally, future CAF-oriented studies of learner-produced written Arabic would likely benefit from the inclusion of additional measurements in each construct of production. For example, the depiction of written fluency could include an operationalization of disfluencies, such as through the coding of scratch-outs and erasures if responses are handwritten, or by using a software program that digitally captures the composition process if responses are typed. Future efforts aimed at investigating the nature of written Arabic may similarly take into consideration
additional factors related to the relative quality of written discourse, such as text organization features (Uysal, 2008) and related groups of textual features (Frigina & Weigle, 2014). Future efforts may likewise benefit from the inclusion of oral data, as well. After all, it is known that the written and oral modalities of linguistic production are associated with divergent types of production. Kormos (2014), for example, showed that L2 learner writing is associated with higher rates of syntactic and lexical complexity than speech produced by the same learners (Kormos, 2014). A multifaceted approach to measuring Arabic CAF that includes oral and written data may illuminate under-explored aspects of the written-oral relationship, as well as provide information regarding the role instruments such as the C-test can play in predicting varied aspects of such production.

The current effort is thus best viewed as an introductory foray into the application of commonly used CAF measurements to the depiction of learner- and NS-produced written Arabic, and as such the constructs accessed by the CAF measurements presented here are explored in a very broad way. Accuracy in particular would benefit from a more fine-grained analysis with respect to written Arabic, as the specific operationalization of this construct has the potential to significantly alter accuracy rates according to various measures such as overall frequency of errors, ratio of error-free units to total units, rates of severe/major errors to minor errors, etc. This construct is also subject to a measure of ideologically charged debate, as the question of what represents “correct” written Arabic has the potential to be highly contentious.

6.6 Implications

The results of this study have a number of implications, including for the teaching of Arabic, for future research efforts at conceptualizing variable CAF performance in written
Arabic texts, and for interpreting the apparent variability of NS written performance. Currently, the impetus to elaborate an empirically supported model of NS and L2 writing at various levels of proficiency is especially critical given that a majority of undergraduate-level learners in the United States report a desire to achieve “professional”-level proficiency in Arabic (Belnap & Nassif, 2011, p. 4).

6.6.1 Implications of variable NS performance

One of the notable findings of the current study is that NS written performance is apparently quite variable according to the selected CAF measures. Furthermore, rather than falling substantially higher than both groups of Arabic learners on the various scales associated with the selected CAF measures, the NS group often appears to be akin to a very-high-performing group of learners. This is true for the measures of fluency, syntactic complexity, and lexical complexity, in which there is substantial overlap between the HP and NS interquartile ranges. The only construct for which this is not the case is accuracy, in which the NS group is almost entirely clustered around the top of the scale, with a mean accuracy rate that approaches 100%. Although a few studies have found systematic differences between learner and NS Arabic production in the oral (Fakhri, 1984) and written (Fadel, 2014; Fareh, 1988) modes, existing knowledge regarding the differential performance of these populations according to the empirical measures of written CAF used here remain extremely scant. These perhaps-surprising results can be interpreted in light of the fact that written MSA represents a register of linguistic production with which NSs can be expected to have varying degrees of familiarity and confidence.

The presence of variation in the writing of Arabic NSs is not a necessarily surprising finding. For instance, Tweissi (1990) found variability in the syntax of L1 Arabic production,
although in this case the source of variation was assumed to be audience, rather than genre or register familiarity. In the current study, a major source of apparent variation in orthographic practices is assumed to be the influence of regional spelling norms. Of the 14 NS participants, 11 were speakers of Saudi Arabic, two of Egyptian, and one of Lebanese. The inclusion of NSs from different dialect regions likely resulted in the introduction of regional variation to the pool of NS responses, for example in the inclusion of region-specific practices for writing the hamza letter.

NS variation likely also stems from the complex tapestry of national educational policies in the Arab world, in which secondary and tertiary education in the same country may be conducted in a variety of languages, and in turn likely leads to varying rates of facility with formal written Arabic. In the current study, all NS participants reported having reached or completed university-level education at the time of data collection, and the reported academic majors of these participants are likewise diverse, although a majority were students of medicine. With one exception, all NS participants report having attended secondary school in Arabic, rather than a foreign language, indicating that this group of writers can reasonably be expected to be familiar with the type of formal writing that occurs as part of a typical Arabic-language educational trajectory. However, in a reflection of the incredibly diverse array of educational systems and formats offered in the Arab world, only two of the final 14 NS participants reported that the language of instruction in their university education was Arabic, and one participant reported attending university in a mixture of Arabic and English. The remaining 11 participants (8 Saudi speakers, 2 Egyptian speakers, 1 Lebanese speaker) report English as the language of instruction in their university level education. This diversity of backgrounds is likely reflected in the varied responses to the questionnaire in which NS participants were asked to rate themselves
against their peers according to their self-perceived L1 (Arabic) writing ability, collectively responding with a mean score of 3.57/5 (SD = 0.76). These responses are significantly lower and more varied than the responses of the 35 Arabic learners to the same question about their L1 (English) writing ability (M = 4.69, SD = .47).

In some respects, NS writing in MSA can be considered a corollary of academic L1 writing in other languages such as English. Just as no population of Arabic speakers can be said to learn MSA as a first language (Al-Wer, 1997), neither is academic discourse the first language of any L1 writer (Tang, 2012). Like MSA, academic writing is a specialized competence which must be acquired—typically through education—by both native and non-native speakers (Hyland, 2016). In the current study, this may have led to the noted high degree of variability in NS writing scores, as well as the relatively wide range of variability in the L1 writing self-assessment. Arabic NSs are known to employ a wide range of codes, registers, and scripts when writing (e.g., Palfreyman & Khalil, 2003; Warshauer et al., 2002), and the current study shows that they are also highly variable according to direct measures of written CAF.

The discussion of literacy and writing in the Arab world is extremely complex, with literacy rates and educational policies varying significantly from country to country. Furthermore, educational and language policies in the Arab world are often highly politicized. Arabic language education is the norm in many countries, but it is alternatively associated with controversial efforts at Islamization or Arabization in others (e.g., see Sharkey, 2012, for a discussion of this issue in Algeria and Sudan). Indeed, the very act of writing in Arabic, as well as the language of education in a given country, has been rendered a locus of conflict in various national struggles over identity, culture, writing. Conceptions of accuracy and “correct” usage are consequently tied into these struggles. Maamouri (1998) writes that “the special focusing on
mistakes and incorrect usage” reflects a notable “concern about the spread of error in the formal Arabic language production which was thought to be detrimental to development” (p. 24).

6.1.2 Implications for TAFL

As noted above, the varied nature of NS production collected in this study indicates that future efforts at teaching and assessing Arabic would benefit from a more “permissive” approach to the conceptualization of written accuracy that is capable of accommodating realistic and expected aspects of NS writing. After all, the fact that several NS participants included colloquial lexical items in their responses to the current study is in keeping with knowledge regarding the diversity of codes, languages and scripts used by Arabic NSs to fulfill a variety of social and communicative purposes in the modality of computer-mediated communication (CMC) (Hallajow, 2016; Bianchi, 2013). In fact, a majority (13/14) of the NS participants in the current study composed their responses via CMC, and this modality may have facilitated the use of the diverse forms discussed above. This is noted in spite of the fact that each prompt was presented to NS participants in MSA. In light of this, the pressing challenge for Arabic language educators might be best formulated not as how to prevent the appearance of colloquial items in the writing of students, but rather how to ensure that such items appear in a contextually appropriate, native-like manner.

However, researchers working with written Arabic may also face the challenge that an entirely descriptive, permissive approach to the conceptualization of written Arabic accuracy has the effect of obscuring varying rates of NS facility with the formal written language, much as an entirely prescriptive approach inevitable obscures the natural diversity of forms and structures in NS writing. El-Essawy’s (2010) study represents a step in this direction, providing a framework
for conceptualizing variation in Egyptian print media, and in turn delineating what could be seen as “acceptable” patterns of colloquial use in one genre of written Arabic.

Arabic practitioners should likewise consider the effect that writing modality has on learner performance, as well. As digital literacy skills continue to gain emphasis and importance in foreign language curricula in coming years, future studies should attempt to interrogate the effect of modality on written CAF. How do the mechanics of typing influence how learners and NSs compose Arabic texts, and how does modality in turn affect how the various structures accessed by CAF are perceived by Arabic-reading audiences? Given that the vast majority of Arabic NSs who participated in this study did so electronically, it stands to reason that the mechanics of electronic communication should be emphasized in an Arabic curriculum that encourages communication and engagement with the Arabic-speaking world outside the classroom.

The results of this study likewise argue for a careful consideration of task type, genre, and complexity in the context of Arabic instruction. In their experimental exploration of the acquisition of Arabic vocabulary by L2 Arabic learners, Golonka et al. (2015) argue that Arabic learners benefit from pedagogical tasks that encourage the expending of high levels of cognitive effort when acquiring novel vocabulary items. The results of the current study expand this, arguing that instructors and curriculum designers likewise consider the effect that task genre likely plays in the use and reinforcement of lexical knowledge among learners. Although lower level learners may, as in the current study, employ more lexically complex language when responding to a prompt framed by images depicting familiar items and activities, higher-level learners may actually produce less lexically complex language in such a condition, possibly because the images serve to restrict possibilities for the use of alternative forms. Given that
cognitive effort is a central factor governing vocabulary acquisition (Hulstijn & Laufer, 2001). Arabic instructors and curriculum designers should strive to encourage this effort through carefully considered pedagogical task design.

The influence of curricular choices on linguistic production is in fact easy to observe in the current study. The frequency of learner use of the item *tadkhīn* “smoking” is certainly evidence of this, as the item is featured prominently and early (in the 9th chapter) in the *Al-Kitaab* textbook series (Brustad et al., 1995, 2005). Curricular designers should therefore carefully weigh the materials and tasks to be used in Arabic programs, given that these choices directly and significantly affect what learners can be expected to do with the language. Learner exposure to novel dialects and registers of Arabic, for example, has been shown to reduce learner anxiety in the classroom (Elkhafaifi, 2005). Practitioners should note that learner writing has been shown to exhibit from qualitative improvements as a result of repeating the same writing task over an extended period of time (Nitta & Baba, 2014) and that repeated exposure over time is an important factor in the acquisition of Arabic vocabulary (Golonka et al., 2015).

The success of LP writers in particular in responding to the various prompts of this study argues for a broader reconceptualization of the role of writing in Arabic curricula. Generally, it appears that LP participants in this study tended to collectively write at above the novice level according to the ACTFL guidelines. The Arabic-specific notes for writers at the Novice Mid level, for example, use “short separate sentences with no specific order,” have “serious communication problems beyond formulaic language,” and are “only aware of semantic value of words; not grammatical function” (ACTFL, 2012). Many LP learners in their second semester of study, however, composed essays that were on par in terms of essay length with many HP and even some NS submissions. Given that such progress can be seen after only one year of
instruction on the part of many LP writers, it stands to reason that a more carefully formulated approach to writing in the Arabic classroom would have the potential to allow even greater gains by lower level writers.

Finally, this study contributes the finding that both the C-test and self-estimates of Arabic ability were much more predictive of variable CAF performance among learners than the commonly used YoS estimate. As noted, YoS is necessarily a rough measure that obscures a great deal of variation in curriculum, instructional philosophy, learner engagement, etc. This corroborates the findings of Rammuny (1975, 1983) who found that YoS was similarly unreliable in predicting learner scores on early versions of the Arabic Proficiency Test among university-level learners. Abdalla and Al-Batal (2011) likewise found that Arabic instruction practices are often highly variable, both within and across levels (p. 15). It may therefore be impossible to determine, for example, how much class time was devoted to writing development the types of texts emphasized for any given learner participant. In light of this, alternative estimates of Arabic proficiency such as the C-test hold great potential for Arabic practitioners seeking to quickly separate Arabic learners into usable groups that correspond to varying proficiency levels. Norris (2006) describes such an effort, presenting the development of a C-test for use in the placement process in a university German program.

Clearly, the writing collected pursuant to this study represents a snapshot in the complex and multifaceted writing developmental processes undergone by all leaner and NS participants. Indeed, empirical research has shown that L2 writing ability is characterized by the emergence of peaks and valleys when studied longitudinally—Nitta and Baba (2014) found that L2 writers are prone to “self-organizing” their L2 system as they participate in a repeated writing task over the course of an academic year (p. 125). The development of writing in Arabic learners is thus a
complex and highly individualized process that benefits from exposure to rich and varied sources of input. As this study shows, however, the nature of the input to which learners have had access at the time of data collection is extremely difficult to measure. Years of Study was not shown here to be as-reliable of a variable as self-assessments or C-test score, and this may well be due to the fact that individual Arabic courses often quite diverse, even if they occur within the context of a “unified” curriculum.
Chapter 7: Conclusion

This study represents an initial effort to apply direct measures of linguistic CAF to the writing of Arabic learners and NSs. In spite of its limited context and scope, the current study makes a number of contributions that will hopefully inform future research efforts, and it holds a number of important implications for the broader TAFL field. First, it is clear that, within the limited context of the current effort, lexical lemmatization was not a prerequisite for the lexical analysis of Arabic writing samples produced by this population of writers. Second, this study has confirmed a finding of researchers working with writing in other languages, namely that the sentence can be an unreliable unit of analysis for learner writing, and that this may be especially so in Arabic. Third, this study has established that the novel “short-cut” estimate of Arabic proficiency used in this study—the C-test—is significantly related to a number of other independent and dependent variables, such as years of study, self-estimates of Arabic ability, and the various CAF measures examined here. Groupings determined on the basis of C-test score consistently proved to be significantly different according to various measures of CAF. Significantly, the C-test consistently emerged as the best predictor of variation in CAF via multiple regression, with the exception of the fluency construct, in which none of the four selected independent variables proved to be significant contributors to variation. Additionally, variance in NS fluency and lexical complexity rates are shown to be closely related to independent self-estimates of L1 (Arabic) writing ability. Finally, this study has demonstrated that different written task genres elicit measurably different types of production according to CAF measures, and that these effects are often found both in the production of learners and NSs.

In addition to successfully representing empirical aspects of the writing of Arabic learners, the CAF measures presented here likewise contribute to the representation of the varied
written competencies of Arabic NSs. Interestingly, the Arabic NSs who participated in this study collectively show a great deal of overlap with the identified high-performing learners, indicating that the writing habits and abilities of Arabic NSs deserve further inquiry. Furthermore, although Arabic learners who participated in this study collectively rated themselves as excellent English writers, Arabic NSs showed a great deal of variance in their self-evaluated L1 Arabic writing ability. This variance as in turn sufficient to reveal a significant positive correlation with NS lexical complexity scores across their three writing samples, arguing strongly for the inclusion of similar self-evaluations in future studies that collect writing from Arabic NSs.

This study has several implications for the teaching and assessment of Arabic. The NS production collected pursuant to this study complicates the conception of accuracy in written Arabic, supporting a broader operationalization of accuracy that is able to accommodate this diversity. It is also clear that NSs and learners produce written language differently according to task genre or experimental condition, arguing for a reformulation of writing tasks in the Arabic classroom with an eye for sequencing pedagogic tasks. Finally, this study presents a series of measures that have been shown successful at revealing differences between Arabic writers at different levels of proficiency. Future efforts will hopefully unpack and diversify these measures in the pursuit of an empirical, fine-grained conception of Arabic writing as it is practiced by learners actively acquiring Arabic.

The act of composing writing in Arabic is exceptionally complex, and this complexity is equally evident in the writing of learners and NSs. Furthermore, the development of writing ability is highly individualized, and should be carefully considered in the context of L2 Arabic curriculums. A great deal of investigation remains for researchers and practitioners working with written Arabic. It is hoped that future studies will build on the methods and findings presented
this dissertation, both by expanding research efforts to larger and broader populations of Arabic writers, as well as by diversifying the types of analyses applied to their writing. Ultimately, an accurate, empirically based conception of the varied written competencies of Arabic learners and NSs can only have a positive impact on efforts to develop new and more effective materials and methodologies for teaching Arabic.
References


http://www.actfl.org/publications/guidelines-and-manuals/actfl-proficiency-guidelines-2012/arabic/%D8%A7%D9%84%D9%82%D8%B1%D8%A7%D8%A1%D8%A9/arabic-reading-examples


Alanen, R., Huhta, A., Martin, M., Tarmanen, M., Mäntylä, K., Kalaja, P., & Palviainen, Å.
(2012). Designing and assessing L2 writing tasks across CEFR proficiency


Albirini, A. (2013). Toward understanding the variability in the language proficiencies of Arabic

Albirini, A., Benmamoun, E., & Saadah, E. (2011). Grammatical features of Egyptian and
Acquisition, 33*(02), 273-303.

Students. *English Language Teaching, 8*(7), 181.

Alderson, J. C. (1979). The cloze procedure and proficiency in English as a foreign

Press.

English/Arabic. *Babel, 52*(3).


Mos, Maria. (2010). *Complex lexical items*. Utrecht: LOT.


*sīyāsāt/ṭarīḥ it il-kitāba*. (n.d.). In *Wikipedia Misrī*. Retrieved November 11, 2016, from https://arz.wikipedia.org/wiki/%D9%88%D9%8A%D9%83%D9%8A%D8%A8%D9%8A%D8%AF%D9%8A%D8%A7%D8%B3%D9%8A%D8%A7%D8%B3%D8%A7%D8%A7%D8%A8%D9%88%D8%A7%D8%A8%D9%87


Appendix A: Arabic C-Test

The following five texts were selected for inclusion in the final version of the Arabic C-test. Presented below in order of increasingly difficulty, they each contain 25 mutilated items beginning after the first intact sentence of the text. These texts were administered to the 35 Arabic learners who also responded to written prompts pursuant to this study. Test takers are instructed to at least attempt each individual text, although they are allowed to allocate their 30 minutes as they see fit. Following the scoring process, in which each attempt at item restoration is rated as correct or incorrect, test users are provided with a score (out of 125 possible points) for each individual test taker.

Text 1

عائلةي

اسمي حسن، وأنا مصري من مدينة دمياط في شمال مصر. لي عائلة كبيرة في الولايات المتحدة الأمريكية ومصر آنذاك.

لي خاص، تسكن مدينة هيوستن حاليًا يعمل زوجه في شرارة بترول، ولله خمسة أو ستة ، ولدان و3 بنات.

ولدها الكلب يعمل ضاحية في الجامع الأمريكي والصغيرة مختصة في الكيمياء. والبكر الأولى تد الأدب العربي في نيويورك والثانية اسمها ليلو ولد

تتعلم العربية لأن عددها طفل جذيب وتسكن في مدينة دالاس حيث يعمل زوجها والثالثة متخصصة في علوم الكمبيوتر. لي أيضا أبناء خال يسكنون في ولاية...
نيوجيرسي حيث يعملون في مطعم عربي.

الحياة جبران خليل جبران
جبران خليل جبران فيلسوف وشاعر وكاتب ورسام لبناني-أمريكي، ولد في 6 يناير 1883 في بلدة بشري شمال لبنان. مات في نيويورك 10 أب 1931، ويعرف أيضا باسم خليل جبران. ساها وهو صغير مع أشقائه وأمه إسماعيل أمركا عامة. حيث دفنت الفن وكتابته الأدبي. اشتهر في العالم الغربي بكتابه الشهير "النبي".

وُلد جبران لعائلة مسيحية. أمه اسماعيل كاميليا رحمة وكان عمها 30 سنة عند وُلاده. أب جبران، وكانت العائلة فقيرة جداً، ولذلك استطاع الدرس في المدرسة العامة، فدرس الدين واللغة العربية في البيت، وهكذا تعلم التاريخ والعلوم والآداب. بعد ذلك قررت والدته الهجرة إلى أمريكا، وكل من أمه وإخوته سافر إلى مدينة نيويورك في 25 يونيو 1895.

تاريخ جامعة عين شمس المصرية
جامعة عين شمس المصرية تعتبر أقدم ثالث جامعة مصرية، تأسست في يوليو 1950 تحت اسم "جامعة إبراهيم باشا".
بعد ثو 23 يوليو 1952، قرر الحكومة المصرية أن تكّد أسماء كل الجامعات المصرية لـ 23 يوليو 1952.

وهكذا فقد 21 فبراير 1954 تغت اسم الجامعات إلى "هليوبولس".

والتالي ذلك تغت في السنة إسمها الحا "عين شما"، الذي يعتبر "هليوبولس"، التي كان أقدم جامعاً في التا. جامعة "هليوبولس" تأ منذ حوالي 5000 سنة، وكـ عندما شهير واسعة، خص في علم الفلك و الهندسة والطب. ربما المثال الأحسن للمعرفة والمهارة في معلمو "أون" هو "أمحتوب" الكاهن الرئيسي والوزير والتصميم الذي في 2700 قبل الميلاد صمم أول التركيب الكبير للحجارة.

Text 4

السفارة الأمريكية تعلن عن برنامج فولبرايت لأساتذة اللغة الإنجليزية

تعلن السفارة الاميركية عن برنامج فولبرايت لتعليم لغة أجنبية للعام 2015. يقدم هذا البرنامج فرصا للأساتذة الشب وطلاب الدرا العليا لدر اللغة الإنجليزية في إحد الكليات والجامع الأمريكية، وكذ فرصة للع في الو ذاته كأس في مجال تعليم اللغة العربية.

يؤ ثبت العت أيضا المتطل بطلبات الالت بهذا البرنامج من الاختصاص التالية: ع الإنسان، الانص ، الخدمة الاجتماعية، علم النفس، التربوي،
التعاليم في المناطق الريفية، الآداب، اللغات، دراسات المنطقة
Teacher, الفنون، علوم الحاسب، تدريب المعلمين (Area Studies)
، الإدارة التربوية، تكنولوجيا التعليم، العلوم الإنسانية، العلوم
الاجتماعية، وعلم الاجتماع.

Text 5

قوات الأمن المصرية تقتتح جامعة الأزهر بالقاهرة

اتهمت قوات الأمن المصرية جامعة الأزهر في القاهرة حيث كان ينتحر طالب
مؤيدون للرئيس المعزول محمد مرسي. وأطلقت قوات الشرطة الغاز المسيل للدموع
من أجل تفريق المتظاهرين.

وقد قالت السلطات إن الهدف من اقتحام الجا هو
مواقع أعمال الشغف. وأشارت إلى "تمكّن الأجهزة الأمنية
มวล إحكام السيطرة على مبنى走得 لأحد الجامعات
وتآمر رئيس الجامعة ومسؤوليها وموظف المحتجزين ببدأ
". وكان الطلبة الجامعة يتظاهرون منذ أواخر احتجاجا علية
عزل الرئيس المنتخب محمد مرسي، في شهر يوليو/تموز الماضي على يد الجيش،
عقب احتجاجات شعبية واسعة. ويطلب المتظاهرونعودة ما يصفونه بالشرعية.
Appendix B: English Background Questionnaire

The following 5-page questionnaire was administered to Arabic learners during data collection. This questionnaire was composed in Google Forms, although it was printed and administered in paper format to participating Arabic learners.

Language Background
Thank you so much for participating!
* Required

1.

2. What is your student status? *
   Mark only one oval.
   ○ Freshman
   ○ Sophomore
   ○ Junior
   ○ Senior
   ○ MA student
   ○ PhD student
   ○ Other:

3. What Arabic course are you currently taking? 
   (if applicable)

   ...

4. Gender *
   Mark only one oval.
   ○ Female
   ○ Male
   ○ Other
5. What is your major? *


6. Country of birth *


7. Have you studied or otherwise learned other languages besides English and Arabic? If so, please list.


8. Have you ever been to an Arabic-speaking region for the purpose of studying Arabic? *
   Mark only one oval.
   ○ Yes   Skip to question 9.
   ○ No   Skip to question 13.

   Previous Arabic Study Abroad

   9. When did you study Arabic in an Arabic-speaking country?


10. Where?
11. **For how long?**  
*Mark only one oval.*
- One summer or less  
- One semester  
- One year  
- More than one year  
- Other: ____________________________

12. **What type of Arabic did you study?**  
*This is an open-ended question. Please provide your response.*
- i.e. MSA only, MSA and a particular dialect, etc.

**Arabic language ability**

13. **Arabic Proficiency Self Assessment**  
*How do you rate your Arabic ability according to the following scale? Mark only one oval per row.*

<table>
<thead>
<tr>
<th></th>
<th>1 (very low/beginner)</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5 (native/native-like)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reading</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Writing</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Listening</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Speaking</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OVERALL Arabic proficiency</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

14. **How do you rate your writing ability in ENGLISH?**  
*This is an open-ended question. Please provide your response.*
- i.e., when compared to your peers, how good of a writer do you feel you are in English  
*Mark only one oval.*

<table>
<thead>
<tr>
<th></th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>low/poor</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>superior/excellent</td>
</tr>
</tbody>
</table>
Arabic Study

15. How many years have you studied Arabic? *

16. Have you studied Arabic in school in the past at each of the levels listed below? If yes, for how long?
   Mark only one oval per row.

<table>
<thead>
<tr>
<th></th>
<th>No</th>
<th>1 year</th>
<th>2 years</th>
<th>more than 2 years</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elementary school</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Junior high (middle) school</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>High School</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>University/College</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

17. If you marked "other" above, where did you study?

Skip to question 18.

Use of ARABIC
You're almost done!

18. On average, how often did you communicate with native or fluent speakers of Arabic in Arabic in the last year? *
   Mark only one oval.
   ○ never
   ○ a few times a year
   ○ monthly
   ○ weekly
   ○ daily
19. **In the past year, how often did you do the following?** *

*Mark only one oval per row.*

<table>
<thead>
<tr>
<th>Activity</th>
<th>never</th>
<th>a few times a year</th>
<th>monthly</th>
<th>weekly</th>
<th>daily</th>
</tr>
</thead>
<tbody>
<tr>
<td>writing essays in Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>texting in Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>writing personal emails in Arabic (i.e. to friends)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>writing formal emails in Arabic (i.e. to a boss)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>posting to social media in Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reading social media in Arabic</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>watching Arabic TV or movies</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reading Arabic novels</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>reading Arabic news</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

20. **any other activities that you commonly did using Arabic during the last year?**

..................................................................................................................................................
..................................................................................................................................................
..................................................................................................................................................
..................................................................................................................................................
..................................................................................................................................................
Appendix C: Arabic Background Questionnaire

The following questionnaire was administered electronically to L1 Arabic writers who participated in this study.

1. الجنس
   □ أنثى
   □ ذكر

2. المستوى التعليمي
   ما هو أعلى مستوى تعليمي وصلت له؟
   □ البكالوريوس (تدرسه حاليا أو أكملته)
   □ الماجستير (تدرسه حاليا أو أكملته)
   □ الدكتوراه (تدرسه حاليا أو أكملته)

3. التخصص
   أكتب هنا:

4. أكملت دراستك الثانوية بأي لغة؟
   أكتب هنا:

5. أكملت دراستك الجامعية بأي لغة؟
   أكتب هنا:

6. أكملت دراستك العليا بأي لغة؟
   أكتب هنا:

7. لهجتك الأم (مثلًا: سعودية، مصرية، لبنانية، إلخ)
8. اللغات - غير العربية - التي درستها / تتحدثها؟
أكتب هنا:

9. كيف تقيم قدراتك الكتابية باللغة العربية؟
أو، كيف ترى نفسك ككاتب/كاتبة بالعربية بالمقارنة مع نظراك العرب؟
متميز/ممتاز
متوسط
ضعيف
أدنى/سيء

□ 1 □ 2 □ 3 □ 4 □ 5
10. كم مرة قمت بالأعمال التالية خلال السنة الماضية؟

<table>
<thead>
<tr>
<th></th>
<th>أبدًا</th>
<th>بعض المرات في العام</th>
<th>شهريًا</th>
<th>أسبوعيًا</th>
<th>يوميًا</th>
</tr>
</thead>
<tbody>
<tr>
<td>الكتابات الأكاديمية بالعربية</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>الرسائل النصية (SMS) بالعربية</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>الإيميلات الشخصية بالعربية (للأصدقاء مثلا)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>الإيميلات الرسمية بالعربية (للمدير في العمل مثلا)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>الكتابة في وسائل التواصل الاجتماعي بالعربية</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>القراءة في وسائل التواصل الاجتماعي بالعربية</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>مشاهدة الأفلام أو المسلسلات التلفزيونية العربية</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>قراءة الروايات العربية</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>قراءة الأخبار بالعربية</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>