2019-12-01

Verb Usage in Egyptian Movies, Serials, and Blogs: A Case for Register Variation

Michael G. White
Brigham Young University

Verb Usage in Egyptian Movies, Serials, and Blogs:

A Case for Register Variation

Michael G. White

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Deryle Lonsdale, Chair
Mark Davies
Chris Rogers

Department of Linguistics

Brigham Young University

ABSTRACT

Verb Usage in Egyptian Movies, Serials, and Blogs:
A Case for Register Variation

Michael G. White
Department of Linguistics, BYU
Master of Arts

This thesis contributes to the discussion of register variation within Egyptian Arabic by focusing on the usage of verbs in blogs and transcripts of movies and television. Register variation has been extensively researched for English as well as several other languages; yet, the lexical and grammatical features that distinguish registers of Egyptian Arabic have not been analyzed. Several challenges have prevented such an analysis, among them the perceived lack of an automatic annotator and the uncertainty of results. In order to overcome these challenges, two corpora were compiled: one containing texts from blogs and the other transcripts of movies and television shows. With each corpus representing a potential register of the dialect, the verbs in each corpus were lemmatized and semi-automatically annotated for either aspect or mood. The verbs were then counted according to lemma, aspect, and mood in order to determine the extent of variance between the two corpora. The effectiveness of the state-of-the-art automatic annotator was also evaluated by comparing the counts it provided to those produced from corrections of its output. This thesis found that verbs are in fact used differently in the two corpora suggesting register variation and identified potential verbal features characteristic of each register. It also found that the automatic tagger produced counts that lead to the same conclusions as the corrected annotation.

Keywords: register variation, corpus, Egyptian Arabic

ACKNOWLEDGEMENTS

I would like to thank

Brigham Young University for their help in funding CALM

My committee members for their encouragement and support

Dr. Lonsdale for his tireless work in reading and revising this thesis

My parents for encouraging me to live in places many deem dangerous

Sara for making my decision to study linguistics life changing

Table of Contents

List of Tables

Chapter 1: Introduction

The internet has made written Egyptian Arabic much more accessible than in the past. Books, newspapers, academic journals, and government documents are composed in Standard Arabic which differs morphologically, lexically, and syntactically from Egyptian Arabic, making these texts a poor choice for representation of Egyptian Arabic. However, as access to the internet spread, blogs and social media sites became filled with texts written using the syntax, vocabulary, and morphology of Egyptian Arabic.

This provides linguists with a new opportunity to collect large samples of the dialect in written form from a variety of sources on numerous topics. These written samples can be separated into registers and then compared with registers whose primary means of delivery is speech in order to determine how these two modes of language production differ. However, such comparisons require more than just corpora containing both written and spoken Egyptian Arabic. Modern studies of register variation rely on lexically and syntactically annotated corpora. Unfortunately an annotated corpus representative of registers of written and spoken Egyptian Arabic has not been available, preventing studies of Egyptian Arabic register variation.

This thesis seeks to jump-start discussion on the potential variation that exists between registers of spoken and written Egyptian Arabic by comparing the frequency of verbs in a corpus of movie/television transcripts and a corpus of blog posts. Although the examination of a single part of speech cannot capture the true extent of the variance of two corpora, it provides a platform from which to launch an in-depth analysis by answering the preliminary questions concerning register analysis for Egyptian Arabic.

1. Is there enough evidence of register variation between movie/television transcripts and blog posts to warrant a more thorough investigation?

2. Is there an automatic annotator that is accurate enough to aid in a study of register?

3. Are there ways that the automatic annotators can be improved?

In order to answer these questions a new corpus was constructed, and the verbs therein were annotated. This provided the ability to count and compare the frequencies of verbs in a manner similar to previous studies of register variation. The frequency counts of the annotations performed manually were also compared to the counts performed by a computer in order to determine whether the accuracy of the automatic tagger is sufficient for studies of register variation. This comparison also provided insight into how annotation of Egyptian Arabic could be improved.

This thesis will proceed as follows: Chapter 2 provides a closer look into previous studies of register variation and the challenges of conducting such a study for Egyptian Arabic. The differences between the Egyptian dialect and Standard Arabic are also discussed. Chapter 3 discusses the methods for collecting and annotating the corpus. Chapter 4 presents how the transcript and blog corpora differ in regard to the frequency of verbs and compares the counts of the corrected annotations to the non-corrected. Analyses of the data is also conducted in this chapter to answer whether there is enough variance in the two corpora to warrant a more thorough investigation. Chapter 5 discusses annotation and examines the results of an automatic annotator to determine its efficacy and how automatic annotation can be improved. Chapter 6 concludes this thesis by discussing the significance of the results, the limitations of this thesis, and potential future studies. Appendices discuss the challenges to annotating Egyptian Arabic. Appendix A presents the common errors of the automatic annotator and how they were limited by a supplemental computer program. Appendix B explains the rationale behind the lemma orthography used for this thesis.

Chapter 2: Background

Introduction

In this chapter a brief introduction to corpora and register variation will be presented along with some of the challenges in applying such studies to Egyptian Arabic. As this dialect[1] may be unfamiliar to the reader, an introduction to the differences between Egyptian Arabic and Standard Arabic will also be given. After this discussion, it will become clear why the corpora used in this thesis must contain texts comprised of only Egyptian Arabic. Corpora meeting this qualification will be introduced in an effort to justify the selection of the corpus used. This chapter will also introduce relevant literature concerning corpus compilation, annotation, and analysis especially as it pertains to CALM, a new corpus of Egyptian Arabic that will be introduced in Chapter 3.

Corpus-Based Register Studies

This thesis uses a corpus-based approach for determining whether a difference exists between two corpora representing potential registers of written and spoken Egyptian Arabic. A corpus, as defined by some linguists, is a collection of texts or transcriptions gathered for the purpose of conducting an empirical study of language (Hunston 2002; Kübler & Zinsmeister 2015; Teubert 2005). Their use in linguistic studies is not limited to a few branches of the discipline; rather, corpora have become an acceptable resource in nearly every field of linguistics (Teubert 2005).

Their widespread use has caused corpora to assume many different forms. Those seeking to study how language has changed throughout time might use a historical corpus like COHA (Corpus of Historical American English) which contains 400 million words from texts from

[1] This thesis will treat Egyptian Arabic as a dialect of Arabic.

every decade from the early 1800s until the 2000s[2]. To understand how British English is used, researchers may use the 100 million-word BNC (British National Corpus) (Leech 1992). However, depending on its purpose, corpora can also be small. The Hanker Corpus, which was compiled to study the writing of economic students, contains 516,000 words (Mäkinen & Hiltunen 2016).

Although fewer in number than English corpora, many types of corpora are available for Arabic. There are Arabic general language (ArabiCorpus), transcribed speech (CALLHOME Egyptian Arabic), specialized (The Quranic Arabic Corpus), parallel (OPUS), and learner (Arabic Learner Corpus) corpora. With these corpora and others, our understanding of Arabic and how it is used by native speakers has increased (Bentley 2015; Buckwalter & Parkinson 2011; Alasmari, Atwell & Watson 2017; Dickins 2017; Henen 2018; Ismail 2015). However, one area that has been largely overlooked in corpus studies of Arabic is discourse analysis. This caused Ryding to claim that "The field of Arabic as a foreign language urgently needs to attend to the empirical description and analysis of authentic Arabic discourse" (Ryding 2006). This is not to say that Arabic discourse analysis has been completely ignored; however, most research surrounds Arabic diglossia, specifically, the mix of Standard Arabic with other Arabic dialects in speech (Doss 2011). This focus has left other areas of Arabic discourse analysis, like register variation, with limited research.

Over the last few decades, numerous studies for several languages have targeted the identification of features that distinguish one register of the language from another. The results from these studies have sought to help: universities better prepare incoming students to understand the new registers unique to college life (Biber et al. 2006); researchers understand

[2] Davies, Mark. (2010-) *The Corpus of Historical American English (COHA): 400 million words, 1810-2009*. Available online at https://www.english-corpora.org/coha/.

how registers might differ in foreign languages, like Spanish, Korean, and Somali (Biber et al. 2006; Biber & Conrad 2001); newcomers to an occupational field learn the type of speech that is necessary for approval (Ferguson 1983); and corpus linguists determine what registers exist in a language (Gries 2006).

Similar analyses have also been conducted for Arabic. Fakhri (2009) used the framework of Contrastive Rhetoric to investigate the variation between Academic Arabic in the disciplines of the humanities and the law. Johnstone (1990) examined three features of spoken Arabic—repetition, parataxis, and formulaicity—and their use in Arabic expository prose. However, the variation that exists between spoken and written Egyptian Arabic has yet to be fully explored.

Although traditionally not used in published works, Egyptian Arabic is appearing more commonly in the written mode, probably due to the internet. Since written language typically is not simply transcribed speech, we must assume that spoken and written Egyptian differ in the use of lexical and syntactic features. The discovery of such features can help correctly categorize other registers of the dialect and improve understanding of how the dialect is used.

However, there are several challenges that stand in the way of such a study. The first is the creation of a corpus representing special usages in spoken Egyptian. Speech corpora are difficult to create—causing them to be small which can diminish the reliability of the results when the target linguistic feature is not high frequency or the corpus does not represent a specialized domain (Egbert 2019). A second challenge is the definition of written Egyptian Arabic. Since Standard Arabic is used in Egyptian newspapers, academic articles, literature, and government documents should it be considered written Egyptian Arabic?

A third issue is the reliability of the current automatic annotators for Egyptian Arabic. Some methods for performing register analyses rely heavily on lexical and syntactic annotations

(Biber & Conrad 2001). Therefore, automatic annotators are needed for conducting such a study on a large corpus without the study becoming burdensome in terms of money and time. However, Egyptian Arabic annotators are largely untested on corpora that they were not originally trained on. Before we examine the difficulties of determining register in Egyptian Arabic, we will first define register and examine how others have studied it in the past.

Register

Register is defined as the different situations or circumstances in which speech acts occur (Johnstone 2008). Situations and circumstances that cause speakers to change their lexical and grammatical choices are said to belong to different registers. For example, a healthcare professional in speaking with a patient may use words and syntactic constructions that are less frequently used during a casual conversation. Because a change in language use occurs, it can be said that patient-provider conversations belong to a different register than casual conversations (Staples 2016).

In corpus linguistic literature, the term register may also include categories normally attributed to genre. Traditionally, genre is defined as culturally determined categories of formal language production (Johnstone 2008). Genres can include literature, poetry, correspondences, and speeches; however, Biber and Conrad (2001), have shown that romance fiction, science fiction, religious texts, academic prose, and biographies should not only be considered genres of the written language, but also unique registers since the frequency of certain lexical and grammatical features change from genre to genre. However, not all genres are registers. Parody can be an example of this since when done well, it mimics the target register in lexical and grammatical features (Bex 1996).

Collection of Registers

Now that we have defined register, we assert that lexical and grammatical features play a role in differentiating registers. In order to find the features that are prominent in each register, a sample of texts representing the register needs to be collected. But how are texts brought together to form the description of a register if we do not yet know for certain which registers exist?

Biber (1993) gives eight parameters to use in classifying registers. These include the primary channel of delivery, format, setting, addressee, addressor, factuality, purpose, and topic. For him, registers are not broad categories—like speech or literature—made up of texts from random sources. Separating texts according to these parameters will help group texts into their appropriate registers.

Once a set of variables has been identified for a register, the next task becomes collecting enough texts so that the range of that register is adequately represented. Unfortunately, there is no single test to determine whether the texts gathered represent the target register. Even large corpora are not necessarily representative if the texts come from a small number of sources. Sinclair (2004) provides a list of six steps that if followed lead to the creation of a representative corpus; however, nothing is offered in the way of proving the efficacy of these steps. Biber (1993a) offers a series of tests intended to give the corpus creator greater confidence in the representativeness of the corpus; however, these tests can only be applied to large annotated corpora (Leech 2007).

Atkins, Clear, and Ostler (1992) argue that representativeness is nearly impossible to scientifically prove because of the sheer size of each register, allowing for someone to prove that

some feature of the register has been underrepresented. However, this should not lead to a rejection of corpus linguistics or attempts to build a representative corpus; rather, corpus users need to approach corpora knowing their strengths and weaknesses. For Atkins, Clear, and Ostler, it is from user feedback that a clear picture of a corpus's representativeness is formed. Additionally, corpora should be used because even unrepresentative corpora can provide insights into a language. This is not to say that the authors support a move away from representativeness. Like Sinclair and Biber, they propose steps for narrowing down the target register to enhance the probability of creating a representative corpus. However, it is important to them that linguists recognize the weaknesses of a corpus but use it, rather than waiting for the ideal corpus.

That being said, a representative corpus is important for analyses of register since if a register is not representative the variance could be due to factors not associated with register. Macaulay (1990) argues that studies of register can be influenced by the education level of the speaker or author and whether the registers contain texts concerning different subjects. However, Biber and Conrad (2001) found that as long as registers are balanced such factors do not influence the results of the analysis.

Dimensions and Features of Registers

Once texts are collected in a principled way with the aim of representativeness, then lexical, syntactic, and grammatical features in each register are counted and compared across registers. The number of features that can be examined are plentiful, and in previous register analyses for English, the frequency of relative clauses, adverbial clauses, noun clauses, co-ordinate clauses, conjoining phrases, infinitive phrases, passives, existential "there", semantic classes, "that" deletion, pronouns, *wh*-clauses, verb tenses, adjective classes, and vocabulary

distributions were used to demonstrate variation (Macaulay 1990; Biber & Conrad 2001; McCarthy & Handford 2004; Biber et al. 2006; Biber 2006; Staples 2016; Mäkinen & Hiltunen 2016). This list is far from exhaustive but should give a sense of the types of phenomena used in a register analysis.

However, not all studies of register variation compare counts from a large number of features. In an analysis of spoken business English and everyday conversational English, McCarthy & Handford (2004) compared the frequency of word forms in each corpus, therefore completely focusing on lexical features of the texts. They compared individual words and word clusters to determine the types of expressions that are more common in each register. The concordance lines containing those words and clusters were also studied to determine the reason behind their increased use.

Another study which used a limited number of features was done by Hiltunen (2016), who analyzed the use of passives within English academic articles in order to further divide this register. She found that certain disciplines do indeed use more passives and that journal articles use them more than papers written by students.

Regardless of the number of features chosen, there is a need to explain what might be causing the variance in feature frequency beyond a difference in register. Determining that register X contains more passives than register Y helps one know how the registers differ, but it does not provide an answer to the question of what is causing more passives to be used in register X. Multi-Dimensional analyses, which were made prominent by Biber (1988), seek to determine the cause of the differences in feature counts among registers. Just as texts can be divided into registers, registers are made up of combinations of dimensions. The number of dimensions within a language is not set, but eight basic ones were identified by Biber (1993b) for

his corpus of English: information vs. involved, narrative vs. non-narrative, elaborate vs. situated reference, overt expression of persuasion, and abstract vs. non-abstract style. Each dimension contains grammatical, syntactic, and lexical features that occur with higher frequency relative to other dimensions. For example, the narrative dimension in English is characterized by past tense verbs, third person pronouns, public verbs, synthetic negation, and present participle clauses (Biber 1993b). A register of the language that uses these structures with higher frequency is then said to belong to the narrative dimension. A register may have more than one dimension and is, therefore, differentiated by the dimensions attributed to it.

The dimensions themselves can be grouped into two larger dimensions: the oral and the literate dimensions. Registers that contain dimensions that typically occur in spoken registers (i.e. involved and non-narrative) are said to fall within the oral dimension of the language and those with dimensions characteristic of the written language are classified in the literate dimension.

Oral Dimension

In daily life, the most common registers in the oral dimension are made up of spontaneous speech. However, in this thesis, our corpus chosen from the oral dimension is scripted speech from television and movies. The difference between spontaneous speech and the language used in movies cannot be denied. Scripted speech is first written, a process that affords the author time to craft each utterance and edit it until it achieves the desired effect. Utterances made spontaneously do not often reflect this luxury. However, this most common type of spoken language is currently difficult to obtain. The recording and transcribing of conversations from a wide range of individuals is time-consuming and costly.

Despite these challenges, some argue that scripted speech is not an alternative to spontaneous speech. Sinclair (2004, pg. 80) states that scripted speech has "a very limited value in a general corpus, because it is 'considered' language, written to simulate speech in artificial settings." He continues, stating that this kind of speech is "not likely to be representative of the general usage of conversation." Not only are the situations and settings artificial, but language may be inappropriate for the demographics of the characters. Movies and television shows can be the product of a single person creating the dialogue for multiple characters. Therefore, movies could be looked at as long monologues rather than dialogues made up of several different speakers.

These assertions inspired Forchini (2012) to study whether scripted speech in movies differs widely from spontaneous speech. In order to do so, she created a 204,636-word corpus made up of 11 movies and compared its language to that which is contained in the Longman Spoken American Corpus. This study was conducted using the multi-dimensional analysis advocated by Biber. Of the five dimensions, movie language only differed from spontaneous speech in one: Forchini found that the language in movies is more abstract (uses conjuncts, agentless passive verbs, by-passives, passive postnominal modifiers, and inter alia) than spontaneous conversations. However, the fact that the two resemble each other in four of the five dimensions led Forchini to conclude that the language contained in movies does not differ significantly from spontaneous conversations.

This idea is further supported by Brysbaert and New (2009) who claim that the frequency counts derived from subtitles more closely correspond to response times from a lexical decision task, which asks respondents to determine whether a stimulus is a word or nonword. The more frequent a word is, the quicker the respondent's ability to process the word and make a judgment

should be. The frequency list generated from movie subtitles was a better indication of how fast participants would respond than the lists created from corpora that include blogs and books. Therefore, language from movies truly has something different to offer. Not only was movie language found to be very similar to spontaneous speech, but a frequency list produced from it seems to accurately capture frequency data better than other genres.

In another study, Taylor (2004) found significant differences between the script and what the actors actually said. This led him to conclude that there is some degree of spontaneity infused into movies. Although a similar study has not been done for Egyptian films, a quick comparison of the script for the film حسن ومرقص *ħʌsʌn wi murʔosˁ* [3] (Maati 2008) reveals similar trends. As early as the second scene in the movie, the actors begin to stray from the script. The scene is given in Table 2.1. The underlined words in the column marked 'Transcription' are those words that seem to have been added by the actors. The words underlined in the 'Script' column were left out of the movie.

Table 2.1: Differences between a script and transcription as evidenced in *ħʌsʌn wi murʔosˁ*

Script	Transcription
مرقص: انت فاكر انك لما تتجوزها ح تقعد تبص في عينيها طول عمرك؟ الحاجات دي بتدوب مع الوقت وبالعشرة عينيكوا ح تطبع على بعض	مرقص: انت فاكر لما ح تتجوزها هتفضل باصص في عينيها طول عمرك؟ الحاجات دي بتدوب مع الايام وبالعشرة عينيكوا ح تطبع على بعض
جرجس: بصراحة يا بابا اصلها كمان هبلة شوية	جرجس: دي هبلة يا بابا
مرقص: يا ابني انت ما شفتش امك يوم ما اتجوزتها .. الجواز بيعقل!	مرقص: انت ما شفتش امك يوم ما اتجوزتها. الجواز بيعقل!
جرجس: مناخيرها طويلة كمان يا بابا	جرجس: مناخيرها مناخيرها طويلة
مرقص: وانت ايه اللي يخليك تحط مناخيرك في مناخيرها يا جرجس	مرقص: أنا قلت هتحط مناخيرك على مناخيرها
جرجس: مش هتجوزها يا بابا	جرجس: الله مش هتجوزها يا بابا
مرقص: وايه اللي دخل الجواز في المناخير .. انت ما شفتش مناخير امك يوم ما اتجوزتها .. زلومة فيل انما بالعشرة والايام	مرقص: مال الجواز ومال المناخير. انت ما شفتش مناخير امك يوم ما اتجوزتها...زلومة فيل. بس مع الايام والمحبة
جرجس: وهي المناخير بتصغر بعد الجواز يا بابا	جرجس: وهي يعني المناخير هتصغر بعد الجواز

[3] When necessary in this paper, Arabic text is followed by an IPA transcription or by an English gloss.

In nearly every utterance, the actors have strayed from what was scripted. The changes do not alter the overall effect of the utterance, and therefore could represent the way in which the actors themselves would speak rather than the author of the script. If this is the case, it would contradict the idea that entire movies represent the speech of the author. That being said, the actors are still using words that they might not otherwise say, and as pointed out by Sinclair, in situations that are not real. However, these changes do suggest that movie transcriptions could have elements of natural speech that are not present in the written production of the language.

This is supported by Biber and Conrad (2001) who found that colloquial writing often contains several elements that are characteristic of speech in English, Korean, and Somali. The features that characterize TV dramas in their study also characterized conversations but were not present in more formal types of writing. This suggests that some registers of writing, including scripted television shows, can closely resemble spontaneous speech.

Literate Dimension

Like the spoken dialect, written Egyptian Arabic will also be represented in an unconventional way through the use of blogs. The literate dimension encompasses all registers which contain written language. For many languages, this dimension can be made up of literature, articles published in newspapers, academic articles, encyclopedia entries, personal correspondences, and official documents (Biber & Conrad 2001; Biber et al. 2006). However, Egyptian Arabic lacks many of these sub-registers, as it was largely confined to personal correspondences until relatively recently. With the availability of the internet came an increase in the registers of written Egyptian Arabic. This is reflected in many recent books written in Egyptian Arabic

which are simply blogs published as books. Since blogs constitute a large portion of that which is written in Egyptian Arabic, they will be used to represent a written register.

Using internet blogs as a single register raises potential issues for this thesis. As found by Biber, Davies and Egbert (2015), English texts that make up the internet can be categorized into several registers; however, this thesis groups all blogs into one register regardless of content. This is not because I believe that blogs belong to a single register. Here, the blogs are not divided because of the scope of this thesis, which did not allow for the classification of all the texts that make up the blog corpus described later. Also, this thesis is not a full study of register variation. The variations found between the transcript corpus and blog corpus are only intended to suggest that variation does exist. However, a true study of register variation should look further into which registers within the blog corpus are most responsible for these variations, if not all of them.

Egyptian Versus MSA Morphological Differences

To better understand why the register representing the literate dimension should not contain texts written in Modern Standard Arabic, I will discuss the differences between it and Egyptian Arabic. This section will also be important for later discussions on the challenges that automatic annotators face when analyzing parts of speech, particularly verbs. To keep this discussion succinct, I will use 'Arabic' when no distinction between Egyptian Arabic and Modern Standard Arabic is necessary.

The verbs of Egyptian Arabic, like other Arabic parts of speech, are lexical templatic roots which are made up of two to five phonemes; most Arabic roots have only three phonemes. Although the root itself does not have any specific meaning "it communicates the idea of a real-

world reference of general field of denotation" (Ryding 2005). Meaning becomes more concrete as the root is placed in one of fifteen different forms. These forms add phonemes to the verb root, but do not necessarily add a specific meaning (Bjøru 2018). These phonemes can be consonants, long vowels, or short vowels which are represented orthographically as diacritics. Three forms also necessitate the gemination of root consonants. It is important to note that the short vowels and gemination diacritics do not have to be written and as such, are commonly not written.

A single verb root can inflect in multiple forms, but no roots occur in all fifteen (Abdel-Massih, Abdel-Malek & Badawi 2009). Forms XI – XV are extremely rare in MSA and even rarer in Egyptian Arabic (Ryding 2005). To demonstrate how this works, Table 2.2 displays the placement of the root ف [f], ع [ʕ], and ل [l] into the ten most common forms.

Table 2.2: The ten most common verb forms in Arabic

I	فعل
II	فعّل
III	فاعل
IV	أفعل
V	تفعّل
VI	تفاعل
VII	انفعل
VIII	افتعل
IX	افعلّ
X	استفعل

The diacritics and short vowels were intentionally not included in Table 2.2 as the short vowels in MSA verbs can differ from Egyptian verbs (Hassan 2000). In fact, the only diacritic mark that will be of importance in the written representation of Egyptian Arabic in this thesis is the *shadda* or ـّـ which geminates the phoneme it is placed over.

One feature of Arabic verbs is that they cannot be represented separate from PERSON or ASPECT. The morpheme that expresses PERSON also expresses either the PERFECT or IMPERFECT aspect. The terms PERFECT and IMPERFECT will be used instead of PAST and PRESENT because of their traditional use in Arabic linguistics (Aboul-Fetouh 1969). The verbs in Table 2.2 are inflected for 3SG.MASC.PERFECT which is an unmarked form of the verb, and therefore is commonly used to demonstrate the different verb forms (Abdel-Massih, Abdel-Malek & Badawi 2009). There are thirteen different categories for PERSON in Standard Arabic, but only eight in Egyptian Arabic. These are given in Table 2.3.

Table 2.3: The different categories of PERSON in both MSA and Egyptian Arabic

MSA	**EGYPTIAN**
1SG	1SG
1PL	1PL
2SG.MASC	2SG.MASC
2SG.FEM	2SG.FEM
2DUAL	
2PL	2PL
2PL.FEM	
3SG.MASC	3SG.MASC
3SG.FEM	3SG.FEM
3DUAL.MASC	
3DUAL.FEM	
3PL	3PL
3PL.FEM	

The morphemes expressing PERSON/PERFECT are suffixed onto the verb stem, while those expressing PERSON/IMPERFECT are either prefixed or circumfixed onto the stem depending on PERSON. In both MSA and Egyptian Arabic these morphemes are nearly identical. Not including short vowels, only 2SG.FEM and 2SG.PL differ from their MSA counterparts in the PERFECT. In the IMPERFECT, Egyptian verbs employ the morphemes used by MSA to express the SUBJUNCTIVE.

All other morphemes expressing the other aspects, moods, or polarities are affixed onto the PERSON/IMPERFECT or PERSON/PERFECT stems. The notable exceptions to this are the

morphemes that express the IMPERATIVE mood. These morphemes expressing both PERSON and IMPERATIVE are affixed directly onto the verb stem.

As the morphemes move further from the stem, the gap between Egyptian and MSA begins to widen. Although both varieties of Arabic have a morpheme to express the FUTURE, it is not identical in form. Egyptian uses the prefix هـ [h] or حـ [ħ] whereas in Modern Standard Arabic the prefix سـ [s] or the free morpheme سوف *saofʌ* is used. The Egyptian prefixes can also be separated from the verb as an unbound morpheme[4].

Egyptian Arabic speakers not only use a different FUTURE morpheme but have added a morpheme that is used to mark either habitual or progressive action (El-Tonsi 1982). In Egyptian, this aspect is expressed by the morpheme بـ [b] which is prefixed onto the PERSON/IMPERFECT stem. In MSA, these aspects are unmarked and are understood through context (Hassan 2000).

The morphemes expressing polarity are also different in the two Arabic varieties. For MSA, لم læm, لن *lan,* ما *mæ,* or لا *lʌ* are free morphemes expressing negative polarity placed in front of verbs. In Egyptian Arabic, two morphemes are circumfixed onto the verbs to express polarity. These morphemes represent the outer boundary of Egyptian verbs since no other morpheme can attach to them. The affix of the circumfix can be represented as either ما *mæ:* or مـ [m] and the suffix as ش [ʃ] . Therefore, كتبت *kɛtɛbt* 1SG.PERFECT."to write" becomes ماكتبتش *mækɛtɛbtɪʃ* when negated. Although the ش [ʃ] is always bound, the pre-word morpheme does not have to be attached to the verb.

[4] This is reflected in the annotated portion of the blog corpus from CALM in which حـ *ħ* and هـ *h* appeared as a bound future tense morpheme in 125 and 1,645 instances respectively and as unbound morphemes 19 times and 42 times. It must be added that these 19 and 24 instances come from a total of five speakers.

Egyptian Versus MSA Lexical Differences

As in any two varieties of a language, there are lexical variations between MSA and Egyptian. To demonstrate the extent of the differences, Table 2.4 contains the twenty most frequent verbs from the movie portion of the corpus collected for this thesis (named CALM) and their MSA equivalents. Each verb is inflected for 3SG.FEMININE.PAST which in Egyptian is created by suffixing ـِت *ɪt* onto the 3SG.MASC.PAST form of the verb and is created in MSA by similarly suffixing ـَت *at* onto the 3SG.MASC.PAST.

Table 2.4: Examples of the verbal differences among three dialects of Arabic

Rank	Egyptian	MSA	English
1	قالت ʔæ:lɪt	قالت qa:lɛt	she said
2	كانت kæ:nɪt	كانت kæ:nɛt	she was
3	عملت ʕamlɪt	فعلت fɛʕɛlɛt	she did
4	بقت bʌʔɪt	أصبحت ʔʌsˤbaħɛt	she became
5	عرفت ʕɪrfɪt	عرفت ʕarafat	she came to know
6	شافت ʃæ:fɪt	شاهدت ʃæ:hɛdɛt	she saw
7	خدت xʌdɪt	أخذت ʔaxaðat	she took
8	جات gæ:t	أتت ʔatat	she came
9	راحت ra:ħɪt	ذهبت ðɛhɛbɛt	she went
10	خلّت xɛllɪt	جعلت dʒaʕalat	she made
11	حبّت ħʌbbɪt	أحبّت ʔaħabbat	she fell in love
12	جابت gæ:bɪt	أحضرت ʔaħdˤarat	she brought
13	قعدت ʔaʕdɪt	جلست gɛlɛsɛt	she sat
14	سابت sæ:bɪt	تركت tɛrɛkɛt	she left s.th.
15	حصلت	حدثت	she happened

	ħasʔalıt	ħadaθat	
16	طلعت tʔılʕıt	طلعت tʔalaʕat	she ascended
17	مشيت mıʃiyıt	مشت maʃat	she walked
18	لقت laʔıt	وجدت wadʒadat	she found
19	كلّمت kɛllımıt	كلّمت kɛllımat	she talked (transitive)
20	اتكلمت ıtkɛllımıt	تكلمت takallamat	she talked (intransitive)

As illustrated by Table 2.4, the variations between the verbs in these two dialects can be vastly different, as is the case with 3, 4, 6, 8-10, 12-15, and 18. In a study of register variation this can affect the results. A corpus mixed with Egyptian Arabic and MSA will be more lexically diverse than a corpus containing only one of the varieties. As lexical diversity is a measure that we look at in this thesis to determine the extent of variation, results will be more accurate if Egyptian Arabic speech is compared to written Egyptian Arabic.

Egyptian Versus MSA Syntactic Differences

In addition to the morphological and lexical differences, verbs in MSA and Egyptian Arabic are also used differently. Egyptian Arabic allows for verbal sequences which may include up to six verbs (Abdel-Massih, Abdel-Malek & Badawi 2009). However, in MSA, many of these instances require the use of either a verbal noun or complementizer. This causes verbal nouns to be much more common in MSA than in Egyptian. Therefore, a mixed corpus of the two varieties could produce inaccurate frequencies for the total number of inflected verbs used. For example, if verbs are more infrequent in a corpus of written Arabic which includes both varieties when compared to a corpus of spoken Egyptian Arabic, not much is learned since the decrease in the written corpus could be due to the MSA.

The number of verbs used in each variety could also be affected by the difference in use of the ACTIVE PARTICIPLE. In Egyptian Arabic, the ACTIVE PARTICIPLE of some verbs is used to express the PRESENT CONTINUOUS and for others the PRESENT PERFECT (Abdel-Massih, Abdel-Malek & Badawi 2009). However, for MSA the ACTIVE PARTICIPLE is used more adjectivally to describe the noun responsible for doing the action (Ryding 2005). In both varieties, active participles replace verbs, but because of their different roles, one variety could use them more frequently than the other causing a decrease in the occurrence of verbs.

Egyptian Arabic Transcript and Blog Corpora

These differences suggest that results of a register study would be easier to interpret if the corpus contained only Egyptian Arabic. Several Egyptian Arabic film and television transcript corpora have been used in recent studies. Hussein (2016) used a corpus of Egyptian movie transcripts to study the pragmatic and syntactic functions of the Egyptian word كده *kɪdʌ*. The corpus contains 231,542 words from seventeen different films. Production dates for these movies range from 1958-2008 with the majority of words in the corpus coming from movies made before 1990. Although the use of verbs in the spoken language likely did not change from 1990 until the present, all of the written materials that will be used in this thesis were produced in the twenty-first century. Therefore, to keep the language representing the oral and literate dimensions as similar as possible, it would be preferable to find a corpus of transcripts in which the majority of movies and television shows were produced after the turn of the current century.

Such a corpus was used by Sayed (2018) to study the use of the discourse marker معلش *maʕleʃ* "oh well, I'm sorry." This corpus contains transcripts of 76 episodes from the 2017-2018 Egyptian television serial سابع جار *sæ:biʕ ga:r* (*Furthest Neighbor*). One potential weakness to

using this corpus in a register study is that all of the transcripts come from one television show. Even if actors largely rely on their own words, most of the content is produced by a handful of speakers which makes it hard to argue that the corpus is representative of Egyptian television and movies.

The issue of representativeness is also important to choosing a suitable blog corpus. Two general Egyptian Arabic blog corpora are the Arabic Multi-Dialect Text Corpus, (Almeman & Lee 2013) which contains thirteen million words and Yet Another Dialectal Corpus (YADC) (Al-Sabbagh & Girju 2012a) which contains six million. Both were created by performing web searches using dialect-specific words and then scraping the text from the webpages returned by the search engine. The Arabic Multi-Dialect Text Corpus used 139 different words determined to be unique to Egyptian Arabic as the search terms or seeds. The frequency of these words does not seem to have played a role in their choice.

According to Biber, Egbert, and Davies (2015) the frequency of the seeds is important to creating a representative blog corpus. They used the most common trigrams in English in order to prevent any bias in the types of pages returned by Google. Sharoff (2006) uses a similar method to create a general language blog corpus; however, instead of using the most frequent trigrams, 500 words from the target language's frequency list are chosen as possible seeds. Between 5,000-8,000 queries are made using four words taken from the list of 500. Only the top ten URLs are retrieved, and the text from those websites is extracted.

Although no such frequency lists exist for Egyptian Arabic, no documented effort was made by the creators of the Arabic Multi-Dialect Text Corpus to choose frequent words or phrases. The creator of YADC, on the other hand, took measures to create a more representative corpus of the texts available online. The queries contained multiple Egyptian-exclusive function

words, or words which are used more for their grammatical function than their lexical meanings. In English, examples of function words include ‘a’, ‘the’, ‘that’, and ‘of.’ Although Sharoff argues that the use of function words creates greater noise in the results, they do appear in the trigrams used by Biber, Davies, and Egbert. One downside to using function words for the creation of an Egyptian corpus is that many of them are found in several dialects. Although a full list of the function words used for the creation of YADC is not given, three are provided in an example. Two of the three words are عشان *ʕæʃæ:n* ‘because’ and مش *mɪʃ* ‘not’ both of which are found in other Arabic dialects (Qafisheh 1992; Tamis & Persson 2013).

The creators of YADC analyzed the corpus to remove instances of MSA; however, because of the size of the corpus, they were unable to separate the Egyptian texts from those produced in other dialects. Additionally, separating the dialects from each other was not a priority for the corpus designers since it was created to be a “dialect corpus” with the first phase being focused on Egyptian Arabic, but not exclusive to it. Therefore, because not all function words were truly dialect-specific, it is likely that texts written in other dialects have been included in YADC. For the purposes of this thesis, a corpus that contains only Egyptian Arabic is preferable.

Summary

One area lacking in research is that of register variation within Egyptian Arabic, especially of the features that distinguish the spoken form from the written. Such a study could be undertaken with the use of two corpora said to represent different registers within the oral and literate dimensions of the language. To represent the oral dimension, film and television transcripts could be used because of the features that they share with spontaneous speech. The literate

dimension could be represented by the language contained in blogs, since registers traditionally used to represent this dimension are written using MSA. Since MSA differs from Egyptian Arabic orthographically, lexically, morphologically, and syntactically, comparing an MSA corpus and an Egyptian corpus would not increase our understanding of how Egyptians write the dialect. Therefore, the two corpora must be of Egyptian Arabic.

In the next chapter, the corpus used in this thesis will be introduced along with the features that will be compared across the two corpora. The type and method for annotation will also be discussed.

Chapter 3: Methodology

Introduction

Any study of register variation needs a corpus, target linguistic features, and a way to determine the extent of variance. In this chapter, I will address all three of these elements. The CALM corpus will be introduced along with the justification for using only verbs as the target linguistic feature. The different verbal categories and features that will be counted and examined will also be given, in addition to the statistical tests used to confirm the significance of the variance.

The Corpus

Chapter 2 mentioned some of the available corpora that could be used for this thesis. However, each had issues concerning representativeness. Therefore, a new corpus was developed in an attempt to more accurately represent both movie/television transcripts and written material found online. This thesis introduces CALM[5] (*Corpus Al-logha Al-Musriya*, Corpus of Egyptian) a two-million-word corpus of Egyptian Arabic that contains transcripts from 65 movies (655,858 words), 104 episodes from 86 different scripted television programs (396,734 words), and blogs (1,092,442 words). For the purposes of this thesis, CALM was divided into two sub-corpora: a sub-corpus of blogs and a sub-corpus of movie and television transcripts.

Transcript Corpus

The transcripts of CALM make up the largest known collection of transcribed Egyptian movies and television programs. The transcripts were produced specifically for their inclusion

[5] CALM is available for free download at http://linguistics.byu.edu/thesisdata/CALMcorpusDownload.html

into CALM to address the scarcity of available transcripts. This also provided greater freedom to include a wide variety of Egyptian films.

All texts in the transcript corpus come from movies and television programs produced in Egypt and written for Egyptian audiences. Although it would have been quicker and cheaper to build a corpus from subtitles, foreign movies are subtitled using Standard Arabic, rendering them useless to this thesis. Subtitles and movies dubbed into Egyptian Arabic were also avoided because of the potential to employ unnatural language in an attempt to reflect structures of the original language of the movie.

Although no restrictions were placed on the release dates of the movies or TV programs, most are from the year 2000 and later. There are thirteen movies from the twentieth century: three from the 60s, one from the 70s, five from the 80s, and four from the 90s. As for the twenty-first century, eighteen movies are from the first decade, and thirty-five from the next seven years. The TV shows are similarly grouped, with the majority of tv shows coming from the second decade of the twenty-first century. Of the 86 different televisions programs 68 originally aired between the years of 2010-2017 and 14 from the first decade of this century. Unlike the transcripts from the movies, CALM only includes three television shows from the twentieth century—all of which originally aired in the 90s.

No conscious effort was made to choose movies and television based upon genre. Webb and Rodgers (2009) show that the linguistic variation among movie genres is not as great as might be expected. By grouping words together that share a common base into word families, they found that knowing the 3,000 most common word families in English allows one to understand 95% of any film regardless of genre. If Egyptian Arabic is similar to English in this regard, then there is not much change in the language employed by the different genres.

However, this choice may have an effect on this thesis. It is possible that certain verb forms are more common in certain genres. Action movies may contain more imperatives than the other genres, and therefore, a high proportion of action movies in the sub-corpus could skew the frequency of imperatives. Therefore, care was taken to make sure that one genre did not dominate the sub-corpus created for movies and television.

Once a movie was transcribed, it was reviewed for accuracy. All the transcripts were either transcribed or checked for accuracy by a native Egyptian speaker. A second reviewer was used to determine the ability of the reviewers to catch all the mistakes in the transcription. This process was necessary as some reviewers were not able to successfully read a transcript while listening to a movie.

The transcripts are orthographical rather than prosodic or phonetic. This choice was made on the basis that the target audience for this corpus is Arabic students and not necessarily linguists familiar with IPA or other transcription methods. This choice also allows a level of consistency within the corpus since the texts taken from the internet are all written in the Egyptian orthography. However, transcribing each word orthographically is a complicated matter since there is no standard orthography for Egyptian Arabic.

Thompson (2004) suggests that a dictionary be chosen and followed in regards to the spelling conventions. Although there is a dictionary for Egyptian Arabic, only the transcripts—not the texts from the blog corpus—would follow these spelling conventions. As orthographic variety is a feature of written Egyptian, it is preserved in the transcripts as well as the texts from blogs. To impose a standard orthography where none exists would be adding an unneeded layer of artificiality to the corpus. Also, for the purpose of this study, the more orthographic variety found in each register the better. Annotating with only one orthography would make the results

less applicable to other corpora which might contain different orthographical choices. Therefore, no standard orthography was imposed upon the transcribers.

Besides the words spoken by the actors, there are a few additions to the transcripts to aid comprehension. Ellipses are used to mark scene changes and longer pauses. Additionally, each utterance made by a speaker is given a line which is headed by the speaker's name followed by a colon. An example is given in Example 3.1.

Example 3.1: An example of a transcript from the spoken portion of CALM

ايناس: ها هتغديني فين؟
حامد: ايه رأيك لو تعدي عليا في البيت!
ايناس: اوكيه
.....
خالد: مش هتبطل الغتاتة اللي في دم اهلك دي
سامح: يالا قوموا انتوا جايين تنامو
علي: صباعي فقفق يا اخي منك لله (يعني عمل بالونة كده)
عمرو: يا اخي همشي عليها ازاي دي دلوقتي
سامح: بقول لكو ايه قومو يالا اتشطفوا عايز اكل سمك
خالد: لما نلحق نفوق الاول احنا مانمناش ساعتين على بعض يا اخي
سامح: انا هفوقك حالا هفوقك حالا
.....
ايناس: انت اللي ابتديت والبادي اظلم
.....

This portion from CALM illustrates the use of ellipses as an indication of a scene change as well as how each speaker is represented. One familiar with Arabic will also notice that the diacritics have been left out of the transcription. Although diacritics would have increased the value of the corpus and improved the rate of annotation, it was decided to leave them off in order to speed up the transcription process.

Blog Corpus

The other sub-corpus created from blogs will be called the blog corpus. Although blogs can be classified into many different genres (Biber, Egbert & Davies 2015), they will be treated as a single register. Although it is common to find among blogs transcriptions of speeches, movies, television programs, and songs, these texts were not included in the sub-corpus because of the nature of this thesis. The blogs were collected through scraping the internet based upon searches from Bing and Google. As discussed in Chapter 2, the method employed by Biber, Egbert, and Davies (2015) is to use the most common n-grams from a particular corpus as seeds for an internet search. Unfortunately, the most common trigrams in the spoken portion of CALM are dialect-neutral. If this technique were followed, much time would have been spent reviewing each text collected to determine whether it was Egyptian or not.

One possible solution to this problem is the use of Google's advance search feature to limit the results to only pages that originated in Egypt. This option was employed; however, I found that Google was still returning pages that contained other dialects of Arabic. This is likely due to the number of immigrants residing in Egypt especially following the tumult of the Arab Spring. In order to avoid a thorough review of each page, I adopted the approach supported by Sharoff (2006) for choosing seeds and relied on frequent dialect specific words to decrease the chances of dialect mixing.

After determining the seeds that I would search for, I used the program BootCat (Baroni & Bernardini 2004) to find, scrape, and convert webpages written in Egyptian Arabic into text files. The dialect specific terms did not completely eliminate the inclusion of other dialects or Modern Standard Arabic into the corpus. A cursory review of the files was completed to remove those texts. However, some Standard Arabic is contained in CALM because it is interwoven

throughout posts written in Egyptian. However, I removed posts that were completely written in Standard Arabic.

Target Features

As in other studies of register variation, this thesis will investigate register variation by comparing the frequency of features in the transcript and blog sub-corpora of CALM. The features that I chose to compare are verbal frequency, diversity, and keyness. The features are focused on verbs in part because this part of speech is slightly easier to identify and annotate than nouns and adjectives (Al-Sabbagh & Girju 2012a).

They were also chosen because of their widespread use in determining register for English. Staples (2016) used past tense verbs to show that patients' conversations with nurses differs from their conversations with doctors. Friginal (2009) used the diversity of verbs in his analysis of the language used by outsourced call centers. Ferguson (1983) characterized the use of the simple past and the progressive as features of sports announcer talk. Verbs also seem to play a role in distinguishing register in other languages as well (Biber & Conrad 2001; Biber et al. 2006). This thesis, by looking only at verbs, does not imply that an examination of verbs is sufficient to determine register; however, they were chosen as they seem to play an important role in distinguishing register.

To compare verbal frequency, diversity and keyness, each verb in the corpora needed to be identified and assigned a verbal category and a lemma. These annotations allowed me to quickly organize and count all the verbs for comparison. Each verb was categorized as either PERFECT (p), IMPERFECT (i), HABITUAL (h), IMPERATIVE (c), or FUTURE (f). Examples of a verb in each of these categories is given below in Table 3.1. These labels are a slight modification of

the categories used by Arfath Pasha et al. (2014)—who categorized the HABITUAL and FUTURE as subcategories of IMPERFECT—and were chosen because of their traditional use within the field of Arabic linguistics (Aboul-Fetouh 1969). Verbs labeled with 'p' are those verbs that are inflected for the PERFECT. Both the negative and positive IMPERATIVE are included in the label 'c'. The verbs proceeded by the FUTURE and HABITUAL morpheme discussed in chapter 2 were given 'f' and 'h'. All other verbs were given the label 'i'.

Table 3.1: Examples of a verb in each verbal category

imperfect	**تشوف**
	تـ - شوف
	tɪ-ʃu:f
	2SG.IMPERFECT.MASC-see
	you see
perfect	شافت
	شاف - ـت
	ʃæ:f-ɪt
	see-3SG.PERFECT.FEM
	she saw
positive imperative	شوفي
	شوف - ـي
	ʃu:f-i:
	see.IMPERATIVE-2SG.FEM
	see!
negative imperative	ماتشوفوش
	ما-تـ-شوف-و-ش
	mæ-t- ʃu:f-u: ʃ
	NEG-2PL.IMPERFECT-see-2PL.IMPERFECT-NEG
	don't see!
habitual	بيشوفوا
	بـ - يـ - شوف - وا
	b-i-ʃu:f-u
	HABITUAL-3PL.IMPERFECT-see-3PL.IMPERFECT
	they see
future	هنشوف
	هـ - نـ - شوف
	hæ-n-ʃu:f
	FUTURE-1PL.IMPERFECT-see
	we will see

Although verbs in the HABITUAL (h) and FUTURE (f) are IMPERFECT, I felt that they should be given their own categories to facilitate the collection of their counts in this thesis and to make it easier for future users of the corpus to search for them.

The verbs in CALM could justifiably be annotated by one person because of the nature of this project. Verbs are a relatively straight forward part of speech in Egyptian Arabic due to their unique morphology. This leaves very little room for debate as to their true part of speech. In the few instances where the verbal morphology was ambiguous with other parts of speech or verb types, native speakers were consulted. Therefore, most of the errors caused during annotation likely come from a lapse in concentration. As a way to limit the number of this type of error, a computer program, MADAMIRA, was used to identify and label all the verbs in the sub-corpora (MADAMIRA will be introduced fully in Chapter 5 of this thesis). After MADAMIRA assigned each tag, they were all reviewed for accuracy. Therefore, for an error in verb tagging to occur both the computer program and I would have to have made a mistake. Although this undoubtedly happened, less errors were produced than if just one method for tagging was employed.

Lemmatization

In addition to assigning verbal categories, the lemma of each verb was also identified. Every word in Arabic, just as in English, has a corresponding lemma. A lemma in this context is a form of a given word that has been chosen to represent all of the inflectional forms of that word. For example, in English ‘go’ is the lemma for ‘goes’, ‘go’, and ‘went’. Lemmas are important for corpus research and register analysis because they allow for the quick identification and

count of all the different forms of a particular verb. Lemmas are particularly important in Arabic because of the various inflections a single verb can undergo.

For English, the infinitive form of a verb often acts as its lemma; however, as was explained in Chapter 2, Arabic lacks a traditional infinitive form since all verbs are inflected for PERSON and ASPECT. Therefore, the 3SG.MASC.PERFECT is usually chosen to represent the lemma for each verb. This is the system employed by both the Standard Arabic Hans Wehr dictionary (1994) and the Egyptian Arabic Hinds/Bedawi dictionary (1986). For this reason, the 3SG.MASC.PERFECT was chosen as the form of the verb to act as its lemma. How that form of the verb was represented orthographically is beyond the scope of our discussion in this chapter; however, it is explained in greater detail in Appendix B.

The Sub-Corpora for Annotation

Even though only verbs were assigned a part-speech-tag, verbal category, and lemma in the thesis, the transcript and blog sub-corpora of CALM were divided into smaller sub-corpora in order to facilitate annotation. From this point on in the thesis, 'sub-corpus' will refer to sub-corpora of the transcription and blog sub-corpora. These smaller sub-corpora were created because the manual annotation of two million words required more time than was allowed for this thesis. The choice to manually annotate the corpus was made in order to answer question two of this thesis: "Is there an automatic annotator that is accurate enough to aid in a study of register?" One way to answer this question is to compare the annotations assigned manually to those assigned by a computer program. More on this process is discussed in Chapter 5 of this thesis.

The two sub-corpora created for annotation are of comparable sizes. The sub-corpus of transcripts contains 228,399 words including 38,768 verbs. The blog sub-corpus contains 141,318 words and 27,616 verbs. Both sub-corpora were intended to be equal sizes; however, the timetable for this project dictated a smaller number of blogs be annotated.

The transcript sub-corpus was created from both movies (113,163 words) and television shows (115,236 words). To ensure that similar words were taken from movies and television programs, two lists were made: one of all the movie titles in CALM and the other for all the television programs. Each list was divided into sub-groups based upon the number of transcribers. Three transcribers contributed movie transcripts and two worked on television shows. Therefore, the movie list was divided into three subgroups and the TV list was divided into two. This was done to ensure that the different orthographies used in the transcripts in CALM are represented in its sub-corpus, so that annotation could address as many different spelling conventions as possible. Only through the analyses of texts containing a variety of orthographies can confidence be given to annotation accuracy scores.

The movie and television transcripts from these sub-groups were the chosen at random for inclusion into the sub-corpus. The transcripts in each of the lists mentioned above were assigned a number and a simple program created using Python, chose a number at random. The transcript with that number was annotated and then added into the sub-corpus. This process continued until the target number was reached.

The creation of the internet sub-corpus proceeded in a different manner. Rather than divide the internet portion of CALM into texts, I divided it into 109 different 10,000-word chunks. This decision was based upon the difficulty of separating the texts from each other. A number at random was then chosen between 1 and 109. In order to get a wider sample from the

corpus I did not annotate neighboring 10,000-word chunks: there had to be at least 20,000 words separating each annotated chunk. In total 21 different 10,000-word chunks were annotated to reach 141,318 words. The number of chucks selected does not match the total number of words due to a number of factors including the removal of Modern Standard Arabic, duplicates, nonsensical posts, or extremely long posts.

Comparison of Sub-Corpora

Once each verb was assigned a part-of-speech tag, a verbal category, and a lemma, each of these features was counted from each sub-corpus and compared in order to determine whether verbs are used differently. This thesis looks at the overall number of verbs in each sub-corpus, as well as the frequency of each verbal category. The lemmas were counted to determine whether certain verbs are used more frequently in one sub-corpus.

Despite the transcript sub-corpus containing more words, where appropriate the counts from each of the corpora were normalized. Also, the statistical tests used to calculate significance took this difference into account. The two statistical tests used were log-likelihood and Bayes Factors which rely on the Bayesian Information Criterion (BIC).

Log-likelihood was relied upon because of its frequent appearance in corpus linguistics studies (Wilson 2013) and reliability over chi square when dealing with word counts that fall on either end of the frequency spectrum (Dunning 1993; Rayson & Garside 2000). The BIC was added into the analysis as an added guard against the misinterpretation of the meaning of the p-values produced by log-likelihood. Wilson argues that p-values produced by log-likelihood are inappropriate for corpus studies of frequency as they do not reveal the probability of the null hypothesis. He instead prefers the Bayes Factors which "allows the corpus linguist to quantify

explicitly the degree of evidence against the null hypothesis" (Wilson 2013 pg. 8). Therefore, any difference in frequency considered to be statistically significant in this thesis, was found to be so according to both log-likelihood and the BIC.

Summary

In this section, I have introduced CALM, and the different verbal categories that will be used to determine the amount of variance between the two sub-corpora. This variance will be measured by log-likelihood and the Bayes Factors. In the next section, I will present the findings which will show that verbs in each of the sub-corpora are used differently.

Chapter 4: Results and Discussion

Introduction

In this chapter, verb frequency counts from both sub-corpora will be presented in order to determine the extent of variance in the use of this part of speech. The overall frequency of verbs is examined followed by counts from each verbal category. A comparison of the most frequent lemmas and verbal diversity is also included in this chapter. These frequency and verbal diversity counts will then be analyzed in order to determine whether the variation in the use of verbs observed in the two sub-corpora is diverse enough to suggest register variation. This will be done by comparing the results to previous studies of register variation while also showing that the key verbs in each sub-corpus can be explained by features more common to the oral dimension.

Overall Frequency of Verbs and Verbal Categories

When working with corpora and frequency counts it is important to clarify exactly what is being counted. To help with this, the terms 'token' and 'type' are employed. In this thesis, I use 'word token' and 'verb token' to mean that each word and verb was counted toward the overall tally regardless of whether it has been counted previously. The terms 'word type' and 'verb type', on the other hand, describe counts that include each word only once regardless of the number of times it appears. For example, the sentence "Amr ran to the store and I ran to the school" has eleven word tokens and two verb tokens but only one verb type and eight word types since 'ran', 'to', and 'the' are used twice. When determining overall frequency of verbs and verb categories, I will be looking at tokens; however, the numbers presented during my discussion on verbal diversity will be counts of verb types.

With the distinction of types and tokens in mind, I will first look at overall verb frequency in the sub-corpora. This number is calculated by dividing the total number of verb tokens from the total number of word tokens. The two corpora reveal that verbs are used significantly more in blogs than in movies and television. Table 4.1 provides the counts for the number of verb tokens and word tokens in each sub-corpus and reveals that the blog sub-corpus contains nearly three percent more verbs than the transcript sub-corpus.

Table 4.1: Percentage of verb tokens in the Transcript and Blog corpora

	Transcript sub-corpus	**Blog sub-corpus**
Total number of word tokens	228,399	141,318
Total number of verb tokens	38,768	27,083
Percentage of total verb tokens	16.97	19.54

However, a simple count of the total number of words and verbs can be somewhat misleading especially in the transcript sub-corpus. The beginning of each line contains the name of the character who is speaking. This word was added by the transcriber in order to ease the readability of the transcript. If transcripts are to represent spoken language, these extra words should not be included in the word count since they are never uttered in conversation. Removing them causes the total word count to decrease to 208,986 which raises the verb-to-word ratio of the transcript sub-corpus to 18.55. However, even with this reduction in the corpus, the number of verbs remains significantly smaller than that of the blog sub-corpus according to log-likelihood and the BIC. The log-likelihood score is 16.87 ($p < 0.0001$) and the BIC score was 4.11 which suggests positive evidence against the null hypothesis.

Despite the increased use of verbs in the blog sub-corpus, not all verbal categories occur with equal frequency in it. The frequency of the IMPERFECT, PERFECT, IMPERATIVE, HABITUAL, and FUTURE per 100 words in each corpus is shown in Table 4.2.

Table 4.2: Frequency of each verbal category per 100 words in both sub-corpora

Verbal category	% of total words in the transcript sub-corpus	% of total words in the blog sub-corpus
Imperfect	6.56	7.90
Perfect	5	6.4
Habitual	1.83	2.46
Imperative	2.57	1.52
Future	1.44	1.3

The differences for each category in Table 4.2 are significant according to log-likelihood and the Bayes Factor except for the FUTURE category. Even though the blog corpus has a higher frequency of verbs, it does not have a higher frequency of verbs in the IMPERATIVE or FUTURE, and regarding the former, it is used with a higher frequency in the movie corpus. However, some of these differences change when we compare frequency to the total amount of verbs in each corpus rather than the total number of words. This is because of the higher concentration of verbs in the blog sub-corpus, causing counts of the verbal categories taken out of the total number of words to be misleading (Gries 2006). Both frequencies are provided in Table 4.3. When this is done, there are changes in the differences of the IMPERFECT and the FUTURE.

Table 4.3: Frequency of each verbal category per 100 verbs in both sub-corpora

Verbal category	% of total verbs in the transcript sub-corpus	% of total verbs in the blog sub-corpus
Imperfect	38.67	40.45
Perfect	26.86	32.54
Habitual	10.8	12.58
Imperative	15.15	7.8
Future	8.51	6.67

When compared to the total amount of words in each corpus, the use of the IMPERFECT is significantly more frequent in the blog corpus; however, this significance disappears when the frequency is compared with the total number of verbs. Despite a log-likelihood score of 12.88 ($p<0.001$), the Bayes Factor test suggests that the difference is not statistically significant.

The opposite is true of the verbs hosting the FUTURE morpheme. Although the increase in frequency of this verb in the transcript corpus was found to be insignificant when compared to the total amount of words in the corpus, its frequency becomes significant when only the total amount of verbs is taken into account. Therefore, if a verb is going to be used, it has a greater chance of being in the FUTURE in the movie corpus than the blog corpus.

As for the PERFECT and HABITUAL, their differences in frequency are significantly higher in the blog corpus regardless of whether they are compared to the total words or verbs in the sub-corpora. The same situation holds for the IMPERATIVE in the transcript corpus: in both cases its use is found to be significantly higher in the transcript corpus.

The IMPERATIVE is not only used more in the transcript corpus but also represents the largest difference in usage between the two registers. To help explain this further, let us separate the imperatives into four categories: negative IMPERATIVE, positive 2SG.MASC.IMPERATIVE, positive 2SG.FEM.IMPERATIVE, and positive 2PL.IMPERATIVE. By comparing the frequency counts of each category of imperatives across corpora, we can determine whether one register uses more of a certain type of imperative than the other. Luckily imperatives in Arabic are fairly unambiguous when it comes to PERSON; still, ambiguity can occur with verbs that have a ي *i:* or ى *a* as the final consonant. Therefore, any count of imperatives based upon PERSON would need to disambiguate these verbs or exclude them. For this thesis the latter option was chosen. Table 4.4 below shows the number of imperatives removed due to ambiguous form as well as the percentage of the total

imperatives it represents. The table also includes the number of lemmas excluded and their percentage of the total lemmas used as an imperative.

Table 4.4: The number and percentage of imperatives tokens excluded due to ambiguity

Corpus	Number of Imperative tokens excluded	Percentage of total Imperative tokens	Number of lemmas excluded	Percentage of total Imperative lemmas
Transcript	1441	24.53%	85	15.15%
Blog	461	21.51%	48	11.91%

I believe that those ambiguous imperatives can be removed from this analysis without dramatically changing the results. Of the 48 lemmas excluded from the blog corpus, the top five constitute 71% of the total number of imperatives excluded. In the transcript corpus, the top twenty lemmas make up 76% of those excluded. Therefore, despite the exclusion of a relatively large number of lemmas, the vast majority of excluded imperatives are represented by a small number of lemmas. Another factor that led me to continue with this analysis is its intent. Categories of the imperatives will only be compared with the same category in the other corpus. For example, I am not making claims about the frequency of the 2SG.MASC.IMPERATIVE compared with the 2PL.IMPERATIVE in the movie corpus. But rather, I address the number of 2SG.MASC.IMPERATIVE verbs in the transcript sub-corpus compared to the blog sub-corpus.

Instead of comparing the frequencies of the imperatives against the total number of words in the corpus, I used the total number of verbs and the total number of imperatives. The numbers for each category of imperative are given in Table 4.5.

Table 4.5: Frequency of imperatives across sub-corpora

		Transcript sub-corpus	Blog sub-corpus
Negative Imperative	Per 100 verbs	1.1	0.7
	Per 100 imperatives	7.29	9.01
Male Positive Imperative	Per 100 verbs	10.14	5.88
	Per 100 imperatives	66.89	75.83
Female Positive Imperative	Per 100 verbs	3.31	0.69
	Per 100 imperatives	21.87	8.87
Plural Positive Imperative	Per 100 verbs	0.6	0.49
	Per 100 imperatives	3.95	6.3

When analyzing the frequency of imperatives taken out of the total number of verbs, the frequencies in the transcript sub-corpus are all significantly higher than the blog sub-corpus except for the positive 2PL.IMPERATIVE. However, as a percentage of the total imperatives used, the blog sub-corpus uses the positive 2SG.MASC.IMPERATIVE and positive 2PL.IMPERATIVE significantly more than the transcript sub-corpus. The negative IMPERATIVE changes from being used significantly more in the transcript sub-corpus in proportion to the total amount of verbs to having no significant difference when taken from all total IMPERATIVE verbs. Only the positive 2SG.FEM.IMPERATIVE remains more frequent in transcript sub-corpus regardless of what its totals are taken out of.

Lemma Frequency and Verbal Diversity

Another feature used to show register variation is the words that are used in each genre. For this thesis we will just be looking at the most frequent verbs. If movie transcripts and blogs truly represent two different registers, based on what other studies of register have shown, we would expect that some verbs are used more frequently in one register than the other. Due to the small size of the annotated corpora, I will only look at verbs that have at least one lemma from either the transcript or blog sub-corpus with a frequency of over 100. Additionally, as this is a

preliminary look into the verbal variation of the two corpora, a full analysis of all the verbs is beyond the scope of this thesis.

There are seventeen verbs whose frequency is significantly higher in one corpus over the other. These verbs are given in Table 4.6 separated according to the corpus in which they are most frequent.

Table 4.6: A list of high frequency verbs that are unique to each register

More common in transcript corpus			**More common in blog corpus**		
اتفضل	**ɪtfaddˁal**	please, come in	بدأ	**bɛdʌʔ**	to begin
خشّ	**xoʃ:**	to enter	دخل	**dɛxɛl**	to enter
هدي	**hɪdi**	to calm oneself	كتب	**kʌtʌb**	to write
اكل	**ʔækɛl**	to eat	حاول	**ɦæ:wɛl**	to try
ساب	**sæ:b**	to leave	لقى	**lʌʔʌ**	to find
استنى	**ɪstɛnʌ**	to wait	حسّ	**ɦɛss**	to feel
شرب	**ʃirib**	to drink	فتح	**fɑtɑɦ**	to open
مشي	**miʃi**	to walk	رد	**rʌdd**	to respond
			قرأ	**ʔɑrʌ**	to read

The difference in usage was determined to be significant using log-likelihood and Bayes Factor. Words that have a statistically higher frequency in one corpus over another are considered *keywords* by corpus linguists (Hunston 2002). The *keyword* function of AntConc (Anthony 2018) confirms the accuracy of Table 4.6 and it is from the AntConc keyword list that the table has been arranged with the verb at the top being the most characteristic in each sub-corpus. The AntConc keyword list is not given in its entirety because it analyzed all the verbs rather than the 100 most frequent from each sub-corpus.

Keyword lists do not in and of themselves suggest a difference in register, since any two texts from the same genre will use different words. For this reason, those studying register variation look at the semantic classes of the *keywords* rather than the words themselves. Although the number of key verbs provided by Table 4.6 is limited, there is a difference in the types of verbs used. The breakdown of the verbs appears in the analysis section of this chapter.

Having a lemmatized corpus also allows for the ability to compare lexical diversity between the two registers. However, calculating the verbal distribution of each sub-corpus is not as straightforward as the previous formulas used to normalize frequency. This is because verbal types do not represent a linear distribution. According to Biber (2006), the larger a corpus is, the more often words are repeated. Therefore, he suggests a different formula for normalizing lexical diversity counts. The formula he proposes is given below in Equation 4.1.

Equation 4.1: Formula for normalizing verb counts

(# of verb types / square root of total # of verbs) x 1000 = normed # of verb types

Using this formula, the verb types per million for each register were calculated and are given in Table 4.7. The table also includes the percentage of the verbs of each sub-corpus that are diverse. The differences are statistically significant, suggesting that the blog corpus is richer in terms of verb types. Therefore, not only are verbs more common in the blog corpus, but they also appear with greater diversity.

Table 4.7: The verbal diversity (verb type/token ration) for both registers

Corpus	Number of verb types	Verbal diversity	Verb types per million verbs	Verb types per million words
Transcript	2,279	5.88%	11,574	4,768
Blog	2,079	7.53%	12,510	5,530

Summary

In this chapter, the verbs belonging to each verbal category were counted and compared across the sub-corpora. This revealed significant differences in how verbs are used in blogs versus movie/television transcripts. The most significant difference was found in the usage of the IMPERATIVE which is utilized much more in the transcript sub-corpus. Not only do imperatives have a higher frequency in this sub-corpus, but female imperatives show up significantly more

than in the blog sub-corpus. The verb lemmas were also counted and compared, revealing differences in the types of verbs used and their diversity. These differences will now be analyzed and used to answer the question "Is there enough evidence of register variation between movie/television transcripts and blog posts to warrant a more thorough investigation?"

Analysis

In this section, I will show that based on the data collected above there is enough evidence to warrant a wider investigation into the variations that exists between these potential registers. In both sub-corpora of CALM, certain tenses and aspects are more common in one sub-corpus than the other. The frequent use of PERFECT and HABITUAL verbs seems to be a feature of blogs; whereas IMPERATIVE and FUTURE verbs are more frequently used in the movies.

The frequency of verbal aspects and moods is used as a distinguishing feature of register in previous studies. For example, in English the frequent use of the PERFECT has been identified as a feature of narration (Biber & Conrad 2001; Staples 2016). If Egyptian Arabic behaves like English, then the higher frequency of the PERFECT in the blog sub-corpus signals a greater reliance on narration than movies and television, suggesting its texts belong to a different register. The possibility of the Egyptian Arabic PERFECT being a feature of narration is further supported by Biber, Egbert, and Davies (2015) who found that for English most texts on the internet could be classified as narrative followed by informational description/explanation.

Similarly, the frequency of the IMPERATIVE in the transcript sub-corpus could easily be a feature of involved and non-narrative speech and was categorized as such in a multidimensional analysis of Somali (Biber & Conrad 2001). Therefore, the frequency of the PERFECT and IMPERATIVE in the sub-corpora suggests a difference in register based upon the use of narration.

The distinct differences in the use of the HABITUAL and the FUTURE could also be linked to narration but could also be due to features of another dimension of the language found within these two registers. As little is known about the features of each dimension of Egyptian Arabic, a deeper investigation is needed so that the frequency of these verbal tenses and aspects can be put into context.

The subjects of IMPERATIVE verbs also seem to be dependent on register. The data collected from the annotated corpora suggest that those writing blogs do not typically write to an audience of a single female. The number of female positive imperatives in the blog corpus is trivial when compared to the movie corpus. What might be the cause of this disparity? In Egyptian Arabic, the 2SG.MASC is used as the ambiguous 'you' when no specific individual is being addressed. Therefore, we would expect this form of the imperative to be used much more than female imperatives in the literate dimension which does not often address specific individuals. However, this explanation has not been verified and further investigation is needed since many factors could affect this outcome. The degree to which this difference proves register variation depends upon the answers to these questions.

The greater frequency of verb tokens and types in the blog corpus also suggests that Egyptian is used differently in blogs and movies. In a study of the variation in spoken and written academic English, one feature used to differentiate the two registers by Biber (2006) was the total number of words found in each part of speech. Interestingly, he found that verbs were much more frequent in lectures than in papers and journal articles. This, along with other studies of English and Spanish more generally suggest that verbs in these two languages are typically more frequent in the oral dimension (Biber 1999; Biber et al. 2006). However, the opposite

appears to be true for Egyptian Arabic: the blog sub-corpus contains a statistically higher number of verbs than the transcript sub-corpus.

Additionally, the blog sub-corpus contains a greater diversity of verbs, which is consistent with English and Spanish and may be expected since authors have time to think about the words they will use and revise their choices (Biber 2006; Biber et al. 2006). One factor that could have contributed to this is the size of the annotated corpus, but it could also be true that a feature of spoken Egyptian Arabic—like Spanish and English—is a lack of verbal diversity. Therefore, if this pattern holds as more of the corpus becomes annotated, it would constitute further evidence of register variation.

Another factor used in the classification of different registers in a multi-dimensional analysis is semantic classes of verbs (Biber & Conrad 2001; Biber et al. 2006; Biber 2006; Biber 2016; Biber & Egbert 2018). Different registers seem to use verbs of particular semantic classes with higher frequency than other registers. Egyptian Arabic does not seem to be an exception to this either. Although the small size of the annotated corpora precluded any serious look at the semantic classes of the verbs in each corpus, a class of verbs is used more frequently in the transcript corpus than the blog corpus. Verbs that appear frequently in the IMPERATIVE are those verbs considered to be *keywords* in the transcript corpus. Table 4.8 shows the *keywords* in both sub-corpora and the percentage that those verbs occur as imperatives in the transcript sub-corpus. The verbs from transcript sub-corpus are marked with (T) and those from the blog sub-corpus with (B). The verbs' frequency rank in the transcript sub-corpus has also been provided to give perspective of overall usage.

Table 4.8: The verbs most common to each sub-corpus organized according to use in the IMPERATIVE in the transcript sub-corpus

Word	Percentage of use of the IMPERATIVE form	Frequency ranking in transcript sub-corpus	Meaning
اتفضل(T)	**ɪtfɑdˁɑl** 93.6%	14th	please, come in
هدي(T)	**hɪdi** 83.2%	55th	to calm oneself
استنى(T)	**ɪstɛnʌ** 56.8%	33rd	to wait
خشّ(T)	**xoʃ:** 43.4%	47th	to enter
فتح(B)	**fɑtɑħ** 32.3%	69th	to open
ساب(T)	**sæ:b** 31.7%	16th	to leave
مشي(T)	**miʃi** 29.9%	20th	to walk
رد(B)	**rʌdd** 26.7%	51st	to respond
دخل(B)	**dɛxɛl** 20.5%	44th	to enter
قرأ(B)	**ʔɑrʌ** 15.9%	100th	to read
اكل(T)	**ʔækɛl** 11.3%	29th	to eat
كتب(B)	**kʌtʌb** 11.1%	94th	to write
حاول(B)	**ħæ:wɛl** 10.2%	65th	to try
شرب(T)	**ʃirib** 5.6%	50th	to drink
بدأ(B)	**bɛdʌʔ** 2.6%	151st	to begin
حسّ(B)	**ħɛss** .09%	60th	to feel
لقى(B)	**lʌʔʌ** 0%	21st	to find

The top three verbs occur more often as an imperative than any other form. It is, therefore, not surprising to find them more commonly in the transcript sub-corpus. The fourth verb on the list خشّ *xoʃ:* "to enter" is interesting because a similar verb دخل *dɛxɛl* "to enter" also appears. Despite their synonymous meanings they are both characteristic in the other sub-corpus. The high frequency of خشّ *xoʃ:* "to enter" as an imperative might explain this. However, the IMPERATIVE of دخل *dɛxɛl* is also frequently used in the transcript sub-corpus. An interesting study would be to determine the situations and settings where an Egyptian might choose the one imperative over the other.

The highest verb on the list that is characteristic of the blog corpus is فتح *fataħ* "to open." At first this may come as a surprise but when grouped with كتب *kʌtʌb* "to write", ردّ *rʌd:* "to reply" and قرأ *ʔarʌ* "to read" which were also more characteristic of the blog sub-corpus, its context becomes clearer. Just as the latter verbs are related to the writing of blogs, so فتح *fɛtɛħ* is

related to the internet as it is used much like the English word ‘open’ as in ‘open a new tab’ or ‘open the website.’ Its classification as more characteristic of the blog sub-corpus is also a result of the size of the corpus. Several texts in the blog sub-corpus concern banks and the opening of accounts. As more and more of CALM becomes annotated, this verb could disappear from among the verbs found to have a statistically higher frequency in the blog corpus.

The verbs characteristic of the transcript sub-corpus with a low frequency of imperatives are the verbs associated with food and beverage. However, their appearance on this list comes from the higher frequency of situations that occur in the transcript corpus surrounding eating and drinking. Examples of such situations from the transcript corpus that include أكل *ʔækɛl* ‘to eat’ and شرب *ʃirɪb* ‘to drink’ are given below in Table 4.9.

Table 4.9: Examples from the transcript corpus demonstrating the uses of ‘to eat’ and ‘to drink’

ينفع أقعد أكل جنبك؟
yɪnfʕ uʔʕud ʕækɛl gænbɛk?
Can I sit and eat next to you?
متتعبنيش علشان خاطري قوم كل ونام.
mætɪtʕbni:ʃ ʕlʌʃæ:n xɑtˁri ʔu:m kul wɑnæ:m.
Don’t be a pain! For my sake go eat and get to bed.
يا حمامة اتفضل اشرب قهوتك.
yæ: ɦʌmæ:mʌ ɪtfʌddˁɑl ɪʃrʌb ʔɑhwɪtɛk.
Hamama go ahead and drink your coffee.
هو احنا شربنا ايه يا عاطف؟
hu: ɛɦnʌ ʃɪrɪbnʌ ɛi: yæ: ʕɑ:tˁɪf?
What did we drink Atif?

The three verbs with the lowest frequency of imperative forms are all more common in the blog sub-corpus. Although بدأ *bɛdʌ* “to begin” and حسّ *ɦas:* “to feel” are relatively more infrequent in the transcript sub-corpus, لقى *lʌʔʌ* “to find” is not. As the twenty-first most frequent

verb, we would expect to see at least one occurrence of its imperative form. Further investigation of both sub-corpora revealed that لقى *lʌʔʌ* was not even modified by the modal لازم *læ:zɪm* "need, must" which could have been combined with the verb to convey a meaning similar to the IMPERATIVE. Therefore, it is not surprising that this verb is found with a higher frequency in the blog sub-corpus and also suggests that it is used differently than its English counterpart.

The features that have been attributed to each corpus in this chapter are features that have been used in other studies to show register variation. Although the data presented here does not prove variation in these two corpora, it justifies expanding annotation.

Summary

The differences in the verbal frequencies of the two sub-corpora suggest that they belong to different registers of the language. A more thorough investigation is needed to confirm this result; however, the differences found are consistent with differences found in the oral and literate dimensions for other languages. The increased use of the PERFECT in the blog sub-corpus and the high frequency of the IMPERATIVE in the transcript sub-corpus are likely linked to the inclusion of more narration in the former. Higher frequencies in the total number of verbs and their diversity in the blog sub-corpus also point toward register variation. In the next chapter, the possibility of conducting such a study with the use of an automatic annotator will be explored.

Chapter 5: Notes on Annotation

Introduction

In this chapter I will answer the second and third questions proposed in the introduction of this thesis: "Is there an automatic annotator accurate enough to aid in a study of register?" and "Are there ways that automatic annotators can be improved?" The automatic annotator used for this project will be introduced and analyzed in order to determine whether it can be relied upon as the sole source of annotations for future studies of verbs in Egyptian Arabic. A brief introduction to the automatic annotation of Egyptian Arabic will be given followed by a comparison of the counts provided by the automatic annotator and those produced by manual annotation. Following an evaluation of the extent of the differences in annotation counts, suggestions will be given as to how automatic annotation can be improved.

Annotation of Verbs

One of the limits of this thesis is the size of the corpora used for analysis. Although CALM contains over two million words, only 456,798 words were analyzed in this thesis. This is because each word was manually reviewed in order to separate the verbs from the other parts of speech and to assign each verb its proper verbal category and lemma. In corpus linguistics, the process of labeling the contents of a corpus is referred to as linguistic annotation and is done so that target structures can be easily searched, or in the case of register studies, counted (Kübler & Zinsmeister 2015). There are many different types of annotation including lexical, syntactic, semantic, and discourse annotation. They can be applied to a corpus either manually, semi-automatically, or automatically. Manual annotations are performed by humans, while automatic

annotations are done by computer programs. Computer-generated annotations that are checked by a human for accuracy afterwards are called semi-automatic annotations.

The large corpora used for register analyses necessitate the use of automatic annotators since doing so manually would require considerable time and money. A prominent program used for multi-dimensional analyses is the Biber Tagger, which automatically analyzes texts to identify and count the target features (Biber 1988). Unfortunately, this annotator is unable to analyze Egyptian Arabic. In fact, automatic annotators available for Egyptian Arabic are limited in their capabilities, and although more advanced resources exist for Standard Arabic, the morphological and lexical differences discussed above cause MSA annotators to struggle in annotating Egyptian Arabic texts (Maamouri et al. 2014).

However, automatic Egyptian Arabic annotators have come a long way in the past fifteen years. In 2004, a part-of-speech annotator for MSA built by Diab, Hacioğlu, and Jurafsky (2004) was achieving an accuracy rating of 95.49%, whereas an analyzer for Egyptian Arabic was only accurate at a rate of 62.76% (Duh & Kirchhoff 2005). One reason for the disparity in the annotators for the two Arabic varieties was the lack of large corpora or a complete lexicon for Egyptian Arabic on which an annotator could be trained (Habash & Rambow 2006).

In order to help solve this problem, Abo Bakr, Shaalan, and Ziedan (2008) developed an annotator that would translate Egyptian Arabic sentences into MSA and then tag the MSA for part of speech. Those parts of speech would then be applied to the Egyptian words. The annotator was able to successfully convert the Egyptian Arabic to MSA 88% of the time and achieved accuracy ratings for part-of-speech tagging of 85%.

At about the same time, researchers began to develop Egyptian Arabic taggers that did not depend upon MSA. Al-Sabbagh and Girju created a finite-state transducer module (Al-

Sabbagh & Girju 2012b) based upon rules associated with morphology at the word level and a tagger based upon Transformation-Based Learning (Al-Sabbagh & Girju 2012c). The former system was trained and evaluated on language that came from three sources: Twitter, Question/Answer (QA) Pairs, and blogs. The highest reported accuracy among them for POS tagging is 0.907 which was achieved on the QA Pairs data. The latter tagger did not perform as well, achieving accuracy of 0.888 which is to be expected since it was created for analyzing Twitter data.

MADAMIRA (Arfath Pasha et al. 2014), like the annotators created by Al-Sabbagh and Girju, analyzes each word according to the possible morphemes attached to it. It then uses language models to provide the morphological analysis, parts of speech, lemma, and diacritics for each word in a text. Its accuracy score for part-of-speech tagging is 0.923 which is slightly better than the annotators created by Al-Sabbagh and Girju. However, MADAMIRA's ability to provide a lemma for each word makes it valuable tool for register variation studies which often count words based on the lemma to determine the prominent semantic word classes of a given register.

Despite these gains for part-of-speech annotators, two questions still surround Egyptian annotators. The first is whether they are accurate enough to be used without the need for a manual review of the results. The CLAWS tagger, upon which the Biber tagger is based, and the Stanford taggers, record part-of-speech accuracy of 96% for English (Leech & Smith 2000; Toutanova & Manning 2000). Although the 92.3% accuracy of MADAMIRA does not appear to be far behind the accuracy of the CLAWS and Stanford taggers, this gap could affect the counts to such a degree that the results become unreliable.

The second question is whether the accuracy level is maintained when applied to other Egyptian corpora. There is simply a lack of published research which evaluates Egyptian annotators on corpora and registers not used in their training data. Evaluating automatic annotators on new corpora and registers has been shown to decrease their accuracy levels (Tseng, Jurafsky & Manning 2005; Derczynski et al. 2013). This is even more likely to occur in Egyptian Arabic because new corpora and registers may contain orthographic conventions that the annotator is unfamiliar with; thus, causing known words to become unrecognizable by the computer. Egyptian Arabic does not have a standard orthography and the 20,000-word corpus used by the creator of MADAMIRA to determine its accuracy could not have contained the numerous spelling variations that exist in Egyptian Arabic (Arfath Pasha et al. 2014).

Comparison of Annotations

In order to evaluate the effectiveness of MADAMIRA, it was used to tag both sub-corpora twice. One copy of the annotations for both sub-corpora were reviewed by me and when necessary the mistakes were corrected. This manually corrected copy of the annotations was used for the main analysis of the thesis. However, the second, uncorrected copy of annotations was also analyzed in a similar fashion: frequency counts for the verbs, verbal categories, and lemmas were collected. Both the semi-automatic annotations (those which were manually corrected) and automatic annotations (those produced solely by MADAMIRA) were then compared automatically.

It must be noted that the non-corrected annotations were not the raw output from MADAMIRA. As noted above, the HABITUAL and FUTURE verbal categories were added to the manually corrected annotations as well as changes made to the orthographic representation of

some lemmas. In order to be able to fairly evaluate MADAMIRA, these changes were applied to all the copies of the annotations.

Frequency Counts Provided by MADAMIRA

The first measure evaluated was MADAMIRA's ability to correctly identify verbs and their appropriate categories. This was done by determining recall and precision scores for MADAMIRA when applied to the sub-corpora of CALM. In this thesis, precision is the percentage of VERB tags assigned by MADAMIRA that were correct. Recall, on the other hand, is the number of verbs correctly identified out of the total number of verbs in the two sub-corpora. The recall and precisions scores for MADAMIRA when applied to the blog and transcript sub-corpora are provided in Table 5.1.

Table 5.1: Recall and Precision scores for MADAMIRA on the blog and transcript sub-corpora

Corpus	Recall	Precision
Blog	0.92	0.923
Transcript	0.912	0.915

The numbers reported in Table 5.1 are consistent with the accuracy score reported by the creators of MADAMIRA (Arfath Pasha et al. 2014). Therefore, the question raised earlier in this chapter concerning its application to registers not originally trained on has been answered. Although MADAMIRA reports slightly lower precision and recall scores for the transcript sub-corpus, the gap is not substantial. The reason for this difference is largely due to the increased use of proper nouns in the transcript sub-corpus which are often lexically ambiguous with verbs. Therefore, MADAMIRA seems to be able to handle multiple registers with similar levels of accuracy.

This, however, does not answer whether these scores are high enough to perform an accurate analysis of verbal use in the two sub-corpora. Table 5.2 gives the counts for each of the

verb types as annotated by only the automatic tagger. The table also includes the percent change from the counts from the manually corrected annotations. The table refers to the transcript and blog sub-corpora as 'non-gold' because they have only been annotated by MADAMIRA. This is in contrast to the 'gold' sub-corpora whose annotations have been manually corrected and are, therefore, more reliable.

Table 5.2: The frequency counts of each verb type by annotation method

	Non-gold transcript	**Change**	**Non-gold blog**	**Change**
Overall Verbs	38,518	-0.65%	27,296	-1.16%
Imperfect	15,807	+5.44%	11,448	+3.96%
Perfect	11,714	+12.47%	9,343	+3.96%
Command	3,762	-35.97%	1,324	-38.22%
Habitual	3,962	-5.35%	3,301	-4.95%
Future	3,273	-0.82%	1,880	+2.01%

Table 5.2 reveals that although the gold and non-gold counts are not identical the difference is minimal for many of the categories. However, MADAMIRA struggled most identifying IMPERATIVE verbs in both corpora and PERFECT verbs in the transcript corpus. The variation in MADAMIRA's accuracy in identifying PERFECT verbs in the transcript and blog corpora is somewhat surprising. However, this is largely due to the increased use of imperatives and proper nouns in the transcript corpus. Both of these forms can be ambiguous with PERFECT verbs.

Despite the variance in frequency counts for the verbs in the gold and non-gold corpora, all variations which were found to be both significant and insignificant in the gold corpora were likewise significant and insignificant for the corresponding non-gold corpora. This nearly holds for the imperatives as well, except that the automatically annotated corpora do not report a significant difference in the use of the 2PL.IMPERATIVE in the blog corpus. Counts provided by

the automatic tagger for the imperatives are given in Table 5.3. MADAMIRA does not attempt to categorize negative imperatives and therefore, each cell in its row contains 'NA.'

Table 5.3: The number of imperatives as reported by MADAMIRA and percent change compared to the semi-automatically tagged corpora

	Non-gold transcript	**Change**	**Non-gold blog**	**Change**
Negative Imperative	NA	NA	NA	NA
Positive 2SG.MASC.IMPERATIVE	1,455	-40.95%	692	-35.33%
Positive 2SG.FEM.IMPERATIVE	877	-31.75%	179	-9.6%
Positive 2PL.IMPERATIVE	165	-28.88%	89	-34.07%

As nearly all observations about the use of the verbal categories in both sub-corpora were reached using only MADAMIRA, I conclude that the program is accurate enough to evaluate verb usage without the need for manual correction. In the next section we will evaluate whether this success holds for evaluation of verbal diversity and lemma usage.

Lemma Frequency and Verbal Diversity Provided by MADAMIRA

The keyword verb lists provided by MADAMIRA are nearly identical to the lists given in Table 4.8, with the exception of six verbs. The counts provided for اتفضل *ɪtfʌdˀal* 'please/come in', ردّ *rʌdd* 'to respond', حسّ *ħɛs:* 'to feel', فتح *fataħ* 'to open', هدي *hɛdi:* 'to calm oneself', and قرأ *qaraˀ* 'to read' were not accurate enough to reveal that these verbs were used with a statistically higher frequency in one genre when compared to the other. As for the verbal diversity for each genre, the automatic annotator provides data that shows more verbal diversity within the blog corpus. However, this gap is reported as wider than it is. The analysis is provided in Table 5.4.

Table 5.4: The verbal diversity for both registers of the non-gold sub-corpora and percent change compared to the gold corpora

	Non-gold transcript corpus	Change	Non-gold blog corpus	Change
Number of verb types	2,867	+ 25.8%	2,751	+ 32.32%
Verbal diversity	7.44%	+ 26.53%	10.08%	+ 33.86%
Verb types per million verbs	14,608	+ 26.21%	16,651	+ 33.1%
Verb types per million words	5,999	+ 25.82	7,318	+ 32.33%

Based on the verbal lemmas generated by MADAMIRA, the diversity of the transcript sub-corpus is 7.44% while the blog corpus is 10.08%. The hand-annotated corpus also shows that the blog corpus contains more verbal types than the transcript corpus; however, this difference is exaggerated by the counts given by the automatic annotator. Whereas the difference in the verbal diversity between the two registers of the hand-tagged corpus is 1.65 percentage points, it increases to 2.64 when the verbs have been lemmatized by MADAMIRA. Therefore, the automatic annotator is less accurate in terms of recognizing the correct lemma for each verb. As demonstrated by Table 5.2, MADAMIRA can recognize with fairly high accuracy whether a word is a verb and place it in the correct verbal category. However, the task of sorting through the morphology to discover the correct lemma proves to be more difficult. Therefore, verbal studies requiring lemma recognition may not be as accurate when using the automatic annotator.

Role of an Automatic Annotator in Future Studies

As discussed in Chapter 4, the differences in the usage of verbs suggests the need for further investigation into the true nature of the variation among the sub-corpora of CALM. To extend

this study further, more annotation will need to be completed. Without the use of an automatic annotator, this will be a time-consuming endeavor. Therefore, it is important to determine the ability of an automatic tagger to aid in such a study.

By comparing the data provided by the corrected annotations to that of the annotations produced solely by MADAMIRA, I find that the annotations provided by the latter for this thesis would lead a corpus researcher to nearly the same conclusions. The counts for overall verbs and verbal categories varied in every case from the numbers provided by the corrected annotations; however, the variations were not enough to change the results. Therefore, it appears that a tagger achieving recall and precisions scores consistent with that of MADAMIRA would have been reliable enough to perform this study of verbs.

In the blog sub-corpus, with the exception of the IMPERATIVE, MADAMIRA was only off from the total number of verbs in each category by less than five percent. In both sub-corpora, MADAMIRA was consistent with the categories that it over and underrepresented. The IMPERFECT and PERFECT were both overrepresented, and IMPERATIVE and HABITUAL were both underrepresented. The only exception was the FUTURE category which showed an underrepresentation in the transcript sub-corpus and the opposite in the blog sub-corpus. Unfortunately, the percent change was not consistent from one sub-corpus to the other, preventing the ability to accurately guess the actual number of occurrences for a verbal category. However, it seems possible to know whether the number provided by MADAMIRA is higher or lower than the counts generated by a semi-automatically annotated corpus.

Improving Annotation

These limitations, once recognized, could also help improve the accuracy of MADAMIRA. One area which affected MADAMIRA's accuracy was in differentiating proper nouns from verbs. This is one of the factors that led MADAMIRA to overrepresent both the IMPERFECT and PERFECT aspects. It also explains why MADAMIRA achieved a lower precision score on the transcript sub-corpus despite the more mainstream orthography used in its texts.

Throughout the annotation process, it was clear that MADAMIRA struggled with proper nouns; however, the extent of its effect was not realized until an examination of MADAMIRA's misses. Collecting the words which MADAMIRA thought to be verbs and their corresponding lemmas allowed for the production of a ranked list of each lemma according to their number of false positives. This list revealed that in the transcript sub-corpus, seven of the top ten words incorrectly assigned a verb tag were names and titles given to people. This is because these names are orthographically ambiguous with the verbs created from the same root letters.

This problem is also compounded by the lack of any kind of marker designated for proper nouns. English proper nouns are distinguished from common nouns by the capitalization of the first letter, whereas Arabic orthography has no case distinction. Table 5.5 contains proper nouns and titles associate with humans like مدام *mɛdæ:m* 'madam' and مامي *ma:mi:* 'mommy.' The numbers on the right are the number of times the name was incorrectly tagged as a verb in the transcript sub-corpus. The ranking associated with each name corresponds to the list of the most frequent non-verbs tagged as verbs in the transcript sub-corpus.

Table 5.5: Number of proper nouns and titles tagged as verbs in the transcript sub-corpus

Rank	Name in Arabic	Name in IPA	Number of times mistaken for verb
1	يحيي	yæɦiyɑ	497
3	سامح	sæ:mɪɦ	191
5	بيري	bi:ri:	68
6	طلعت	tˁɑlʕɑt	61
7	محسن	muɦsɪn	51
8	مدام	mɛdæ:m	46
9	حكمت	ɦɪkmɛt	44
10	رفعت	rɑfʕɑt	43
11	هاشم	hæ:ʃɪm	40
12	مامي	mɑ:mi:	36
13	آمين	æ:mi:n	36
17	بت	bɪt	26
22	توحة	tu:ɦɑ	24
27	بيرلا	bi:rlɑ	20
29	تريز	tɛrɑi:z	18

These seventeen word forms represent 1,239 of the 3,290 words that were incorrectly assigned the tag VERB in the transcript sub-corpus. This is 37.6% of all the false positives and yet does not represent all of the names thought to be verbs. Names also occur in the blog sub-corpus, but much more infrequently. Of the top thirty false positives in the blog corpus, only eight were names totaling 223 occurrences which constitutes only 10.5% of the total number of false positives.

The largest contributor to the increase in names in the transcript sub-corpus is its format: the beginning of each line contains the name of the character who is talking. It would be possible to change the format of the scripts and remove the character markers before annotation; however, this does not change the fact that annotating Arabic names poses a significant challenge to taggers in general. There are many ways that a tagger could potentially disambiguate names from verbs like looking at the position of the word in question in the sentence or by analyzing the parts of speech of the words surrounding it. However, looking

through the annotated portions of the transcript sub-corpus reveals another simpler solution for some of the names.

Consider the most common false positive in the transcript corpus يحيي *yæħiya* which can either be a name or a verb conjugated for 3SG.MASC.IMPERFECT. In both annotated sub-corpora, there are 522 instances of this word form as the noun and five instances of a verb using this same three letter root. The verbs are given in Table 5.6.

Table 5.6: The verbs in the annotated sub-corpora that share a root with the proper noun يحيي

يحيو هم *yæ-ħai:y-u-hom* 3PL.IMPERFECT-greet-3PL.IMPERFECT-3PL.ACC *He greets them*
بيحييه *biħai:yi:* HAB-3SG.IMPERFECT-greet-3SG.ACC *He greets him*
احييك *ʔaħai:yi:k* 1SG.IMPERFECT-greet-2SG.ACC *I greet you*
احييك *ʔaħai:yi:k* 1SG.IMPERFECT-greet-2SG.ACC *I greet you*
أحياني *ʔaħi:yæ:ni:* 3SG.IMPERFECT.give new life.1SG.ACC *He gave me knew life*

However, the word form يحيي does not appear as a verb in the annotated corpus. If we expand our search of a verb with the form يحيي beyond the annotated sub-corpora to include all of CALM, there is only one occurrence in over two million words of the lexeme يحيي as a verb. It appears in the blog corpus in a phrase quoted from the Quran من يحيي العظام وهي رميم *mæn yuħi:-l-ʕɪzˁa:ma wahiya rɛmi:m* "who gives life to these decayed bones." Despite this ambiguity,

it seems more efficient to train the annotator to recognize يحيي as a noun rather than a verb since CALM suggests that it overwhelmingly appears as a name. If the tagger were programmed to tag every instance of يحيي as a noun, it would only incorrectly tag one word rather than the 1,359 it would miss if applied to all of CALM.

This also highlights the importance of training annotators on multiple registers since the annotated blog sub-corpus only contained one instance of the name يحيي and four verbs derived from the same root, in contrast to the transcript sub-corpus which contained 521 occurrences of the name and only one instance of the verb. Therefore, an annotator trained only on the blog corpus would favor the verb over the noun.

A similar condition exists for the second word in Table 5.5 سامح *sæ:*mɦ which can either be the 3SG.MASC.PERFECT or 2SG.MASC.IMPERATIVE form of the verb 'to forgive' in addition to being a name. Of the 54 instances of this verb in both annotated portions of CALM, there are four occurrences of the 3SG.MASC.PERFECT and nineteen of the 2SG.MASC.IMPERATIVE; however, all of the verbs in these forms also have the direct object morpheme suffixed onto them. Since proper nouns do not share affixes with verbs, none of the verbs are ambiguous with the name. In fact, in the entire movie corpus of CALM, there are 712 instances in which the name سامح is orthographically identical to a verb. In all of these cases, not a single one is the verb. One instance is an active participle and the rest are the name. Even if we include in our search بسامح which could either be 1SG.HABITUAL 'forgive' or 'with Samih', only a single instance is returned; however, it is the name and not the verb.

Table 5.7: Number of occurrences of ambiguous lexemes as a verb and a proper noun in the transcript corpus

	As a Noun	As a Verb
سامح	711	0
يحيي	1409	1
رفعت	138	5

A similar trend occurs with the name رفعت *rɪfʕæt* which is formally identical with the verb رفعت *rɛfʕat* '1SG/2SG.MASC."raised." Of the 143 instances of رفعت only five were the verb. However, not all proper nouns behave in this manner. The word form حكمت *ħkmt* which is the ninth most mis-tagged lexeme, appears in the transcript corpus as a verb just as often as the noun. Therefore, collecting a list of names and training an annotator to tag those word forms as nouns would not necessarily improve accuracy. However, this thesis has demonstrated that some proper nouns are not as ambiguous as previously thought, and therefore, annotation can be improved as these word forms are tagged as nouns.

Another way to easily improve annotation would be to study the verbs that MADAMIRA struggles to lemmatize correctly. This would help give more accurate results when creating *keyword* lists and studying the lexical diversity of the corpora. As demonstrated earlier, each lemma appears in multiple inflected forms. This can sometimes cause MADAMIRA to assign erroneous lemmas to verbs; however, MADAMIRA is consistent in its errors. Therefore, incorrect lemmas that share a form refer back to the same verb. This allows for the automatic correction of these lemmas.

The only difficulty is matching the mistakes with the actual lemmas. For example, MADAMIRA, as run on my computer, returned جهب *ghb* as a lemma. However, neither the Hans Wehr MSA dictionary nor the Hinds/Bedawi dictionary includes an entry for a word with those root phonemes. Through annotation, it became clear that whenever جهب *ghb* was assigned as a lemma, MADAMIRA should have been assigning جاب *gæ:b* for the verb meaning "to bring." Table 5.8 provides a list of verbs whose lemmas were incorrectly assigned as in the example with جاب *gæ:b*. The percentages correlate to the number of lemmas incorrectly

assigned for that verb in each of the transcript and blog sub-corpora. Therefore, the second row in the table should be read as MADAMIRA incorrectly lemmatized the verb جاب *gæ:b* 97% of the time in the transcript sub-corpus. The last column in the table is the number to which the percentages translate. The second column on the left provides the most common lemmas incorrectly assigned to the verb forms.

Table 5.8: Verbs frequently given an incorrect lemma by MADAMIRA

Correct lemma	Incorrectly assigned lemma	Transcript	Blog	Number of missed lemmas
جاب ***gæ:b*** **'to bring'**	جهب	97%	92%	765
حسّ ***ħɛs:*** **'to feel'**	أحسّ، حسس، حشّ، أحشّ	89%	88%	222
غسل ***ɣɛsɛl*** **'to wash'**	أغاث، غسّل	65%	46%	19
ساب ***sæ:b*** **'to leave behind'**	سيّب، ثيّب، هسيبي، سبّن، سيب، سبة، سابّ، أثاب	22%	22%	131
طلع ***tˤalaʕ*** **'to ascend'**	أطلع، اطّلع	11%	15%	59
لقى ***lʌʔʌ*** **'to find'**	لقي، تلاقى، لقيا، لاقى	6%	10%	56

The information from which Table 5.8 is drawn is valuable to improving annotation because it provides the most missed lemmas along with the incorrect lemmas that are frequently used for each verb. Using this data, a supplemental program aimed at correcting such mistakes, along with others was created during the annotation process and is discussed in greater detail in Appendix A.

This usefulness of the supplemental program in a study of register variation is seen in its ability to improve MADAMIRA's recall and precision for part-of-speech tagging. Table 5.9 includes the recall and precision scores for the blog sub-corpus from Table 5.1, as well as the scores when the corpus is annotated by both MADAMIRA and the supplemental program.

Table 5.9: Improvement to MADAMIRA's recall and precision by the supplemental program when applied to the blog sub-corpus

	Recall	**Precision**
MADAMIRA	0.920	0.922
MADAMIRA+ Supplemental	0.922	0.944

Although recall is minimally affected, precision increases from 0.922 to 0.944. This increase is sizeable and gives hope that Egyptian Arabic annotators will be able to tag part of speech as effectively as annotators for English. This increase provided by the supplemental program could be discounted since the program itself was written specifically for the blog sub-corpus; however, its changes to MADAMIRA focus on high frequency words which appear in any Egyptian Arabic text. Also, this comparison was done on the blog sub-corpus, which means that the increase is not due to a large number of proper nouns being correctly identified. This increase suggests that even when applied to another corpus, the supplemental program would improve the precision of MADAMIRA.

Summary

This chapter demonstrated that although automatic annotation of Egyptian Arabic is not perfect, a program that achieves accuracy at the same levels as MADAMIRA would have been accurate enough to perform this analysis on the verbs of CALM without manual corrections. The areas in which MADAMIRA struggled were discussed and ways to improve automatic annotation were

also presented. This chapter should give hope that automatic annotation of Egyptian Arabic is becoming more and more reliable with simple fixes that bolster accuracy.

Chapter 6: Conclusions

Introduction

In this chapter, answers to the questions of this thesis will be reviewed, in addition to a discussion of its limitations. The final section will introduce possible topics for future research based upon the findings.

Answers to the Research Questions

This thesis conducted a look into the use of verbs in two potential registers of Egyptian Arabic in order to answer the questions set forth at the beginning of this thesis.

1. Is there enough evidence of register variation between movie/television transcripts and blog posts to warrant a more thorough investigation?
2. Is there an automatic annotator that is accurate enough to aid in a study of register?
3. Are there ways that the automatic annotators can be improved?

The results show that there is significant variance in the usage of verbs in the two sub-corpora. These differences are consistent with variations found between other registers in previous multi-dimensional analyses. Therefore, there is enough evidence to warrant a more thorough investigation into how these two corpora differ. These results also lay the groundwork for future studies by providing a description of some of the dimensions of Egyptian Arabic based upon empirical data. If multi-dimensional analyses are to be conducted on Egyptian Arabic, a clearer understanding of the features of each dimension is needed. This thesis is a start to organizing those features.

Despite the challenges to annotating Egyptian Arabic, including a non-standard orthography, the automatic tagger was able to produce results that were not significantly

different from those produced through a process of manual correction of mistakes. This suggests that an annotator achieving an accuracy of 92% is sufficient for a study of verbs. More studies are needed to confirm this; however, this thesis should engender more trust in MADAMIRA and other annotators that achieve similar results.

It was also discovered through this thesis that some proper nouns thought to be ambiguous with verbs were not so. This will help improve annotation efforts especially with corpora of movie and television transcripts which include the frequent use of names. This coupled with other steps aimed at improving annotation accuracy did in fact raise the precision score of MADAMIRA. This should similarly be encouraging for those waiting for the production of a more accurate annotator for Egyptian Arabic.

Limitations

Although this thesis demonstrates the differences in the use of verbs in the blog and transcript sub-corpora of CALM, it is not without limitations. The blog corpus is composed of several registers of the language. To determine the true nature of the difference between the transcript corpus and the blog corpus, the latter should be further divided into distinct sub-registers.

Another limitation due to the scope of this thesis was the annotation of only one part of speech. This is especially important for the discussion on the effectiveness of MADAMIRA. Although the automatic tagger was found to provide results consistent with manually corrected annotations, it is unknown whether the tagger would perform as well on other parts of speech.

Future Study

Although transcripts of movie and television were used to represent speech, this is a claim that is still being debated. Therefore, the variation found between the two corpora could simply be the variation that exists between dialogue and narration which could both be sub-registers within the register of written Egyptian Arabic. A possible way to answer this question could be to compare written dialogues in the blog corpus to the dialogues in the transcript corpus.

A future study could take a closer look into the disparity in the frequency of verbs. This difference could be due to something as simple as the need to describe movement since it must be described in written material. However, it could also represent a change in style which could use more verbs that take complements or use fewer active participles.

Another question that can be explored concerns the increased use of the HABITUAL in the blog sub-corpus. Is this verbal mood more common in narrative language or is it the feature of another dimension represented in the blogs? Further investigation into the use of this mood can improve curriculum of Egyptian Arabic as native English speakers who often struggle to use it with native-like accuracy can receive clearer guidance and instruction.

As these questions are answered, our understanding of the use of Egyptian Arabic in speech and writing will improve. Such studies will also lead to more comprehensive explanations of grammar that take into account the differences that occur in both the oral and literate dimensions. Although many are waiting for an improvement in automatic annotation tools and quality of corpora, this thesis has demonstrated that much can still be learned from the materials that are currently available.

References

Abdel-Massih, Ernest T., Zaki N. Abdel-Malek & El-Said Badawi. 2009. *A Reference Grammar of Egyptian Arabic*. Washington, D.C.: Georgetown University Press.

Abo Bakr, Hitham, Khaled Shaalan & Ibrahim Ziedan. 2008. A Hybrid Approach for Converting Written Egyptian Colloquial Dialect into Diacritized Arabic. *The 6th International Conference on Informatics and Systems, INFOS 2008, The Special Track on Natural Language Processing*, 27-33. Cairo: Faculty of Computers and Information, Cairo University.

Aboul-Fetouh, Hilmi M. 1969. *A Morphological Study of Egyptian Colloquial Arabic*. The Hague: Mouton.

Alasmari, Jawharah, E. Atwell & J. Watson. 2017. Using the Quranic Arabic Corpus for Comparative Analysis of the Arabic and English Verb Systems. *International Journal on Islamic Applications in Computer Science and Technology*, 5(3). 1–8.

Alian, Marwah & Arafat Awajan. 2018. Arabic Tag Sets. *Proceedings of SAI Intelligent Systems Conference*, 592–606. Springer.

Almeman, Khalid & Mark Lee. 2013. Automatic Building of Arabic Multi Dialect Text Corpora by Bootstrapping Dialect Words. *Proceedings of 2013 First International Conference on Communications, Signal Processing, and their Applications*, 1–6. Sharjah, UAE: Institute of Electrical and Electronics Engineers (IEEE).

Al-Sabbagh, Rania & Roxana Girju. 2012a. YADC: Yet Another Dialectal Arabic Corpus. *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC '12)*, 2882–2889. Istanbul: European Language Resources Association.

Al-Sabbagh, Rania & Roxana Girju. 2012c. A Supervised POS Tagger for Written Arabic Social Networking Corpora. (Ed.) J. Jancsary. *11th Conference on Natural Language Processing (KONVENS)* 5. 39–52. Vienna, Austria: ÖGAI.

Anthony, L. 2018. *AntConc (Version 3.5.7)*. Tokyo, Japan: Waseda University. http://www.laurenceanthony.net/software.

Arfath Pasha, Mohamed Al-Badrashiny, Mona T. Diab & Ahmed El Kholy. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC '14)*, 1094-1101. Reykjavik: European Language Resources Association.

Atkins, Sue, Jeremy Clear & Nicholas Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7(1). 1–16.

Baroni, Marco & Silvia Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. *Proceedings of the Fourth Language Resources Evaluations Conference (LREC '04)*, 1313-1316. Lisbon, Portugal: European Language Resources Association.

Bentley, Randell. 2015. *Conditional Sentences in Egyptian Colloquial Arabic and Modern Standard Arabic: A Corpus Study*. Brigham Young University MA Thesis.

Bex, Tony. 1996. *Variety in Written English*. New York: Routledge.
Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.
Biber, Douglas. 1993a. Representativeness in Corpus Design. *Literary and Linguistic Computing* 8(4). 243–257.
Biber, Douglas. 1993b. The Multi-Dimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings. *Computers and the Humanities* 26(5/6). 331–345.
Biber, Douglas. 1999. *Longman Grammar of Spoken and Written English*. England: Harlow.
Biber, Douglas. 2006. *University Language: A Corpus-Based Study of Spoken and Written Registers*. Philadelphia: John Benjamins.
Biber, Douglas. 2016. Using Multi-Dimensional Analysis to Explore Cross-Linguistic Universals of Register Variation. *Genre- and Register-Related Discourse Features in Contrast*. Philadelphia: John Benjamins.
Biber, Douglas & Susan Conrad. 2001. Register Variation: A Corpus Approach. *The Handbook of Discourse Analysis*, 175-196. Massachusetts: Blackwell Publishers.
Biber, Douglas, Mark Davies, James K. Jones & Nicole Tracy-Ventura. 2006. Spoken and Written Register Variation in Spanish: A Multi-Dimensional Analysis. *Corpora* 1(1). 1–37.
Biber, Douglas & Jesse Egbert. 2018. *Register Variation Online*. New York: Cambridge University Press.
Biber, Douglas, Jesse Egbert & Mark Davies. 2015. Exploring the Composition of the Searchable Web: A Corpus-Based Taxonomy of Web Registers. *Corpora* 10(1). 11–45.
Bjøru, Øyvind. 2018. Transitivity Alternations in the Derived Verbal Stems in Arabic. Presented at Jil Jadid 2018. Austin, TX.
Brysbaert, Marc & Boris New. 2009. Moving Beyond Kučera and Francis: A Critical Evaluation of Current Word Frequency Norms and The Introduction of a New and Improved Word Frequency Measure for American English. *Behavior Research Methods* 41(4). 977–990.
Buckwalter, Tim & Dilworth Parkinson. 2011. *A Frequency Dictionary of Arabic: Core Vocabulary for Learners* (Routledge Frequency Dictionaries). New York: Routledge.
Derczynski, Leon, Alan Ritter, Sam Clark & Kalina Bontcheva. 2013. Twitter Part-of-Speech Tagging for All: Overcoming Sparse and Noisy Data. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2013)*, 198–206.
Diab, Mona, Kadri Hacioğlu & Daniel Jurafsky. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. *Proceedings of HLT-NAACL 2004: Short Papers*, 149–152. Association for Computational Linguistics.
Dickins, James. 2017. The Pervasiveness of Coordination in Arabic, with Reference to Arabic>English Translation. *Languages in Contrast* 17(2). 229–254.
Doss, Madiha. 2011. Arabic and the Media Ḥāl id-Dunyā An Arabic News Bulletin in Colloquial (’Āmmiyya). *Arabic and the Media*, 123–140. Leiden, The Netherlands: Brill.
Duh, Kevin & Katrin Kirchhoff. 2005. POS Tagging of Dialectal Arabic: A Minimally Supervised Approach. *Proceedings of the ACL Workshop on Computational Approaches to Semitic Languages*, 55–62. Association for Computational Linguistics.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19(1). 61–74.
Egbert, Jesse. 2019. Corpus Design and Representativeness. *Multi-Dimensional Analysis*, 2–42. London: Bloomsbury.
El Dik, Dena & Emad Iskander. 2019. *Yalla!:Let's Learn Egyptian Colloquial Arabic Verbs*. Cairo: The American University in Cairo.
El-Tonsi, Abbas. 1982. *Egyptian Colloquial Arabic: A Structure Review*. Vol. 1. Cairo, Egypt: American University in Cairo.
Fakhri, Ahmed. 2009. Rhetorical Variation in Arabic Academic Discourse: Humanities versus Law. *Journal of Pragmatics* 41(2). 306–324.
Ferguson, Charles A. 1983. Sports Announcer Talk: Syntactic Aspects of Register Variation. *Language in Society* 12(2). 153–172.
Forchini, Pierfranca. 2012. *Movie Language Revisited: Evidence from Multi-Dimensional Analysis and Corpora*. Bern: Peter Lang.
Friginal, Eric. 2009. *Language of Outsourced Call Centers: A Corpus-Based Study of Cross-Cultural Interaction*. Philadelphia: John Benjamins.
Gries, Stefan Th. 2006. Exploring Variability within and between Corpora: Some Methodological Considerations. *Corpora* 1(2). 109–151.
Habash, Nizar, Ramy Eskander & Abdelati Hawwari. 2012. A Morphological Analyzer for Egyptian Arabic. *Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology*, 1–9. Association for Computational Linguistics.
Habash, Nizar Y. & Owen C. Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for The Arabic Dialects. *The 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL*, 681-688. Sydney: Association for Computational Linguistics.
Hassan, Gadalla. 2000. *Comparative Morphology of Standard and Egyptian Arabic*. LINCOM EUROPA.
Henen, David. 2018. *"ya" between Vocative and Non-Vocative Use in Egyptian Film Language A Corpus Analysis: Pragmatic Functions and Formal Features*. Egypt: American University in Cairo.
Hiltunen, Turo. 2016. Passives in Academic Writing: Comparing Research Articles and Student Essays across Four Disciplines. *Corpus Linguistics on the Move*, 132–157. Boston: Brill Rodopi.
Hinds, Martin & El-Said Badawi. 1986. *A Dictionary of Egyptian Arabic*. Beirut: Librairie Du Liban.
Hunston, Susan. 2002. *Corpora in Applied Linguistics* (Cambridge Applied Linguistics). 4th edn. Cambridge: Cambridge University Press.
Hussein, Mona. 2016. *Propositional and Non-Propositional Functions of /Keda/ in the Language of Egyptian Film*. Egypt: American University in Cairo.
Ismail, Ahmad. 2015. *ṭab asta'zen ana ba'a: A Corpus-Based Study of Three Discourse Markers in Egyptian Film Language*. Egypt: American University in Cairo.

Johnstone, Barbra. 1990. Orality and Discourse Structure in Modern Standard Arabic. In Mushira Eid (ed.), *Perspectives on Arabic Linguistics*, 215–233. Amsterdam: John Benjamins.
Johnstone, Barbra. 2008. *Discourse Analysis*. Malden, MA: Blackwell.
Jurafsky, Daniel & James H. Martin. 2009. *Speech and Language Processing*. New Jersey: Pearson Prentice Hall.
Kübler, Sandra & Heike Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora*. New York: Bloomsbury.
Leech, Geoffrey. 2007. New Resources, or Just Better Old Ones? The Holy Grail of Representativeness. *Corpus Linguistics and the Web*. New York: Rodopi.
Leech, Geoffrey Neil. 1992. 100 Million Words of English: The British National Corpus (BNC). *Language Research,* 28(1): 1-13.
Leech, Geoffrey & Nicholas Smith. 2000. *Manual to Accompany the British National Corpus (Version 2) with Improved Word-Class Tagging*. Lancaster: UCREL.
Maamouri, Mohamed, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash & Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. *Ninth International Conference on Language Resources Evaluation (LREC '14)*, 2348–2354. Reykjavik: European Language Resources Association.
Maati, Yousef. 2008. *ħʌsʌn wi murʔosˁ*. Cairo, Egypt: Al-Dār al-Masriya al-Lubnāniya.
Macaulay, Marcia. 1990. *Processing Varieties in English: An Examination of Oral and Written Speech across Genres*. UBC Press.
Mäkinen, Martti & Turo Hiltunen. 2016. Creating a Corpus of Student Writing in Economics: Structure and Representativeness. *Corpus Linguistics on the Move*, 59–84. Boston: Brill.
McCarthy, Michael & Michael Handford. 2004. "Invisible to Us": A Preliminary Corpus-Based Study of Spoken Business English. *Discourse in the Professions: Perspectives from Corpus Linguistics*, vol. 16, 167–201. John Benjamins Publishing.
Qafisheh, Hamdi A. 1992. *Yemeni Arabic Reference Grammar*. Kensington, MD: Dunwoody Press.
Rayson, Paul & Roger Garside. 2000. Comparing Corpora Using Frequency Profiling. *Proceedings of the Workshop on Comparing Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics*, 1–6. Hong Kong University of Science and Technology (HKUST).
Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge, England: Cambridge University Press.
Ryding, Karin C. 2006. Teaching Arabic in the United States. *Handbook for Teaching Arabic Language Professionals in the 21st Century*, 13–20. 2nd edn. New York: Routledge.
Sayed, Mukhtar. 2018. *maʕleš maʕleš: A Corpus-Based Study on the Discourse Marker maʕleš*. Egypt: American University in Cairo.
Sharoff, Serge. 2006. Creating General-Purpose Corpora Using Automated Search Engine Queries. *Wacky! Working Papers on the Web as Corpus,* 63–98. Bologna: GEDIT.
Sinclair, John. 2004. Corpus Creation. In Geoffrey Sampson & Diana McCarthy (eds.), *Corpus Linguistics: Readings in a Widening Discipline*, 78–84. New York: Continuum.

Staples, Shelly. 2016. Identifying Linguistic Features of Medical Interactions: A Register Analysis. *Talking at Work*. London: Palgrave Macmillian.

Tamis, Rianne & Janet Persson (eds.). 2013. Sudanese Arabic-English; English-Sudanese Arabic: A Concise Dictionary. SIL International.

Taylor, Christopher John. 2004. The Language of Film: Corpora and Statistics in the Search for Authenticity. Notting Hill (1998), a Case Study. *Miscelánea* 71–86.

Teubert, Wolfgang. 2005. My Version of Corpus Linguistics. *International Journal of Corpus Linguistics* 10(1). 1–13.

Thompson, Paul. 2004. Spoken Language Corpora. *Developing Linguistic Corpora: A Guide to Good Practice*, ed. M. Wynne. Oxford: Oxbow Books: 1-16.

Toutanova, Kristina & Christopher D. Manning. 2000. Enriching the Knowledge Sources Used in a Maximum Entropy Part-of-Speech Tagger. *Proceedings of the 2000 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora: Held in Conjunction with the 38th Annual Meeting of the Association for Computational Linguistics,* vol. 13, 63–70. Association for Computational Linguistics.

Tseng, Huihsin, Daniel Jurafsky & Christopher Manning. 2005. Morphological Features Help POS Tagging of Unknown Words across Language Varieties. *Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing*, 32-39. Jeju Island, Korea: Asian Federation of Natural Language Processing.

Webb, S. & M. P. H. Rodgers. 2009. The Lexical Coverage of Movies. *Applied Linguistics 30(3).*407–427.

Wehr, Hans. 1994. سبت. (Ed.) J Milton Cowan. *A Dictionary of Modern Written Arabic*. Urbana, IL: Spoken Language Services.

Wilson, Andrew. 2013. Embracing Bayes Factors for Key Item Analysis in Corpus Linguistics. In M Bieswanger & A Koll-Stobbe (eds.), *New Approaches to the Study of Linguistic Variability*, 3–9. Frankfurt: Peter Lang GmbH.

Appendix A: Supplemental Annotator

Resolving Ambiguities

Although MADAMIRA facilitated the tagging of the verbs, it—as with all other automatic taggers—is not completely accurate. The errors made were largely due to lexical ambiguity, morphological ambiguity, or unfamiliar verbs. These mistakes were corrected manually and with a supplemental program that I created to fix the high frequency errors. This appendix discusses the frequent errors and the ways in which some were addressed with the supplemental program.

Egyptian Arabic contains many word forms that are lexically ambiguous especially when diacritic marks are not included. A small sample of ambiguous forms is given in Table A.1. The majority of the lexical ambiguities had to be corrected manually but some were resolved with the two programs used for annotation.

Table A.1: Words that are lexically ambiguous in Egyptian Arabic

قوله	*ʔullu* "tell him"	*qaulu* "his saying"
كلمته	*kɛlimtu* "I talked to him"	*kilmitu* "his word"
درس	*dɛrɛs* "he taught"	*dɛrs* "lesson"
قتل	*qatal* "he killed"	*qatl* "killing"

MADAMIRA tended to resolve cases of lexical ambiguity by assigning the word the tag 'NOUN'. For many instances this was appropriate as with the noun حب *ħobb* 'love' and its various inflected forms that could be ambiguous with the inflections of the verb حب *ħabb* "3SG.MASC.PAST.love" or "2SG.MASC.IMPERATIVE.love". These various forms are given in Table A.2. Of all 131 instances of this word in the annotated sub-corpora, 20 were a verb and 111 a noun. The results per sub-corpus are given in Table A.3.

Table A.2: Different conjugations of the verb حب *ħabb* introducing ambiguity with noun forms

	As a Noun	**As a Verb**
حب	"love"	3SG.MASC.PAST.love or 2SG.MASC.IMPERATIVE.love
حبي	"my love"	2SG.FEM.IMPERATIVE.love
حبنا	"our love"	3SG.MASC.PAST.love.1PL
حبك	"your love"	3SG.MASC.PAST.love.2SG
حبكو	"y'all's love"	3SG.MASC.PAST.love.2PL
حبه	"his love"	3SG.MASC.PAST.love.3SG.MASC
حبها	"her love"	3SG.MASC.PAST.love.3SG.FEM
حبهم	"their love"	3SG.MASC.PAST.love.3PL

Table A.3: How often an ambiguous form of حب was either a noun or a verb in each sub-corpus

Corpus	**حب as noun**	**حب as verb**
Blog	66	5
Transcript	45	15

However, MADAMIRA also assigned the 'NOUN' tag to ambiguous lexemes which are predominately used in the sub-corpora as verbs. This type of error was one of the reasons for creating a secondary program to analyze MADAMIRA's annotations. The ambiguous lexemes that were frequently used as verbs were identified during the manual correction of MADAMIRA's annotations. Three such verbs were: روح *ru:ħ*, قوم *ʔu:m*, and خد *xʌd*.

The noun روح *ru:ħ* means "spirit" while the verb containing those same three letters can either be 2SG.MASC.IMPERATIVE.go, 2SG.MASC.IMPERATIVE."return home", or 3SG.MASC.PERFECT. "return home". Out of the total occurrences of the ambiguous forms from both sub-corpora, the verb was used just slightly more than the noun. In the transcript sub-corpus, the verb occurred

53% of the time whereas in the blog corpus, 58% of all the instances were verbs. The exact numbers are provided in Table A.4.

Table A.4: How often an ambiguous form of روح was either a noun or a verb in each sub-corpus

Corpus	روح as noun	روح as verb
Blog	24	33
Transcript	79	99

This suggests that the choice to tag روح *ru:ħ* as a noun would not lead to many more errors than its tagging as a verb. However, if the word in front of روح *ru:ħ* is taken into consideration, this word becomes even less ambiguous. In the transcript sub-corpus, of the 79 instances where روح *ru:ħ* was not a verb, 60 were proceeded by the vocative morpheme يا *yæ:*. This ratio is not as high in the blog corpus with only 11 of the 24 instances of the non-verb روح *ru:ħ* being proceeded by يا *yæ:*. Therefore, even though this word is ambiguous there are ways for an automatic annotator to resolve the ambiguity.

Two other verbs whose tags were easily corrected by the supplemental program were قوم *ʔu:m*, and خد *xʌd*. The noun قوم *ʔu:m* means "people, nation" or it could mean 2SG.MASC.IMPERATIVE."stand". خد *xʌd* as a noun means "cheek" and as a verb it means 2SG.MASC.IMPERATIVE."take" or 3SG.MASC.PERFECT."take". In both the transcript and blog sub-corpora the nouns were rarely used. Their frequencies are given in Table A.5. Those word forms were used to express the verbs 98% of the time. The higher frequency of the verb form over the noun form was not known at the beginning of the annotation process, but as it became apparent, it was written into the supplemental program to change these part-of-speech tags from 'NOUN' to 'VERB'.

Table A.5: How often قوم *ʔu:m* and خد *xʌd* were a noun and a verb in each sub-corpus

Corpus	قوم as noun	قوم as verb	خد as noun	خد as verb
Blog	0	36	3	31
Transcript	1	73	2	208

In addition to mistakes due to lexical ambiguity, there were errors caused by morphological ambiguity. One of the causes of this ambiguity was the circumfixes ما *mæ:* and ش *ʃ* which typically express polarity on verbs but can be applied to several prepositions in Egyptian. Therefore, مفيش *mæfi:ʃ* "there is not", ملكش *mælɛkʃ* "you don't have", and ماله *mæ:lu:* "what's wrong with him" were all tagged as verbs despite being prepositions. Luckily, prepositions belong to a closed class of words, facilitating the ease of correcting this error with the supplementary program.

Another common mistake that was fixed automatically was the tagging of proper nouns as verbs. As discussed in greater detail elsewhere in this thesis, proper nouns in Arabic do not have any distinguishing features and some of them are lexically ambiguous with certain verb forms. The need to correct the mis-tagging of names as verbs arose from the transcripts which start each new line with the name of the character speaking. In some cases, names were repeated more than sixty times in a given movie. Therefore, the creation of the supplemental program facilitated tagging a large number of transcripts.

Words unfamiliar to the annotator also lead to incorrect tags on the verbs. Besides morphology and context, inclusion of a particular word in the automatic annotator's training set also aids in assigning the correct tag to a word. Each time the word appears in the training data, more data is generated to be used in calculating the probability for its correct tag. If the program has not seen the word previously and does not carry distinguishing morphology the annotator likely backs off to a predetermined tag which is usually "NOUN". One way to eliminate this problem is to train the annotator on an enormous tagged corpus. However, regardless of size, the annotator is bound to run into new words.

An example of this can be found with the word دهولتوا *dʌhwɛltu:* 2PL.PERFECT.“throw into confusion”. Although there is morphology suffixed onto the verb that identifies it as a verb, MADAMIRA was unable to correctly identify it, tagging it instead as a noun. This verb is not a high frequency verb, occurring only once in the sub-corpora. It is likely that MADAMIRA did not encounter this word in its training set which contained 27,000 verbs (Habash, Eskander & Hawwari 2012). The sub-corpora contained 66,384 verb tokens, and therefore likely contained many verbs not in the training set.

Even if a verb was in the training data, the morphological feature combinations could obscure the true identity of the word. This occurred with the word ماتبستيش *mætbæsti:ʃ* 2SG.FEM.PERFECT.PASSIVE.“kiss”. MADAMIRA correctly identified the active and positive forms of this word, but the alternations caused by the morphology made it difficult for the word to be recognized. For this reason, the tag “NOUN” was incorrectly assigned. The correction of the errors due to unfamiliarity were almost exclusively done manually. This was also the case for correcting the verbal-category tags.

The assigning of verb category (IMPERFECT, PERFECT, HABITUAL, FUTURE, and IMPERATIVE) is typically straightforward; however, there are a few verb forms that are ambiguous. The command form of verbs for 2SG.MASC and 2SG.PL often resemble other verb forms, especially 3SG.MASC.PERFECT and 3PL.PERFECT. When an imperative begins with a consonant cluster, an alef [ا] is added to break up this cluster causing the IMPERATIVE to become identical to 1SG.IMPERFECT. Table A.6 shows a few examples where this is the case.

Table A.6: Examples of ambiguity that exists among verbal categories

وصل	2SG.MASC.CAUSATIVE.IMPERATIVE.arrive	3SG.MASC.CAUSATIVE.PERFECT.arrive
اشتغل	2SG.MASC.IMPERATIVE.work	3SG.MASC.PERFECT.work 1SG.IMPERFECT.work
ادخل	2SG.MASC.IMPERATIVE.enter	1SG.IMPERFECT.enter

The example of اشتغل shows that there is also ambiguity between 3SG.MASC.PERFECT and 1SG.IMPERFECT when the verb form begins with an alef. This type of ambiguity was corrected manually.

Lemmatization

The lemmatization of verbs also provided opportunities for automatically correcting mistakes. One such mistake was the assignment of different lemmas to a single verb. For example, MADAMIRA assigned one of twelve lemmas for the verb اخد *æxad* "to take" depending upon its conjugation. The different lemmas are given in Table A.7.

Table A.7: The different lemmas assigned to the verb خد *xad* "to take" by MADAMIRA

1	أَخَد
2	أَخَذ
3	اَخَذ
4	خَد
5	اخدّ
6	خُدَّة
7	آخِذ
8	وَخِّد
9	أَخْذ
10	آخَذ
11	أَخِّذ
12	خِدْن

This table was included not to demonstrate any weakness of MADAMIRA, but rather to show that inflectional morphology and orthography can cause problems for even the most

advanced annotators. Some of these lemmas—6,7, and 10—were intended for nouns demonstrating that many verbs are ambiguous with nouns. Lemmas 1 and 5; 2 and 9; and 7 and 10 are identical when the diacritics are removed. As discussed above, because diacritics are infrequently used to disambiguate lemmas they can be removed, limiting the number of lemmas needing correction.

Part of the process of correcting the lemmas was done automatically with the supplemental program. By the end of annotation, this program contained 252 lines of code dedicated to lemmas. More could have been done to automate this process; however, before any lemma was changed automatically, the effect of the change had to be taken into account. Those lemmas assigned to verbs on the basis that it was a noun could not always be automatically changed if the noun was commonly used. Since this thesis only focused on verbs, I did not want to write code that would affect the other parts of speech. This was done with the hope that those using the supplemental program on their corpora could do so knowing that the accuracy of the other parts of speech would be affected minimally.

The effect of orthography on lemma assignment should also be mentioned here. MADAMIRA exceeded expectations in handling spelling variation for many verbs. This is not to say that errors did not occur because of orthography; however, MADAMIRA was robust enough to allow for spelling variations. The error that I would like to highlight occurs when orthographical choices cause one verb to become ambiguous with another. I will use the verb قعد *ʔæʕɛd* "to sit" to demonstrate this point. The first root letter for the root is *qaaf* [ق], which in Standard Arabic is pronounced [q] but in Cairene Egyptian Arabic is typically pronounced as [ʔ]. However, the glottal stop can also be represented with a *hamzaa* [ء] which orthographically is seated on top of an *alef* when it is not surrounded by a closed vowel such as [i], [ɪ], [u], or [ʊ]. If

one is writing based on pronunciation, then "he sat", which is commonly written as قعد could be written اعد reflecting the choice to replace [q] with [ʔ].

In the IMPERFECT, speakers tend to pronounce this verb without the glottal stop, opting instead to simply geminate the *ayn* [ع] which is pronounced [ʕ] (El Dik & Iskander 2019). Therefore, "he sits" can be pronounced يعّد *yuʕ:ud* rather than يقعد *yuʔʕud*. Despite this common pronunciation, the vast majority of instances of the verb قعد are written with the *qaaf* [ق]. In the blog sub-corpus there are 8 instances in the PERFECT out of 115 where the *qaaf* is represented with an *alef*. In the IMPERFECT, neither a *qaaf* nor an *alef* are used in the verb 5 of its 84 instances.

Unfortunately for the automatic annotator, when this verb is spelled without the *qaaf*, it becomes ambiguous with other verbs. In addition to meaning "he sat" اعد can also mean "I count" or "he prepared." In the IMPERFECT, when the *qaaf* is not written, 2SG.FEM.IMPERFECT."sit", تقعدي, becomes تعدي which is ambiguous with 2SG.IMPERFECT."pass", 2SG.FEM.IMPERFECT."count", and 2SG.FEM.IMPERFECT."prepare". Even without the verb قعد, these verbs are ambiguous; however, an undefined orthography for the dialect, can add more ambiguity. Table A.8 shows MADAMIRA's assignment of part of speech and lemma to the verb قعد when it is written without a *qaaf*. The column of the left shows the verb as it appeared in the blog sub-corpus.

Table A.8: How the verb قعد was tagged by MADAMIRA when it appeared without a *qaaf*

Word in Corpus	**Part of speech**	**Lemma**
واعدنا	verb	عاد
اعد	verb	عدّ
اعدت	verb	اعد
واعدت	verb	عدّ
اعدت	verb	عاد
يعد	verb	عدّ
تعد	verb	عدّ
وأعد	verb	عدّ

تعدي	verb	تعدّى
عدت	noun	عدّة
واعد	verb	عدّ

MADAMIRA handles the ambiguity well: only the lemmas تعدّى, عاد, and عدة are inappropriate, in addition to عدّ being assigned as the lemma for واعدت which it could not be. It is interesting to note that the lemma عدّ appears in the blog sub-corpus only seven times which also happens to be the number of times قعد was ambiguous with عدّ. This suggests that for annotating Egyptian blogs it might be productive to spend time determining whether surrounding words can be used to disambiguate these two verbs.

In some instances, the surrounding words and context cannot be used by human annotators to disambiguate two verbs. Above, the example of the Form I verb صرخ *sarax* and the Form II verb صرّخ *sarrax* was used to show that not all Form II verbs have a causative meaning. Both verb forms have the meaning of "to scream." Another verb like this is the verb ظبط *zˤʌbʌtˤ* "to adjust, order, apprehend" and ظبّط *zˤʌbːʌtˤ* which Hinds and Bedawi define as "intensive of *zˤʌbʌtˤ*." These verbs were easily disambiguated for the transcript sub-corpus since the recordings could be used. However, because diacritics are typically not written, this process was difficult for the blog sub-corpus. Context simply could not be used to determine whether the author intended the use of the Form I or the Form II. The native speaker consultants simply could not assign a lemma for these verbs with a high degree of certainty. Therefore, the lemma assigned to them by the automatic annotator is the lemma that they were given except in cases that the annotator assigned a lemma outside of Form I or II. Future studies concerning the disambiguation of these verbs should not consult the blog sub-corpus. The transcript sub-corpus is a better corpus for that task.

Even though the consultants were unable to help in the case of these two verbs, they were heavily relied upon in the disambiguation of many other verbs. Without them, there could be serious questions about the reliability of the annotations since they were completed by a non-native Egyptian Arabic speaker. It is standard to have annotations completed by more than one annotator and then to have the two annotations compared. This is done with the knowledge that annotation is difficult and that mistakes are bound to occur due to either the difficulty of language or lapses in concentration (Kübler & Zinsmeister 2015).

Appendix B: Lemmatization of Egyptian Arabic

Introduction

This appendix explains the rationale behind the lemma orthography used in the sub-corpora of CALM. Although this process is more straightforward for Standard Arabic, it is a little more challenging for Egyptian Arabic due to non-standardized orthography. I will justify the choices that I made as I did not adhere strictly to Standard Arabic orthography to represent the lemmas, nor did I choose to represent all lemmas phonetically as is done in other treatments of Egyptian Arabic.

Lemmatization Differences

Determining how Egyptian Arabic is to be represented orthographically is a contentious subject that I wished to avoid in this thesis; however, because lemmas had to be created, it is one that I am forced to wade into. There are two main ways to represent Egyptian Arabic. The first is to orthographically represent all words shared by both Egyptian Arabic and Standard Arabic using Standard Arabic orthography. The words that are unique to Egyptian Arabic are represented with an orthography that matches pronunciation. The Hinds/Bedawi (1986) dictionary for Egyptian Arabic choses a different path and represents the lemmas according to Egyptian pronunciation. Therefore, there are no lexemes with the letters ث [θ] or ذ [ð] and in some words the ض [dˁ] has been replaced with د [d].

This latter approach can be problematic since some high frequency verbs are loan words from Standard Arabic and, therefore, are expected to be written as they appear in Standard Arabic. For example, the verb حذّر *ħʌð:ar* "he warned", which is represented with this spelling in Standard Arabic, is found in the Hinds/Bedawi dictionary under حزّر *ħʌz:ar*. This would be similar to a dictionary of American English replacing the lexeme for "water" with "wader" as the

latter represents how the word is pronounced. However, there are definite advantages to using this orthography in Egyptian Arabic: chief among them is that learners who have only heard the words in conversation will be able to look them up in the corpus without knowing the Standard Arabic spelling.

The one disadvantage to both systems is that neither reflects the reality of how Egyptians tend to represent their language. Therefore, the lemmas in CALM do not strictly follow Standard Arabic orthography or that put forth in the Hinds/Bedawi dictionary. Instead I based the spelling on the most common orthography as found in the corpus. Therefore, for some words the lemmas are similar to those in the Hinds/Bedawi dictionary, though for others it follows the Standard Arabic spelling. In Table B.1, four words for which I chose to represent the lemmas following Standard Arabic spelling are presented. The numbers in the table correspond to the total occurrences in the blog sub-corpus for each spelling.

Table B.1: The total occurrences in the blog sub-corpus for spelling variants of four words

MSA spelling	Blog sub-corpus frequency	Bedawi/Hinds dictionary spelling	Blog sub-corpus frequency
حذّر	13	حزر	1
أثّر / تأثر	10	أسّر / تأسر	0
ضحك	112	دحك	0
اتضايق	30	اتدايق	3

Each of the verbs in Table B.1 appear in the blog corpus with a higher frequency written following MSA conventions. However, there were instances where the Bedawi/Hinds spelling was more common. A sample of these verbs and their frequencies in the blog sub-corpus is given in Table B.2.

Table B.2: The total occurrences in the blog sub-corpus for the two spelling variants

MSA spelling	Blog sub-corpus frequency	Bedawi/Hinds dictionary spelling	Blog sub-corpus frequency
كذب	4	كدب	16
ملأ	4	ملى	14
أخذ	28	اخد	427

Therefore, based on overall frequency as found in the blog corpus, some of the lemmas are represented following MSA orthography and others following the orthography presented in Hinds/Bedawi. Some may argue that the choice of mixing both MSA and Egyptian spelling for the lemmas is inconsistent, making searching the corpus difficult for those who are not familiar with the dialect and how Egyptians choose to write it. However, this can be solved by programing a search function to be aware of these alternations and return the results for a certain lemma regardless of spelling.

One feature adopted from the Bedawi/Hinds dictionary is the inclusion of an *alef* at the onset of all lemmas belonging to Forms V and VI. Therefore, the lemma for the Form V verb "to talk" is represented as اتكلم instead of تكلم which is how the verb form is represented in MSA. This choice was again based upon the native contributors to the blog corpus who included the *alef* when inflecting the verb for the PERFECT. Since the lemma is based upon the conjugation for the PERFECT, it seemed appropriate to add the *alef* onto the lemma as well.

These choices for lemma form and spelling required me to make changes to the lemmas provided by MADAMIRA as its creators represented them differently. For the most part, the lemma for was the same: both MADAMIRA and I represented the lemmas as the 3SG.MASC.PERFECT for each verb form. However, MADAMIRA does not view the passive verbs as a verb form. Therefore, even though passive verbs in Egyptian are placed into the Form V

pattern, MADAMIRA chose the lemma based upon which verb form was made passive and included another layer of annotation marking the verb as passive. This is because both Form I and II of a verbal root that can be made passive and both look identical as the short vowels are not represented orthographically. MADAMIRA's extra layer of annotation makes it clear just which verb form is being made passive. It also makes clear which verbs have a non-passive Form V.

MADAMIRA's lemmas also have diacritics which help disambiguate the verbs further. Some roots have multiple Form I verbs with different meanings, and the only way to distinguish them is through short vowels. A list of the ones encountered during annotation is given in Table B.3.

Table B.3: A list of verbs whose meanings are distinguished not by form but by short vowels

قِلِق	ʔɪlɪʔ	to become worried	قَلَّق	ʔɛlɛʔ	to cause to worry
تِعِب	tɪʕɪb	to become tired	تَعَب	taʕab	to tire, wear out
بِعِد	bɪʕɪd	to become distant	بَعَد	baʕad	to take away, remove, to distance
شِبِع	ʃɪbɪʕ	to become satiated	شَبَع	ʃabaʕ	to satiate
قِرِف	ʔɪrɪf	to be disgusted, sickened	قَرَف	ʔaraf	to repel, sicken
وِقِف	wɪʔɪf	to come to a stop	وَقَف	waʔaf	to suspend
ثِبِت	sɪbɪt	to become immobile	ثَبَت	sabat	to prove
هَدى	hadа	to set on the right path	هَدى	hadа	to give a gift
خَرَج	xarag	to go out, leave	خَرَج	xarag	to put outside

هان	hæ:n	to become insignificant, of little value	هان	hæ:n	to insult, humiliate

The length of this list demonstrates that for the majority of verb lemmas in the sub-corpora, short vowels were not needed. In fact, the last three verbs in the table demonstrate that even the short vowels may not be enough to disambiguate the lemmas. In the last three rows, there are two Arabic verbs in each cell. The word to the right of the slash is the PERFECT and to the left is the IMPERFECT. In the last three rows, the two PERFECT verb forms are identical which means that the lemmas would be identical since the lemma is taken from the 3SG.MASC.PERFECT. Therefore, more than diacritics was needed to disambiguate these verbs from each other. MADAMIRA solved this problem by adding numbers next to the lemmas to distinguish them.

The diacritics and the numbers are helpful in disambiguating the different verb lemmas; however, in order to finish the task of annotation in a timely manner, the numbers and diacritics were largely removed. This is because they were frequently incorrect causing much more correction than was necessary. All diacritics except for a *shadda* over the second root letter in the lemma were removed. This is because this *shadda* distinguishes the Form I from the Form II verbs. However, removing the short vowels made the lemmas of the verbs in Table B.3 ambiguous; therefore, the short vowels were replaced during annotation. For the verbs in the last three rows, a "_2" was added to the lemmas for "to give a gift", "to put outside", and "to insult, humiliate." This is because the short vowels in the perfect are the same "to set on the right path", "to go out/leave", and "to become insignificant".

Summary

Although lemma orthography is rather straight forward for Standard Arabic, it can be more complicated for Egyptian Arabic due to the lack of standard spelling conventions. For this thesis, lemma orthography was chosen based upon frequency: the most frequent spelling in the blog sub-corpus for each verb became the lemma. In some instances, lemmas were identical to the orthography found in dictionaries of Standard Arabic; however, in other cases, the lemmas reflected pronunciation. Additional changes to the orthography of the lemmas includes the removing of all short vowels and diacritics except from Form II verbs and for those verbs with more than one verb per form.