# Verbs in Egyptian Arabic: a case for register variation

Michael Grant White
Department of Linguistics, Brigham Young University
Provo, Utah, USA 84602
mgrantwhite@gmail.com

Deryle W. Lonsdale
Department of Linguistics, Brigham Young University
Provo, Utah, USA 84602
lonz@byu.edu

## Abstract

The limited availability of Egyptian Arabic (EA) corpus resources, especially speech corpora, has left open opportunity for research into such dialect phenomena as register. In this paper we introduce a new two-million-word EA corpus, CALM. We perform a register analysis on EA between two subcorpora of CALM (i.e. Movies and Blogs), showing several features that vary between the two. A discussion follows about how annotation was carried out automatically, how it was hand-corrected, and what the prospects are for carrying out similar studies using CALM.

## 1 Introduction

The advent of the internet has made written Egyptian Arabic much more accessible than in the past. Traditional sources of written language like books, newspapers, academic journals, and government documents are composed in Modern Standard Arabic which differs morphologically, lexically, and syntactically from Egyptian Arabic. However, as access to the internet spreads, so does the appearance of written Egyptian Arabic on blogs and social media sites. Collections of digital texts provide linguists with a new opportunity to collect large samples of the written dialect from a variety of sources on numerous topics.

One active area of corpus linguists involves the identification and characterization of registers (Gries, 2006), a type of language use that is determined by the situation or circumstance in which the speech act occurs (Johnstone, 2008). Situations that cause speakers to change their lexical and grammatical choices are said to belong to different registers. Some overlap exists between the notions of register and genre (Biber and Conrad, 2001), but a fine-grained distinction between the two is not necessary for this discussion.

Modern studies of register variation rely on annotated corpora. Unfortunately, an annotated corpus containing both written and spoken Egyptian Arabic is not available preventing studies of Egyptian Arabic register variation.

To help explore this problem, this paper introduces a new two-million word corpus of Egyptian Arabic. We perform a preliminary analysis of variation between two registers found in the corpus, on the basis of partial annotation, namely that of verbs. We compare verb frequency, lexical diversity, and other phenomena between two subcorpora and confirm a quantitative difference between two putative registers: Movies and Blogs. We show that, although the examination of a single part of speech cannot capture the true extent of the variance of two registers, it provides a platform from which to launch an in-depth analysis by answering the preliminary questions concerning register analysis for Egyptian Arabic.

We also offer comments on the use of automatic tools for annotation, and mention various types of post-processing that help improve the annotations for corpus analyses like the one discussed here.

## 2 Background

A corpus is a collection of texts or transcriptions gathered for the purpose of conducting an empirical study of language (Hunston, 2002; Kübler and Zinsmeister, 2015), and they have become a valuable resource in nearly every field of linguistics (Teubert, 2005). Corpora assume many different forms and sizes in accordance with their intended

purpose. Varying amounts of corpus content are available across different languages, dialects, and text types.

In this paper we focus on Arabic corpus linguistics and associated annotation and analysis. Several types of corpora are available for Arabic: examples include ones for general language (ArabiCorpus[1]), transcribed speech (CALLHOME Egyptian Arabic and MGB-3) (Canavan et al., 1997; Ali et al., 2017), specialized (The Quranic Arabic Corpus[2]), parallel (OPUS[3]), and learner corpora (Arabic Learner Corpus) (Alfaifi and Atwell, 2015). With these corpora and others, our understanding of Arabic and how it is used has increased (Buckwalter and Parkinson, 2011; Bentley, 2015; Ismail, 2015; Alasmari et al., 2017; Dickins, 2017; Henen, 2018). One area that has been largely overlooked in Arabic corpus studies is discourse analysis, especially for learner Arabic (Ryding, 2006) and for different dialects.

### 2.1 Register

Research on registers in a corpus targets the identification of features that distinguish one register from another. Biber (1993) gives eight parameters to use in classifying registers: the primary channel of delivery, format, setting, addressee, addressor, factuality, purpose, and topic. Separating texts according to these parameters helps establish appropriate registers.

Once texts are collected in a principled way with the aim of representativeness, feature-based register analysis can begin. Each register has characteristics or dimensions of language associated with it (Biber, 1993) based on pertinent lexical, grammatical, and syntactic features. For example, the narrative dimension in English is characterized by past tense verbs, third person pronouns, public verbs, synthetic negation, and present participle clauses. By comparing the frequency of the features in different types of texts, the dimensions in which they fall can be determined. The dimensions for each text type are then taken as characteristic of the register to which the texts belong.

In this paper we deal with two primary registers: the oral and the literate. In daily life, the most common registers in the oral dimension come from spontaneous speech. However, this type of data is currently costly to collect and transcribe. Often corpora contain scripted speech from television and movies to represent oral language. Spontaneity suffers somewhat: since scripted speech is first written, it 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.

This has led to debate within the corpus community, with some asserting that scripted language reflects artificial settings created by the same author, and hence may not be completely realistic (Sinclair, 2004). On the other hand, others have found only minimal differences between movie language and spontaneous speech (Taylor, 2004; Brysbaert and New, 2009; Forchini, 2012). This issue becomes particularly interesting in Egyptian Arabic.

Written registers are very frequently used in corpus studies. Sample text types that fall within the written registers include literature, newspaper articles, academic articles, encyclopedia entries, personal correspondence, and official documents (Biber and Conrad, 2001; Biber et al., 2006). With its widespread availability of texts, the internet is also a source for written register data. Biber et al. (2015) found that English texts from the internet can be categorized into several registers: narrative, information description/explanation, opinion, interactive discussion, how-to/instructional, informational persuasion, lyrical, spoken, and hybrid.

### 2.2 Arabic corpus analysis

The oral/written distinction and corpus registers in general is more complex and nuanced for Arabic's diglossic situation: the standarized version of the language, MSA, is used in many written situations, whereas a wide array of dialects is used for everyday spoken language. Collecting, annotating and analyzing corpus data is hence more complicated and much remains to be done in examining the variation that exists between spoken and written Arabic. Examples of such work include Fakhri's (2009) investigation of the variation between academic Arabic in the disciplines of the humanities and the law. Johnstone (2008) examined Arabic expository prose and identified the use of three features—repetition, parataxis, and formulaicity—which are typically associated with spoken language.

Several Egyptian Arabic film and television transcript corpora have been used in recent stud-

[1] See http://arabiCorpus.byu.edu.
[2] See http://corpus.quran.com.
[3] See http://opus.nlpl.eu.

ies. Hussein (2016) used a corpus of Egyptian movie transcripts to study the pragmatic and syntactic functions of the Egyptian word[4] كده kɪdʌ. This 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 pre-1990, which makes the content somewhat dated for contrasting it with recent content such as internet texts.

Such a corpus was used by Sayed (2018) to study the use of the discourse marker معلش maʕleʃ. This corpus contains transcripts of 76 episodes from the 2017-2018 Egyptian television serial سابع جار sæ:biʕ ga:r. One potential weakness to using this corpus in a register study is that all of the transcripts come from one television show. Most of the content is produced by only a handful of speakers/characters, calling into question representativeness.

The issue of representativeness is also important in choosing a suitable blog corpus. Two general Egyptian Arabic blog corpora are the Arabic Multi-Dialect Text Corpus (Almeman and Lee, 2013), which contains thirteen million words and Yet Another Dialectal Corpus (YADC) (Al-Sabbagh and Girju, 2012b) 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.

In building a corpus with web searches, the frequency of the seeds is important for representativeness (Sharoff, 2006; Biber et al., 2015). No such frequency lists exist for Egyptian Arabic, and 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. One downside to using function words is that many of them are found in several dialects (Qafisheh, 1992; Tamis and Persson, 2013). For our purposes, a corpus that contains only Egyptian Arabic is preferable.

## 3 Introducing CALM

Because of the need for a sizable corpus of Egyptian Arabic language, we collected and annotated a new corpus designed in an attempt to more accurately represent both oral and written language. This paper introduces CALM (**C**orpus **al**-**L**ogha al-**M**usriya, Corpus of Egyptian) a two-million-word corpus of Egyptian Arabic. CALM contains transcripts from 65 movies (comprising 655,858 word tokens), 88 scripted television programs (396,734 word tokens), and internet texts (1,092,442 word tokens). Some of the content has been annotated, as described in this paper, and annotation is ongoing. The corpus is available via download[5].

For the purposes of this paper, two subcorpora were extracted from CALM: a subcorpus of movie/television transcripts, and a subcorpus of internet texts.

### 3.1 The transcript subcorpus

The transcripts of CALM make up the largest known collection of transcribed Egyptian Arabic movies and TV programs produced in Egypt and written for Egyptian audiences. In other languages a quicker and cheaper method to build a comparable corpus would be from subtitles, but in Arabic foreign movies are subtitled using MSA. Only movies and programs popular to Egyptian audiences were selected for transcription based on the belief that they contain more mainstream language and are written by those who are able to skillfully mimic everyday speech.

Note that, as mentioned earlier, some debate exists about whether movie transcripts truly represent spontaneous speech (vs. the author's creative voice). A comparison of a script in CALM from the Egyptian film حسن ومرقص fiʌsʌn wi murʔosˁ (Maati, 2008) with the movie transcripts reveals several instances where actors stray from the script, both omitting words from the script and adding their own content spontaneously.

Most movies and TV programs are from the year 2000 and later. No conscious effort was made to choose movies and TV based upon genre, or to balance the content across genres. However, care was

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

[5]See http://linguistics.byu.edu/thesisdata/CALMcorpusDownload.html.

taken to make sure that one genre does not dominate the subcorpus created for movies and TV.

Once a movie was selected for inclusion in the corpus, it was transcribed and then reviewed for accuracy by a native Egyptian speaker. A second reviewer was used to determine the ability of the reviewers to catch all of 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.

### 3.2 The blog subcorpus

The other content in CALM was created from internet texts and will be called the blog subcorpus. Although internet texts can be classified into many different genres (Biber et al., 2015), in this paper they will be treated as a single register. We exclude internet texts that contain transcriptions of speeches, movies, television programs, and songs. Some of the blog texts were collected from the internet based upon seeded n-gram searches via Bing and Google, as discussed in the previous section, though this time relying on frequent dialect-specific words to decrease the chances of dialect mixing. We also used BootCat, a do-it-yourself web-to-corpus text conversion pipeline (Baroni and Bernardini, 2004), to find, scrape, and convert other webpages written in Egyptian Arabic into text files. A cursory review of the files was completed to remove non-EA texts that were returned by the process. However, some MSA is contained in CALM because it is interwoven throughout posts written in Egyptian; posts completely written in MSA, though, were removed.

EA exhibits numerous orthographic, lexical, morphological, and syntactic differences from MSA that will be familiar to many Arabists (El-Tonsi, 1982; Hassan, 2000; Ryding, 2005; Abdel-Massih et al., 2009). Even the representation of lemmas (base forms, or dictionary citation forms) and their orthography varies across EA dictionaries, necessitating a custom representation for CALM annotation. A discussion of these is beyond the scope of this paper, but an extensive list of the ones relevant to CALM corpus creation and annotation is available (White, 2019, forthcoming).

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 it differs from Egyptian Arabic lexically, syntactically, and morphologically, 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 EA.

Once these corpora have been decided upon, the next step is to determine the features that should be counted and compared. These features can be large or relatively few in number. In the next section, we perform a feature-based examination of two subcorpora from CALM which, we will show, represent two different registers of EA.

## 4 Case study: verb register variation

To assure tractability for this study, two subcorpora were created from CALM: (1) the Movies subcorpus, consisting of transcriptions from movies (113,163 word tokens) and television shows (115,236 word tokens), for a total of 228,399 word tokens including 38,768 verbs; and (2) the Blogs subcorpus, containing 141,318 word tokens and 27,616 verbs. While not exactly balanced, they are of reasonably comparable size.

For this study only the verbs were annotated, in part because they are slightly easier to identify and annotate than nouns and adjectives (Al-Sabbagh and Girju, 2012a), and because of their widespread use in determining register (Ferguson, 1983; Friginal, 2009; Staples, 2016). Table 1 gives sample annotations for several verbal features.

In this section, then, we perform a register variation analysis on verb features to characterize the two dimensions of content in CALM: oral versus literate (or spoken versus written), as represented by the Movies and Blogs subcorpora, respectively.

The first step was to annotate each verb in the two subcorpora. Once each verb was assigned a part-of-speech tag, a verbal category, and a lemma, each of these features were counted from each subcorpus and compared in order to determine whether verbs are used differently. Since the size of the two subcorpora was not exactly the same, counts from each were normalized. We used two

| | |
|---|---|
| IMPERFECT | ممكن تدلني على حد يديني عنوانه |
| PERFECT | مراد لو كان عايز يتجوزني فعلا مكنش ادى ودنه لسارة |
| IMPERATIVE Positive | طب وانا ذنبي ايه يا شوقي بيه ادیني التمثالين بتوعي وحاجتك جاية في الطريق |
| IMPERATIVE Negative | متدنيش الوش الكئيب ده. كفاية البومة اللي عندي في البيت |
| HABITUAL | فولتارين! من امتى بندي حقن فولتارين احنا |
| FUTURE | على كل حال أنا مش هديهم اي كلمة إلا لما أسمع رأيك بقى |

Table 1: Sample feature-based verb annotations

statistical tests to compute significance: (1) log-likelihood because of its frequent appearance in corpus linguistics studies (Wilson, 2013); and (2) the Bayesian Information Criterion (BIC) for its reliability over chi square when dealing with word counts that fall on either end of the frequency spectrum (Dunning, 1993; Rayson and Garside, 2000).

Verbs are more common in the Blogs subcorpus than in Movies (see Table 2). However, not all verbal categories (IMPERFECT, PERFECT, IMPERATIVE, HABITUAL, and FUTURE) in Blogs occur more frequently than in Movies. Table 3 shows the category differences, all of which are statistically significant except for FUTURE. 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 Blogs, causing counts of the verbal categories taken out of the total number of words to be misleading (Gries, 2006).

| Total | Movies | Blogs |
|---|---|---|
| # of words | 228,399 | 141,318 |
| # of verbs | 38,768 | 27,083 |
| % of verbs | 16.97 | 19.54 |

Table 2: Totals and percentages of verbs

| Category | % of Words | | % of Verbs | |
|---|---|---|---|---|
| | Movies | Blogs | Movies | Blogs |
| Imperfect | 6.56 | 7.90 | 38.67 | 40.45 |
| Perfect | 5 | 6.4 | 26.86 | 32.54 |
| Habitual | 1.83 | 2.46 | 10.8 | 12.58 |
| Imperat. | 2.57 | 1.52 | 15.15 | 7.8 |
| Future | 1.44 | 1.3 | 8.51 | 6.67 |

Table 3: Frequency of verbal categories

When factored by the total amount of words in each subcorpus, IMPERFECT is significantly more frequent in Blogs; however, this significance disappears when the frequency is compared with the total number of verbs. The opposite is true of the verbs hosting the FUTURE morpheme. Although the comparative frequency of this verb in Movies was insignificant when compared to the total words, its frequency becomes significant when compared only to verbs. PERFECT and HABITUAL are significantly higher in Blogs by both comparisons.

The IMPERATIVE represents the largest difference in usage between the two subcorpora. We investigate further by separating the imperatives into four categories: negative IMPERATIVE, positive 2SG.MASC.IMPERATIVE, positive 2SG.FEM.IMPERATIVE, and positive 2PL.IMPERATIVE.

Instead of comparing the frequencies of the imperatives against the total number of words in the corpus, we compared them to the total number of verbs. The numbers for each category of imperative are given in Table 4.

| Type | Per 100 . . . | Movies | Blogs |
|---|---|---|---|
| Negat. | verbs | 1.1 | 0.7 |
| | imperatives | 7.29 | 9.01 |
| Male Posit. | verbs | 10.14 | 5.88 |
| | imperatives | 66.89 | 75.83 |
| Female Posit. | verbs | 3.31 | 0.69 |
| | imperatives | 21.87 | 8.87 |
| Plural Posit. | verbs | 0.6 | 0.49 |
| | imperatives | 3.95 | 6.3 |

Table 4: Frequency of imperatives across subcorpora

The frequencies of IMPERATIVEs in Movies are all significantly higher than in Blogs except for the positive 2PL.IMPERATIVE. However, the table reveals that as a percentage of the total imperatives used, the Blogs uses the positive 2SG.MASC.IMPERATIVE and positive 2PL.IMPERATIVE significantly more than Movies. Negative imperative use is significantly greater in Movies compared to all verbs, but not significantly greater when compared to

all IMPERATIVE verbs. Only the positive 2SG.FEM.IMPERATIVE remains more frequent in Movies regardless of its comparison set.

Another feature used to prove register variation is lemma frequency across verbs from each register. Among the verbs that occur in either subcorpus with a frequency of over 100, seventeen verbs show a frequency significantly higher in one register over the other (see Table 5). Table 6 illustrates

| More common in... | | | |
|---|---|---|---|
| Movies | | Blogs | |
| اتفضل | please, come in | بدأ | to begin |
| خشّ | leave | دخل | to enter |
| هدي | to calm oneself | كتب | to write |
| أكل | to eat | حاول | to try |
| ساب | to leave | لقي | to find |
| استنى | to wait | حسّ | to feel |
| شرب | to drink | فتح | to open |
| مشي | to walk | رد | to respond |
| | | قرأ | to read |

Table 5: Contrastive distribution of high-frequency verbs across registers

percentage of IMPERATIVE verb forms in each subcorpus for these seventeen verbs.

| Word | | % Imperat. | Meaning |
|---|---|---|---|
| اتفضل | M | 93.6 | please, come in |
| هدي | M | 83.2 | to calm oneself |
| استنى | M | 56.8 | to wait |
| خشّ | M | 43.4 | to enter |
| فتح | B | 32.3 | to open |
| ساب | M | 31.7 | to leave |
| مشي | M | 29.9 | to walk |
| رد | B | 26.7 | to respond |
| دخل | B | 20.5 | to enter |
| قرأ | B | 15.9 | to read |
| أكل | M | 11.3 | to eat |
| كتب | B | 11.1 | to write |
| حاول | B | 10.2 | to try |
| شرب | M | 5.6 | to drink |
| بدأ | B | 2.6 | to begin |
| حسّ | B | 0.09 | to feel |
| لقى | B | 0 | to find |

Table 6: IMPERATIVE usage of verbs most common to each subcorpus (M=Movies, B=Blogs)

Having a lemmatized corpus also permits comparison of lexical diversity between the two registers. Using the Biber (2006) formula for normalizing lexical diversity counts, statistics for each register are given in Table 7. All differences displayed are statistically significant, suggesting that Blogs is richer in verb types as well as in verb diversity.

| | Types | Diversity | Types/1M Verbs | Types/1M Words |
|---|---|---|---|---|
| Movies | 2,279 | 5.88 | 11,574 | 4,768 |
| Blogs | 2,079 | 7.53 | 12,510 | 5,530 |

Table 7: Diversity of verbs

### 4.1 Reflections

The data reported above provide enough evidence to warrant a wider investigation into the variations that exist between these potential registers. In English, use of the PERFECT has been identified as a feature of narration (Biber and Conrad, 2001; Staples, 2016). If Egyptian Arabic behaves like English, then the higher frequency of the PERFECT in Blogs signals a greater reliance on narration than in Movies, hence forming different registers. The possibility of the Egyptian Arabic PERFECT being a feature of narration is further supported by Biber et al. (2006), who found that most English internet texts could be classified as narrative.

Similarly, the frequency of the IMPERATIVE in Movies could easily be a feature of involved and non-narrative speech as found in Somali (Biber and Conrad, 2001). Therefore, the frequency of the PERFECT and IMPERATIVE in both subcorpora suggests a difference in narrative-based register separation. Distinct differences in HABITUAL versus FUTURE could also be linked to narration, but possibly some other feature. 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 variance of the frequencies of the verbal aspects and moods further supports register variation as one corpus alone cannot be used to generate a description of how verbs are used. In both subcorpora IMPERFECT is used more than any other verb aspect or mood followed by PERFECT. If this description were based solely on Movies, IMPERATIVE would be the third most common verb form. However, this is not true of Blogs. Therefore, the omission of one subcorpus from analysis would skew the description of the dialect: both need to be taken into account when producing a description of the language.

The subjects of IMPERATIVE verbs also seem to differ across register. The number of female positive imperatives in Blogs is trivial when compared to Movies. However, many factors independent of register could affect this result. Further investigation is needed to determine whether this difference can be used to indicate register variation.

Greater frequency of verb tokens and types in Blogs in EA is also interesting. Biber (2006) found, for academic English, that verbs were much more common in spoken academic registers (e.g. lectures vs. journal articles); features associated with verbs were also found to be characteristic of the oral dimension of Spanish and English more generally (Biber, 1999; Biber et al., 2006). However, the opposite appears to be true for Egyptian Arabic: Blogs contains a statistically higher number of verbs than Movies does.

Blogs also contains a greater diversity of verbs, which is consistent with English and Spanish; this may be expected as authors have time to think about the words they will use and revise their choices (Biber, 2006; Biber et al., 2006). In this study our differences in EA verb usage (i.e. number of verbs and their variety) suggest that the language contained in Blogs and Movies is different. In theory, both are written and revised; therefore, the difference in the diversity of verbs cannot be due to the fact that one of the registers is written. 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.

## 5 Notes on annotation

Linguistic annotation is the process by which additional linguistic information is added to a corpus in order to facilitate quantitative analyses of corpus content and user queries (Kübler and Zinsmeister, 2015). Manual annotations are performed by humans, automatic annotations are done by a computer program, and automatic annotations that are checked by a human for accuracy are called semi-automatic annotations.

Automatic annotators available for EA are somewhat limited, and although more resources exist for MSA, the morphological and lexical differences cause MSA annotators a challenge in annotating EA texts (Maamouri et al., 2014). In 2004, a part-of-speech annotator for MSA was achieving an accuracy 95.49% (Diab et al., 2004) versus a contemporary analyzer for EA with an accuracy at 62.76% (Duh and Kirchhoff, 2005). One reason for the disparity was the lack of large corpora or a complete lexicon of EA for annotator training (Habash and Rambow, 2006).

Abo Bakr et al.'s (2008) annotator translated Egyptian Arabic sentences into MSA and then tagged the MSA for part of speech, which would then be applied back to the Egyptian words. Conversion of the Egyptian Arabic to MSA was successful 88% of the time, and overall accuracy ratings for tokenization and part-of-speech tagging for EA were 90% and 85% respectively.

Al-Sabbagh and Girju (2012a) created an Egyptian Arabic tagger that did not depend upon MSA. Originally trained on three language types (Twitter, QA Pairs, and blogs), its highest reported F-measure among them for POS tagging is 0.907 (QA Pairs), though it had less success on blogs (with an F-measure of 0.888).

MADAMIRA (Arfath Pasha et al., 2014) analyzes each word according to the possible morphemes attached to it. It then uses language models to provide a morphological analysis, part-of-speech, lemma, and diacritics for each word in a text. Its accuracy score for part-of-speech tagging is 0.923. MADAMIRA's ability to provide lemmatization makes it valuable tool for register variation studies.

One issue regarding annotators involves whether they are accurate enough to be used without the need for a manual review of the results. Another annotator issue is how well accuracy persists when annotating texts in a different domain from the training set. There is an apparent lack of published research on using MADAMIRA in this cross-domain fashion. MADAMIRA was trained on transcripts of speech (Habash et al., 2012), but the literature is less clear about the register of Egyptian Arabic on which it was evaluated (Arfath Pasha et al., 2014). We would expect the average accuracy of MADAMIRA to shift either up or down when applied to other registers of the dialect as this phenomenon has been found in other languages (Tseng et al., 2005; Derczynski et al., 2013).

Numerous tagsets are available to use for Arabic part-of-speech tagging (Arfath Pasha et al., 2014;

Alian and Awajan, 2018). A modified version of the tagest employed by MADAMIRA was used for CALM annotation. Verbs were annotated as such, even in the presence of pronominal object suffixes and prepositional proclitics. Additionally, a second layer of annotation was applied to all verbs to indicate certain verbal categories. MADAMIRA divides verbs into three groups: imperfect (i), perfect (p), and command (c). In CALM, two more categories were created from the imperfect category. Although verbs in the HABITUAL (h) and FUTURE (f) are IMPERFECT, these were promoted as separate categories for ease of searching. Another change to MADAMIRA's annotations in CALM is the identification of negative imperatives and their inclusion into the "command" category. (The default tagset collapses imperfect verbs and negative imperatives into one class.)

Annonation of CALM also includes a few other adjustments to MADAMIRA's output: (1) MADAMIRA does not view the passive verbs as a verb form but adds an extra layer of annotation; these were folded into the basic verb paradigm in CALM. (2) Slight differences in lemmatization involved clarification by adding short vowels where necessary.

Overall MADAMIRA performed relatively well in annotating Movies and Blogs from CALM, and a combination of post-processing, both manual and automatic, made corrections when necessary. Hereafter we refer to raw annotations as "non-gold", and corrected annotations as "gold". Table 8 shows both the non-gold and gold statistics for the content shown earlier in Table 7.

| | Types | Divers. | Types/1M Verbs | Types/1M Words |
|---|---|---|---|---|
| Non-gold | | | | |
| Movies | 2,867 | 7.44 | 14,608 | 5,999 |
| Blogs | 2,751 | 10.08 | 16,651 | 7,318 |
| Gold | | | | |
| Movies | 2,279 | 5.88 | 11,574 | 4,768 |
| Blogs | 2,079 | 7.53 | 12,510 | 5,530 |

Table 8: Diversity of verbs (non-gold and gold)

Table 9 gives the counts for each of the verb types as annotated by the automatic tagger (the "non-gold" annotations) and after human correction (the "gold" annotations), and the percent change between the two annotation types. In all cases, the cross-register differences in verb usage that were significant in the gold subcorpora also held in the non-gold subcorpora. This nearly holds for the imperatives as well, except that the non-gold corpora do not report a significant difference in the use of the 2PL.IMPERATIVE in Blogs. As explained earlier, the automatic tagger does not attempt to categorize negative imperatives. For that reason, each cell in its row contains 'NA'.

For verb diversity measures, MADAMIRA data are nearly identical to the lists generated by hand (i.e. those in Table 5) except for six verbs whose counts were not accurate enough to reveal the statistically significant register differences. Regarding comparative verbal diversity, MADAMIRA scores diversity in Movies at 7.44% and in Blogs at 10.08% (a difference of 2.64%) whereas manual correction yields a difference in diversity of only 1.65%.

In conclusion, the annotations produced solely by MADAMIRA would have led researchers to nearly the same conclusions as those reached above with hand-corrected annotated data. 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. Except for the IMPERATIVE category, MADAMIRA's total number of verbs in each category in Blogs changed by less than 5% after hand-correction. In both subcorpora, MADAMIRA was consistent with the categories that it over- and under-represented: 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 Movies and the opposite in Blogs.

One difficulty MADAMIRA had was in differentiating proper nouns from verbs, a challenge since Arabic has no capital letters. IMPERFECT and PERFECT were overrepresented due to misclassification, precision was lower on Movies, and recall on proper nouns suffered. In Movies, 7 of the top 10 words incorrectly tagged as verbs were actually names and titles given to people. Seventeen word forms represent 1,239 of the 3,290 recall errors of this type, comprising 37.6% of all the false positives. Names in the Blogs subcorpus were also problematic; in the top 30 false positives there, 8 were names (totaling 223 occurrences). This type of ambiguity only accounts for 10.5% of the total number of false positives, though, likely due to lower use of personal names in the Blogs.

| | Movies | | Blogs | |
|---|---|---|---|---|
| | Non-Gold | % Change | Non-Gold | % Change |
| All 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 |
| Imperatives | | | | |
| Neg. | NA | NA | NA | NA |
| Pos. 2SG.MASC | 1,455 | -40.95 | 692 | -35.33 |
| Pos. 2SG.FEM | 877 | -31.75 | 179 | -9.6 |
| Pos. 2PL | 165 | -28.88 | 89 | -34.07 |

Table 9: Effect of hand correction for frequency counts

Overall MADAMIRA performed relatively well in annotating the verbs in Movies and Blogs from CALM. However, in order to achieve higher accuracy for this paper, the annotations were manually reviewed and corrected. Throughout the process of manual correction, high-frequency errors made by MADAMIRA became apparent and a supplemental Python post-processor was developed to target these mistakes. This program was able to boost MADAMIRA's precision score from 0.922 to 0.944. Although the post-processor was able to reduce the number of corrections needed, every automatically assigned annotation was manually reviewed. Details are discussed elsewhere (White, 2019, forthcoming).

## 6 Conclusions and future work

This paper discussed the need for an Egyptian Arabic corpus of spoken language transcripts and introduced CALM, a new two-million word corpus of spoken EA. It also conducted an analysis into the use of verbs in two potential registers of EA.

The results show significant variance in the usage of verbs in Movies versus Blogs. These differences are consistent with variations found between other registers in previous multidimensional analyses. These results also lay the groundwork for future studies by providing a description of some of the dimensions of EA based upon empirical data.

We also showed that in spite of the challenges in annotating Egyptian Arabic, an automatic tagger was able to produce results that were not appreciably different from those produced through a process of manual correction.

The scope of this work was to show how a non-trivial subset of CALM could serve as data for a register analysis. It was limited in several ways, all of which can be extended via further research. First, a finer distinction into register types (especially blog subtypes) could be enacted, as has been done for other languages. In addition, this work involved annotations based on only one part of speech (i.e. verbs), whereas other categories could serve for similar analyses once annotations are available. Third, given the ongoing debate about whether transcripts of scripted speech can be used to represent speech, more study should ascertain how exactly dialogue and narration are characterized for register in EA. Finally, insight could be sought concerning the frequent use of the HABITUAL in Blogs. Is this due to the narrative dimension, or some other one represented in Blogs? Answers to this question can inform curricula for Egyptian Arabic learners, who often find this feature difficult.

## References

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

Hitham Abo Bakr, Khaled Shaalan, and Ibrahim Ziedan. 2008. A Hybrid Approach For Converting Written Egyptian Colloquial Dialect Into Diacritized Arabic. In The 6th International Conference on Informatics and Systems (INFOSYS 2008).

Rania Al-Sabbagh and Roxana Girju. 2012a. A Supervised POS Tagger for Written Arabic Social Networking Corpora. In Proceedings of the Conference on Natural Language Processing (KONVENS), pages 39–52.

Rania Al-Sabbagh and Roxana Girju. 2012b. YADC: Yet another Dialectal Arabic Corpus. In Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC), pages 2882–2889.

Jawharah Alasmari, E. Atwell, and 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.

Abdullah Alfaifi and Eric Atwell. 2015. Arabic learner corpus.

Ahmed Ali, Stephan Vogel, and Steve Renals. 2017. Speech recognition challenge in the wild: Arabic MGB-3. In 2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU), pages 316–322.

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

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

Mohamed Al-Badrashiny Arfath Pasha, Mona T. Diab, and Ahmed El Kholy. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. In Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC).

Marco Baroni and Silvia Bernardini. 2004. BootCaT: Bootstrapping Corpora and Terms from the Web. In Proceedings of the 4th Langauge Resources Evaluations Conference (LREC), page 1313.

Randell Bentley. 2015. Conditional Senteces in Egyptian Colloquial Arabic and Modern Standard Arabic: A Corpus Study. Master's thesis, Brigham Young University.

Douglas Biber. 1993. Representativeness in Corpus Design. Literary and Linguistic Computing, 8(4):2430–257.

Douglas Biber. 1999. Longman Grammar of Spoken and Written English. Harlow, England.

Douglas Biber. 2006. University Language: A Corpus-based Study of Spoken and Written Registers. John Benjamins, Philadelphia.

Douglas Biber and Susan Conrad. 2001. Register Variation: A Corpus Approach. In The Handbook of Discourse Analysis. Blackwell Publishers, Massachusetts.

Douglas Biber, Mark Davies, James K. Jones, and Nicole Tracy-Ventura. 2006. Spoken and Written Register Variation in Spanish: A Multidimensional Analysis. Corpora, 1(1):1–37.

Douglas Biber, Jesse Egbert, and Mark Davies. 2015. Exploring the Composition of the Searchable Web: A Corpus-Based Taxonomy of Web Registers. Corpora, 10(1):11–45.

Marc Brysbaert and 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.

Tim Buckwalter and Dilworth Parkinson. 2011. A Frequency Dictionary of Arabic: Core Vocabulary for Learners. Routledge, New York.

Alexandra Canavan, George Zipperlen, and David Graff. 1997. CALLHOME Egyptian Arabic Speech. Linguistic Data Consortium Web Download, LDC97S45. Philadelphia, PA.

Leon Derczynski, Alan Ritter, Sam Clark, and Kalina Bontcheva. 2013. Twitter Part-of-speech Tagging for All: Overcoming Sparse and Noisy Data. In Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), pages 198–206.

Mona Diab, Kadri Hacioğlu, and Daniel Jurafsky. 2004. Automatic Tagging Of Arabic Text: From Raw Text To Base Phrase Chunks. In Proceedings of HLT-NAACL 2004: Short papers, pages 149–152. Association for Computational Linguistics.

James Dickins. 2017. The Pervasiveness of Coordination in Arabic, with Reference to Arabic$gt$English Translation. Languages in Contrast, 17(2):229–254.

Kevin Duh and Katrin Kirchhoff. 2005. POS Tagging Of Dialectal Arabic: A Minimally Supervised Approach. In Proceedings of the acl Workshop on Computational Approaches to Semitic Languages, pages 55–62. Association for Computational Linguistics.

Ted Dunning. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19(1):61–74.

Abbas El-Tonsi. 1982. Egyptian Colloquial Arabic: A Structure Review, volume 1. American University In Cairo, Cairo, Egypt.

Ahmed Fakhri. 2009. Rhetorical variation in Arabic academic discourse: Humanities versus law. Journal of Pragmatics, 41(2):306–324.

Charles A. Ferguson. 1983. Sports announcer talk: Syntactic aspects of register variation. Language in Society, 12(2):153–172.

Pierfranca Forchini. 2012. Movie Language Revisited: Evidence from Multi-Dimensional Analysis and Corpora. Peter Lang, Bern.

Eric Friginal. 2009. Language of Outsourced Call Centers: A Corpus-based Study of Cross-cultural Interaction. John Benjamins, Philadelphia.

Stefan Th. Gries. 2006. Exploring variability within and between corpora: some methodological considerations. Corpora, 1(2):109–151.

Nizar Habash, Ramy Eskander, and Abdelati Hawwari. 2012. A morphological analyzer for Egyptian Arabic. In Proceedings of the Twelfth Meeting of the Special Interest Group on Computational Morphology and Phonology (SIGPHON), pages 1–9. Association for Computational Linguistics.

Nizar Y. Habash and Owen C. Rambow. 2006. MAGEAD: A Morphological Analyzer And Generator For The Arabic Dialects.

Gadalla Hassan. 2000. Comparative Morphology of Standard and Egyptian Arabic. LINCOM EUROPA.

David Henen. 2018. "ya" between Vocative and Non-Vocative Use in Egyptian Film Language A Corpus Analysis: Pragmatic Functions and Formal Features. American University in Cairo, Egypt.

Susan Hunston. 2002. Corpora in Applied Linguistics. Cambridge University Press, Cambridge.

Mona Hussein. 2016. Propositional and Non-Propositional Functions of /Keda/ in the Language of Egyptian Film. American University in Cairo, Egypt.

Ahmad Ismail. 2015. ṭab asta'zen ana ba'a: A corpus-based Study of Three Discourse Markers in Egyptian Film Language. American University in Cairo, Egypt.

Barbra Johnstone. 2008. Discourse Analysis. Blackwell, Malden, MA.

Sandra Kübler and Heike Zinsmeister. 2015. Corpus Linguistics and Linguistically Annotated Corpora. Bloomsbury, New York.

Mohamed Maamouri, Ann Bies, Seth Kulick, Michael Ciul, Nizar Habash, and Ramy Eskander. 2014. Developing an Egyptian Arabic Treebank: Impact of Dialectal Morphology on Annotation and Tool Development. In Proceedings of LREC, pages 2348–2354.

Yousef Maati. 2008. fiʌsʌn wi murʔosʕ. Al-daar Al-Masriya Al-lubnaniya, Cairo, Egypt.

Hamdi A. Qafisheh. 1992. Yemeni Arabic Reference Grammar. Dunwoody Press, Kensington, MD.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In Proceedings of the workshop on Comparing corpora, volume 9, pages 1–6. Association for Computational Linguistics.

Karin C. Ryding. 2005. A Reference Grammar of Modern Standard Arabic. Cambridge University Press, Cambridge, England.

Karin C. Ryding. 2006. Teaching Arabic in the United States. In Handbook for Teaching Arabic Language Professionals in the 21st Century, pages 13–20. Routledge, New York.

Mukhtar Sayed. 2018. maʕleš maʕleš: A CORPUS-BASED STUDY ON THE DISCOURSE MARKER maʕleš. American University in Cairo, Egypt.

Serge Sharoff. 2006. Creating general-purpose corpora using automated search engine queries. In Baroni and Bernardini, editors, WaCky! Working papers on the Web as Corpus, pages 63–98. Gedit.

John Sinclair. 2004. Corpus Creation. In Geoffrey Sampson and Diana McCarthy, editors, Corpus Linguistics: Readings in a Widening Discipline, pages 78–84. Continuum, New York.

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

Rianne Tamis and Janet Persson, editors. 2013. Sudanese Arabic-English; English-Sudanese Arabic: A Concise Dictionary. SIL International.

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

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

Huihsin Tseng, Daniel Jurafsky, and Christopher Manning. 2005. Morphological features help POS tagging of unknown words across language varieties. In Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing.

Michael Grant White. 2019, forthcoming. Verb usage in Egyptian Arabic: a case for register variation. Master's thesis, Brigham Young University.

Andrew Wilson. 2013. Embracing Bayes Factors for key item analysis in corpus linguistics. In M. Bieswanger and A. Koll-Stobbe, editors, New Approaches To The Study Of Linguistic Variability, pages 3–9. Peter Lang, Frankfurt.