Annotating corpus data for a quantitative, constructional analysis of motion verbs in Modern Standard Arabic

Dana Abdulrahim
University of Bahrain
darahim@uob.edu.bh

Abstract

This article proposes an annotation method of corpus data for the purposes of providing a constructionist account of lexical behavior. The lexical items in question are seven verbs of motion in Modern Standard Arabic that pertain to the events of COME (*atā*, *ǧā'a*, *ḥaḍara*, and *qadima*) and GO (*ḏahaba*, *maḍā*, and *rāḥa*). The tag set selected for the annotation of the COME and GO data frames consists of morphosyntactic tags that characterize verb usage as well as semantic tags that aim to highlight the semantic component of, for instance, adverbial and adpositional phrases that accompany the verb. I will briefly demonstrate the analytical potential of such data frame by discussing the various kinds of statistical tests such data frame is designed to undergo, as a means of better understanding lexical behavior in context, and, eventually, arriving at a better understanding of lexical and constructional choices made by native speakers of Arabic, as demonstrated in corpora.

1 Introduction

The core tenets of constructionist theories of language claim that the basic unit of linguistic organization is a construction. According to Croft and Cruse (2004:257), a construction "consist[s] of pairings of form and meaning that are at least partially arbitrary", where 'meaning' is referred to as the conventionalized function of a construction. This conventionalization of a construction's meaning/function includes not only the literal meaning of an utterance, but also the discourse situation of that utterance, as well as any pragmatic implication conveyed by that utterance (Croft and Cruse, 2004).

The concept of a 'construction' in cognitive approaches to grammar, therefore, relates to both the idiomatic portions of language, where the meaning of an utterance is not predictable from the component parts of which it consists (e.g. *raining cats and dogs*), as well as the co-occurrence of any two (free or bound) morphemes that reflect general morphosyntactic structures and where the meaning of an utterance is fully predictable from its component parts (e.g. *I need to sleep*). Such view of grammar postulates that "the interaction of syntax and lexicon is much wider and deeper than the associations of certain verbs with certain complements" (Bybee, 2010:77), and that a considerable part of our linguistic knowledge consists of conventionalized expressions, or constructions (Langacker, 1987).

In light of these constructionist assumptions, therefore, the behavior of a lexical item is best understood in its context of use and not in isolation, an idea that stretches back decades (cf. Firth, 1957). This includes not only lexical collocates, but also the entire morphosyntactic frame that hosts a lexical item. All these elements contribute to the composed or conventionalized meaning/function expressed by a particular linguistic item. In order to examine lexical behavior, therefore, we need to move beyond single semantic, morphological, or syntactic properties of a lexical item and scrutinize the entire lexico-syntactic frame in which it occurs. Increasingly, this is done through examination of corpus data. The availability of corpora facilitates and motivates such highly contextualized analytical approach, since corpora provide a large amount of naturally occurring, contextualized uses (as opposed to the reliance on introspective and elicited data that may not reflect actual language usage at all). Moreover, corpora provide large amounts of linguistic data, which allows the researcher to conduct extensive quantitative analyses of the phenomenon in question.

In Modern Standard Arabic, the existence of several verbs denoting the motion events COME (*atā*, *ǧā'a*, *ḥaḍara*, and *qadima*) and GO (*ḏahaba*, *maḍā*, and *rāḥa*) provides an excellent case study for a constructionist, corpus-based examination of the features that characterize the usage of supposedly near-synonymous lexical items. In Abdulrahim (2013) I have argued that the four COME verbs – as well as the three GO verbs – can be interchangeably used in contexts where the event depicts a strictly deictic and physical motion event, as in (1) and (2).

Proceedings of the EMNLP 2014 Workshop on Arabic Natural Langauge Processing (ANLP), pages 28–38,
October 25, 2014, Doha, Qatar.

(1) أتت / جاءت / حضرت / قدمت جدتي إلى المطار

atā / ǧā'a / ḥaḍara / qadima.PERF.3SG.F	grandmother.CL.1SG.GEN
came	my grandmother

ALL	ART=airport
to	the airport

'My grandmother came to the airport'

(2) ذهب \ مضى \ راح الأب إلى مركز الشرطة

ḏahaba / maḍā / rāḥa.PERF.3SG.M	ART=father-NOM	ALL
went	the father	to

station	ART=police
station	the police

'The father went to the police station'

However, these verbs diverge greatly in their metaphorical and idiomatic uses, in addition to showing idiosyncratic patterns of lexico-syntactic behavior. For instance, The sentence in (1) would not admit all four verbs when the aspect inflection on the verb is changed. To illustrate, in (3), if we hold all constructional features constant and change verb inflection from perfective to jussive, this results in a preference for *atā* and *ḥaḍara* by native speakers of Arabic over *ǧā'a* or *qadima*.

(3) لم تأت / تجيء / تحضر / ؟تقدم جدتي إلى المطار

NEG	***atā / ǧā'a / ḥaḍara / qadima***.JUSS.3SG.F
did not	come

grandmother.CL.1SG.GEN	ALL	ART=airport
my grandmother	to	the airport

'My grandmother did not come to the airport'

The above example stresses the need to investigate features of the lexico-syntactic construction that each COME verb most typically associates with. For these purposes, I have adopted the methodological approach outlined in Gries (2006), Gries & Divjak (2009), and Gries & Otani (2010) for a constructionist description of the Behavioral Profile of a lexical item. I have also employed logistic regression – namely polytomous logistic regression, outlined in detail in Arppe (2008) – as a statistical method that models lexical or constructional choices as a function of a wide range of contextual features.

The quantitative approach to lexical analysis presented in this paper involves constructing a data frame for every lexical item under study, in which numerous corpus concordance lines are individually marked-up for an extensive set of linguistic features (morphological, syntactic, and semantic). This includes, for examples, specific elements pertaining to verb morphology, phrase structure, as well as the different elements that co-occur with the verb in specific constructions. Such data frame can undergo numerous exploratory and multi-variate statistical tests as a means of zeroing in on the kinds of constructions associated with the verbs in question.

2 The corpus

In order to construct the multi-variate data frame for the analysis of motion verbs in Arabic, I chose ArabiCorpus (arabicorpus.byu.edu) as the source for data. ArabiCorpus is a free online corpus developed by Dilworth Parkinson at Brigham Young University. As of October 2012, the corpus contains around 146,000,000 word tokens from different written and spoken genres. At the time of data collection (Fall 2010) the corpus contained around 69,000,000 word tokens. Additional MSA as well as pre-modern texts have been added to the corpus since the beginning of 2011. Texts included in ArabiCorpus almost exclusively belong to the written genre, save for a small sub-corpus of spoken Egyptian Arabic. The written genres covered in ArabiCorpus vary from newspaper writing, pre-modern writing, modern literature, to nonfiction. These genres are represented in the corpus in varying proportions with newspaper writing accounting for over 90% of the total size of the entire corpus. News articles included in this sub-section of ArabiCorpus cover issues from 1996 to 2010 and are extracted from periodicals published in different parts of the Arab world (North Africa, Egypt, Arabian Gulf, the Levant, etc.). For this study, the MSA subcorpora that were queried for COME and GO uses are related to newspaper, modern literature, and nonfiction writing. As expected, most examples returned from corpus queries were in fact drawn from the newspaper genre.

ArabiCorpus is not tagged for parts-of-speech (POS) which makes the search for particular grammatical categories a daunting task. Nevertheless, it can be easily queried using regular expressions. This study targeted very specific inflected forms of the MSA verb: perfective, imperfective, jussive, subjunctive, and imperative; and excluded participial forms (e.g. active participle) and nominal forms (e.g. verbal nouns). Relying on orthographic regular expressions, therefore, proved to be the most efficient method

for querying these particular forms. What may complicate any search in an Arabic written corpus is the lack of short vowels which are indicated by certain diacritics written over or underneath a letter. Naturally, it was necessary to filter corpus returns manually to eliminate any irrelevant forms that may have been returned in the corpus search.

Despite the time-consuming nature of such corpus querying steps, ArabiCorpus proved to be a reliable and rich source for contextualized language uses. Add to that the fact that even though the online interface of ArabiCorpus only displays 10 words before and after the KWIC (key word in context), the researcher can still retrieve the entire text hosting that sentence. Therefore when the analysis or the annotation requires going beyond the 10 word window to examine the entire phrase structure, it is possible to retrieve such information from ArabiCorpus. Another added benefit of using the online interface of ArabiCorpus is the ease of downloading all returned hits of a certain lexical item or construction to be viewed in a spreadsheet, which was a major step in the collection of data for this research.

3 COME and GO data frames

In Abdulrahim (2013) I proposed a quantitative (as well as a qualitative) analysis based on the construction of a data frame for each one of the seven verbs of motion. Each individual data frame is typically composed of a large number of corpus concordance lines (500 concordance lines) involving the KWIC (i.e. verb under investigation) in its natural context of use. Every concordance line is thoroughly examined and tagged for a wide range of morphosyntactic and semantic features. These constructional features include the syntactic structure that hosts the verb, the patterns of verbal inflections for every instance of verb use (e.g. subject number, person, and gender, as well as other morphosyntactic aspects of the Arabic verb), the semantic properties of other components of the construction (e.g. semantic properties of the subject), as well as the inclusion or exclusion of phrases, lexical items, or clitics denoting a starting point of the event (SOURCE), a terminal point of the event (GOAL), etc.

Such a heavily annotated dataset has the potential of being explored statistically in multiple ways via simple frequency count methods as well as complex multi-variate statistical modeling. Such quantitative approach to analyzing corpus data aims to define the specific characteristics of the constructions associated with the various meanings and functions of each MSA COME and GO verb involved in this study. In the following section I will elaborate on the selection of these contextual features for the annotation of COME and GO data frames.

3.1 Selection of contextual features and the annotation of corpus data

The first step for creating a multi-variate data frame is to generate a list of features or variables which are relevant to the motion event schemas in questions and which reflect the morphosyntax of Modern Standard Arabic. Along the lines of Gries's study on the polysemy of the English verb *run* (2006), Gries and Divjak's (2006) investigation of Russian verbs of TRY, as well as Gries and Otani's (2010) analysis of the synonymy and polysemy of adjectives of size in English, I developed a large set of morphological, syntactic, and semantic features that reflect the usage of MSA motion verbs.

The variable set includes nominal variables (multiple levels) and binary variables (YES/NO values indicating absence or presence of feature). Table 1 shows the different categories of variables subsumed under morphological, syntactic, and semantic variables. In Appendix A, I provide examples and illustrations of the different annotations of levels within semantic variables taken from the actual data frame. [1]

Morphological variables	***Levels***
TENSE	PRESENT, PAST, FUTURE, IRREALIS (non-finite forms)
ASPECT	SIMPLE, HABITUAL, PROGRESSIVE, PERFECT, INCHOATIVE, NON-FIN (non-finite forms)
MORPHOLOGICAL ASPECT AND MOOD OF THE VERB	IMPERFECTIVE, PERFECTIVE, SUBJUNCTIVE, JUSSIVE, IMPERATIVE
SUBJECT PERSON	1^{ST}, 2^{ND}, 3^{RD}
SUBJECT NUMBER	SINGULAR, DUAL, PLURAL
SUBJECT GENDER	FEMININE, MASCULINE, NIL (for 1^{st} person inflections)

[1] The data frame was, in fact, coded for more variables than the set laid out in Table 4, such as the different morphosyntactic realizations of GOAL, SOURCE, MANNER, etc., as well as certain recurring lexical elements (e.g. adverbs, adverbial uses, and other lexical items). These additional variables did not form part of the quantitative analysis in Abdulrahim (2013). Nevertheless, they are of some interest and proved to be useful for a qualitative analysis of MSA motion verbs.

Syntactic variables	*Levels*
TRANSITIVITY	YES, NO
INTERROGATIVE	YES, NO
NEGATIVE	YES, NO
PREPOSITIONAL PHRASE	YES, NO
LOCATIVE ADVERB PHRASE	YES, NO
ADVERBIAL PHRASE	YES, NO
SERIAL VERB CONSTRUCTION	YES, NO

Semantic variables	*Levels*
SUBJECT CATEGORY	ACTIVITY, ANIMAL, ATTRIBUTE, BODY, COGNITION, COMMUNICATION, CONTENT (of a document/speech), DEMONSTRATIVE, DUMMY SUBJECT, EVENT, GROUP, HUMAN, LOCATION, NOTION, OBJECT/ARTIFACT, SENSE, STATE, SUBSTANCE, TIME
GOAL PHRASE	YES, NO
SOURCE PHRASE	YES, NO
MANNER PHRASE	YES, NO
SETTING PHRASE	YES, NO
PATH PHRASE	YES, NO
PURPOSIVE PHRASE	YES, NO
COMITATIVE PHRASE	YES, NO
TEMPORAL PHRASE	YES, NO
DEGREE PHRASE	YES, NO

Table 1. A selection of variables GO and COME corpus hits were coded for.

As mentioned earlier, the primary motivation for this set of 23 linguistic features/tags has been the lexico-syntactic properties of deictic motion event schemas in MSA. For instance, a deictic motion event is likely to include a phrase specifying a GOAL and/or a SOURCE of the motion event. In addition, it may include MANNER of motion and the inclusion of a COMITATIVE phrase (i.e. accompaniment by an object/individual in the GO or COME event). Each verb usage was also coded for the semantic category of the sentential subject or, conceptually speaking, the moving entity involved in the motion event. These categories include HUMAN, OBJECT or ARTIFACT, and also more abstract/non-physical entities such as EVENT, COMMUNICATION (e.g. a statement), COGNITION (e.g. an idea), etc. As for the morphosyntactic features selected for tagging motion verbs, these reflect the inflectional properties of the MSA verb (MORPHOLOGICAL ASPECT AND MOOD, NUMBER, PERSON, and GENDER) as well as TENSE and ASPECT. The variable labeled TRANSITIVITY, only pertains to certain uses of COME verbs in MSA where COME verbs can appear in transitive constructions in which the direct object is the GOAL of the motion event.

Text genre was not considered a variable since, as I mentioned earlier, the majority of the annotated 3,500 corpus hits belong to the genre of newspaper writing. Results obtained from examining this data frame should, therefore, be considered as mostly reflective of the usage of COME and GO verbs in newspaper writing. Sentence (4) is an example of a contextualized verb use annotated for the features listed above.

(4) وهي تمضي بسرعة في مؤامراتها

CONJ=PP	***maḍā***.IMPF.3SG.F	INST=speed
and she	goes	quickly

LOC	conspiracies-CL.3SG.F.GEN
in	her conspiracies

'And it's [i.e. Israel] quickly going ahead with its conspiracies'

VERB	*maḍā*	**TENSE**	PRESENT
ASPECT	SIMPLE	**MORPH_ASP/ MOOD**	IMPERFECTIVE
SUBJ_NUM	SINGULAR	**SUBJ_PER**	3RD
SUBJ_GEN	FEM	**SUBJ_CAT**	GROUP
INTEROG	NO	**NEGATION**	NO
SVC	NO	**PP**	YES
LOC_ADV	NO	**ADVERBIAL**	YES
GOAL	NO	**SOURCE**	NO
MANNER	YES	**SETTING**	YES
PATH	NO	**PURPOSIVE**	NO
COMITATIVE	NO	**TEMPORAL**	NO
DEGREE	NO		

4 Statistical analyses

A wide range of statistical tests can be applied in order to explore the data frames described above for various purposes.[2] For instance, we can simply run the COME and GO data frames through mono-variate exploratory tests such as chi-square tests as a means of zeroing in on the distribution of contextual elements per each GO and COME verb. This kind of analysis would constitute a first step towards identifying divergence in usage patterns associated with each MSA motion verb. This preliminary step further motivates and justifies the examination of interaction patterns among the contextual features, as

[2] See Hastie, T., et al (2009), and Agresti (2002) – among others – for comprehensive discussions on statistical tests that can be applied to multi-variate data frames.

well as the identification of clusters of features that are closely tied to certain verb uses. A multivariate analysis eventually facilitates the identification of prototypical uses of each verb as well as the less prototypical uses. In the following I will briefly discuss three types of statistical methods that can be applied to the MSA COME and GO data frames: (i) chi-square test; (ii) cluster analysis; and (iii) polytomous logistic regression analysis.[3]

4.1 Chi-square tests

The primary purpose of subjecting the COME and GO data frames to chi-square test of independence is to examine whether the distribution of the different levels of variables (tags) do not vary as a function of verb (null hypothesis), or, if they actually do vary as a function of verb (alternative hypothesis). For instance, if we examine the occurrence of a GOAL phrase per each GO verb, would the distribution of variables be the same or different across the three verbs. To test this hypothesis –where we have an independent variable (verb) and a dependent variable (GOAL) –we can run a chi-square test on variable distribution. Table 2 shows the observed frequencies for the occurrence/absence of a GOAL phrase per each GO verb, while Table 3 shows the expected frequencies calculated by the command chisq.test()$expected in R (www.r-project.org).

VERB	GOAL - YES OBS. FREQ.	GOAL - NO OBS. FREQ.
ḏahaba	298	202
maḍā	32	468
rāḥa	1	499

Table 2. Observed values for the variable GOAL by GO verb.

VERB	GOAL - YES EXP. FREQ.	GOAL - NO EXP. FREQ.
ḏahaba	110.3333	389.6667
maḍā	110.3333	389.6667
rāḥa	110.3333	389.6667

Table 3. Expected values for the variable GOAL by GO verb.

[3] See Abdulrahim (2013) for further description of the properties and applications of these statistical analyses on the MSA COME and GO data frames.

The calculated Pearson's *chi*-square test for the distribution given in Table 4 proved to be quite significant: X^2 = 277.1034, *df* = 6, *p-value* < 2.2e-16. This indicates that the distribution the variable GOAL for each GO verb deviates highly significantly from the expected distribution.

We can also examine the cell-wise divergences from a uniform distribution for this particular contingency table by conducting a standardized Pearson's residual (discussed in Agresti 2002: 81; Arppe, 2008: 83-84). These test statistics can either be retrieved in R by using the command chisq.test()$std or by running the function chisq.posthoc(), which is part of the statistical package {polytomous} developed by Antti Arppe (2012). Table 4 contains the calculated values, which indicate whether the observed co-occurrence frequency reported in each individual cell is significantly *more* or *less* than expected.[4] The chisq.posthoc()function presents an easier way to interpret these figures, in that it assigns +/–/0 values for each cell, which can be interpreted as insignificant (0), significantly more than expected (+), or significantly less than expected (–).

VERB	GOAL - YES	GOAL - NO
ḏahaba	24.78665 (+)	-24.78665 (–)
maḍā	-10.34611 (–)	10.34611 (+)
rāḥa	-14.44053 (–)	14.44053 (+)

TABLE 4. Standardized Pearson's residuals for the occurrence of GOAL by GO verb.

As discussed earlier, these exploratory tests constitute a first attempt at understanding the distributional patterns of selected variables among the different verbs. Such mono-variate methods undoubtedly set the stage for the more complex multi-variate analyses that will follow and to which I turn next.

4.2 Cluster analysis

Clustering methods can help us examine the joint effect on the overall verbal behavior for each verb in the GO and COME verb set. One such method is referred to as hierarchical agglomera-

[4] Typically, the standardized Pearson's residual value is significantly higher than what is expected when it is > 2.0, and significantly lower than expected when the value is < -2.0 (Arppe, 2008).

tive cluster analysis (explained more in detail in Gries, 2006; Diviaj and Gries, 2006; Gries and Otani, 2010, among others). Generally speaking, this clustering method groups together the lexical elements that are most similar to one another and, at the same time, the ones that are highly dissimilar to other elements in other clusters. Therefore, what we expect to see from this statistical method is a clustering dendrogram that shows us which COME or GO verbs overlap in their usage as opposed to the ones with which they hardly share any characteristics.

This method requires generating a table that lists relative frequencies (or proportions) of co-occurrence values of dependent variables per independent variable (the GO and COME verbs under study). A similarity/dissimilarity matrix is first computed, followed by computing a cluster structure based on a specific amalgamation rule.[5] The resulting cluster structure can then be visually represented in a dendrogram. The calculations involved in the different stages of hierarchical agglomerative cluster analysis have been made easier to conduct using BP 1.01 script, a program written by Stefan Gries (2009) for R. This R-based script uses a host of statistical methods required in the stages mentioned above. It initially generates a co-occurrence table of relative frequencies of the different levels (IDTAG-LEVELS) within variables (IDTAGs).[6] Table 5 shows a sample of such output table generated by BP 1.01 for the distribution of TENSE by COME verb.

IDTAG-LEVEL	*atā*	*ḥaḍara*	*ğā'a*	*qadima*	
FUT	0.028	0.076	0	0.002	columns
IRR	0.188	0.126	0.022	0.022	sum
PAST	0.162	0.694	0.97	0.966	to
PRES	0.622	0.104	0.008	0.01	1.0

Table 5. Sample of a co-occurrence table generated by the BP 1.01 script for the variable (IDTAG) TENSE by COME verb.

The BP 1.01 script returns a comprehensive table with similar values for all dependent by independent variable co-occurrences that have been fed into the script. This particular table can be subjected to a number of tests including the hierarchical agglomerative cluster analysis. For this particular clustering technique I relied on the (dis)similarity metric 'Canberra', and 'Ward' as the amalgamation rule that computes a cluster structure.[7]

The dendrogram in Figure 1 shows two major divides between the four verbs that the hierarchical agglomerative cluster analysis deemed significant. The first cluster (on the left) formed in this analysis appears to group the verbs *atā* and *ğā'a* together, while the other cluster groups *ḥaḍara* and *qadima* together. Here, we find that the AU *p*-value (Approximately Unbiased) – which is a probability measure computed through multi-scale bootstrap resampling – for the cluster containing *ḥaḍara* and *qadima* is calculated to approximate 87%, while the AU *p*-value for the cluster of *atā* and *ğā'a* is 82%. Again, this does not necessarily imply that *ḥaḍara* and *qadima* are highly similar, but that they are very dissimilar from *atā* and *ğa'a* in their usage. Indeed, subsequent mutli-variate as well as qualitative analyses showed that *atā* and *ğā'a* shared more usage patterns than they did with the other COME verbs.

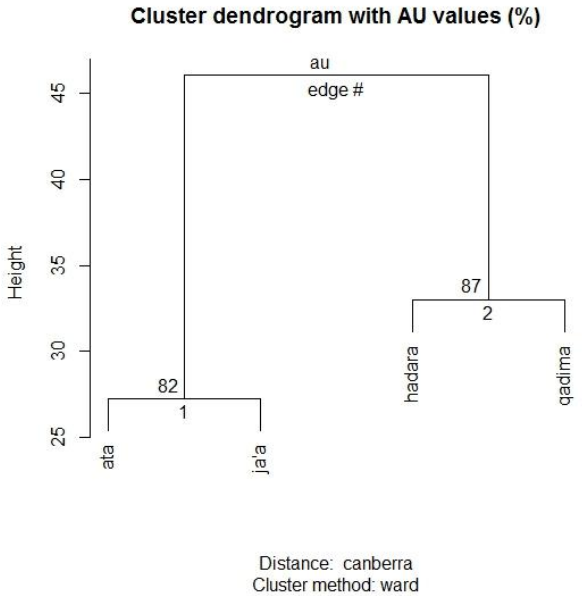

Figure 1. Dendrogram based on the COME multivariate data frame.

4.3 Polytomous logistic regression

Another multi-variate analysis that can be applied to this kind of data frame for the purpose of examining patterns of variable interaction is logistic regression. Polytomous logistic regression analysis (explained in detail in Arppe,

[5] An amalgamation rule is what determines whether or not two items are sufficiently similar in order to be linked or clustered together.

[6] The idea of an ID tag was introduced by Atkins (1987) in her work on *danger*, where she examined collocates, colligations, POS, as well as other characteristics of the key word. An ID tag was therefore used to refer to the individual contextual features co-occurring with the keyword.

[7] For a detailed description of this clustering method see Gries (2009), pp 306-319.

2008), in particular, applies advanced algorithms in order to determine the relative effects of multiple predictor variables (the tags/contextual features) on the choice of more than two outcome variables (i.e. the four COME verbs and the three GO verbs). Generally speaking, logistic regression would estimate probabilities of the occurrence of each COME or GO verb given a particular context of use, and is therefore compatible with the view that linguistic choices are probabilistic rather than categorical (Bresnan, 2006; Arppe, 2007, 2008, 2009; among others). In a nutshell, polytomous logistic regression estimates variable parameters which can be interpreted "naturally" as *odds* (Harrell 2001). In other words, it determines the extent to which the existence of a variable (i.e. feature/tag) in the context increases (or decreases) the *chances* of a particular outcome (i.e. verb) to occur, with all the other explanatory variables being equal.

To illustrate, we can conduct such analysis on the annotated COME data frame. The first step is to select a set of variables to include in the model. Note, however, that the binary and nominal variables listed in Table 1 need to be converted into the form of logical variables in order to be included in the logistic regression model.[8] The selection of these variables relies, first of all, on a mono-variate analysis (such as the inspection of standardized Pearson's residuals) as a means of figuring out which variables seem to have explanatory potential as opposed to those that do not. A second criterion for variable selection relies on inspecting pair-wise association patterns between variables. That is to say, we need to examine the extent to which certain variables have a high rate of co-occurrence, as a means of reducing collinearity in the regression model. The model listed in (5) includes 30 logical variables as the independent variables and the COME verb as a dependent variable.

(5) **VERB ~** TENSE.FUT + TENSE.PAST + TENSE.PRES + ASPECT.HAB + ASPECT.SIMPLE + MORPH_ASP.MOOD.SUBJN + TRANSITIVITY.YES + SUBJ_NUM.PL + SUBJ_PER.1ST + SUBJ_PER.3RD + SUBJ_GEN.FEM + SUBJ_CAT.ACTIVITY + SUBJ_CAT.COMMUNICATION + SUBJ_CAT.DEMONSTRATIVE + SUBJ_CAT.EVENT + SUBJ_CAT.GROUP + SUBJ_CAT.INDIVIDUAL + SUBJ_CAT.STATE + SUBJ_CAT.TIME + NEGATION.YES + PP.YES + LOC_ADV.YES + ADVERBIAL.YES + GOAL.YES + SOURCE.YES + MANNER.YES + SETTING.YES + PURPOSIVE.YES + COMITATIVE.YES + TEMPORAL.YES

The overall *accuracy* rate calculated for this model is 0.845. The *accuracy* measure (Menard, 1995: 28-30; Arppe, 2008: 129-132) corresponds to the number of times the model assigned the highest probability estimate to the actually observed verb in a given annotated context. We can also examine the individual accuracy rates per verb as a means of zeroing in on which particular verb(s) the model was more successful in predicting. We can now examine the probability estimates that the polytomous logistic regression analysis assigns to each of the COME verbs per annotated context (4 verbs * 2,000 sentences). These estimated probabilities range from very high values (approaching 1.00) to very low values (approaching 0.00) and any values in between, depending on the set of predictors (i.e. contextual features) present in a particular context of use. We can illustrate with sentences (6) and (7) which are extracted from the original data frame. In (6) the verb received an almost categorical probability estimate, while in (7) the verb received a less categorical probability estimate. It is also possible to examine the set of contextual features that each sentences was tagged for and which were used as predictor variables in the logistic regression model

(6)

atā = 0.022 *ḥaḍara = 0.000* ***ǧā'a = 0.978*** **(observed)** *qadima = 0.000*	**contextual features used (in the model):** TENSE.PAST + ASPECT.SIMPLE + SUBJ_PER.3RD + SUBJ_CAT.DEM + LOC_ADV.YES + SETTING.YES

جاء ذلك خلال تصريحات أدلى بها الوزير خورشيد

ǧā'a.PERF.3SG.M	DEM	ADV
came	that	during

statements	declare.PERF.3SG.M
statements	declared

INST=CL.3SG.F	ART=minister	Khurshid
by it	the minister	Khurshid

'This came during statements that the minister Khurshid made'

[8] Every level of variable is turned into an individual (logical) variable with the levels TRUE/FALSE indicating whether this variable has or has not been observed in the context of use. For instance, the variable TENSE has four levels: PRESNT / PAST / FUTURE / NIL. When turned into logical variables, we end up with four different variables (TENSE_PRESENT, TENSE_PAST, TENSE_FUTURE and TENSE_NIL), the presence or absence of which is indicated by TRUE or FALSE.

(7)

atā = 0.199 ***ḥaḍara = 0.137*** **(observed)** *ǧā'a = 0.247* *qadima = 0.416*	**contextual features used (in the model):** TESNE.PAST + ASPECT.SIMPLE + SUBJ_PER.3RD + SUBJ_CAT.HUMAN + PP.YES + LOC.ADV.YES + MANNER.YES + COMITATIVE.YES

وقد حضر الأب علي الفور ومعه عددا من زملائه الأطباء

CONJ=DM	***ḥaḍara***.PERF.3SG.M	ART=father	LOC
and already	came	the father	on

ART=immediately	CONJ=COM-CL.3SG.M	number
the immediately	and with him	number

ABL	colleagues-CL.3SG.M.GEN	ART=doctors
of	his colleagues	the doctors

'And the father came immediately with a number of his physician colleagues'

The sentence in (6) can be considered as a prototypical use of the verb *ǧā'a*. In (7), however, note that the verb which received the highest probability estimate was not the actually observed verb in that context. Nevertheless, all four verbs were assigned more-or-less equal probability estimates. This may indicate that this is one context of use in which the four COME verbs can be used interchangeably. Relying on my native speaker intuition, substituting the observed verb with the other COME verbs in (7) does not raise any red flags, especially since this particular contexts of use indicates physical motion of a HUMAN agent, as I discussed earlier in this paper.

Of course, not all predictions made by the model were accurate. Among the sentences for which a single verb received a very high probability estimate, a number of instances in which the predicted verb was not the observed verb were found. Such sentences proved to be worthy of scrutiny due to the fact that some of these "mis-predictions" were associated with less typical uses of the verb that was actually observed in context. For instance, in (8), the verb *qadima* was the verb observed in context, yet the model chose *ḥaḍara* instead as the verb that was most fitting in that context.

(8)

atā = 0.022 ***ḥaḍara = 0.962*** **(predicted)** *ǧā'a = 0.005* ***qadima = 0.011*** **(observed)**	**contextual features used (in the model):** SUBJ_PER.3RD + SUBJ_CAT.HUMAN + ADVERBIAL.YES + GOAL.YES + MANNER.YES + TRANSITIVITY.YES

وكان علي بن عبد الله إذا قدم مكة حاجا أو معتمرا عطلت قريش مجالسها

wa=kāna	*'ali bin 'abdillah*		*iḏā*
CONJ=be.PERF.3SG.M	Ali Bin Abdullah		COND
and was	Ali Bin Abdullah		if

qadima	*makka-ta*	*ḥāǧǧan*	*aw*
qadima.PERF.3SG.M	Mecca-ACC	pilgrim	CONJ
he came	Mecca	pilgrim	or

mu'tamiran	*'aṭṭalat*
pilgrim	suspend.PERF.3SG.F
minor.pilgrim	suspended

qurayš	*maǧālisa-ha*
Quraysh	meetings-CL.3SG.F
Quraysh	its meetings

'When Ali bin Abdullah used to come to Mecca on a pilgrimage Quraysh would suspend its meetings'

Interestingly, this particular usage of *qadima* in (8) can be found in a specific genre, that of historical narrative. While *atā, ǧā'a*, and *ḥaḍara* may all appear in transitive constructions in MSA, *qadima* normally does not. It is, however, used in transitive constructions to signal a shift in register, as in the example in (8). Since such pattern of use occurs less frequently than the general overall usage of *qadima*, the model assigns *ḥaḍara* instead as the most plausible verb choice for such context. Careful inspection of "mis-predictions" such as the above is, therefore, an important step to identify the less typical uses of verbs, as well as to decide whether the variable set chosen for the model has or has not been effective in accounting for verb usage. The probability estimates calculated for the GO data frame did not yield such satisfying results, and did not necessarily agree with my native speaker's intuition. In Abdulrahim (2013: 101) I attributed such findings to the set of variables that GO verbs were coded for in the data frame (which, more or less, resembled the variable set COME verbs were coded for). More specifically I suggested that the data frame should include more lexical or collocational variables.

5 Conclusions

The methodological approach to lexical analysis, described here, represents a departure from traditional, compartmentalized treatments of the Arabic verb. In this paper, I have adopted a construction-based approach that considers various aspects of language (morphology, syntax, semantics, etc.) as equally responsible for defin-

ing the behavior of a linguistic item. The creation of a 500-row data frame per verb has allowed us to probe into the frequency and distribution facts regarding the usage of seven highly frequent motion verbs in MSA. Moreover, the annotation of each corpus return for a wide range of contextual and semantic features offered the possibility of foregrounding the most prototypical aspects of use for each verb, as well as highlighting shared patterns of usage among the near-synonymous verbs in a set. In this paper, I have argued that the value of constructing a data frame of this type lies in developing more sophisticated lexico-syntactic frames of linguistic items, in that it allows us to extract preferred profiles of the lexical or constructional items under study.

Thankfully, there is a wide range of statistical tests that have made the examination of and search for lexico-syntactic patterns in large data frames easier and more manageable. These statistical tests vary from simple, mono-variate exploratory test to complex and multi-variate predictive models. Each one of the three statistical analyses discussed in this paper serves to highlight a particular aspect of variable distribution and variable interaction and, thus, helps us understand the complexity of the relationship between the near-synonymous COME and GO verbs. Generally speaking, the application of such statistical tests to large, multi-variate data frames helps us examine the particular linguistic features that characterize lexical and constructional choices, which may have direct applications in natural language generation. In addition, the identification of prototypical and marginal uses of verbs – discussed particularly in 4.3 – can possibly contribute to developing readability assessment of texts for learners of MSA.[9]

Finally, lexicographic treatments of the highly frequent motion verbs discussed in this paper, as exhibited in bilingual and, mostly, monolingual dictionaries, range from almost adequate to completely mis-representative descriptions of the major and minor senses of these verbs (Abdulrahim, 2013). Many monolingual dictionaries follow a traditional and highly ideological system of lexical representation whereby archaic uses of a lexical item are foregrounded and little attention is paid to more contemporary uses. The quantitative (and qualitative) analysis of a data frame such as the ones described here can help tease apart the different idiosyncratic uses for each of the seven motion verbs as well as identify the most and the least prototypical uses. One of the practical applications of such a data frame, therefore, is to create extensive, usage-based dictionary entries that are more representative of contemporary language use and that would be useful for the native speaker of the language, the language learner, and the language researcher.[10]

Acknowledgments

I would like to thank Professors Antti Arppe, John Newman, and Sally Rice at the University of Alberta for their constant guidance and feedback throughout the various stages of the creation of this data frame, the application of statistical tests, and the linguistic analysis. I would also like to thank the ANLP 2014 reviewers of this paper for their encouraging feedback and insights.

References

Abdulrahim, Dana. 2013. A corpus study of basic motion events in Modern Standard Arabic. Unpublished doctoral dissertation. University of Alberta. Edmonton, Alberta. To be found on http://hdl.handle.net/10402/era.33921

Abdulrahim, Dana. (submitted). Quantitative approaches to analyzing COME constructions in Modern Standard Arabic. To appear in A. Hardie, A. T. McEnrey (eds.), *Arabic Corpus Linguistics*.

Agresti, Alan. 2002. *Categorical data analysis* (2nd ed.). Hoboken: John Wiley & Sons, Hoboken.

Arppe, Antti. 2007. Multi-variate methods in corpus-based lexicography: A study of synonymy in finnish. *Proceedings from the Corpus Linguistics Conference (CL2007)*, Birmingham, United Kingdom.

[9] I would like to thank an anonymous reviewers of this paper for pointing out these particular applications of the statistical methods discussed here.

[10] See Abdulrahim for three samples of suggested usage-based dictionary entries for the COME verb *atā*: (1) a *corpus-illustrated* dictionary entry that elaborates on the existing (bilingual) dictionary entries of the verb by supplementing relevant corpus examples for each verb sub-sense or usage; (2) a minimalist sub-sense frequency-based dictionary entry that orders the verb entries according to the frequency of occurrence of the overall general usage (physical, metaphorical, etc); and (3) a *usage-based* dictionary entry for *atā* that is directly based on the quantitative and qualitative analyses conducted in her study (2013: 243-249)

Arppe, Antti. 2008. Univariate, bivariate and multivariate methods in corpus-based lexicography - a study of synonyms. (PhD, University of Helsinki). *Publications of the Department of General Linguistics,* 44.

Arppe, Antti. 2009. Linguistic choices vs. probabilities - how much and what can linguistic theory explain? In: Featherston, Sam & Winkler, Susanne (eds.) *The Fruits of Empirical Linguistics 1*, pp 1-24. Berlin: de Gruyter.

Atkins, Beryl. T. S. 1987. Semantic ID tags: Corpus evidence for dictionary senses. Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary, pp17–36.

Bresnan, Joan. 2006. Is syntactic knowledge probabilistic? Experiments with the English dative alternation. *Pre-Proceedings of the International Conference on Linguistic Evidence,* Tübingen, Germany. (pp. 2-4).

Bybee, Joan. 2010. *Language, usage and cognition*. Cambridge: Cambridge University Press.

Croft, William. & Allan. D. Cruse. 2004. *Cognitive linguistics*. Cambridge: Cambridge University Press.

Divjak, Dagmar. S., & Stefan T. Gries. 2006. Ways of trying in Russian: Clustering and comparing behavioral profiles. *Corpus Linguistics and Linguistic Theory, 2*(1), 23-60.

Firth, John. R. 1957. A synopsis of linguistic theory, 1930-1955. In J. R. Firth. 1968. *Selected Papers of J. R. Firth 1952-1959* (pp. 168-205). London: Longman.

Gries, Stefan Th. 2006. Corpus-based methods and cognitive semantics: The many meanings of *to run*. In S. Th. Gries, & A. Stefanowitsch (Eds.), *Corpora in cognitive linguistics: Corpus-based approaches to syntax and lexis,*pp 57-99. Berlin/New York: Mouton de Gruyter.

Gries, Stefan Th. 2009. *Statistics for linguistics with R: A practical introduction*. Berlin: De Gruyter Mouton.

Gries, Stefan Th. & Dagmar S. Divjak. 2009. Behavioral profiles: A corpus-based approach towards cognitive semantic analysis. In Vyvyan Evans, & S. S. Pourcel (Eds.), *New directions in cognitive linguistics*, pp 57-75. Amsterdam/Philadelphia: John Benjamins.

Gries, Stefan. Th. & Naoki Otani. 2010. Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal, 34*, pp 121-150

Harrell, Frank E. 2001. *Regression modeling strategies: With applications to linear models, logistic regression and survival analysis*. New York: Springer-Verlag.

Hastie, Trevor, Robert Tibshirani, & Jerome Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction*. New York: Springer.

Langacker, Ronald. 1987. *Foundations of cognitive grammar, vol. I: Theoretical prerequisites*. Stanford, Calif.: Stanford University Press.

Menard, Scott. 1995. *Applied logistic regression analysis*. Thousand Oaks: Sage Publications.

Data sets and R scripts

Abdulrahim, Dana. (forthcoming). Arabic come and go. A micro-corpus of 3500 occurrences of seven motion verbs in MSA (*atā*, *ğā'a*, *ḥaḍara*, *qadima*, *ḏahaba*, *maḍā*, and *rāḥa*), annotated for a wide range of morphological, syntactic, and semantic variables

Arppe, Antti. 2012. polytomous: Polytomous Logistic Regression for fixed and mixed effects. R package version 0.1.4.

Gries, Stefan. Th. (2009). BehavioralProfiles 1.01. A program for R 2.7.1 and higher.

Appendix A. Examples for annotation for semantic variables

variable	sample of annotation
SUBJECT CATEGORY:	
ACTIVITY	هجوم 'attack', عمليات 'operations', تصويت 'voting'
ANIMAL	جواد 'horse', كلب 'dog'
ATTRIBUTE	كرم 'generosity', شهرة 'fame'
BODY	عيون 'eyes', قدم 'foot'
COGNITION	تفكير 'thought', خيال 'imagination'
COMMUNICATION	سؤال 'question', تقرير 'report'
CONTENT (of a document/speech)	جاء في البيان *ğā'a*.PERF.3SG.M LOC ART=statement 'came in the statement…', جاء في الرسالة *ğā'a*.PERF.3SG.M LOC ART=letter 'came in the statement…',

DEMONSTRATIVE	جاء ذلك *ǧā'a*.PERF.3SG.M LOC DEM 'that came…', جاء هذا *ǧā'a*.PERF.3SG.M LOC DEM 'this came…', etc.
EVENT	إجتماع 'meeting', ندوة 'symposium', قمة 'summit'
GROUP (representing humans collectively)	اليابان 'Japan', المنتخب 'varsity', الحكومة 'the government'
HUMAN	الأولاد 'the boys', البابا 'the Pope'
LOCATION	موقع 'location', المدن 'the cities'
NOTION	الأذية 'harm', مصدر 'source'
PHYSICAL OBJECT/ARTIFACT	منح 'grants', القمح 'wheat'
SENSE	صوت 'voice/sound'
STATE	الموت 'the death', مرحلة 'phase'
SUBSTANCE	حرائق 'fires', مطر 'rain'
TIME	موسم'season', الغد 'tomorrow'
GOAL PHRASE: YES	مساعداتنا تذهب **إلى الشيشان** aid.CL.1PL.GEN *ḏahaba*.IMPF.3SG.F **ALL ART=Chechnya** 'Our aid goes to Chechnya'
SOURCE PHRASE: YES	الهجرات الجنوبية التي قدمت **من الهند** ART=immigrations ART=southern RP *qadima*.PERF.3SG.F **ABL ART=India** 'The southern immigrations that came from India…'
MANNER PHRASE: YES	هذه الجهود لم تذهب **هدرا** DEM ART=efforts NEG *ḏahaba*.JUSS.3SG.F **vain.ADV** 'These efforts weren't in vain'
SETTING PHRASE: YES	بل تأتي **في اطار مخطط شامل** CONJ *atā*.IMPF.3SG.F **LOC frame plan comprehensive** 'It, however, comes as part of a comprehensive plan'
PATH PHRASE: YES	خسارة أتت **على رأسمال البنك** deficit *atā*.PERF.3SG.F **LOC capital ART=bank** 'A deficit that destroyed the bank's capital'
PURPOSIVE PHRASE: YES	ذهبت **لزيارته** وسألته *ḏahaba*.PERF.1SG **PURP=visit.CL.3SG.M.ACC** CONJ=ask.CL.3SG.M.ACC 'I went to visit him and asked him'
COMITATIVE PHRASE: YES	برنامجكم لم يأت **بجديد** show.CL.3PL.M.GEN NEG *atā*.PERF.3SG.M **COM=new** 'Your show did not come up with anything new'
TEMPORAL PHRASE: YES	أذهب لتناول أيس كريم **في أي وقت** *ḏahaba*.IMPF.1SG PURP=have.VN ice cream **LOC any time** 'I go to have ice cream at any time'

DEGREE PHRASE: YES	تأتي **دائما** عبر عمليات السطو المنتظم *atā*.IMPF.3SG.F **ADV** LOC operations burglary ART=organized 'Comes always through operations of organized burglary'