RiCL Research in Corpus Linguistics

Object and subject Heavy-NP shift in Arabic

Emad Mohamed
Suez University / Egypt

Abstract – In order to examine whether Arabic has Heavy Noun Phrase Shifting (HNPS), I have extracted from the Prague Arabic Dependency Treebank a data set in which a verb governs either an object NP and an Adjunct Phrase (PP or AdvP) or a subject NP and an Adjunct Phrase. I have used binary logistic regression where the criterion variable is whether the subject/object NP shifts, and used as predictor variables heaviness (the number of tokens per NP, adjunct), part of speech tag, verb disposition (ie. whether the verb has a history of taking double objects or sentential objects), NP number, NP definiteness, and the presence of referring pronouns in either the NP or the adjunct. The results show that only object heaviness and adjunct heaviness are useful predictors of object HNPS, while subject heaviness, adjunct heaviness, subject part of speech tag, definiteness, and adjunct head POS tags are active predictors of subject HNPS. I also show that HNPS can in principle be predicted from sentence structure.

Keywords – Heavy NP Shift, Arabic, corpus, syntax, logistic regression

1. INTRODUCTION

Heavy NP shift (Kimball 1973) occurs in English when a heavy or long direct object NP occurs in the final position in the clause, separated from the verb by something like a prepositional phrase. An example is the English sentence from George Orwell's *Animal Farm* (cited by Stallings et al. 1998) *Snowball had found in the harness-room an old green tablecloth of Mrs. Jones's*, where the delayed object *an old green tablecloth of Mrs. Jones's* is obviously longer and heavier than the prepositional phrase *in the harness-room*. While Arabic seems to have a similar phenomenon, the picture is more complicated. An Arabic translation of the *Animal Farm* example, whether we maintain the shifted order (example (1a)) or use an unshifted one (in (1b), should be acceptable:

(1a) Shifted Arabic rendering of the *Animal Farm* sentence[1]

وكان	سنوبول	قد	وجد	في	غرفة	العدة	غطاء	منضدة	قديما	يخص	السيدة	جوز
and-was	Snowball	really	found	in	room	the-tool	cover	table	old	belong-to	Mrs	Jones
	Subject		Verb		PP					Object		

(1b) Unshifted Arabic rendering of the *Animal Farm* sentence

وكان	سنوبول	قد	وجد	غطاء	منضدة	قديما	يخص	السيدة	جوز	في	غرفة	العدة
and-was	Snowball	really	found	cover	table	old	belong-to	Mrs	Jones	in	room	the-tool
	Subject		Verb				Object				PP	

[1] Although Arabic is a right-to-left language, the examples go from left to right.

Traditional Arabic grammarians, whose work was mostly descriptive, state that the natural order of the sentence in Arabic is Verb > Subject > Object, following Sibawayh (1988: 35), who claims that the SVO order is also "good and frequent Arabic, and it seems that they put first what is of importance and concern to them although they [ie. the subject and object] are both important". This stylistic view of Arabic word order has iterated in the case of the prepositional phrase and the adverb as well, but while there is a preferred, or default, VSO order, the position of adverbs and prepositional phrases in the sentence does not seem to have a strong preference. Hassan (1974: 245) writes that "an adverb needs to be attached to a head, but the head does not have to precede it", and then cites several examples with the adverb in disparate positions in the sentence. Hassan (1974: 444–445) and Ali (2011) share the same view about prepositional phrases, but Anis (1978: 234) explains that when the subject is long, in terms of the number of words it comprises, then it comes after the other, shorter, head dependents. This may have been the first mention of HNPS in Arabic, and while it was concerned with the subject, it can also apply to the object in relation to the prepositional phrase and the adverb.

This also highlights another problem with HNPS in Arabic. Arabic has a freer word order than English, and while in English the problem is usually whether the object precedes, or follows, the PP, in Arabic prepositional phrases and adverbs can potentially be placed anywhere around the head verb, the subject, and the object. HNPS may then be the case when the Verb-dependent PP precedes the subject in a VSO sentence, or the object in SVO sentences.

There does not seem to be any (corpus-based) work on HNPS in Arabic, whereas English HNPS has received some attention. Stallings et al. (1998) state that speakers find shifted structures with a direct short object NP to be awkward or ungrammatical. Object length, complexity, or the difference between the length of the object and the PP were also found to play a role in HNPS (Ross 1967; Kimball 1973; Hawkins 1994; quoted in Stallings and McDonald 2011). Others maintain that HNPS is a result of the given-new information processing system, and since longer objects contain more new information, they are shifted to the end (Firbas 1964). There is also the tradition of sentence comprehension which attributes HNPS to constraints placed on the speaker to accommodate the needs of the hearer, who finds it difficult to process large objects in the middle of the sentence (Chomsky and Miller 1963; Wasow 1997), or to facilitate planning and production (Arnold et al. 2000). Stallings et al. (1998) examined the argument structure of verbs and found that those verbs that can potentially take a sentential object and double-object verbs are more likely to trigger HNPS, which would explain that HNPS is also a verb property. While none of these studies investigates Arabic, I expect that (at least some of) these findings may be universal and I will examine how much these affect word order in Arabic. Furthermore, while the phenomenon is called Heavy Noun Phrase Shifting, it is affected by much more than NP-internal factors, and heaviness is just but one of them.

The outline of the article is as follows: in Section 2 I specify the research questions, while in Section 3 I present the data, methods, and definitions used in this study. In turn, Section 4 contains the results and discussion. Finally, Section 5 offers some conclusions and directions for future research.

2. RESEARCH QUESTIONS

In this study I examine a number of linguistic features with the purpose of finding whether they affect the order in which the object and subject NPs occur relative to the other material (PP/Adverb) in the sentence. While many pragmatic features may be of interest (eg. focus and speaker/writer preferences), I limit my variable set to those that can be extracted from a syntactically parsed corpus. The following features represent the predictor variables in a logistic regression model:

1. Part of speech tag. The object can *inter alia* be a common noun, a proper noun, or a pronoun. The object grammatical category could be a factor in deciding whether to shift. In multi-token objects, the object POS tag is taken to be that of the object head word. The same holds true for the subject. Adjuncts, on the other hand, can be either prepositional phrases or adverbs. For the sake of simplicity, I will use the term adjunct for both PPs and adverbs, and I will disambiguate between the two categories only when necessary.

2. Object/Subject definiteness. The object/subject may take on one of three definiteness features: (a) definite, where the noun is prefixed by the definite article *Al*; (b) indefinite, where no *Al* is attached; and (c) construct, where we have a noun-noun compound with a possession meaning; eg. *kitab-ul-Talib* ('student's book'). Construct nouns are semantically definite. Object/subject definiteness may be related to the distinction between given and new information, definiteness indicating already known (given) information. It may be the case then that new (possibly indefinite) information will have priority over given information, and thus a definite object will be more likely to shift.

3. Object/subject number. The object/subject may be plural, dual, or singular.

4. Pronoun in object/subject. Is there a pronoun among the object/subject dependents? The assumption here is that since pronouns need referents, the presence of a pronoun may be an incentive for the object to stay nearer its head verb. While my corpus does not contain pronoun reference resolution information, I assume that most pronouns will refer back to a previously mentioned entity.

5. Verb disposition. HNPS may be a verb property inasmuch as some verbs favor NP shifting while others do not. Stallings et al. (1998) found that verbs that have a history of NP shift are more likely to favor shifting structures. The history of a verb was measured in terms of the verb argument structure, more specifically with regard to: (a) whether the verb can take a sentence complement as well as an NP complement. Those S/NP verbs were found to be more likely to favor shifting than verbs that take only NP complements. While these verbs could be adjacent to their S-complements, this typically changes when the verb also governs a PP or an adverb, as in *Mary said in a loud voice that Bill would sing* and *Mary learned yesterday that she would be allowed to go hiking*; (b) whether the verb takes two objects, as is the case with *give* and *grant*; and (c) whether the verb can be used in a verb particle structure (eg. *My mother brought the question up* vs. *My mother brought up the question of whether we should go on vacation this year*). Only (a) and (b) will be considered in this article since Arabic lacks the verb particle structure. For this feature, I have manually created a list of 13 Arabic double-object verb lemmas, and another list of 43 verb lemmas that can take S complements.

6. Pronoun in PP. This is similar to 4 above, but entails that the PP will be closer to the head if it contains a pronoun.

Using the aforementioned factors as predictor variables and the shift status (whether the NP is shifted or not) as a criterion variable, I seek to answer the following questions:

- is there a statistically significant relationship between the criterion variable and the predictor variables, taken as a set? More specifically, is there a statistically significant relationship between (a) shift status and (b) the variables that constitute the models of Arabic HNPS?
- what is the nature of the relationship between the criterion variable and the predictor variables? Of the variables that constitute the factor model, which variables display logistic regression coefficients that are significantly different from zero? What is the sign of each of these coefficients?
- what is the strength of the relationship between the criterion variable and the predictor variables taken as a group? More specifically, what is the strength of the relationship between (a) shift status and (b) the factors of the factor model, taken as a group? When the logistic regression equation is used to classify instances into groups under the criterion variable, how accurate are these classifications?[2]

3. Data and methods

3.1. Important definitions

HEAVINESS. The term 'heaviness' can be understood in terms of length (how many tokens the object NP has; Kimball 1973), of NP complexity (the depth of the NP and its constituents; Ross 1967), or of the relative length of the NP compared to the other material ruled by the verb (such as a PP) (see, for example, Hawkins 1994). In a corpus analysis, Hawkins found that NP shift did not happen much until the object was four words longer than the remaining material in the sentence. In this study I use the number of tokens as a measure of heaviness for both the NP and the PP (or adverb). Due to the morphological nature of Arabic, I use tokens rather than words. A blank-space delimited word in Arabic can consist of multiple tokens. For example, the word *lmnAzlhm* (English 'to their homes') consists of a preposition *l*, a noun *mnAzl*, and a possessive pronoun *hm*. While I discuss heaviness in general terms here, in the actual analysis there will be several heaviness predictors: object heaviness, subject heaviness, PP heaviness, and adverb heaviness.

ARABIC. By 'Arabic' I mean Modern Standard Arabic (MSA henceforth). MSA is the official language of newspapers, books, TV news, and most informational data in the Arabic-speaking countries, but the language spoken by Arabs is a continuum of dialects (Classical Arabic <> Regional dialects). The use of MSA in this study is not optimal since it is mainly a written language and thus reflects a high degree of planning, but it is the only variety of Arabic for which there is a syntactically parsed corpus, a necessity for conducting such research. Until spoken varieties of Arabic can be accurately parsed, MSA is the only option for corpus-based syntax studies.

HNPS. For the purposes of this study, I treat HNPS for both subjects and objects. The following definitions are relevant:

– Object HNPS occurs when, in an SVO clause, a PP or an AdvP intervenes between the verb and the object where the same verb governs the subject, the object, and the PP/AdvP. The VOS order is excluded since the object in VOS clauses is almost always a pronoun. Object pronouns in Arabic are realized as verbal suffixes, and nothing can thus intervene between the verb and its object. VSO clauses are also excluded since the subject already intervenes between the verb and the object.

---

[2] It is worth noting here that the since logistic regression is binary, in the actual analysis I do not deal with a factor like subject POS, but rather with binary factors of the different POS tag values. For example, if the factor *subject POS* can assume the values NOUN, PRONOUN, and PROPER_NOUN, one then ends up with three binary factors: *IS_SUBJECT_A_NOUN*, *IS_SUBJECT_A_PROPER_NOUN*, *IS_SUBJECT_A_PRONOUN*. I use only two of these in the analysis, treating the third as a baseline, in order to combat multicollinearity, or the dummy variable trap (Dormann et al. 2013). Using this method, I ended up having thirteen variables for the subject HNPS experiment (a 13-factor model) and ten variables for the object HNPS experiment (a 10-factor model).

– Subject HNPS occurs when, in a VS clause, a PP or an AdvP intervenes between the verb and the subject where the same verb governs the subject and the PP/AdvP.

3.2. Data

The data for this study come from the Prague Arabic Dependency Treebank (henceforth PADT; Hajič et al. 2004). The PADT encodes multiple levels of linguistic information for Modern Standard Arabic and is based on newswire text. I do not use the PADT directly but depend instead on the version that was prepared for the CoNLL-X dependency parsing shared task (Buchholz and Marsi 2006). The CoNLL-X version encodes the information in columnar format (shown in Table 1), which makes it easy to extract the needed information. Table 2 uses Buckwalter's (2002) encoding for Arabic, which I use throughout.

#	Token	Lemma	POS	Linguistic Information	Head	Function
1	taboda>u	bada>-a	V	Mood=I|Voice=A|Person=3|Gender=F|Number=S	0	Pred
2	AlHamolapu	Hamolap	N	Gender=F|Number=S|Case=1|Defin=D	1	Sb
3	taEAwunAF	taEAwun	N	Case=4|Defin=I	1	Obj
4	maEa	maEa	P	_	3	AuxP
5	AljamoEiy~Ati	jamoEiy~ap	N	Gender=F|Number=P|Case=2|Defin=D	4	Obj
6	Al>aholiy~api	>aholiy~	A	Gender=F|Number=S|Case=2|Defin=D	5	Atr

Table 1. A sentence in the CoNLL format

Table 1 shows that the sentence has six tokens numbered from 1 to 6. The first word *taboda>u* has the lemma *bada>a* and the Part of Speech V. The linguistic information column (Mood=I|Voice=A|Person=3|Gender=F|Number=S) indicates *inter alia* that the verb is in the indicative mood and that it is in the singular form. The head column shows that it is headed by token 0, which is the imaginary head of every tree, and the function column states that it functions as a head. If we scroll down to tokens 2 and 3, we can see that their head is token 1, and we have a situation in which a verb governs both a subject and an object. In this example, the order is Verb-Subject-Object which seems to be the canonical order in (written) Arabic (see Section 1 above).

To study Object NP shifting in Arabic, I have extracted 293 SVO clauses from the PADT, each of which has a transitive verb that governs a subject, a noun object, and a PP or an adverb. Objects that are pronouns have been excluded from the analysis since pronominal objects obligatorily get reduced to suffixes on the verb, and no element can intervene between a verb and its pronominal object. This small dataset shows that the majority of the examples do not show any shifted NP structures (78.8%). The adjunct phrase precedes the object in 21.2 percent of the cases. For subject HNPS, the data set comprises 2,893 examples with Verb Subject PP. Out of these, 340 (11.75%) have shifted subject NPs.

3.3. Logistic regression for hypothesis testing

Logistic regression is a statistical procedure that allows researchers to examine the relation between a dichotomous outcome variable and one or more predictor variables (Hatcher 2013: 316). It is related to multiple regression but differs from it in that the outcome to be predicted is categorical rather than numerical. The procedure is typically used to answer questions about the nature, significance, and relations between predictor variables, as a set as well as individually, and the outcome categorical variable whose levels are assigned likelihoods of occurrence in the process. Logistic regression is iterative and uses maximum likelihood estimation to find the model that best fits the data. Logistic regression is commonly used in Natural Language Processing (NLP) predictions, although it is commonly known in the NLP community as Maximum Entropy Classification. I use logistic regression for my hypothesis testing. For the actual calculations, I use a Python pipeline consisting of Pandas,[3] Statsmodels,[4] and Scikit-learn (Pedregosa et al. 2011).

[3] http://pandas.pydata.org
[4] http://statsmodels.sourceforge.net

4. RESULTS AND DISCUSSION

4.1. Object HNPS

For object HNPS, we have run the logit regression with 10 predictor variables, as shown in Table 2. The Pseudo R-squ. for the model is 0.4084. Only two factors, object size and adjunct size, have been found to be statistically significant, and are grayed out in Table 2. Table 2 shows that the logistic regression model is significant, which means that at least one of the predictor variables can explain (part of) the variation in the criterion variable (object shift status). Only the object size and adjunct size, which were previously introduced as heaviness, have been found to be statistically significant at $p < 001$.

Factor	Coef	std err	Z	P>|z|	[95.0% Conf. Int.]	
Object Size	0.1356	0.028	4.861	0.000	0.081	0.190
PP Size	-0.7973	0.174	-4.581	0.000	-1.138	-0.456
Verb Disposition	1.8418	1.240	1.485	0.137	-0.589	4.272
Pronoun in PP	0.4556	1.244	0.366	0.714	-1.982	2.893
Pronoun in Object NP	-0.2775	1.632	-0.170	0.865	-3.476	2.921
Object POS Pronominal	-0.0757	0.741	-0.102	0.919	-1.528	1.377
PP POS = Adverb	0.1958	0.487	0.402	0.688	-0.759	1.151
Object Definiteness = Indefinite	2.1002	1.192	1.762	0.078	-0.235	4.436
Object Definiteness = Construct	1.8511	1.512	1.225	0.221	-1.112	4.814
Object Number = Plural	1.7672	0.957	1.847	0.065	-0.108	3.642
intercept	-1.8067	1.363	-1.326	0.185	-4.478	0.864

Table 2. Object HNPS results: Observations = 293, Pseudo R-squ. = 0.4084, Log-Likelihood = -89.454, LL-Null= -151.21, LLR p-value = 9.795e-22

Object size is positively correlated with shifting, while adjunct size is negatively correlated. As an object NP increases in size, so does its chance of being shifted, and as an adjunct increases in size, the chance of the object NP shifting becomes smaller. This is in line with the results on English (eg. Stallings et al. 1998). The average size of object NPs in the data set is 5.4 tokens with a minimum of 1, a maximum of 50, and a standard deviation of 8.14. Shifted NPs have an average size of 12 tokens, with a standard deviation of 12.23. Unshifted object NPs have an average size of 3.7 tokens with a standard deviation of 5.4. The average size of adjuncts in the data set is 6.1 with a minimum of 1, a maximum of 54, and a standard deviation of 7.3. Shifted adjuncts have an average size of 2.35 tokens, with a standard deviation of 1.15 and a maximum length of 5 tokens. Unshifted adjuncts have an average size of 7.1 and a standard deviation of 7.97.

The relationship between objects and adjuncts is shown in Figures 1 and 2. In Figure 1, we can see that unshifted adjuncts dominate the top part of the figure as they generally have more tokens than unshifted object NPs, although there are cases in which this does not hold true. The relation is generally fuzzier than that depicted in Figure 2, where shifted NPs are obviously dominant over shifted PPs.

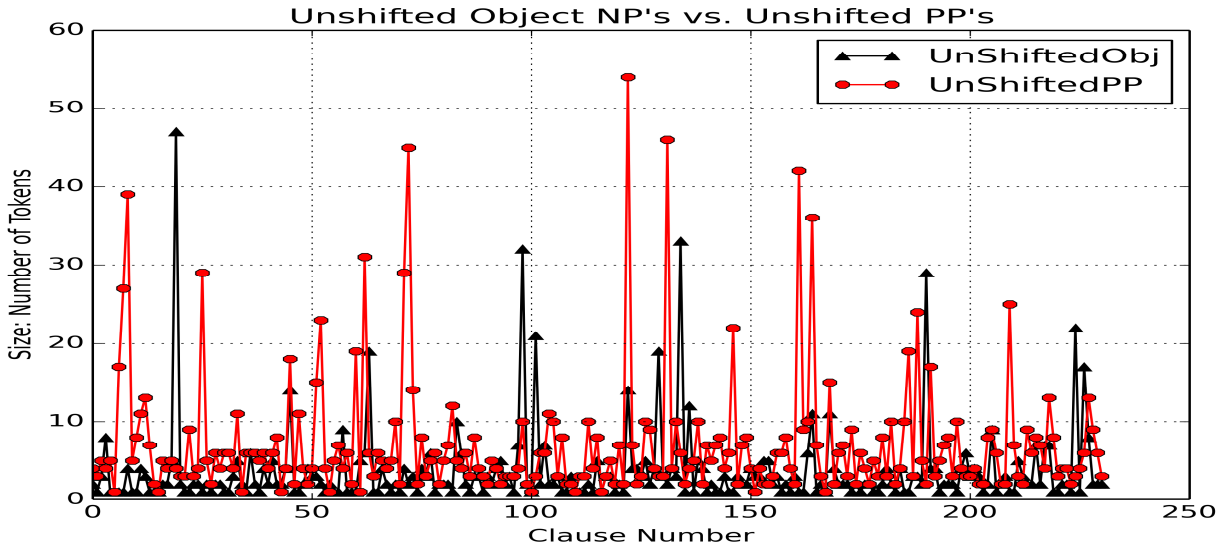

Figure 1. Unshifted objects and unshifted PPs

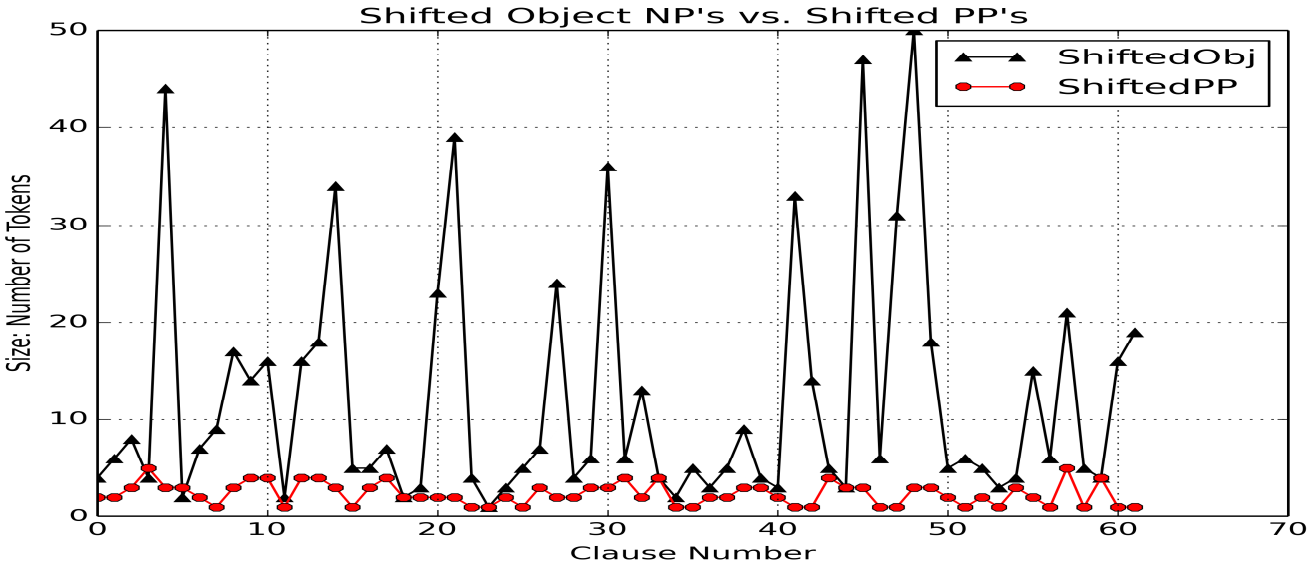

Figure 2. Shifted objects and shifted PPs

In order to estimate how each factor contributes to shifting, I use the odds ratio (henceforth OR). The OR estimates the multiplicative change in the odds of membership in the targeted group for every 1 unit of increase in the predictor variable (Wright 1995: 243; Tabachnick and Fidell 2007: 461–462). In this paper, the Adjusted OR, which controls for the other predictor variables (Huck 2004), is used. Table 3 lists the Adjusted OR scores for the individual predictors along with their 95 percent confidence intervals. Hatcher (2013: 330–331) sets these criteria for interpreting the Adjusted OR:

– OR < 1: the predictor variable is negatively related to the criterion variable, after controlling for the other variables. As scores on this predictor variable increase, subjects are less likely to be assigned to the category coded as 1.

– The no-effect value for an OR is 1: a score of 1 means that the predictor variable is not related to the dichotomous outcome variable after controlling for the other predictors. (Haddock et al. (1998) provide guidelines in which OR scores close to 1 represent weak relationships between the predictor and the criterion, a score greater than 3 may be interpreted as a strong positive relationship, and a score of less than 0.33 may be interpreted as a strong negative relationship.)

– OR > 1: after controlling for the other predictor variables, this predictor variable is positively related to the outcome variable. As scores on this predictor variable increase, subjects are more likely to be assigned to the event category, which is coded as 1 (the shifted class).

Factor	95% Confidence Interval		Odds Ratio
Object Size	1.084260	1.209496	1.145168
Adjunct Size	0.320337	0.633732	0.450564
Verb Disposition	0.554998	71.689923	6.307755
Pronoun in PP	0.137812	18.048961	1.577136
Pronoun in Object NP	0.030934	18.559426	0.757702
Object POS Pronominal	0.216871	3.963344	0.927111
PP POS = Adverb	0.467995	3.161014	1.216281
Object Definiteness = Indefinite	0.790234	84.415688	8.167505
Object Definiteness = Construct	0.328957	123.236961	6.367080
Object Number = Plural	0.897708	38.179081	5.854370

Table 3. Odds ratios for object HNPS

Object size thus has a weak positive relationship with object HNPS, while adjunct size has a moderate relationship. While other factors may have high absolute OR values, their confidence intervals cross 1, which is the no-effect value, and are thus not contributing much to the model.

One way of evaluating a logistic regression model is through classification. In classification, we check how successful the model is in predicting whether language users will shift their NPs in light of the presence of some features (eg. adjunct size and object POS). I use classification for the same purpose here. For evaluation, I use precision, recall, and the *f-1* score. Manning et al. (2009: 192) define precision (P) as the fraction of retrieved documents that are relevant, and recall (R) as the fraction of relevant documents that are retrieved. Precision is then *true positives / (true positives + false positives)* and recall is *true positives / (true positives + false negatives)*. The *f-1* score is the harmonic mean of precision and recall and is computed as *F = 2 x (precision x recall) / (precision + recall)*.

Table 4 presents the classification results for object NP shift status and shows that the logistic regression model performs differently on the negative class (detecting when the NP is not shifted) than on the positive class.

Class	Precision	Recall	f1-score	Support
no-shift	0.88	0.96	0.92	231
shift	0.79	0.53	0.63	62
avg/total	0.86	0.87	0.86	293

Table 4. Object HNPS classification

The results show that while the model is performing well, the phenomenon cannot be fully explained by the logistic regression model, which leaves room for some other (possibly non-syntactic) explanation.

4.2. Subject HNPS

A total of 2,893 examples with verbs, subject NPs, and adjuncts have been used, of which 340 (11.75%) are shifted (ie. the adjunct intervenes between the verb and its subject in a VS order). Table 5 presents the logistic regression values.

Factor	Coef	std err	Z	P>|z|	[95.0%	Conf. Int.]
Subject Size	0.1238	0.012	9.937	0.000	0.099	0.148
Adjunct Size	–0.7618	0.064	–11.911	0.000	–0.887	–0.636
Verb Disposition	0.1217	0.295	0.412	0.680	–0.457	0.700
Verb Takes 2 Objects	0.1072	1.104	0.097	0.923	–2.056	2.270
Pronoun in PP	0.2572	0.200	1.287	0.198	–0.134	0.649
Pronoun in Subject NP	–0.1177	0.236	–0.500	0.617	–0.580	0.344
Subject POS Pronominal	–1.4427	0.436	–3.309	0.001	–2.297	–0.588
Subject is Proper Noun	–0.5085	0.236	–2.154	0.031	–0.971	–0.046
Adjunct = Adverb	–2.0456	0.207	–9.879	0.000	–2.451	–1.640
Subject Definiteness = Indefinite	–0.4702	0.225	–2.086	0.037	–0.912	–0.028
Subject Definiteness = Construct	–0.2981	0.236	–1.265	0.206	–0.760	0.164
Subject Number = Plural	–0.0304	0.212	–0.143	0.886	–0.446	0.386
intercept	0.7096	0.282	2.520	0.012	0.158	1.261

Table 5. Subject HNPS results: Observations = 2,893, Pseudo R-squ. =0.2978, Log-Likelihood = –735.30

Unlike object NP shifting, which was affected by only two factors, subject NP shifting seems to be affected by five factors: subject size, adjunct size, subject POS, adjunct POS, and Subject Definiteness:

Subject and adjunct size. Subject size and adjunct size work in opposite directions, but the latter seems to have a stronger effect. Judging by the absolute coefficient value, adjunct size is a more important predictor of subject HNPS than the size of the NP subject itself. The effect size of the subject size (measured in terms of OR) is weak, while the effect size of the adjunct size is moderate (Haddock et al. 1998).

Subject POS tag. Out of the 2,893 subjects, 80 percent are common nouns, 14 percent are proper nouns, and 6 percent are pronominal subjects. Compared to common nouns, proper nouns and pronouns do not prefer being shifted. Proper nouns have an OR of 0.61, which is a moderate negative contribution, while pronominal subjects have an OR of 0.24, which is a very strong negative contribution. According to these ORs, pronominal subjects are four times less likely to shift than non-pronominal subjects, and proper nouns subjects are about twice less likely to shift than non-proper nouns in the absence of all other factors.

Adjunct POS tag. Adjuncts can either be prepositional phrases or adverb phrases. Adverbs constitute 15.23 percent of all PPs, and they are less likely to engage in a shifted structure. With an OR of 0.13, they have a very strong negative effect as they are about eight times less likely to engage in a subject HNPS structure than prepositional phrases.

Subject Definiteness. An indefinite subject is less likely to shift than a definite one. Its OR is 0.62, which is a moderate effect at best. Indefinite subjects constitute 30 percent of all subjects, while construct state subjects, which are semantically definite, constitute 11.7 percent. The rest of the subjects are definite.

The rest of the predictor variables are not significant and their ORs are very close to 1, indicating a very small effect size. The OR values of all the subject factors, as well as their confidence intervals, are listed in Table 6.

Factor	95% Confidence Interval		Odds Ratio
Subject Size	1.104508	1.159797	1.131815
PP Size	0.411815	0.529158	0.466814
Verb Disposition	0.633455	2.013488	1.129360
Verb Takes 2 Objects	0.128008	9.679992	1.113155
Pronoun in PP	0.874159	1.913256	1.293248
Pronoun in Subject NP	0.560158	1.410658	0.888927
Subject POS Pronominal	0.100543	0.555349	0.236297
Subject is Proper Noun	0.378611	0.955305	0.601405
PP is Adverb	0.086174	0.194031	0.129307
Subject Definiteness = Indefinite	0.401705	0.972025	0.624874
Subject Definiteness = Construct	0.467737	1.177759	0.742213
Subject Number = Plural	0.639943	1.470486	0.970066
intercept	1.170881	3.530268	2.033107

Table 6. Subject Odds Ratios

Predicting subject HNPS (Table 7) seems to be harder than predicting object HNPS in spite of the abundance of examples in the subject function. The model suffers in the shift class recall with a poor performance of 0.19, although its performance on the non-shift class is near optimal.

Class	Precision	Recall	f1-score	Support
no-shift	0.90	0.99	0.94	2553
shift	0.67	0.19	0.29	340
avg/total	0.87	0.89	0.87	2893

Table 7. Predicting subject HNPS

4.3. General discussion

In the foregoing sections, I have presented two models that seek to explain the position of the object NP and the subject NP in relation to the adjunct headed by the same verb, and I am now in a better position to answer the questions raised in Section 2.

There is a significant relationship between the criterion variable (shift status) and the general seven-factor model comprising heaviness, verb disposition, POS tag, definiteness, number, pronoun in NP, and pronoun in adjunct. A significant relationship between the model and the criterion means that at least one of the predictor factors in the model can explain (part of) the variation in the criterion variable, but there are differences between subject HNPS and object HNPS.

As regards object HNPS, only the size factor turned out to be significant. The object model pseudo R-squared is 0.41, which means that the model can explain 41 percent of the variation in the data set. Although the data set is small, it has been shown that the logistic regression model can predict the structure with an *f1*-score of 0.86. This could mean that the traditional syntactic wisdom that Arabic adverbs and prepositional phrases can occur anywhere in the sentence is not entirely true. While there are no hard and fast rules, there are possibly constraints that limit the placement options.

The results on subject HNPS are similar but less compelling. The pseudo R-squared for the subject model is 0.30, which suggests that the subject model is less explanatory than the object one. Also, while the object model has only two significant/effective factors, in the subject model the grammatical category of the adjunct and the subject also have a role to play. When the subject is a pronoun or a demonstrative, the subject is four times less likely to shift, and when the subject is a proper noun, it is about twice less likely to shift too. When the verb governs an adverb along with the subject NP, the subject NP is five times less likely to shift than when the verb governs a prepositional phrase. Definiteness also plays a (small) role in subject HNPS: with an OR of 0.62, the effect is negative and moderate.

5. Conclusion

In this article I have carried out a study of HNPS in Arabic for both subjects and objects. The models presented can explain part of the variation in the data, but the study is not without limitations. One problem in this study is that the data presented are not taken from naturally occurring language but from planned written text. This may pose a problem to use the findings of this study in sentence processing research. Another problem is that the dataset is limited in size due to the unavailability of enough syntactically parsed corpora in Arabic. One possible solution for the data sparseness issue is to use automatically parsed data, but there does not seem to be an Arabic parser that produces automatic annotation that is correct enough for such research. This is a correlational study limited by the information that can be extracted from parsed text, and many factors that can affect HNPS, such as new versus information and subject and object animacy, are difficult to model correctly using only these data. None of these problems is easy to fix, but I hope more data will be made available in the future.

For further research, I will extend the analysis to Classical Arabic to see whether the current models can work for this old variety, which is still used as part of the Arabic continuum.

References

Ali, Fadl-Allah Elnour. 2011. Fronting and pre-posing in Arabic. *Journal of Science and Technology* 12/2. (فضل الله النور علي: ظاهرة التقديم والتأخير في اللغة العربية مجلة العلوم والتقانة. مجلد 12 - العام 02 -2011 للعلوم والتكنولوجيا- جامعة السودان.) <http://www.sustech.edu/staff_publications/20130606062042714.pdf>

Anis, Ibrahim. 1978. *The secrets of the Arabic language*. Sixth edition. Cairo: The Anglo-Egyptian Bookshop. (إبراهيم أنيس. من أسرار اللغة. الطبعة السادسة. مكتبة الأنجلو المصرية. مصر1978.)

Arnold, Jennifer E., Thomas Wasow, Anthony Losongco and Ryan Ginstro. 2000. Heaviness vs. newness: the effects of complexity and information structure on constituent ordering. *Language* 76/1: 28–55.

Buchholz, Sabine and Erwin Marsi. 2006. CoNLL-X shared task on Multilingual Dependency Parsing. In *CoNLL-X '06 Proceedings of the Tenth Conference on Computational Natural Language Learning*. Stroudsburg, PA: Association for Computational Linguistics, 149–164. <http://www.aclweb.org/anthology/W06-2920>

Buckwalter, Tim. 2002. Arabic morphological analyzer version 1.0. Linguistic Data Consortium (LDC) Catalogue Number: LDC2002L49.

Chomsky, Noam and George A. Miller. 1963. Introduction to the formal analysis of natural languages. In R. Duncan Luce, Robert R. Bush and Eugene Galanter eds. *Handbook of mathematical psychology. Vol. 2*. New York: Wiley, 269–321.

Dormann, Carsten F., Jane Elith, Sven Bacher, Carsten Buchmann, Gudrun Carl, Gabriel Carré, Jaime R. García Márquez, Bernd Gruber, Bruno Lafourcade, Pedro J. Leitão, Tamara Münkemüller, Colin McClean, Patrick E. Osborne, Björn Reineking, Boris Schröder, Andrew K. Skidmore, Damaris Zurell and Sven Lautenbach. 2013. Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Journal of Ecography* 36/1: 27–46.

Firbas, Jan. 1964. On defining the theme in functional sentence perspective. *Travaux Linguistique de Prague* 1: 267–280.

Haddock, C. Keith, David Rindskopf and William R. Shadish. 1998. Using odds ratios as effect sizes for meta-analysis of dichotomous data: a primer on methods and issues. *Psychological Methods* 3/3: 339–353.

Hajič, Jan, Otakar Smrž, Petr Zemánek, Jan Šnaidauf and Emanuel Beška. 2004. The Prague Arabic Dependency Treebank: development in data and tools. In *Proceedings of the NEMLAR International Conference on Arabic Language Resources and Tools*, 110–117.

Hassan, Abbas. 1974. *Comprehensive Arabic grammar*. Fifteenth edition.(عباس حسن: النحو الوافي. دار المعارف. الطبعة الخامسةعشرة.)

Hatcher, Larry. 2013. *Advanced statistics in research. Reading, understanding, and writing up data analysis results*. Michigan: Shadow Finch Media LLC.

Hawkins, John A. 1994. *A performance theory of order and constituency*. Cambridge: Cambridge University Press.

Huck, Schuyler W. 2004. *Reading statistics and research*. Fourth edition. Boston: Pearson.

Ibn Malik, M. 1260. *Al-Khulāsa Al-Alfiyya* (*Arabic grammar in 1000 verses*). <http://shamela.ws/index.php/book/9904>

Kimball, John. 1973. Seven principles of surface structure parsing in natural language. *Cognition* 2: 15–47.

Manning, Christopher D., Prabhakar Raghavan and Hinrich Schütze. 2009. *An introduction to information retrieval*. Oxford: Oxford University Press.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Édouard Duchesnay. 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Ross, John R. 1967. *Infinite syntax*. Norwood: Ablex.

Sibawayh, Amr ibn Bahr. 1988. *The book of Arabic grammar*. Third edition. Cairo: Maktabat al-Khānjī. (عمرو بن عثمان بن قنبر الحارثي الملقب بسيبويه. الكتاب. تحقيق عبد السلام محمد هارون. الطبعة الثالثة (1988، القاهرة: مكتبة الخانجي.).

Stallings, Lynne M. and Maryellen C. MacDonald. 2011. It's not just the 'Heavy NP': relative phrase length modulates the production of Heavy-NP Shift. *Journal of Psycholinguistic Research* 40: 177–187.

Stallings, Lynne M., Maryellen C. MacDonald and Padraig O'Seaghada. 1998. Phrasal ordering constraints in sentence production: phrasal length and verb disposition in Heavy-NP Shift. *Journal of Memory and Language* 39: 392–417.

Tabachnick, Barbara G. and Linda S. Fidell. 2007. *Using multivariate statistics*. Fith edition. Boston: Pearson.-

Wasow, Thomas. 1997. Remarks on grammatical weight. *Language Variation and Change* 9: 81–105.

Wright, Raymond E. 1995. Logistic regression. In Laurence G. Grimm and Paul R. Yarnold eds. *Reading and understanding multivariate statistics*. Washington DC: American Psychological Association, 217–244.