Language Variation and Change, **25** (2013), 255–285.
© Cambridge University Press, 2013 0954-3945/13 $16.00
doi:10.1017/S0954394513000173

Subject expression and discourse embeddedness in Emirati Arabic

JONATHAN OWENS
Bayreuth University

ROBIN DODSWORTH
North Carolina State University

MARY KOHN
University of North Carolina, Chapel Hill

ABSTRACT

Since Prince (1981) and Givón (1983), studies on discourse reference have explained the grammatical realization of referents in terms of general concepts such as "assumed familiarity" or "discourse coherence." In this paper, we develop a complementary approach based on a detailed statistical tracking of subjects in Emirati Arabic, from which two major categories of subject expression emerge. On the one hand, null subjects are opposed to overt ones; on the other, subject-verb (SV) is opposed to verb-subject (VS). Although null subjects strongly correlate with coreferentiality with the subject of the previous clause, they can also index more distant referents within a single episode. With respect to SV vs. VS, morpholexical classes are found to be biased toward one or the other: nouns are typically VS, pronouns SV. We conclude that the null subject variant is the norm in Emirati Arabic, and when an overt subject is appropriate, lexical identity biases the subject into SV or VS order, generating word order as a discourse-relevant parameter. Overall, our approach attempts to understand Arabic discourse from a microlevel perspective.

INTRODUCTION

The relationship between the form of referring expressions and discourse structure has been approached in various ways. For Prince (1981) and Chafe (1994), the focus was on individual referring lexemes, defined in terms of assumed familiarity (Prince, 1981) or identifiability and activation level (Chafe, 1994; also Lambrecht, 1994). For Givón (1983), describing and explaining overall text coherency is the larger goal, for which he uses several quantified measures applied to a given referent (e.g., lookback, persistence). These measures, in turn,

We would like to thank the reviewers for their helpful and challenging comments.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

typically correlate with hierarchically arranged grammatical structures; a longer lookback value, for instance, is associated with a topicalized (L-dislocated) noun phrase (NP). By contrast, Ariel (1994, 2000) and Levinson (1987) follow a "selectional approach," explaining why in discourse given lexemes and/or grammatical constructions are chosen (Ariel, 1994) or how they are interpreted (Levinson, 1987). Levinson (1987) interpreted the processes by which the reference is deduced of different classes of lexemes (null, pronouns, nouns), using a modified account of Gricean implicatures. Ariel (1994:32), in the context of her accessibility theory, explained the choice of a much larger range of lexical classes than Levinson (1987) did according to the three principles of informationality, rigidity, and attenuation. The more a referring expression has of each of these properties, the more likely it is a full noun will be used.

In this paper, we define the broad grammatical and pragmatic environments for the three possible realizations of Emirati Arabic subjects—subject-verb (SV), verb-subject (VS), and null—in verbal clauses, on the basis of analyses of two independent spoken corpora. Similar to Prince (1981) and Chafe (1994), the focus of analysis is on the discourse status of the individual lexeme. Using both standard logistic regression and conditional inference trees, we will argue that the data support a lexically biased, discourse-pragmatic account, with separate factors governing the distribution of null vs. overt subjects, on the one hand, and the distribution of SV vs. VS clauses, on the other. We specifically suggest that the lexical bias is manifested in an inherent weighting of lexical items to occur in SV or VS order. One hallmark of our approach is to reduce interpretive categories such as "importance, topicality, continuity" or "shared knowledge" to a poststatistical phase of the analysis. The study is based initially on the statistical analysis of the referential status of the subject interpreted in terms of categories of coreference.

SUBJECT REALIZATION IN SPOKEN ARABIC

There are relatively few corpus-based studies of subject-verb realization on spoken Arabic. Holes (2009), on a closely related Gulf Arabic dialect, specifically noted that VS tends to be used for new referents and in narratives predominantly, but SV is realized by nonspecific subjects and often signals background information in narratives. Owens, Young, Rockwood, Mehall, and Dodsworth (2010) and Owens, Dodsworth, and Rockwood (2009) look respectively at null subject verbal clauses and at SV/VS order. The first study shows that Levinson's (1987) Gricean-derived principles make correct predictions in general about the distribution of null subjects; however, exceptions are so numerous that a multifactorial account is necessary. The latter study concludes that word order in Emirati Arabic is pragmatically, not syntactically determined.

In a short study, using a rich set of discourse status categories, Gundel, Bassene, Gordon, Humnick, and Khalfaoui (2010) classified a referring expression as being in focus, activated, familiar, uniquely identifiable, referential, or type identifiable.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

They use a mixture of corpus and introspective data. However, their intuitions of Tunisian Arabic (Gundel et al., 2010:1782, e.g., 24) appear to rule out a class of subjects that are frequent in our corpus, namely null subjects whose referents have been evoked in discourse but do not occur in the previous clause. Dahlgren (1998) and Brustad (2000) are further studies, though there are none, to our knowledge, whose explicit goal has been to reconcile the discourse status of both overt and null subject clauses in spoken Arabic.

THE CORPORA

Our data consist of two sets of transcribed Emirati Arabic conversations. The first set, "oral," was collected by Dr. Bill Young in Al-Ain, United Arab Emirates. The second set is a codification of Al-Rawi (1990), a transcription, translation, and commentary on nine texts collected in Dubai in the mid-1980s (see Hoffiz, 1995, for a description of this dialect).[1] The texts were annotated for various parameters using Microsoft Word, and a PERL script was written to import the texts into a spreadsheet. Statistical analyses were carried out using R (R Development Core Team, 2010).

The Al-Rawi (1990) texts are monologues. Apparently, Al-Rawi (1990) set up a topic, usually about traditions and practices in Dubai, or asked for a folktale (Al-Rawi, 1990:6), and the interlocutor was left to describe or narrate. All speakers are women, eight in all. What are termed oral, because the audio in this case is available, consist of two interviews conducted in Al-Ain, UAE, and one telephone conversation. Of the interviews, one is an interview between Bill Young (an American) and an Emirati, and the other also has Bill Young as a participant, though most of the conversation is between two Emiratis. Bill Young's utterances are excluded from the analysis. The telephone conversation is from Appen (2006), provided via the Linguistic Data Consortium, and appears to involve two women from Al-Ain.

The data, thus, consist of three independent samples: the Al-Rawi (1990) collection from the 1980s, the two recordings by Bill Young, and the Appen phone conversation. The phone conversation is too short (six minutes) to provide enough tokens for independent analysis, so it has been combined with the Bill Young data. The Al-Rawi (1990) collection is adequately large for independent analysis and is so treated. Although a larger corpus is desirable, the consistent results obtained between the two independent samples are indicative of a solid basis for further analysis.

CODING METHODS

All subjects accompanied by verbal predicates were coded for clause type—null subject, SV, or VS—and a set of grammatical and pragmatic variables. SV and VS clauses are those in which a subject NP occurs as a free morpheme, in addition to the obligatory bound inflectional morphology on the verb. In all

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

varieties of Arabic, including Emirati, all finite verbs are marked inflectionally for person, number, and gender. A partial paradigm illustrating these basic inflectional contrasts for the imperfect of the verb *yiktib* 'he writes' is as follows.

	SG	PL
1	a-ktib	ni-ktib
2M	ti-ktib	ti-ktib-uun
2F	ti-ktib-iin	ti-ktib-an

a- = 1.SG, ni- = 1.PL, ti- = 2, -iin = F.SG, -uun = M.PL, -an = F.PL.

These inflections occur both with overt subjects—nouns, pronouns, pronominals—and without overt subjects. Hence, if an overt subject occurs, there will, to phrase the issue neutrally, be two exponents of subject: the first, an overt subject, and the second, the subject inflection. Schematically, our coding practice can be illustrated as follows in (1a–1c).

(1) a. *zaad-it* *il-ifluus*
Increase-3F DEF-money
'The money has increased.'

b. *il-ifluus* *zaad-it*
DEF-money increase-3F
'The money has increased.'

c. Ø *zaad-it*
Ø increase-3F
'It (F) has increased.'

In (1a/1b) there is an overt subject: (1a) is coded as a VS token and (1b) as SV. (1c) has no overt subject and, hence, is coded as a null subject clause. Note that in all of (1a–1c) the form of the verb is identical, the 3FSG form of the verb *zaad*; the 3FSG suffix *-it* (perfect verb) has no bearing on the subject coding as it is identical throughout. In (1a/1b), *-it* cross-references the subject *il-ifluus*, whereas in (1c) *-it* cross-references a nominal referent outside of the clause. When there is an overt subject, there will be two formal exponents of subject, one the overt subject itself, the other the verbal inflection that occurs in all imperfect and perfect verbs. In the case of null subjects, there will be one formal exponent of subject, the verbal inflection. The concept of null subject encodes the assumption that the referent of the null-subject verb is to be found in the wider discourse or situational environment.

How the verbal inflection is to be accounted for theoretically, whether in terms of agreement, pro-drop, pragmatic inferencing, or other means, is a question outside the scope of the present paper. The only systematic contrast relevant here is that between the overt and the null (Ø) subject and the pragmatic factors governing the choice of one or the other.

To illustrate our coding from actual textual examples, in (9) below, 2MSG is referenced twice, once with the overt pronoun *inta*, once in the verb prefix *ti-*. In

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

(10), 2MSG is referenced only by the verb prefix *ti-* (*tə-mill*). Given that in both cases, the verb is marked by the same inflectional element (*ti-*), our basic *contrast*—overt vs. null—makes no reference to the verbal inflection. In our system, clauses were coded as SV or VS only if an overt subject NP was present.

We first discuss excluded tokens and then the coding methods for the independent variables.

Exclusions

Several disparate classes of tokens were excluded from the analysis because they exhibit categorical word order. Question word subjects, for instance (*šu* 'what?'; *mnuu* 'who?') are always SV, as are subject nouns modified by relative clauses. Although the current study deals only with verbal predicates, nonverbal predicates are nearly always either predicate–S or S–predicate. Existential expressions (*fi, bi* 'there is/are', *mub* 'there isn't') and "have" existential expressions with *ʕind* (lit. 'at') as predicate are categorically VS, whereas equational, attributive, and locational predicates are categorically SV.[2] We also excluded what we consider to be grammatically determined forms. Emirati Arabic, along with virtually all varieties of contemporary Arabic, have a number of subordinate contexts, for instance, complements of verbs such as *raad* 'want' and *gidir* 'be able', which require subordinate verbs to be in the imperfect form. These carry over all the subject categories—person, number, gender—from the matrix verb, so they have no independent choice of subject. They are excluded from the current analysis, as in the underlined verb in (2).

(2) *Ø* ***ni****-gdar* ***nə****-rjaʕ* (*ni-gdar tə-rjaʕ)
we-can we-return (we-can you-return)
'We can return.' (AR, Oral)

Finally, we excluded all relative clauses, as they have a fixed relation of coreference to the antecedent. Coreferential subjects in the relative clause are, in any case, categorically null and hence excluded, whereas nonsubject coreference, in our view, requires independent study before being included in the present analysis.

Independent variables

The independent variables investigated here are those which other, quantitatively based studies of word-order variation in discourse analysis have used (e.g., Givón, 1983; Prince, 1981; see also Ariel, 1988, 2000; Owens et al., 2009:28–30).

Person, number, and gender. Verbs in Emirati Arabic are inflected for person, number, and gender, yielding up to 10 distinct categories, in two basic paradigms, the imperfect verb and the perfect.

Morpholexical class. The initial categories were pronoun, quantifiers and numerals, personal name, indefinite noun, definite noun, and possessed noun. Possessed nouns have the canonical shape (in our data) "possessed N

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

il-possessor" and are included among the definite nouns because the definiteness of the possessed phrase is determined by the definiteness of the possessor.[3] Proper nouns were also combined with definite nouns. Indefinite nouns are unmarked with the definite article *il-*. Quantifiers and numerals include forms such as *kiθiir* 'many', *xamsa* 'five', *kill* 'all', and *ħadd* 'someone, anyone'. The category of pronoun includes both personal pronouns and demonstrative pronouns. In the Al-Rawi (1990) data, there are only 15 tokens of numerals and quantifiers, and they are combined with the pronouns.

Previous clause. This variable encodes whether the previous clause is SV, VS, or null subject and is designed to index the degree to which the realization of a clause is "primed" by the previous clause. This was only coded for in the oral data.

Independent adverbial or conditional clause. Holes (2009:62) reported that in his Bahraini Arabic data, subjects in dependent adverbial or conditional clauses are almost invariably VS. Because we do have variation in our data, we coded for adverbial or conditional clauses. This variable is binary (yes/no) (see (15), for example).

Complement of gaal *'say'.* Our previous analysis has shown that sentential complements of *gaal* overwhelmingly occur as SV, as in (3).

(3) Ø ***gaal*** *l-ii,* ***ana*** *a-ruuħ*
Ø said to-me I I-go
"**He$_1$** said to me, '**I**'m$_1$ going'. (A., Oral)

This and the previous variable turn out to be weakly significant, but as they play a small overall role in the analysis, they will be left out of the discussion, other than when they elucidate our disharmonic patterns, as will be discussed later (see n. 9).

Previous mention. An important aspect of the analysis has to do with the referential relationships among subjects. Drawing heavily upon Prince (1981), we have distinguished four categories.[4]

1. *Same as previous.* This is a relation of coreference. The subject's referent is the same as the referent of either the immediately preceding subject (4a) or of a nonsubject argument of the preceding verb (4b). In addition, we have included here subjects whose referent is included in the referent of the preceding subject; typically in this case, the preceding subject is a quantifier, for example, *il-waaħid* 'someone, an individual,' which is specified in the following clause, or when the preceding subject is plural, and the following subject indexes one of its class of referents (4c). In the following examples, boldface marks the coreferential relations.

(4) a. Ø ***i****-yiib-****uun*** *maaʔ,* Ø ***i****-biiʕ-****uun*** *ʕala* *l-meena*
3-bring-PL water 3-sell-PL in DEF-harbor
'**They$_1$** brought water and **they$_1$** sold it in the harbor.' (Al-Rawi, 1990:102)

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

b. *yoom* *Ø ʕarris-**ó***, *wu* *Ø **xađđ***[5] *bint xaal-a*
when married-they.him and took daughter uncle-his
'When they had married **him$_1$** and **he$_1$** married his cousin (daughter of maternal uncle)' (Al-Rawi, 1990:110)

c. *yoom* *Ø saar-**aw*** *fi* *ṭ-ṭiriig,* *Ø baaʕ-**at*** *bint ʕamm-a*
when went-they on DEF-road sold-she daughter uncle-his
'When **they** were going, **she** sold his cousin.' (Al-Rawi, 1990:110)

In (4c), the referent of *she* is one of the two referents included in the preceding *they*.

2. *Available in discourse*. The subject's referent is different from the preceding subject's referent, but it has been previously mentioned in the discourse (textually evoked, in Prince's [1981] taxonomy). In (5), the subject *ṭəwaari* was classified as "available," that is as having been already introduced in previous discourse.

(5) *aħyaanan* *bi-yii-hum* ***ṭəwaari***
sometimes FT-come-them emergency duty
'Sometimes emergency duty is imposed on them.' (lit. 'comes to them')

In fact, this idea of emergency duty (in the military) was introduced as a lexeme, *ṭəwaari*, 17 clauses earlier. Even in this mention, however, *ṭəwaari* was classified as available, as it is clear that *ṭəwaari* is a subclass of the general idea of *mənaawabaat*, which in turn was introduced 12 clauses earlier:

(6) *he* *aħyaanan* *ti-jii-na* ***mənaawab-aat***
ya sometimes 3F-come-us duty-FPL
'Ya, occasionally we'd have rotational duty.' (A, Oral)

3. *New referent*. The referent is entirely new at this point in the discourse, as in the case of *mənaawab-aat* in (6). New referents here include both "brand new" and "unused" items in Prince's (1981) taxonomy. In (7), the first subject is considered a new referent ("plural passive," see later discussion), whereas the second, which has the same referent, is classified as "same as previous" (i.e., same referent as previous subject). Once a referent has been established, any further reference to this subject will not be "new," but rather either same as previous or available.

(7) *haay* *Ø i-samm-uun-a* *~~ḍ-ḍ~~ahar,* *dəwaam* *i~~ḍ-ḍ~~ahar* *Ø i-samm-uun-a*
this.F Ø 3-call-PL-it.M DEF-afternoon shift DEF-afternoon 3-call-PL-it
'This they call it (it is called) "afternoon", "the afternoon shift" they call it.' (Ar, Oral)

The value "available in discourse" in the previous mention collapses what is, in theory, an infinity of values, that is, all potential same references that occurred in the previous discourse. As an empirical issue, as a reader rightly points out, it is not self-evident that all "lookback" distances should code-reference in the same way, and that our binary cut could mask significant internal variation. To test this issue, the first 600 records of the oral data containing an "available in discourse" value were coded for the three available subvalues: occurs 1–2 clauses, 3–5 clauses, or 6+ clauses back. A chi-square test running the revised available values

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

against the three clause types—null, SV, or VS—as well as null vs. combined SV/VS (i.e., overt nominal subjects) showed no significant difference, nor did another test of the same data running available vs. the different grammatical classes defined in this paper.[6] Numerically, null values do show a gradual decrease as one moves backward, but so do overt referents, and even at the maximum 6+ measure, there are as many null tokens as there are overt. In the discussion section, we will briefly suggest a discourse basis for this overall lack of difference.

4. *Situational*. The subject's referent is situationally given, corresponding to the "situationally evoked" category in Prince (1981) and exemplified in (10). We should point out that first- and second-person values—the prototypical situational values—which are part of indirect discourse and hence do not reference the interlocutors in the immediate speech situation, are not included among the situationals.

Coding for previous mention is straightforward for the vast majority of tokens. Not all are unambiguous, however. Here we discuss three challenges in coding for previous mention, in order to elucidate our coding decisions and to indicate where, with a bigger corpus, further fine-tuning might be appropriate.

Two of the challenges concern new referents. Null referents are prototypically anything but "new" (e.g., Ariel, 1988; Givón, 1983:18). In Emirati Arabic, however, it is common for a new referent to be introduced with a null third-person masculine plural subject, 'they say', 'they do', etc., as in (7).[7] We think of these as "plural passives," as an appropriate translation in English could indeed be a passive. In all such tokens, the use of the plural passive is not cued by any previous tokens. On this basis, they are considered new. Impersonals are another source of null new subjects (see Owens et al., 2009:47–48, for more details). In (8), the subject of *ikuun* is classified as indexing a new referent.

(8) Ø *i-kuun* *arbaʕt* *ayyaam*
3MSG-be four days
'It should be four days.' (Oral)

A third challenge in coding for previous mention pertains to the situational category. As already noted, within previous mention, we distinguish between first- and second-person subjects used as a part of a narrative, and those used to refer to situationally given referents, especially the interlocutors. Situationally given referents encompass several well-profiled subtypes. A question-answer pair, for instance, often involves situationally given referents of first- and second-person subjects, as in (9).

(9) A: *kam* *ti-dfaʕ* *inta*
how much you-pay you
AR: *ma* *ʕaarif* *kam* *taðkərt-a*
not know how much ticket-its
A: 'How much do you pay?'
AR: 'I don't know how much a ticket is.' (Oral)

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

Situationally given referents also appear in discourse markers such as *fihimte* 'did you get it' or *ʕəraft keef* 'you follow how?' (lit. 'you know how?'), which do not call for an answer. They also include generic referents (10).

(10) *sitt* *saaʕaat* *ti-kaffi,* *ma* ***Ø tə****-mill* *min* *id-darraaja*
six hours 3F-suffice, not you-bore from DEF-motorcycle
'Six hours are enough. **You** don't get bored with a motorcycle!' (AR, Oral)

Here instead of a second-person singular subject, a very similar meaning is obtained with a plural passive, or by the use of an indefinite quantifier, *il-waaħid* 'someone, a person, an individual' (cf. Owens et al., 2009:40–47, for further discussion).

Given a larger corpus, these special instances of new referent and situational could be coded as separate categories. This, in fact, is what we did for some categories. For instance, in one statistical run, we had two situational variables, one consisting of question-answer pairs as in (10), the other, all other situationals, and in this case, the results were all but identical to those presented here.

STATISTICAL RESULTS

Analysis of oral data

We begin with the oral data. The variables and token counts for the oral data appear in Table 1.[8] The majority of clauses (72%) are null, followed by SV (22%) and VS (6%).

The token counts for previous mention in Table 1 indicate that new referent is the only category not dominated by null subjects. In the other categories, all of which correspond to "old" referents to varying degrees, null subjects are the majority, followed by SV. VS scarcely appears in the situational category, which is consistent with our earlier observation (Owens et al., 2009) that the primary, but not sole, function of VS is to introduce new referents. Table 1 also shows that the relationship between clause type and previous mention is far from categorical. In particular, new referents are as likely to be introduced in null clauses as in VS clauses; conversely, VS accounts for 3.5% of the "in previous clause" category.

With respect to the relationship between clause type and previous clause, the question is whether a priming effect is visible, such that clauses that are preceded by SV (for example) are more likely to be SV clauses. We find some evidence of a priming effect: SV clauses occupy a larger share of the SV category of previous clause than they do in the null category of previous clause, and VS clauses are relatively dominant in the VS category of previous clause. Nevertheless, null clauses are the majority in all three categories of previous clause.

Morpholexical class shows the pattern described in Owens et al. (2009): Indefinite nouns are VS-dominant, and all of the other non-null categories are predominantly SV. Definite/possessed/proper nouns show a relatively high

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

TABLE 1. *Variables and token counts*

	Null	SV	VS	Total
Previous mention				
New referent	60	78	36	174 (10%)
Available in discourse	256	108	24	388 (23%)
In previous clause	818	152	36	1006 (58%)
Situational	112	37	7	156 (9%)
Previous clause				
Null	854	221	64	1139 (66%)
SV	277	114	19	410 (24%)
VS	115	40	20	175 (10%)
Morpholexical class				
Def./proper/poss. N	( )	77	37	114 (7%)
Indefinite N	( )	15	33	48 (3%)
Pronoun	( )	244	26	270 (16%)
Numeral/quantifier	( )	39	7	46 (3%)
Null	1246	( )	( )	1246 (72%)
Person/number				
1SG	193	93	6	292 (17%)
1PL	215	56	10	281 (16%)
2SG	210	34	3	247 (14%)
2PL	33	6	3	42 (2%)
3MSG	308	117	46	471 (27%)
3FSG	52	36	25	113 (7%)
3PL	235	33	10	278 (16%)
Complement of *gaal* 'say'				
Yes	19	12	2	33 (2%)
No	1226	363	101	1690 (98%)
In adverbial clause				
Yes	21	8	3	32 (2%)
No	1225	367	100	1692 (98%)
Total	1246 (72%)	375 (22%)	103 (6%)	1724

proportion of VS. Person/number must be considered in association with morpholexical class, as all first- and second-person subjects are pronouns. In addition, we expect third-person subjects to occur as VS more often than first- and second-person subjects do because third-person subjects are more likely to index new referents. In general, we expect person/number to be a significant predictor of clause type, but only via its interaction with other factors.

A serious obstacle to statistical assessment of these trends is the collinearity among the first four variables in Table 1. The results of chi-squared tests (Table 2) indicate that every pair of variables in this set is significantly correlated. For this reason, we first explore the relationships among the variables visually and then compare separate statistical models to assess their relative effects on clause type.

We first look at the relationship between morpholexical class and previous mention. We expect a complex interaction with respect to clause type; for

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

TABLE 2. *Results of chi-squared tests between independent variables*

Pair of variables	Significance level
Previous mention/previous clause	$p < .001$
Previous mention/morpholexical class	$p < .001$
Previous mention/person/number	$p < .001$
Previous clause/morpholexical class	$p < .01$
Previous clause/person/number	$p < .05$
Morpholexical class/person/number	n/a

example, indefinite nouns, but not pronouns, are expected to introduce new referents, and so indefinite nouns are predicted to favor VS due to their association with the new referent category of previous mention. The question, then, is whether morpholexical class has an effect on clause type independently of its interaction with previous mention. Figure 1 gives an initial look at the data. There are three rows of jittered data: the top row corresponds to VS clauses, the middle row to SV clauses, and the bottom row to null clauses. Every individual character represents one clause. The bottom row of data consists entirely of the null morpholexical category (dark squares) because this row corresponds to the null clause type. As expected, there is a strong and complex interaction between previous mention and morpholexical class. In particular, overt situational referents are always pronouns, with the exception of one numeral/quantifier. In addition, indefinite nouns predominantly index new referents, though indefinite nouns do occur in the other previous mention categories.

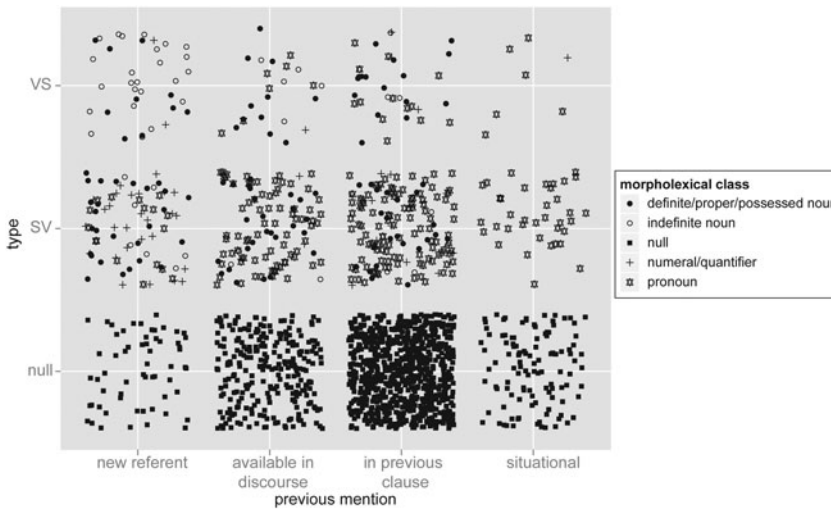


FIGURE 1. Relationship among clause type, previous mention, and morpholexical class.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

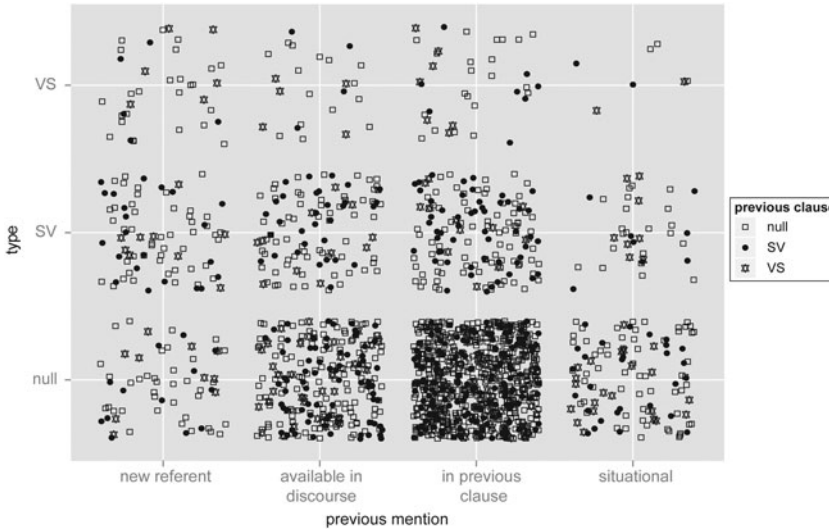


FIGURE 2. Relationship among clause type, previous mention, and previous clause.

Yet Figure 1 also offers evidence that morpholexical class has an independent effect on clause type, in addition to its interaction with previous mention. First, pronouns are predominantly SV, and they are always SV when they introduce new referents. Numerals/quantifiers are also predominantly SV in the new referent category. By contrast, indefinite nouns are predominantly VS. Definite/proper/possessed nouns are distributed more evenly across SV and VS clauses in all the previous mention categories in which they appear.

We next consider the relationship among previous mention, previous clause, and clause type, as shown in Figure 2. We observed above that some "in previous clause" referents occur as SV or VS, rather than the expected null. We now ask whether these are primed by preceding SV and VS clauses. Figure 2 shows that these clauses are as likely to have null previous clauses as they are to have SV or VS previous clauses. That is, they are not obviously primed by the previous clause. Inspection of Figure 2 then leads to the same conclusion for the SV and VS clauses in the available in discourse category. In general, Figure 2 reveals no clear interaction between previous mention and previous clause, which suggests, first, that the null clauses occurring in the "in previous clause" category do not result from priming effects. Second, it suggests that any effect of previous clause on clause type does not result from its correlation with previous mention.

Finally, we consider the relationship between person/number and previous mention (Figure 3). All categories of person/number are predominantly null. However, the set of possible referents of first- and second-person subjects is, by definition, more restricted than the set of possible third-person referents. Accordingly, the great majority of new referents occur as third-person subjects, especially 3MSG forms. This may (but does not necessarily) account for the

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

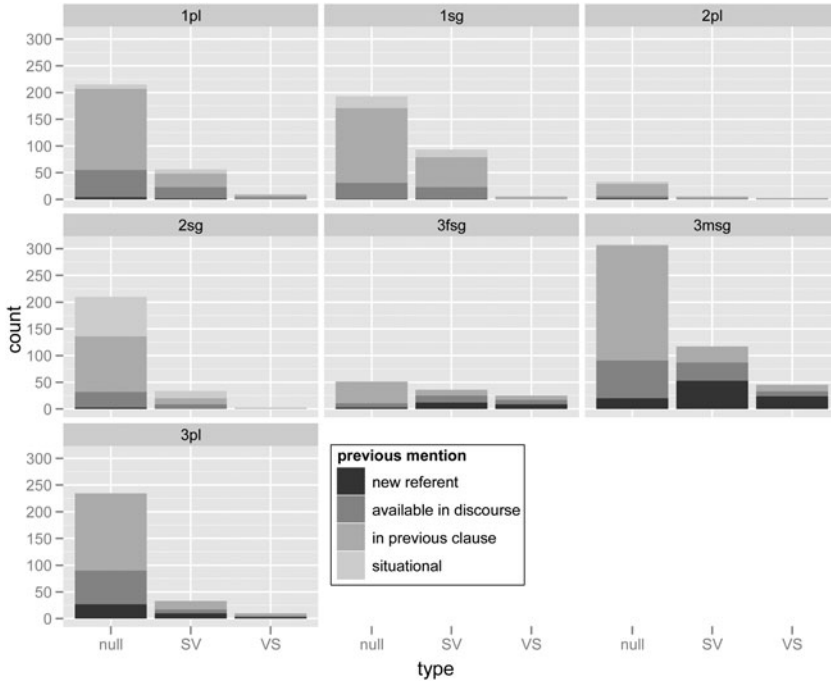


FIGURE 3. Relationship among clause type, previous mention, and person/number.

extreme rarity of first- and second-person VS forms. Situational referents belong exclusively to first- and second-person subjects and, for that reason, are always pronouns unless they are null.

The following statistical analysis assesses the effects of each of the independent factors on clause type. Collinearity among the variables—other than *gaal* complement and adverbial clause—precludes building regression models with any two of them together. Therefore, separate regression models will be compared to determine which factors best account for the variation in clause type. We fitted two groups of logistic regression models: one group for null vs. overt clauses, and one group for SV vs. VS.

Oral results: Null vs. overt

Three logistic regression models were fitted in R, each with one of the three factors of previous mention, previous clause, or person/number.[9] Morpholexical class was not used because null clause types correspond, by definition, to the null category in morpholexical class. Table 3 shows estimates, significance levels, Akaike's information criterion (AIC), and deviance residuals for each model. Positive estimates indicate greater likelihood of overt as opposed to null subjects. The null deviance and *df* are 2035 and 1722, respectively.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

TABLE 3. *Results of three logistic regression models: Null vs. overt clause is the response variable*

Factor	Estimates	AIC	Residual deviance (*df*)
Previous mention (reference level = in previous clause)	New referent: 2.11[a] Available: .812[a] Situational: .55[b]	1884	1872 (1717)
Previous clause (reference level = null)	SV: .38[b] VS: .47[b]	2027	2018 (1718)
Person/number (reference level = 3MSG)	1SG: –.05 1PL: –.55[b] 2SG: –1.1[a] 2PL: –.76 3FSG: .79[a] 3PL: –1.1[a]	1950	1932 (1714)

Note: [a]$p < .001$, [b]$p < .01$. Token counts appear in Table 1.

A comparison of AIC values indicates that the model with previous mention is the best fit for the data. Analysis of deviance indicates that this model is significantly better than the other two models ($p < .001$ in both cases). This model shows that all categories of previous mention significantly favor overt realizations, relative to in previous clause. The strongest effect, that is, largest estimate, corresponds to new referents, and the weakest effect corresponds to situational referents, which, given their inherent availability, can be considered the oldest. In general, the longer the referent has been in discourse, the less likely the subject is to be overt.[10]

However, all three models show significant effects and significant reduction from the null deviance. To assess whether previous clause and person/number significantly affect clause type independently, and not just because they correlate with previous mention, we implemented conditional inference trees in R (cf. Hothorn, Hornick, & Zeileis, 2006; Tagliamonte & Baayen, 2012). Conditional inference trees use a binary recursive splitting algorithm to find maximally homogeneous subgroups of data. For our purposes, its main advantage over logistic regression is that it is less affected by collinearity among factors. Therefore, it can show the ways in which the three factors operate together (or not) to determine clause type. Rather than using the binary clause type variable (null vs. overt) that was used in the logistic regressions, the conditional inference tree looks at all three levels (null, SV, VS) simultaneously.

The highest node in Figure 4 corresponds to the difference between new referents and the other three categories of previous mention. This distinction is the strongest in the analysis. Lower nodes represent significant effects that operate *within* either the set of new referents or the set of non-new referents. For example, among new referents only, third-person singular subjects are significantly more likely to be realized as overt than are all other categories of person/number. In the same way, there is a significant distinction between overt

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

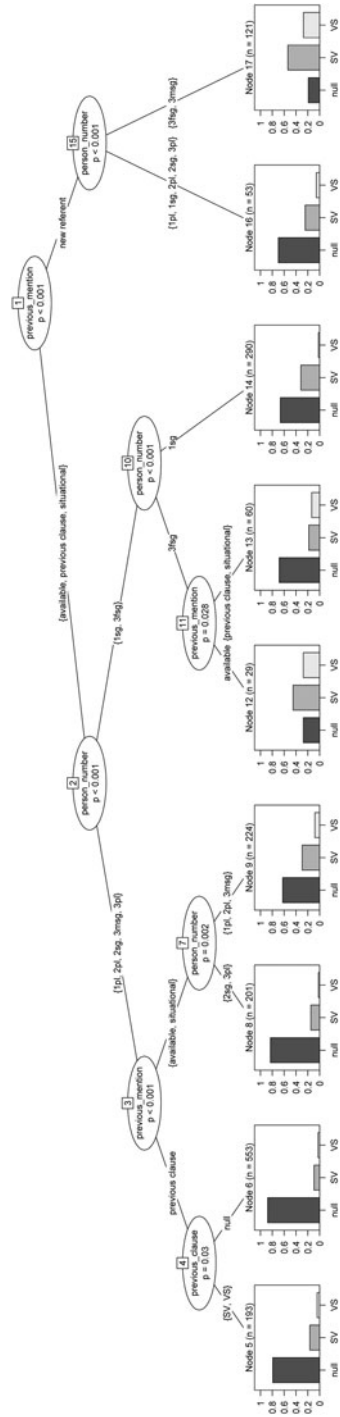


FIGURE 4. Conditional inference tree for null, SV, and VS clauses.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

and null previous clauses (far left of tree), but the distinction is embedded within three higher nodes that restrict the relevant set of tokens. Overall, the tree suggests that previous mention accounts for the most variation in the corpus, followed by person/number, followed by previous clause. This is consistent with the logistic regression results (Table 3), but with the added information that person/number, in particular, is a predictor in its own right.

Oral results: SV vs. VS

The analysis of the two overt clause types, SV and VS, differs from the preceding analysis only insofar as morpholexical class can be used as an additional factor. Table 4 summarizes the results of four logistic regression models whose response variable is SV vs. VS clauses. Positive estimates indicate greater likelihood of VS clauses. The null deviance and *df* are 498 and 477, respectively.

Both the AIC and the residual deviance indicate that the model with morpholexical class is the best fit for the data. Definite and indefinite nouns significantly favor VS relative to pronouns, with indefinite nouns showing the stronger effect. In contrast to the null vs. overt results, previous mention is not a strong predictor here. Nevertheless, the significant effects of all four factors in Table 4 may be either independent effects or the products of those factors' correlations with morpholexical class. Again, we generate a conditional inference tree to explore potential simultaneous effects, this time with four factors (Figure 5).

Figure 5 is much simpler than Figure 4, selecting only morpholexical class to define maximally homogeneous subgroups. The most important distinction lies between indefinite nouns and all other morpholexical classes. Within the latter set, there is a distinction between definite nouns and the remaining two

TABLE 4. *Results of three logistic regression models: SV vs. VS clause type is the response variable*

Factor	Estimates	AIC	Residual deviance (*df*)
Previous mention (reference level = in previous clause)	New referent: .65[a] Available: –.08 Situational: –.25	501	489 (472)
Previous clause (reference level = null)	SV: –.58[a] VS: .53	498	488 (473)
Person/number (reference level = 3MSG)	1SG: –1.82[b] 1PL: –.79[a] 2SG: –1.5[a] 2PL: .25 3FSG: .57 3PL: –.23	475	457 (469)
Morpholexical class (reference level = pronoun)	Definite noun: 1.52[b] Indefinite noun: 3.04[b] Numeral/quantifier: .53	425	413 (472)

Note: [a]$p < .05$, [b]$p < .001$. Token counts appear in Table 1.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

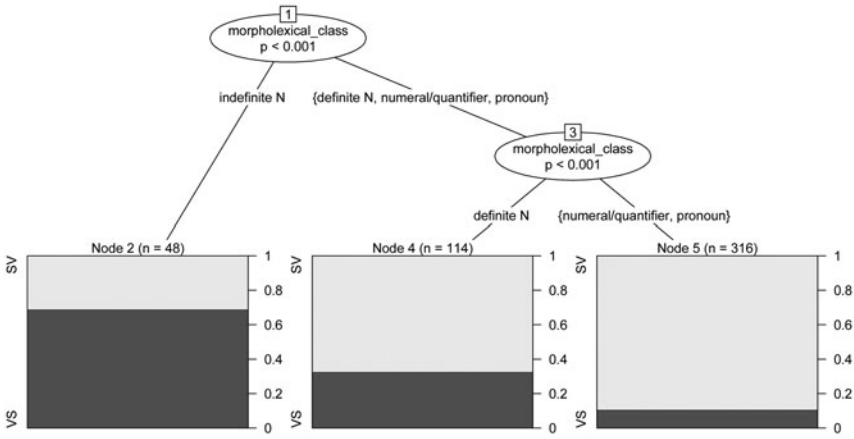


FIGURE 5. Conditional inference tree for SV vs. VS clauses.

morpholexical categories. VS, in other words, occurs primarily with indefinite nouns, followed by definite nouns. It is tempting to attribute this pattern to referential newness—as indefinite nouns are most likely to index new referents—but Figure 5 does not identify previous mention as a significant factor. (As we will show, previous mention is also not a significant factor in the Al-Rawi, 1990, SV/VS analysis.)

In summary, the distinction between null and overt subjects is primarily a function of previous mention. Roughly, the longer the referent has been in discourse, the more it will favor a null realization. Among new referents, third-person subjects are more likely than first- or second-person subjects to be overt, according to Figure 4 (see especially node 15). By contrast, the variation between SV and VS is driven by morpholexical class: Indefinite and then definite nouns favor VS, and this is not a function of their link to new referents.

Analysis of Al-Rawi (1990) corpus

We now turn to the analysis of the Al-Rawi (1990) data in order to compare them with the oral results. Although the Al-Rawi (1990) corpus offers a useful testing ground for the conclusions derived from the oral corpus, it has some nontrivial differences related to its greater proportion of narrative discourse. As shown in Table 5, SV and VS account for roughly equal shares of the Al-Rawi (1990) data, whereas SV outnumbered VS by more than 3:1 in the oral corpus. The large number of VS tokens results in a different previous mention distribution: the new referent category has more VS tokens than either null or SV, and the available in discourse category has more VS than SV tokens. There are also prohibitively few situational tokens, again because the Al-Rawi (1990) corpus consists primarily of narrative discourse rather than canonical conversation. The situational category is hereafter removed from the analysis. Another difference is

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

TABLE 5. *Variables and token counts, Al-Rawi (1990) corpus*

	Null	SV	VS	Total
Previous mention				
New referent	20	22	33	76 (8%)
Available in discourse	241	41	52	334 (36%)
In previous clause	478	23	9	510 (55%)
Situational	3	2	0	5 (1%)
Morpholexical class				
Def./proper/poss. N	( )	35	79	114 (12%)
Indefinite N	( )	3	6	9 (1%)
Pronoun	( )	48	10	58 (6%)
Null	742	( )	( )	742 (80%)
Person/number				
1SG	37	17	1	55 (6%)
1PL	96	2	0	98 (11%)
2SG	23	2	1	26 (3%)
2PL	13	0	1	14 (2%)
3SG	323	56	71	450 (49%)
3PL	250	11	21	282 (30%)
Complement of *gaal* 'say'				
Yes	37	14	4	55 (6%)
No	705	74	91	870 (94%)
In adverbial clause				
Yes	32	2	14	48 (5%)
No	710	86	81	877 (95%)
Total	742 (80%)	88 (10%)	95 (10%)	925

that in the Al-Rawi (1990) corpus 3MSG and 3FSG subjects have identical distributions with respect to clause type, and so they are represented as a single "3SG" category. Further, the few numeral/quantifier tokens were combined with pronouns, so morpholexical class has four levels rather than five. Finally, previous clause, the weakest factor in the analysis of the oral data, is not coded here.

With these differences in mind, we turn to a visual analysis of the relationships among the collinear factors of previous mention, morpholexical class, and person/number. Figure 6 parallels Figure 1. Each character represents one clause in the Al-Rawi (1990) corpus, and the characters are jittered for readability. The predominance of VS relative to the oral corpus appears in the new referent and available in discourse categories. As in the oral corpus, pronouns are predominantly SV. However, whereas the few pronouns indexing new referents in the oral corpus were categorically SV, here they are categorically SV in the in previous clause category. From this difference, we conclude tentatively that in Emirati Arabic generally, pronouns strongly favor SV, even if they are not categorically SV in any previous mention category. Another similarity with the oral corpus is the tendency for indefinite nouns to favor VS to a greater extent than definite nouns. The indefinite noun > definite noun > pronoun hierarchy for VS holds in the Al-Rawi (1990) corpus. Figure 6 also shows that previous

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

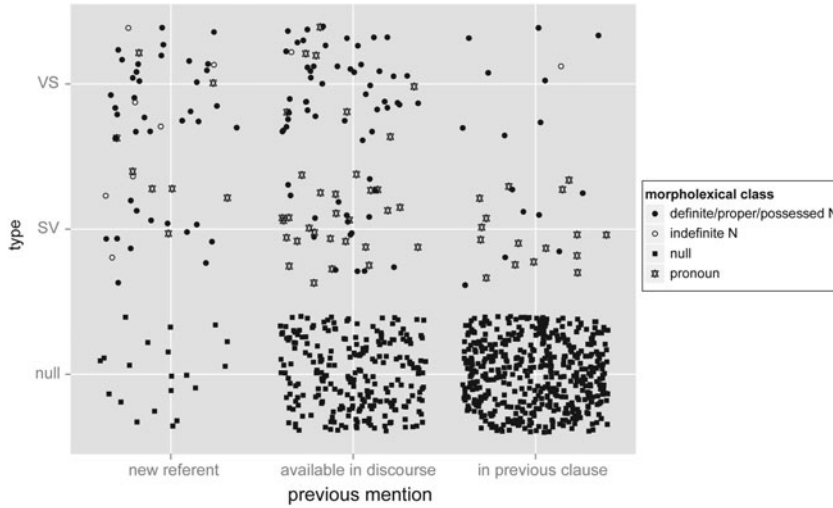


FIGURE 6. Clause type, previous mention, and morpholexical class, Al-Rawi (1990) corpus.

mention remains a strong overall predictor of the null/overt distinction, with newer referents favoring overt subjects.

Figure 7 parallels Figure 3, showing the relationship among clause type, previous mention, and person/number. Again, all first- and second-person tokens are pronouns, and so a correlation between person/number and morpholexical

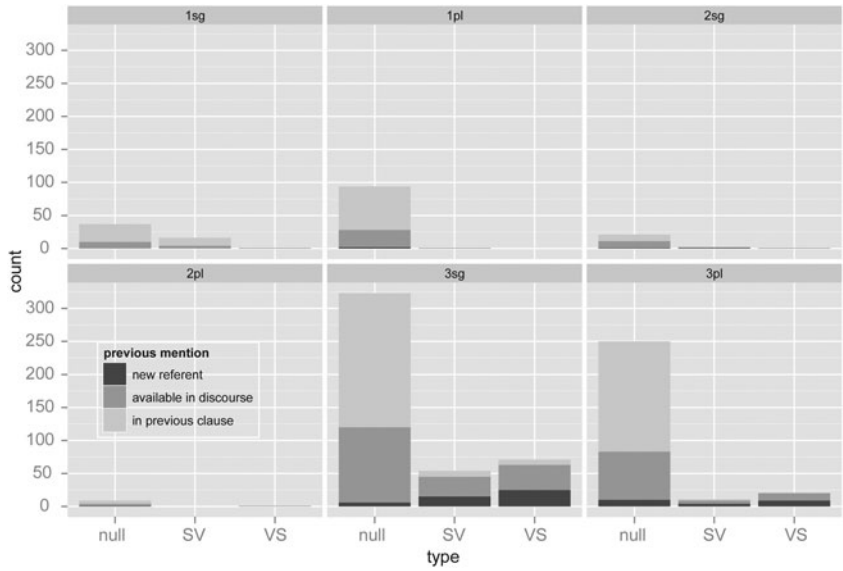


FIGURE 7. Clause type, previous mention, and person/number, Al-Rawi (1990) corpus.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

class is unavoidable. Figure 7 shows that the increase in VS tokens relative to the oral corpus is entirely contained within third-person forms. This could be a consequence of third-person forms being realized as definite and indefinite nouns, as in the oral corpus, though statistical analysis will suggest otherwise.

The statistical analysis of the Al-Rawi (1990) data is isomorphic with the analysis of the oral data. Two groups of logistic models were fitted, one with null vs. overt as the response variable and one with SV vs. VS as the response variable. The collinear factors—previous mention, morpholexical class, and person/number—were used in separate runs. Morpholexical class was not used in the null vs. overt runs, as all categories of this variable are categorically null or overt.

Al Rawi results: Null vs. overt

Table 6 shows estimates, significance levels, AIC, and residual deviance for each model. Positive estimates indicate greater likelihood of overt as opposed to null subjects. The null deviance and degrees of freedom are 912 and 919, respectively.

The AIC and residual deviance figures confirm that previous mention is a stronger predictor of null vs. overt clauses than person/number is. However, because person/number also shows significant effects, we once again generate a conditional inference tree to explore the possible interacting effects of previous mention and person/number (Figure 8). The tree shows that person/number does predict clause type, but only within two separate categories of previous mention, after the higher-level distinction between new referents and older referents. The person/number effects are somewhat different from those in the oral data. Here, 3SG and 1SG subjects are more likely to be overt than other categories, but only among the non-new referents. This tree is simpler than the parallel one for the oral corpus, but the results are similar in that previous mention shows the highest-level effect, followed by person/number.

Al-Rawi results: SV vs. VS

The analysis of the two overt clause types, SV and VS, differs from the preceding analysis only insofar as morpholexical class can be used as an additional

TABLE 6. *Results of three logistic regression models with null vs. overt clause type as the response variable, Al-Rawi (1990) corpus*

Factor	Estimates	AIC	Residual deviance (*df*)
Previous mention (reference level = in previous clause)	New referent: 3.74[a] Available: 1.77[a]	722	712 (915)
Person/number (reference level = 3SG)	1SG: –.05 1PL: –3.7[a] 2SG: –1.2 2PL: –1.6 3PL: –1.1[a]	844	828 (912)

Note: [a]$p < .001$. Token counts appear in Table 5.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

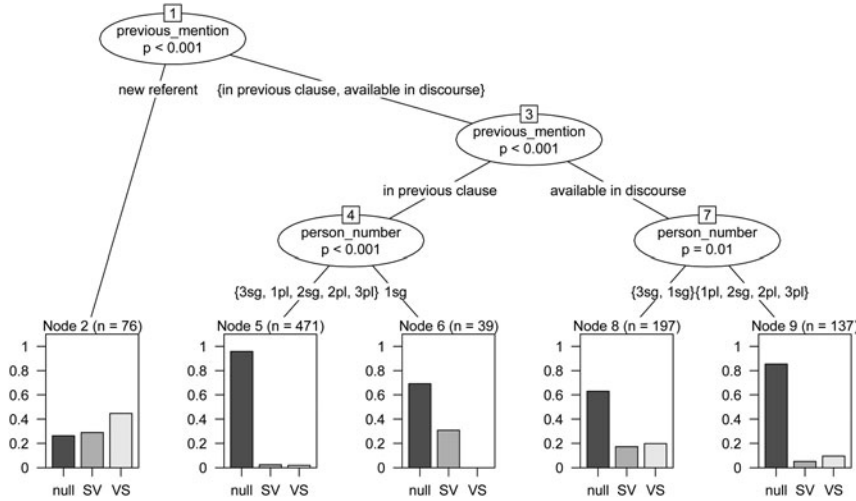


FIGURE 8. Conditional inference tree for null, SV, and VS in Al-Rawi (1990) corpus.

factor. Table 7 summarizes the results of three logistic regression models of which the response variable is SV vs. VS clauses. Positive estimates indicate greater likelihood of VS clauses. The null deviance and *df* are 250 and 180, respectively.

Again, the results replicate those of the oral corpus, showing that morpholexical class has the strongest effect on the SV/VS distinction. Definite and indefinite nouns are more likely to occur as VS than are pronouns. We generate a conditional inference tree to determine whether the significant effects of the other factors arise from their correlation with, or in interaction with, morpholexical class. The resulting tree (Figure 9) is again simpler than the parallel tree from the analysis of the oral data, but it shows the same basic effect:

TABLE 7. *Results of three logistic regression models with SV vs. VS clause type as the response variable, Al-Rawi (1990) corpus*

Factor	Estimates	AIC	Residual deviance (*df*)
Previous mention (reference level = in previous clause)	New referent: 1.3[a] Available: 1.1[a]	237	227 (176)
Person/number (reference level = 3MSG)	1SG: –2.6[a] 1PL: –15.7 2SG: –.65 2PL: 16.03 3PL: .30	235	219 (173)
Morpholexical class (reference level = pronoun)	Definite noun: 2.16[b] Indefinite noun: 2.07[c]	208	198 (174)

Note: [a]$p < .05$, [b]$p < .001$, [c]$p < .01$. Token counts appear in Table 5.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

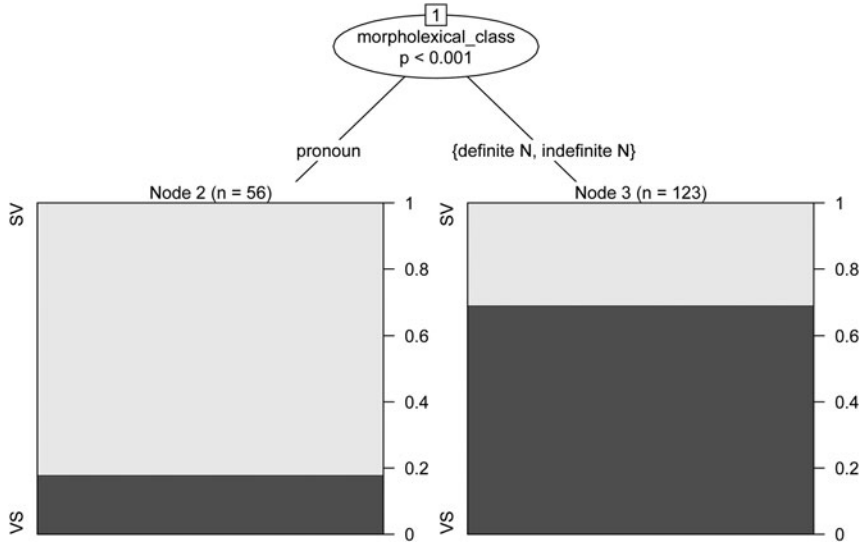


FIGURE 9. Conditional inference tree for SV and VS, Al-Rawi (1990) corpus.

Morpholexical class alone determines SV vs. VS. However, whereas the highest-level distinction in the tree from the oral analysis is between indefinite nouns and all other morpholexical categories, the Al-Rawi tree distinguishes pronouns from nouns. That is, both analyses show pronouns significantly favoring SV relative to nouns, but in the oral data, there exists a stronger distinction between indefinite nouns (which favor VS) and all other morpholexical categories.

In summary, the Al-Rawi results are completely consistent with the oral results insofar as the null vs. overt distinction is governed by previous mention with secondary person/number effects, whereas the SV/VS distinction is governed by morpholexical class.

In this context it can be noted that a reviewer suggested, first, that animate and inanimate subject referents may be realized differently and second, that the discourse status of our plural passives (cf. example (7)) is problematic in the context of the mental representation of generic subjects. Inanimate referents account for only 6% of subjects in the oral corpus and 3% in the Al-Rawi (1990) corpus.[11] Although they are significantly more likely to be realized as VS than as SV or null in both corpora, omitting inanimates from the analysis does not substantively change any of the results. Concerning plural passives, no significant differences arose when they were excluded in a separate statistical run (see n. 7).

DISCUSSION

We highlight two isomorphic results from the statistical analyses of the two independent samples. First, the overarching variable accounting for the

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

patterning of null vs. overt subjects is previous mention, with the presence of an identical referent in the previous clause significantly favoring a null subject. Secondly, morpholexical class is the dominant predictor of SV vs. VS, with definite and indefinite nouns favoring VS relative to pronouns.

SV, VS, and the generation of categorical meaning

In classical treatments such as Givón (1983), Prince (1981), and Chafe (1994), various discourse properties, including the expression of subject, are bundled under a general parameter. For Givón (1983), this is "topic continuity," for Prince (1981) "assumed familiarity," for Chafe (1994:83), the subject as "starting point," and for Ariel (1988, 1994, 2000), "accessibility". Within a Givónian framework, the discourse property of "continuity" is assumed, and continuity measures are defined. A key measure is "lookback," which indicates, for a given nominal token, how many clauses have intervened since the referent's last mention. Different grammatical structures typically have different characteristic lookback scores. Unstressed pronouns, for instance, tend to have a previous referent not more than two clauses back, whereas [demonstrative + noun] may have a distance of 8 clauses or more (Brown, 1983:328). If greater distance corresponds to lesser continuity, then an unstressed pronoun signals a higher degree of discourse continuity than does a [demonstrative + noun]. Although Givón (1983) does not formulate his hierarchies in terms of independent and dependent variables, continuity can be taken to be a dependent variable: for example, if you want to signal a high degree of continuity, use an unstressed pronoun.[12] Equally, the morpholexical categories can be viewed as the dependent ones. If you use an unstressed pronoun, then you signal a high degree of continuity.

What our results for Emirati Arabic show, however, is that the choice between SV and VS is to a large degree predictable independent of referential continuity: pronouns and quantifiers will be SV, nouns will be VS. Word order *depends on* morpholexical class. Following this basic observation, it is a straightforward inference that lexical classes have as one of their inherent properties a weighting to occur in either of the two possible word orders. This approach thus extrapolates away from the discourse conditions under which a given word order will be chosen and focuses on the properties of lexical items themselves. This perspective can be contrasted with that in Owens et al. (2009), where it was suggested that the two word orders in Emirati Arabic, SV and VS, had broad, categorically contrastive meanings: SV represented "availability," whereas VS represented "presentation", either the presentation of individual referents, and/or the presentation of a chain of events or instructions. This interpretation followed in the tradition of associating specific *constructions* with discourse-relevant meanings. In this study, we retain this pragmatic account but view it from a different angle. Specifically, we suggest that the contrasting discourse pragmatic meanings of SV and VS emerge from the pragmatic properties of the morpholexical classes that tend to occur in

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

either of these orders. In other words, the fact that SV corresponds to availability follows from the fact that pronouns, the class that dominantly realizes SV, tend to refer to entities that have been introduced previously or are situationally given. The association of VS with presentational discourse inheres in the meaning of definite and indefinite nouns insofar as they typically introduce new entities (cf. Chafe, 1994:80).

Correlations between morpholexical class and informational categories are common. English itself has a marked VS word order. In Birner's (1994:245) formulation, the element preposed to the inverted subject must not be information-newer than the inverted subject.

(11) Also complimentary is red and white wine.

A number of languages have, like Emirati Arabic and English, a tendency to place newer information after older information, the traditional theme-rheme distinction (cf. Jacennik & Dryer, 1992, for Polish and general discussion). However, Givón (1983:26) noted that Biblical Hebrew has presentational order as SV, and Tomlin and Rhodes (1992:119) noted that in Ojibwa in general new information is on the left (also Mithun, 1992:39). In Brazilian Portuguese, VS order marks a backgrounded, not a discourse-prominent subject (Naro & Votre, 1999), which aside from existentials is SV (Ferrari, 1990). Still, in all of these cases it is clear where in the clause new/old information is found, and this is often bifurcated between SV and VS order.

Looking at the issue from the morpholexical perspective, there is ample evidence for discourse-functional SV/VS variation to imply choice of morpholexical class. In Brazilian Portuguese, backgrounded VS subjects tend to be pronouns or quantifiers, whereas SV is realized by evoked full nouns. In English, unstressed pronouns cannot even occur as an inverted subject, which follows from their inherently evoked status.

(12) *Around the corner came he.

A number of languages with variable SV/VS word order have categorical word order for certain lexical categories. In Klamath, referential indefinites for instance, are always SV (Meyer, 1992:173), whereas definite subjects are either SV or VS. In the Central Chadic language of Wandala (Frajzyngier, 2012:9, 181, 193), independent subject pronouns precede the (non-reduplicated) verb, but nominal subjects follow. In the neighboring language of Glavda (Buba & Owens, 2007), existential subjects and focused subjects occur in VS order but pronouns in the nominal verb paradigm are obligatorily SV (see Owens et al., 2009:41, for a brief survey).

Underlying these observations is the fact that nouns prototypically introduce text-external information into discourse. Pronouns, on the other hand, are typically anaphoric.[13] These inherently contrastive discourse-lexical properties

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

TABLE 8. *Prototypical functions of null and overt subjects, relative to previous mention*

Previous mention	Null	Overt
New	Mainly as impersonal/plural passive	Normal discourse function
Available or situational	Under special circumstances	Normal discourse function
In previous clause	Normal discourse function	Under special circumstances

are mirrored in Emirati syntax by their inherent propensity to occur in VS vs. SV clauses, respectively.[14]

Null subjects, harmony and disharmony

Integrating null subjects, the basic issue is different from that pertaining to SV/VS. Here the question can be termed a selectional/interpretive one, pertaining only to the interpretation of third-person subjects. From the perspective of the speaker, the decision is when to select a null or an overt subject, and from the listener's, what the referent of a null subject is. Although in the majority of cases a null subject is selected/interpreted when the following clause has the same referent as the current one, there are an appreciable number of exceptions. By the same token, there are a good number of instances where an overt subject has the same referent as the preceding one, again against the dominant statistical trend. Table 8 above summarizes the realization of null and overt subjects relative to our basic referentiality categories.

The "normal" functions are those that logistic regression found to be most strongly associated with null or overt. We will term these the *harmonic* values. The "under special circumstances" cells we term *disharmonic*. The two primary disharmonic cells are overt subjects coreferential with the previous subject and null non-coreferential subjects. Turning first to the former type, there are 188 of these in the oral data (Table 1). A closer look at these tokens reveals two dominant grammatical and pragmatic environments. First, 33 of these tokens occur in dependent clauses that are either complements of *gaal* 'say' (see (3)) or complements of an elided *gaal*.

To gauge the discourse status of the remaining overt subjects that are coreferential with the previous subject, a review of about half the sample was made, the complete Al-Rawi (1990) texts and the first 800 clauses in oral. Among these, 13 overt coreferential subjects are verbatim repetitions of the previous clause, including a number of false starts. Of the remainder, 31 occur in instances of "discourse hiatus," our cover term for a diverse set of discourse junctures that typically are marked with overt subjects. Representative types of discourse hiati are as follows.

A topic of discussion ends, and another begins with an overt subject having the same referent as the previous subject (Owens et al., 2010:38). In (13), the verb *nəgʕad* ends an episode, the same speaker begins the next episode with the same subject, but with an overt one (overt subjects italicized in English, underlined in Arabic):

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

(13) AR *Ø nə-gʕad hinaak baʕd ma Ø ir-ruuħ, Ø nə-rjaʕ Ø nə-gʕad*
we$_1$-stay there still not we$_1$-go, we$_1$-return *Ø we$_1$-stay*
AR *fa marṛa, Ø rəħ-na bi siyaakəl, ana wə rəbiiʕ-i*
and once Ø went-we$_1$ by bicycle, I and friend-my
AR 'We'd continue staying there, we wouldn't go (home), we'd return and we$_1$'d just stay.'
AR 'And then one time, *we*$_1$ went by bicycle *I and my friend*$_1$' (Oral)

A highpoint in a scene is frequently marked by an overt subject in VS word order, even when the previous subject is coreferential:

(14) *Ø yi-gbiḍ$_1$ il-ʔarnab wu yi-ðbaħ-a ir-rayyaal$_1$*
Ø he-grabs DEF-rabbit and he-slaughter-it DEF-man
'He$_1$ grabs the rabbit and *the man*$_1$ kills it' (Al-Rawi, 1990:155)

An independent clause precedes, with an overt subject, followed by a dependent clause, which sets the scene for the continuation of the narrative, with an identical overt subject:

(15) *bi-l-baačir ʕaad yaa ir-rayyaal$_1$, yoom yaa ir-rayyaal$_1$ gaal*
on-DEF-tomorrow so came DEF-man when came DEF-man said
'The next day the man came again. When *the man* came he said . . .' (Al-Rawi, 1990:121)

Other coreferential overt functions include:

- The overt subject is rhetorically emphatic.
- In one case, speaker A interrupts speaker AR. AR answers A with a null subject, then continues the story with an overt coreferential subject.
- The previous clause is a relative clause. The following (independent) clause's subject is coreferential with the subject of the relative clause, and it is overt.
- A coreferential overt subject closes a narrative.
- A coreferential overt subject marks a transition from background to main storyline

The list is heterogeneous—in some cases, the repetition functions to highlight an event within a story, in others it is determined by the speech situation, an interpolation by an interlocutor for instance. A conspiracy of factors converge to produce the coreferential overt subject. However, in all cases, the hiatus introduces a newness residing in the shift in discourse perspective, not in a new referent. From this vantage point, the use of overt in previous clause subjects is consistent with the conclusion that overt subjects favor new referents. We surmise that the subject is an item always available to be "phoneticized" to mark a discourse hiatus when necessary.

We now turn to the second disharmonic type: null subjects whose referents are either situationally given or available in discourse, but are not identical to the referent of the previous subject. In Owens et al. (2010:45), it is shown that the

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

likelihood of this type of null subject correlates positively with the number of tokens of the referent in a discourse episode.[15] That is, the more a referent is used in a discourse context, the higher the likelihood there is that it will be realized as null. This point can be illustrated here with one instance from the oral data, an example easily replicable in many episodes both in the oral data and in Al-Rawi (1990).

In the longest of the three texts in the oral sample, the lion's share of the conversation is controlled by AR and A, and on the whole it is conversational. However, there are longer episodes in it where one or the other of the interlocutors tells a story about an aspect of their life experience. In one of these, there are 210 clauses. The participants in the story are an unnamed taxi driver, the narrator AR, his friend, and a policeman. Only the narrator and the policeman have significant roles. The policeman is named by an overt noun only twice, once when he is introduced, *waaħid min iš-šurṭa* 'one of the police', and once when he is reintroduced after the narrator and his friends absent the scene and return, finding the policeman, where he is identified as *nafs ir-rayyaal* 'the same man'. This same referent occurs four times as *ana* 'I', each time introduced by *gaal* 'say', and 21 times as a null subject. Eight of the null subjects are non-coreferential with the previous subject (the second null disharmonic type), the longest separation from previous reference being eight clauses. We see here that the relatively weak correlation of null/overt subject with distance from referent discussed earlier (see n. 6) is a reflection of the fact that once referents have been established in an episode, they will be marked by null subjects, even at relatively long lookback distances.

To continue the parallel to English, this same point is in evidence in Prince's (1981) oral narrative (see n. 12). In that story as well, the main protagonist is introduced by a noun only once, at the beginning of the story ("Jan"), and in all subsequent subject references is referred to with an unstressed pronoun, "she" (or "I"). The very last reference to she is also with an unstressed pronoun "she", though it is separated from the last reference by 12 clauses, and by 14 tokens of an intervening "she$_2$."

If we stipulate simply that Emirati Arabic avoids overt subjects with discourse-available referents, except in cases of discourse hiatus and in clausal complements of *gaal*, then this second disharmonic type becomes harmonic. There is no independent tendency in narratives to reintroduce a referent with an overt subject after a separation of several clauses from the referent's last mention (a la Chafe's, 1994, semi-active). Cases of discourse hiatus are quite frequent and diverse; if not predictable in the current corpus, they are at least explicable in discourse-functional terms and certainly amenable to better characterization with a larger corpus. The disharmonic cells in Table 8 can be thus seen to be subject to conditions that render them harmonic against these factors.

CONCLUSIONS

Without abandoning the goals and analytical categories of the important work on discourse carried out in particular in the 1980s and 1990s, we have attempted to

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

refocus this research within a more explicitly statistical framework, integrating interpretive categories to an outcome based on a classification of subjects in terms of coreference. A contrast emerges between null and overt subjects, based on coreferentiality with the previous subject. This major contrast, in turn, evokes analytic challenges of its own. In particular, if the coreferentiality contrast is the normal or harmonic case, the task of defining the large and heterogeneous class of disharmonic cases, is identified as a locus for future research. By the same token, our treatment does permit focusing on fundamental interpretive issues, one of which concerns the distribution of SV and VS clauses.

Here we would like to advance three arguments in favor of recognizing an inherent, discourse-pragmatic coding of referring expressions. The first argument is simply that our interpretation is directly compatible with the empirical observations. To a highly significant degree, nouns and pronouns/pronominals are distributed in complementary VS and SV clause structures.

Second, assuming that nouns and verbs are inherently weighted for their occurrence as subject helps account for the fast and constant processing of clauses. Impressionistically, there is no "cost" (contrary to Chafe, 1994; Lambrecht, 1994) in using a null vs. overt or SV vs. VS structure in terms of hesitations or other physical measures. We advance this as a testable hypothesis. Postulating an inherent weighting would help explain the constant speed of delivery across subject types and dovetails with the idea that individual lexical entries contain rich statistical and collocational information (see n. 14).

The third argument derives specifically from the oral results. In the case of a sequence of non-coreferential subjects, the speaker chooses either a null subject or an overt subject. As shown, non-coreferentiality in this case favors an overt subject, regardless of SV/VS order (see Table 3, Figure 4). What is selected preferentially is an entity, a nominal lexeme, not a word order. Furthermore, because a subject can also be null, what is selected is an entity, not merely the abstract category "subject." From this, we conclude that what is targeted is a nominal subject, that is, a pronoun, pronominal, or noun.

NOTES

1. All but the text on p. 169 was included in the analysis.

2. As a percentage of total predicates in a typical text, the existentials and equatives are a relatively small proportion. In Owens et al. (2010:26–27), figures are presented on texts from different Arabian peninsular varieties, including the ones analyzed here, which in slightly under 80% of all predicates are verbal. Thus, on a statistical basis alone the current study deals with the core of spoken Arabic discourse. One interesting category not included in the study is the active participle (numerically hardly 4% of all predicates in the statistics just cited), which marks number and gender like a verb, but not person.

3. Benmamoun and Choueiri (2013) described the possessive construction as the merger of two nouns, possessed and possessor, into a single lexical unit. The definiteness feature of the unit is sensitive to the second noun.

4. Our previous studies included "same/different subject," a binary variable encoding whether the subject is morphologically the same as or different from the immediately preceding subject. This variable was included in our earliest models, but it was generally not significant.

5. The suffix *-ó* is a portmanteau for "3PL subject" and "3MSG object." The following verb *xađđ* is the unmarked 3MSG subject form.

6. For instance, the chi-square distribution for null vs. overt (combined SV + VS) is as follows:

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

	Null	Overt	Total
1–2	28	11	39
3–5	16	11	27
6 or more	7	7	14
Total	51	29	80

Note: $\chi^2 = 2.47$, $df = 2$, $p > .29$.

When SV and VS are coded as separate values, the significance drops even more.

7. This is quite common in spoken Arabic; for example, Moroccan Arabic has a 3MSG "passive" identical to 3MSG active (Maas, 2011:181). Three readers point out interpretive issues of considering the plural passive as "new." This is clearly a significant issue. Here we simply note three salient points, two of them utilitarian. First, these subjects do not obviously fit among any other class: same as previous, previously mentioned, or situational. Second, merely from a statistical perspective, there are not enough tokens of them to include them as a parallel class to other previous mention categories. Third, we did perform statistical runs excluding these tokens, which resulted in no difference at all on the overall contrastive categories vis-à-vis their inclusion.

8. It is interesting to note that the figures in Table 1 differ markedly from those presented in Ariel (2000:13–15) for Hebrew which were also based on corpus data drawn from a number of studies. In Hebrew, there appears to be a far higher tendency, for instance, for Ø-marked subjects to appear with first- or second-person verbs as opposed to third-person verbs. In our data, third-person plurals have the highest tendency toward Ø marking, whereas 1SG and 3FSG are about the same proportionally as regards Ø vs. overt. We therefore hesitate to draw broad typological conclusions on the basis of even a range of texts and to generalize to such essentialized terms as "Semitic languages." Though we have argued for the historical Semitic relevance of the data summarized here (Owens & Dodsworth, 2010), we hesitate to generalize the observations even to contemporary spoken Arabic at large, given the different ways of expressing subject-verb word order among different varieties.

9. The adverbial clause and *gaal* complement variables were also included in all models and were weakly significant in most cases, but they will not be discussed further.

10. An anonymous reviewer suggests that two important dimensions define discourse. One is "related to the text measure referential (anaphoric) distance. But it is *not* about distance per-se (a mere heuristic), but rather about mental accessibility in either one of the three representational systems in the mind/brain Focal attention, working memory, or episodic memory." The second "can be measured (again, a mere heuristic) in terms of cataphoric persistence, but it's again *not* about persistence per-se but rather about *importance/topicality* of the referent that is, how prominent episodic-memory representation (of the referent) is the text-producer trying to induce the text-receiver to build." Leaving aside methodological issues in linking cognitive and discourse categories, we feel that the issue of topicality and cataphoric persistence is problematic. There are two prominent reasons. First, cataphoricity as a measurement (Givón, 1983) is an analyst's concept and has no psychological reality. Cataphoric persistence measures "the number of recurrences of the referent in the subsequent ten clauses" (Givón, 1988:248). Clearly this count can be realized only on a post hoc basis, only after the clauses have been recorded. Except perhaps in the extreme case of memorized stories, however, speakers do not plan 10 clauses in advance (Levelt, 1989:258), let alone determine in these 10 clauses how many tokens of the referent will be inserted, and where.

Second, the idea of "subsequent 10 clauses" is ill-defined. Consider the following hypothetical example, shortened to five clauses, of a referent introduced then sporadically reactivated as indicated. Patterns such as the following are common in our data: N_1 pro_1 pro_1 X (different referent) X pro_1. From a Givónian perspective (1983, 1992), presumably, the initially introduced N_1 has a persistence score of 3. However, this is too low, unless the claim is made that the only coreference relation that exists is that between a N_1 and pro_1, not between $pro_1 + pro_1$. In fact, this hypothetical example has a score of 6: N is repeated in the subsequent three pros, the first pro in the subsequent two, and so on. If the discourse basis of the persistence measure itself is questionable, certainly one should pause before trying to link it to cognitive categories. The way sketched here of *representing* cataphoric persistence complements the observation in Owens et al. (2010:45) that the longer a referent persists in discourse, the more likely it is to be null.

11. Looking further afield, the major animacy effect resides in the subject vs. object parameter. Whereas subjects are overwhelmingly animate, objects are overwhelmingly inanimate, a contrast worthy of further investigation.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

12. This interpretation serves to focus the independent and dependent variables in Givón's (1983) approach. There are conditions of appropriate usage of any of the relevant structures that add other dependent variables.

Prince's (1981) "assumed familiarity" has a similar interpretation to "continuity," though she only implicitly matches morphosyntactic structures with different degrees of "availability." In Prince's (1981:238–242) analysis of a text, the "evoked" pronouns are in all likelihood unstressed. Applying a Givón-like lookback count to the main protagonist in the story gives a score of 1.4, which is identical to referent in the previous clause in most instances. The English unstressed subject pronouns thus appear analogous to the Emirati null subject.

13. There are of course exceptions (Prince, 1981:224, 232); see Ariel (1988) for general discussion.

14. This perspective has been recently proposed in a different context by Erker and Guy (2012:552) in a variationist study of Spanish in New York City. They observe that the occurrence of null/overt subject pronouns in various morphogrammatical contexts is better differentiated and better profiled with high-frequency than it is for low-frequency verbs. They argue that the postulation of person-based constraints are applicable only, or to a greater degree, among high-frequency verbs, and that establishing these generalizations "require[s] lexical representations that are rich enough to store information about frequency and collocations."

15. See Owens et al. (2010:30–33) for characterization of "episode" in these texts.

REFERENCES

Al-Rawi, Rosina. (1990). *Studien zum Arabischen Dialekt von Abu Daby*. Heidelberg: Julius Groos.

Appen Pty Ltd. (2006). Gulf Arabic conversational telephone speech. Philadelphia: Linguistics Data Consortium. Available at: http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2006T15.

Ariel, Mira. (1988). Referring and accessibility. *Journal of Linguistics* 24:65–87.

_____. (1994). Interpreting anaphoric expressions: A cognitive versus a pragmatic approach. *Journal of Linguistics* 30:3–42.

_____. (2000). The development of person agreement markers: From pronouns to higher accessibility markers. In M. Barlow & S. Kemmer (eds.), *Usage-based models of language*. Stanford: Center for the Study of Language and Information. 197–260.

Benmamoun, Elabbas, & Choueiri, Lina. (2013). The syntax of Arabic from a generative perspective. In J. Owens (ed.), *The Oxford handbook of Arabic linguistics*. Oxford: Oxford University Press. 115–64.

Birner, Betty. (1994). Information status and word order: An analysis of English inversion. *Language* 70:233–259.

Brown, C. (1983). Written English. In T. Givón (ed.), *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam: Benjamins. 317–341.

Brustad, Kristen. (2000). *The Syntax of spoken Arabic*. Washington, DC: Georgetown University Press.

Buba, Bello, & Owens, Jonathan. (2007). Glavda morphology. In A. Kaye (ed.), *Morphologies of Asia and Africa*. Winona Lake, IN: Eisenbrauns. 641–676.

Chafe, Wallace. (1994). *Discourse, consciousness, and time*. Chicago: University of Chicago Press.

Dahlgren, Sven-Olof. (1998). *Word order in Arabic*. Göteberg: Acta Universitatis Gothoburgensis.

Erker, Daniel, & Guy, Gregory. (2012). The role of lexical frequency in syntactic variability: Variable subject personal pronoun expression in Spanish. *Language* 88:526–557.

Ferrari, Lilian V. (1990). Distribution and function of word order variation in Brazilian Portuguese. *Journal of Pragmatics* 14:649–666.

Frajzyngier, Zygmunt. (2012). *A grammar of Wandala*. Berlin: Mouton.

Givón, Talmy (ed.). (1983). *Topic continuity in discourse: A quantitative cross-language study*. Amsterdam: Benjamins.

_____. (1988). The pragmatics of word-order: Predictability, importance and attention. In M. Hammond, E. Moravcsik & J. Worth (eds.), *Studies in syntactic typology*. Amsterdam: Benjamins. 243–284.

_____. (1992). The grammar of referential coherence as mental processing instructions. *Linguistics* 30:5–56.

Gundel, Jeanette, Bassene, Mamadou, Gordon, Bryan, Humnick, Linda, & Khalfaoui, Amel. (2010). Testing predictions of the givenness hierarchy framework: A crosslinguistic investigation. *Journal of Pragmatics* 42:1770–1785.

Hoffiz, Benjamin. (1995). *Morphology of U.A.E. Arabic, Dubai dialect*. Ph.D. thesis, University of Arizona.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press

Holes, Clive. (2009). Word order and textual function in Gulf Arabic. In J. Owens & A. Elgibali (eds.), *Information structure in spoken Arabic*. London: Routledge. 61–74.
Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *J Comput Graph Stat* 15:651–674.
Jacennik, Barbara, & Dwyer, Matthew. (1992). Verb-subject order in Polish. In D. Payne (ed.), *Pragmatics of word order flexibility*. Amsterdam: Benjamins. 209–242.
Lambrecht, Knud. (1994). *Information structure and sentence form*. Cambridge: Cambridge University Press.
Levelt, Willem. (1989). Speaking: From intention to articulation. Cambridge: MIT Press.
Levinson, S. (1987). Pragmatics and the grammar of anaphora: A partial pragmatic reduction of binding and control phenomena. *Journal of Linguistics* 23:379–434.
Maas, Utz. (2011). *Marokkanisches Arabisch: Die Grundstrukturen*. Munich: Lincom.
Meyer, Karen. (1992). Word order in Klamath. In D. Payne (ed.), *Pragmatics of word order flexibility*. Amsterdam: Benjamins. 167–192.
Mithun, Marianne. (1992). Is basic word order universal? In D. Payne (ed.), *Pragmatics of word order flexibility*. Amsterdam: Benjamins. 15–61.
Naro, Anthony, & Votre, Sebastiao. (1999). Discourse motivations for linguistic regularities: Verb/subject order in spoken Brazilian Portuguese. *Probus* 11:75–100.
Owens, Jonathan, Dodsworth, Robin, & Rockwood, Trent. (2009). Subject-verb order in spoken Arabic: Morpholexical and event-based factors. *Language Variation and Change* 21:39–67.
Owens, Jonathan, Young, Bill, Rockwood, Trent, Mehall, David, & Dodsworth, Robin. (2010). Explaining Ø and overt subjects in spoken Arabic. In J. Owens & A. Elgibali (eds.), *Information structure in spoken Arabic*. London: Routledge. 20–60.
Owens, Jonathan, & Dodsworth, Robin. (2010). Stability in subject-verb word order: From contemporary Arabian Peninsular Arabic to Biblical Aramaic. *Anthropological Linguistics* 51:151–175.
Prince, Ellen F. (1981). Toward a taxonomy of given–new information. In P. Cole (ed.), *Radical pragmatics*. New York: Academic Press. 223–255.
R Development Core Team. (2010). R: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing. Available at: http://www.R-project.org/.
Tagliamonte, Sali, & Baayen, Harald. (2012). Models, forests, and trees of York English: Was/were variation as a case study for statistical practice. *Language Variation and Change* 24:179–202.
Tomlin, Russel, & Rhodes, Richard. (1992). Information distribution in Ojibwa. In D. Payne (ed.), *Pragmatics of word order flexibility*. Amsterdam: Benjamins. 117–136.

https://doi.org/10.1017/S0954394513000173 Published online by Cambridge University Press