PERSPECTIVES ON ARABIC LINGUISTICS XV
AMSTERDAM STUDIES IN THE THEORY AND HISTORY OF LINGUISTIC SCIENCE

General Editor
E. F. KONRAD KOERNER
(Zentrum für Allgemeine Sprachwissenschaft, Typologie und Universalienforschung, Berlin)

Series IV – CURRENT ISSUES IN LINGUISTIC THEORY

Advisory Editorial Board

Lyle Campbell (Christchurch, N.Z.); Sheila Embleton (Toronto); Brian D. Joseph (Columbus, Ohio); John E. Joseph (Edinburgh) Manfred Krifka (Berlin); E. Wyn Roberts (Vancouver, B.C.); Joseph C. Salmons (Madison, Wis.); Hans-Jürgen Sasse (Köln)

Volume 247

Dilworth B. Parkinson and Samira Farwaneh (eds.)

Perspectives on Arabic Linguistics XV
Papers from the Fifteenth Annual Symposium on Arabic Linguistics, Salt Lake City 2001.
PERSPECTIVES ON
ARABIC LINGUISTICS XV
PAPERS FROM THE FIFTEENTH ANNUAL
SYMPOSIUM ON ARABIC LINGUISTICS,
SALT LAKE CITY 2001

Edited by
DILWORTH B. PARKINSON
Brigham Young University

SAMIRA FARWANEH
University of Arizona

JOHN BENJAMINS PUBLISHING COMPANY
AMSTERDAM/PHILADELPHIA
CONTENTS

Editorial Note

Introduction
  Dilworth B. Parkinson

I. Sociolinguistics

Discourse Particles Revisited: The Case of Wallahi in Egyptian Arabic
  Mustafa A. Mughazy

Theories of Code Switching in the Light of Empirical Data from Egypt
  Reem Bassiouney

Translating Arabic Speech Act Expressions
  Rudolf Reinelt

II. Syntax

Relative Clauses in Syrian Arabic: Two Reconstruction Problems
  James Darrow

Interpretability, Feature Strength, and Impoverished Agreement in Arabic
  Mark S. LeTourneau

III. Corpus Linguistics

Evolution of MSA, the Case of Some Complementary Particles
  Mark Van Mol
CONTENTS

NP Structure Types in Spoken and Written Modern Standard Arabic (MSA) Corpora
Sameh Al-Ansary 149

Comparing Frequencies of Lexical Productions in Arabic Words
Stephen Taylor 181

Future Variability: A Corpus Study of Arabic Future Particles
Dilworth B. Parkinson 191

Index of Subjects 213
EDITORIAL NOTE

On March 2-3, 2001, the Fifteenth Annual Symposium on Arabic Linguistics was held at the University of Utah in Salt Lake City, Utah. The symposium was sponsored by the Arabic Linguistics Society and the University of Utah. Samira Farwaneh was the local organizer of the symposium.

A total of twenty-four papers were presented at the symposium; of these, nine are published in this volume. The papers presented at the symposium were selected on the basis of an anonymous review of abstracts submitted to the Program Committee. The papers included in the volume were further reviewed by the editors before their final acceptance for publication.

The transcription of all Arabic materials in the body of the papers follows the International Phonetic Alphabet or standard equivalents. The Arabic emphatics, however, are represented by a dot underneath the symbol, and long vowels as sequences of two vowels. The transliteration of Arabic names and titles follows accepted formats, with some simplification in the use of diacritics.

The preparation and printing of the final manuscript was done using facilities available at Brigham Young University. I am indebted to Aaron Schofield, Spencer Penrod, and Tessa Hauglid, who served as assistant editors for this volume.
INTRODUCTION

Dilworth B. Parkinson

The papers in this volume deal with various topics in Arabic Linguistics, and focus on three specific areas: sociolinguistics, syntax, and corpus linguistics. The papers are data-oriented, and focus both on new theoretic insights and new techniques of data analysis.

Section I includes three papers in the area of sociolinguistics. Mughazy looks at a (less-typical) non-oath use of the oath phrase wallaahi in Egyptian Arabic. He demonstrates several differences between the oath and non-oath usages, and then shows how the non-oath usage functions as a mitigating politeness or face-saving device. Perhaps his most interesting result is that some of the usages that are thus mitigated are surprising from a western, or even a ‘theoretical’ point of view, and can only be understood in a local Egyptian context of what is and what is not face threatening and what needs to be mitigated.

Bassiouney’s topic is diglossic code-switching. Using data from Egypt, she looks at various theories of code-switching, concentrating on the Matrix language model, and finds that her Egyptian data poses problems for all current proposals. She wonders if diglossic and bilingual code-switching are indeed examples of the same phenomenon, and gives suggestions for modifying the theory to account for her data.

Reinelt looks at ‘speech act expressions,’ words that refer to specific speech acts (like ‘to say,’ ‘to declare’), in the Quran, and in German and English translations of the same to demonstrate some aspects of translation theory and efficacy. He finds a major difference between what might be termed ‘academic’ and ‘informative’ translation styles.

Section II contains two papers on syntax. Darrow looks at reconstruction problems in Syrian Arabic relative clauses, contrasting
semantic interpretation with a copy-theoretic approach. Syrian Arabic raises certain problems for any account, but ‘discomfort notwithstanding,’ he shows that a copy-theoretic approach provides a reasonable account of the data.

LeTourneau looks at what he calls ‘impoverished’ number agreement in Standard Arabic, specifically in relation to Form VI verbs, and tweaks a number of minimalist assumptions to account for the data.

Section III is a set of four papers dealing with the general area of corpus linguistics. Corpus linguistics in general includes two main orientations: 1) scholars coming from the computer science area, interested in using corpora to test computational models of the language, and 2) scholars coming from a more traditional linguistics, using corpora to analyze particular characteristics they have identified as problems for whatever reason. The Taylor paper comes from the first category. The author describes his attempt to build a lexical analyzer for Arabic, which turns out to be several orders of magnitude more complex than a similar program for English would be. He then presents the results of a count of frequencies of various verbal forms, involving the categories of person, tense, gender and number.

The other three papers come from the second orientation. Al-Ansary uses a computational model in association with both a spoken and written corpus to compare NP structures in Spoken and Written Standard Arabic. There is a large body of interesting work comparing spoken and written English, but I believe that this is the first look at this topic for Arabic.

Van Mol utilizes a corpus of Spoken Standard Arabic from three Arabic countries to investigate to what extent the language is uniform on a syntactic level across the Arabic speaking world. Besides giving an excellent discussion of the problems and pitfalls of developing such a corpus, he arrives at some highly interesting results, showing that although most of the items investigated were identical in the various countries, there were differences.

Parkinson’s paper follows a similar tack, using a large corpus of newspapers from various countries to look at the variability in future particle use, both across countries and across genres.
SOCIOLINGUISTICS
DISCOURSE PARTICLES REVISITED
THE CASE OF WALLAHI IN EGYPTIAN ARABIC

Mustafa A. Mughazy
University of Illinois at Urbana-Champaign

1. Introduction
The Egyptian Arabic oath phrase *wellaahi* (by God) is often used in utterances that do not have the illocutionary force of issuing vows or threats -- the typical uses of *wellaahi*. When used in non-oath utterances, it is often described as a ‘dummy word of hesitation’ or ‘speech filler’, used to reserve one’s turn in a conversation or to give the speaker time to organize his/her thoughts (Piamenta, 1979). Most studies on Arabic politeness and religious formulas, such as Farghal & Borini (1997) and El-Sayed (1989), have focused on sentential formulas rather than discourse particles. The arguments presented in this paper describe two distinct non-oath uses of *wellaahi* that have different distribution patterns and trigger different implicata about the speaker’s beliefs with regard to politeness and power relations, in particular the lack of power to effect events in a given context by relegating that power to God.

The oath *wellaahi* is a prepositional phrase that is composed of the prefix preposition *wa* (by), which is usually referred to as *waaw al-qasam* (oath *wa*), the noun phrase *allaah* (God), and the genitive case marking suffix *i*. In terms of pragmatic function *wellaahi* is a member of a set of interchangeable oath phrases that includes *winnabi* (by the Prophet), *wiqpigil* (by the Bible), *wiladra* (by the Virgin) and many others that have the same morphological structure. In a given speech event, the use of one of these phrases rather than any other is motivated by the speaker’s religious beliefs and his/her view of whether declaring

---

1 I am very grateful for the invaluable comments and suggestions I have received from Professor Devin Stewart and Rajeshwari Pandharipande. All errors and inadequacies are solely mine.
these beliefs and the degree to which they are observed help achieve their goals. Otherwise, a more neutral oath phrase can be used such as *wiḥyaatak* (by your life) and *wilṣiiš wilmalḥ* (lit. by the bread and salt ‘by our friendship’), among several others.

Although *wallaahi*, when used in non-oath utterances, has the same morphological structure, it has different functions and patterns of distribution, and it cannot be substituted with any other oath phrase. In addition, *wallaahi* always receives the main sentence stress in oath utterances but is never stressed in non-oath ones. Oath utterances can be made emphatic by modifying the noun phrase *allaah* (God) with the adjective *il-‘a(iim* (almighty), but such adjunction is not possible in non-oath uses. Another difference is that the genitive case marking morpheme is optional in oath structures (i.e., *wallaah* instead of *wallaahi*), but it is obligatory in non-oath uses. These differences prompt the claim that non-oath *wallaahi* is a frozen or fixed expression best described as a discourse particle.

Crystal (1980) asserts that discourse particles in general have been ignored in descriptions of grammars partly because they were viewed as meaningless interjections that lack interesting patterns of distribution and because they occur mainly in face-to-face interactions. Schourup (1982) and Carlson (1984) argue that the use and distribution of discourse particles are rule governed, and that they are used to convey social and pragmatic meaning. Since particles typically do not change the truth value or the propositional content of an utterance, speakers have the choice whether or not to use a given particle depending on their perception of how using (or not using) that particle helps achieve their goals, and how an addressee would interpret the motivation behind that use, assuming that the use of linguistic forms is rational and goal-oriented behavior (Green, 1982).

Levinson (1997:162) describes discourse particles as hedges that speakers use to communicate their beliefs about the extent to which they adhere to the maxims of Grice’s Cooperative Principle. For example, when responding to the question “Where are my glasses?” a speaker

---

2 Many devout Moslems would not use any oath phrase other than *wallaahi*, as oath phrases that do not make direct mention of God are considered inappropriate and against religious teaching.

3 Requests seem to be an exception to this generalization, since any oath phrase can be used. However, in such instances the request is an extension of the oath use i.e., beseeching.
would say, “Well, they are not here”. Using the discourse particle ‘well’ indicates that the speaker does not fully adhere to the maxim of Quality: ‘agents will not say what they believe to be false or what they do not have adequate evidence for’, which is a basic assumption that individuals use to interpret others’ linguistic behavior rather than a moral rule of felicity or politeness (Green, 1996). The same analysis applies to politeness particles assuming that politeness is a universal phenomenon derived from Grice’s maxims despite the differences in the cultural backgrounds that define politeness in a given community (Green, 1996, and Brown & Levinson, 1978). Therefore, failing to use discourse particles appropriately, especially politeness particles, leads to the undesired consequences of misinterpreting speakers’ intentions and actions (Svartvik, 1980).

2. The Distribution Patterns of Wallaahi

One of the difficulties encountered when analyzing the different uses of wallaahi is its ambiguity, as illustrated in example (1), which can be used to mean either (1a) or (1b). In these examples it is how wallaahi is interpreted that determines the illocutionary force of the utterance i.e., whether it is an emphasized statement (1a) or a polite refusal, say to a request for a loan (1b). However, such structures can be easily disambiguated by using the structural constraints on the occurrence of oath wallaahi, which can occur in either the utterance initial or final position, as in example (2). When used in non-oath utterances, wallaahi occurs only initially, as in example (3), because otherwise it is uninterpretable, as in (4).

(1) wallaahi mašandišṣ̌ fikus
   a. wallaahi (I swear) I do not have money.
   b. wallaahi (I am afraid) I do not have money.

(2) wallaahi ma-ṣaxatt-iš haaga (wallaahi)
   I swear I did not take anything (I swear).

(3) wallaahi ana baqtariḥ in iḥna nišṭiki.
   By God, I suggest that we complain.

(4) ana baqtariḥ in iḥna nišṭiki wallaahi!
   I suggest that we complain, by God.
The function of \textit{wallaahi} when used as an oath phrase is to signal the speaker’s commitment to the propositional content of the utterance i.e., the speaker is emphasizing his/her strict adherence to Grice’s maxim of Quality. Therefore, the oath phrase is used only with declarative sentences that have the illocutionary force of stating or threatening. Only in utterances where \textit{wallaahi} cannot be used as an oath particle (e.g., directives and questions), can it occur in the utterance final position, as in example (5) below.

(5) \textit{iftah il-baab wallaahi} \\
\hspace{1cm} Open the door \textit{wallaahi}. (polite directive)

Another distinction between the two uses of \textit{wallaahi} is based on their functional patterning with two structurally distinct types of negation in Arabic: descriptive or sentential negation, which is formed by attaching the prefix \textit{ma-} and the suffix \textit{-i}ṣ to the verb, and metalinguistic negation, which is expressed by preceding the verb with the negative operator \textit{mi}ṣ. Descriptive negation is a truth functional type of negation that is used to deny the propositional content of an utterance, as in example (6) below, where \textit{wallaahi} is used to emphasize the negation. The other type of negation is not truth functional and is used to object to a previous utterance on any ground including its form (phonology, morphology, etc.) or to an implicature or presupposition triggered by that utterance (Horn, 1985). Hence, it is usually followed by a rectification that provides grounds for the objection (Horn, 1989). The non-oath \textit{wallaahi} can be used only with metalinguistic negation, as in example (7) below, where \textit{wallaahi} has a meaning similar to the English sentential adverb ‘actually’. The function of metalinguistic negation is to deny what the addressee said earlier (Geurts, 1998), which is a face-threatening speech act that requires mitigation. Inappropriate use of \textit{wallaahi} with negation, which is detected by intonation patterns and the morphological structure of the verb, results in the addressee’s inability to recognize the mitigation or interpret the negation.

(6) \textit{wallaahi ana ma-a-xatt-i}ṣ ik-kitaab \hspace{1cm} (\textit{wallaahi} is stressed) \\
\hspace{1cm} (I swear) I did not take the book.

(7) \textit{wallaahi, ana mi}ṣ a-xatt ik-kitaab? ana istalafi tuh \hspace{1cm} (\textit{wallaahi} is not stressed) \\
\hspace{1cm} \textit{wallaahi} (actually) I did not ‘TAKE’ the book, I borrowed it.
3. **Wallaahi and Politeness**

Having distinguished between the oath use of *wallaahi* and the non-oath ones in terms of distribution, this section describes the function of *wallaahi* as a politeness marker. The claim that *wallaahi* can be used as a politeness marker is motivated by the observation that it is used preceding potentially face-threatening speech acts such as those in examples (8)-(13) below that occurred in conversations between individuals of equal social status and power differential. A speaker would calculate the potential effects of his/her up-coming utterance on the addressee and decide whether these effects aid achieving his/her goals or not. If the utterance is viewed as likely to offend the addressee, and consequently hinder achieving these goals, *wallaahi* is used as a forewarning that is intended to be interpreted as to mitigate the undesired effects.

(8) wallaahizintazzlaazimti2tiki (suggestion)  
*wallaahi* you must complain.

(9) wallaahizintazzHurrzti'millizintaz'awzu (offering options)  
*wallaahi* it is up to you to do whatever you want.

(10) wallaahizruddz'a-t-t-ilifoon (directive)  
*wallaahi* answer the phone.

The lack of *wallaahi* makes the utterances in examples (8)-(10) on-record face-threatening acts i.e., intended to make the addressee uncomfortable. For example, the utterance in (10) is a command, which is toned down by the use of *wallaahi* to a polite directive or a request. In the examples in (11)-(13), the use of *wallaahi* not only makes these utterances less face-threatening, but also it changes the speech act. For example, without *wallaahi* the utterance is (11) would be interpreted as an accusation, and the one in (12) as a threat or a demand⁴. Example (13) is particularly interesting because the use of *wallaahi* changes the implicature of the utterance, as it indicates that the speaker disagrees with the addressee, yet the issue is still negotiable, whereas the same utterance

---

⁴ The categorization of the speech act is dependent on the addressee’s interpretation of the speaker’s intentions. For example, whether the utterance in (12), without *wallaahi*, is a threat or a demand is determined by the addressee. In either case, the use of *wallaahi* blocks against such interpretations.
without *wallaahi* would be interpreted as a final rejection rather than a disagreement.

(11) *wallaahi* inta kunt yaltaan  
*wallaahi, you were wrong.*  
(criticizing)

(12) *wallaahi* ana ṣaawiz filuusi bukra  
*wallaahi I want my money back tomorrow.*  
(request)

(13) *wallaahi* ana mīṣ muwaaffi  
*mīṣ-l-iqtiraah da*  
*wallaahi I disagree to that suggestion.*  
(disagreeing)

Lakoff (1973) describes three rules that speakers follow to achieve politeness in discourse. The first of these rules is ‘do not impose’, where imposition is characterized as impeding others’ actions or acting in a manner that violates their autonomy and desires, provided that the social statuses and power relations between the interlocutors are unequal. For an individual to be polite, according to this rule, is to ‘avoid, mitigate, or ask permission or apologize for making the addressee A do anything which A does not want to do’, (Green, 1996:148). The second rule is to offer options, and it applies in contexts where the speakers are of similar social status yet not intimates or close friends. According to this rule, a speaker would execute his/her speech acts in a way that reflects respect for others’ autonomy and at the same time give leeway for having one’s opinions or requests denied without risking anyone’s face being threatened or lost. Lakoff’s third rule is to promote feelings of camaraderie in close relationships. Here, formal politeness is avoided because it presupposes a distance between the interlocutors, a consequence that is undesired in such contexts.

Assuming that these rules are generalizations based on observations of natural language behavior rather than prescriptive rules of etiquette, they are expected to account for the uses of *wallaahi* in the examples listed above. Following Lakoff’s characterization of politeness, *wallaahi* is a formal politeness marker, as it rarely occurs in interactions between family members or intimates, where using it would be a violation of politeness Rule 3. Lakoff’s rule 1 accounts for the use of *wallaahi* in the speech acts that involve varying degrees of imposition such as those illustrated earlier.
The speech act illustrated in (9), repeated below as (14) is particularly problematic for Lakoff’s characterization of politeness, especially to Rule 2: ‘offer options’. This particular speech event took place in a context where the social distance between the interlocutors was small, yet they were not intimates. However, the speaker chose to mitigate offering options (or avoiding imposition) by using the formal politeness marker wallaahi indicating that he assumed that offering options is an act of imposition.

(14) wallaahi inta hurr tišmil illi inta ūawzu (offering options)
    wallaahi it is up to you to do whatever you want

The contradiction between offering options and not imposing them stems from the inadequacy of Lakoff’s model to account for the speaker’s belief that offering options can, in itself, be interpreted as an act of imposition. Many speakers of Egyptian Arabic would interpret offering options as an indication that the speaker is not keen on the addressee’s well being and would place the burden of making a decision on the addressee; therefore, offering options requires mitigation. Another problem with Lakoff’s view is that asymmetrical power relations are a prerequisite for Rule 1, while all the utterances in examples (8) - (13) occurred in contexts of equal status and power. Therefore, using politeness markers is not a consequence of having power or the lack of it, but is a rational behavior that achieves speaker’s objectives.

Another view of politeness, introduced by Brown and Levinson (1987), is based on Grice’s universal account of the rationality of communicative acts and the universal notion of face. Face is defined as the individual’s claimed public self-image, which consists of two complementary aspects: negative face, which is “the basic claim to territories, personal preserves, rights to non-distraction—i.e., to freedom of action and freedom from imposition”, and positive face, which consists of “the self-image or ‘personality’ (crucially including the desire that this self-image be appreciated and approved of) claimed by interactants”, (Brown and Levinson, 1987:61). According to this view, politeness is the outcome of individuals’ mutual awareness of each other’s face needs and their ability to calculate the effects of their actions, including speech acts, on others’ face.
Brown and Levinson describe two types of politeness strategies: negative politeness and positive politeness. Negative politeness is the set of acts that are geared toward satisfying the addressee’s negative face, and it is characterized by “self-effacement, formality and restraint, with attention to H’s (the hearer or addressee) self-image, centering on his wants to be unimpeded” (Brown and Levinson, 1987:70). While negative politeness is achieved by refraining from imposition and violation of the addressee’s autonomy, positive politeness is oriented toward satisfying the addressee’s positive face needs of being approved of and appreciated, which is achieved by promoting feelings of camaraderie and mutual respect.

Despite the similarities between Lakoff’s view of politeness and that of Brown and Levinson, Green (1996) observes two crucial differences. First, Brown and Levinson attempt to characterize the factors that would make one politeness strategy more appropriate than others in certain contexts such as power differential and social distance, which is the extent to which an individual identifies with others (Hudson, Detmer & Brown, 1995). Second, Brown and Levinson predict that in contexts where one speaker has more power than the other(s), they would use different politeness strategies. The factors that characterize face needs and what constitutes politeness behavior are culturally determined, and even within the same culture a speaker would choose the politeness strategy he/she sees best suited to aid achieve his/her objectives. For example, an employer can use either positive or negative politeness strategies to refuse an employee’s request for a leave although the employer has enough power to make a bald on-record speech act of refusal. The choice between making that refusal by apologizing (negative politeness) or by praising the employee and showing how the company cannot do without him/her (positive politeness) depends on the employer’s judgments with regard to the consequences of the use of these strategies.

The view of politeness proposed by Brown and Levinson can account for the use of *wallaahi* as a politeness marker with a few reservations. The particle *wallaahi* is used preceding speech acts that potentially threaten the addressee’s negative face, which is defined culturally. For example, suggestions and offering options are culturally viewed in Egypt as involving a certain degree of imposition; therefore, speakers prefix their utterances with *wallaahi* to mitigate the impositions and to reduce the potential threat to the addressee’s negative face. On the other hand,
speech acts that involve disagreement or criticism are interpreted as to threaten the addressee’s positive face. Consequently, speakers use wallaahi to mitigate that threat. Therefore, the use of wallaahi cannot be characterized as a positive or negative politeness strategy. It is a general politeness strategy that is used to reduce the potentially undesired effects of utterances irrespective of the kind of face needs that are violated.

Another complication for the view of Brown and Levinson is that they assume that producing a face-threatening act violates only the addressee’s face needs. However, they do not recognize reflexive face, where, as in the Egyptian cultural context, threatening an addressee’s face results in threatening the speaker’s face as well, in particular his/her positive face (Mughazy, 2000). For example, when disagreeing with an interlocutor, the speaker is risking being perceived as an aggressor who violates feelings of solidarity, hence wallaahi would be used to preserve the speaker’s as well as the addressee’s face. In other words, the speaker’s face needs are not necessarily separate from those of the addressee. Therefore, the notion of reflexive face makes it difficult to establish a positive correlation between the power-differential among the interlocutors and the politeness strategies they use.

A more general view of considerate behavior or politeness is that of Green (1996), where it is defined as to consist of any behavior (including speech acts) that is intended to be interpreted as motivated by the intention to make the addressee feel good (emotionally), or to keep him/her from feeling bad. This view, although similar in the basic assumptions to that of Brown and Levinson, avoids the cultural or idiosyncratic differences in what constitutes polite use of language or the cultural values of face and power. According to this view, politeness strategies, such as the use of wallaahi, are not organized or ranked hierarchically, and the appropriateness of the use of a given strategy is a mere reflection of the speaker’s calculation of how helpful that strategy might be in achieving her goal. Therefore, Green’s view can better account for politeness as a universal phenomenon, as it leaves all intervening variables open to variation according to the culture and the speakers involved.

As a discourse particle that cannot stand as an utterance by itself, non-oath wallaahi does not communicate any independent semantic propositions. Its use invites the implicature of the speaker’s attempt to be polite by mere convention. It is an example of a ‘short-circuited’ implicature, where the addressee does not have to calculate or derive the
implicature from discourse (Morgan, 1978). In other words, *w allaahi*, when used as a politeness marker, instead of an oath phrase, has lost its religious meaning. Using it does not reveal any of the speaker’s religious beliefs in a manner similar to that of the English politeness marker *please*, that native speakers do not immediately associate with the verb ‘to please’ meaning to make someone happy.

The use of *w allaahi* as a politeness particle stems from the interlocutors’ knowledge of a cultural system of expectations known as *’asham*. This system involves a set of rights and obligations that regulate relationships between individuals who are not complete strangers or intimates. For example, an employee criticizing his coworker would prefix his utterance with *w allaahi* to remind the addressee that he is expected to accept well-intended criticism, as long as it is mitigated i.e., not intended as to offend the addressee. In addition, an employee requesting his employer to give him a raise would prefix the utterance with *w allaahi* to remind the addressee of the speaker’s expectation to have that request granted (out of generosity). Therefore, *w allaahi* is not only used to mitigate face threats but also to justify the imposition.

4. **Wallaahi and Power**

Most accounts of politeness assume a direct correlation between power and politeness in that power relations determine whether politeness is required or not and which politeness strategy is most appropriate. Brown and Levinson (1987) view power as a set of stable social valuations that can be attributed to individuals (e.g., by belonging to an advantaged social class) as well as to the roles individuals assume and are accepted by others (e.g., being an expert in some field or specialization). Moreover, they acknowledge that situational variables such as momentary weaknesses, personality traits and alliances, can change power relations or even override the stable social valuations. Adapting a Gricean perspective, they assert that speakers assess power relations in every speech event, so as to perform their speech acts in a manner that would help achieve their goals without being hindered as a consequence of misjudging others’ power.

For Brown and Levinson (1987:76), power is a crucial factor in determining the politeness strategies to be used in any context, as they measure the effect of a face-threatening speech act according to the formula: \( W_x = D(S,H) + P(H,S) + R_x \). According to this formalism the
weightiness of a face-threatening act \( W_x \) is the sum total of the values for the social distance \( D \) between the speaker \( S \) and the hearer \( H \), the power differential \( P \) of the hearer over the speaker, and the rating of imposition \( R \) according to the culture(s) of the interlocutors, assuming that these variables can be quantified on a scale of 1 to \( n \). For example, when both social distance and the rating of imposition are small and constant i.e., the interlocutors know each other and the speech act is of limited imposition, it is the power differential of the addressee over the speaker that determines the politeness strategy to be used, as in (i) and (ii) below, where (i) is most appropriate for an employee who acknowledges his/her employer’s power, whereas (ii) is most appropriate for an employer reinforcing his/her power.

i. Excuse me, Sir. Would it be all right if I smoke?

ii. Mind if I smoke?  (Brown & Levinson, 1987:80)

The effects of power relations on politeness, as described above, predict that a speaker with more power than the addressee would not use \textit{wallaahi} in contexts where the social distance and the rating of imposition are small, since \textit{wallaahi} has been characterized as a formal politeness marker. However, in examples (15a) and (15b) the two interlocutors perform face-threatening speech acts, namely a request and a refusal, and yet both use \textit{wallaahi}, despite the asymmetrical power relations. The arguments presented below aim to illustrate how \textit{wallaahi} in example (15b) is used to reduce the face threat by indicating the speaker’s lack of power- not the social or institutional power over the addressee but the power to affect events. In Brown and Levinson’s terms, \textit{wallaahi} in (15b) is used to reduce the \( P \) value to minimize the \( W \) value.

\begin{enumerate}
  \item (15a) \textit{wallaahi} kun ‘aawiz ‘agaaza.  \( ^5 \)  \\
        \textit{wallaahi} I wanted a leave.
  \item (15b) \textit{wallaahi}, ma‘adarš adeek ‘agaaza.  \( \text{refusal} \)
        \textit{wallaahi} I cannot give you a leave.
\end{enumerate}

\( ^5 \) The use of the past tense in this utterance is another way of achieving politeness, as it helps avoid interpreting the utterance as a demand. Using both \textit{wallaahi} and the past tense achieve a higher level of politeness than just one of them.
In example (15b) the speaker (an employer) has more institutional power over the addressee (an employee), yet due to the speaker’s perception of the social distance between him and the addressee, he realizes that the addressee expects that request to be granted in compliance with the social system of rights and obligation tasamm discussed earlier. The speaker’s goal is to refuse granting the request without causing the addressee loss of face, which could consequently cause him loss of reflexive face. Moreover, the speaker is aware that it is his power to grant or reject the request that is the cause of the potential face threat. Therefore, he chooses to waive that power by using the particle wallaah before performing the act, thus asserting that the decision to turn down the request was not the result of his power over the addressee but the lack of power over events and relegating that power to the unquestionable divine power of God; hence no motivation for loss of face.

Interestingly, wallaah is often used in speech acts that do not usually require politeness markers such as consoling, complaining, and delivering bad news. In these cases wallaah cannot be substituted with any of the other politeness markers such as winnabi (by the Prophet) or lamu?axza (excuse me) i.e., in these cases wallaah is not a politeness marker per se. Therefore, the uses of wallaah as a politeness marker and as a power marker can overlap, as in (15b), where politeness is achieved by declaring lack of power, or it can be used independently as a power marker, as in (16). In example (16) the speaker was a teacher telling one of her students about the results of an exam. Of course both interlocutors are aware of the fact that it is the teacher who has the power to assign grades, yet the teacher decides to avoid evoking this power and prefers to use wallaah to decline any responsibility for the results.

(16) wallaah innati-ga bta$tak mi$ kuwayvisa
   wallaah your results are not good.

The speech act in example (17) was performed by a parent complaining about her son’s slow progress at school to one of her son’s teachers. By using wallaah she is inviting the implicature that her son’s problem is not caused by her own negligence and at the same time she is blocking against the implicature that it is the teacher’s fault; it is God’s
will or fate. The lack of *wallaahi* in utterances such as those in examples (16) and (17) is usually interpreted as stressing the speaker’s power as an agent that is capable of effecting events, and hence being responsible for the consequences.

(17) *wallaahi* ilwalad miš maaši kuwayyis fi lmadrasa
Wallaahi the boy is not doing well at school.

Ng and Bradac (1993:7) discuss cases where speakers deny their own power if they think that declaring their possession of such power impedes achieving their goals, and they refer to this process as ‘depoliticizing’ power, which they define as to “camouflage it [an influence message] as something else, and by doing so, communicators render their influence attempts more palatable to the targets of influence and at the same time lessen their accountability”. Therefore, *wallaahi* is one conventional linguistic tool speakers of Egyptian Arabic use to ‘depoliticize’ the messages of their potentially face-threatening acts, where the mentioning of God reminds the addressee(s) that individuals do not have full control over events. It is necessary to point out that in examples (15b) and (16) the institutional power of one speaker over the other is not denied, and that the use of *wallaahi* would not affect the valuation of that stable type of power differential.

The uses of *wallaahi* discussed above support Fairclough’s claim that there are two distinct types of power relations: ‘the power over’, which is the power individuals have over each other due to their positions in the social hierarchy, and ‘the power to’, which is the power individuals have over events in a given context (Fairclough, 1989). Of course the two types can overlap, because an individual’s power over events can (but not always) be a consequence of one’s social power, but in certain contexts speaker intentionally suspend or deny one type of power and maintain the other, especially since there are the linguistic tools such as *wallaahi* that can communicate and maintain the distinction.

5. **Conclusion**

The arguments presented in this paper support the claim that discourse particles such as non-oath *wallaahi* are not meaningless speech fillers that have no patterns of distribution. The use of *wallaahi*, both as a
politeness marker and as a power marker, is determined by speakers' beliefs about the effects of their utterances, and whether these effects help or hinder achieving their goals. The choice to use *wallaahi* is governed by the social system of expectations that dictates when politeness is required and when expressing power is not desired. In other words, the use of *wallaahi* is an example of how individuals use language to satisfy or manipulate the requirements of cultural norms.

**REFERENCES**


1. Introduction

Egypt is considered a diglossic community. According to Ferguson (1959), diglossia is the existence of two varieties H (high) and L (low) each with a different function. H is the more respected, “highly codified variety”, while L is used in conversations, and informal situations. Diglossia in Egypt has been examined by quite a number of linguists (cf. Badawi, 1973; Mitchell, 1986; Mejdell, 1996). Linguists concluded that speakers can alternate between their use of H and L in the same stretch of discourse, but apart from very few studies (like Eid, 1982), no one has examined the syntactic rules that govern this switching between H and L. Moreover, there are even fewer studies that attempt to explain syntactic constraints on diglossic switching as part of code switching. Code switching has been defined by Gumperz as “the juxtaposition within the same speech exchange of the passages of speech belonging to two different grammatical systems or subsystems” (1982:59). Therefore code switching is not restricted to switching between different languages (Myers-Scotton, 1998b:18; Mejdell 1999:226). The need for studies that examine diglossia within the framework of code switching is urgent. Walters (1996:181) for example recognises that the mixing between H and L that occurs in the Arab world can be studied in relation to theories of code switching, especially in relation to Myers-Scotton’s theory of Matrix language.

“Work that considers the nature of diglossic switching in the light of Myers-Scotton’s model of the grammatical structure of code switching would also be welcomed (...) Arabic will provide a fertile ground for testing and refining her
model even as her model provides the most comprehensive framework to date for analysing code switching as a formal problem." (1996:193)

In this paper I examine the occurrence of three morphosyntactic variables, the negation system in ECA (Egyptian colloquial Arabic) and MSA (modern standard Arabic), the system of deixis in ECA and MSA, and the ECA aspeutal marker b- in contemporary monologues in Egypt in relation to code switching. I try to answer the following two questions:

- What are the syntactic and morphological possibilities of mixing between MSA and ECA in relation to these variables?
- How do these possibilities contribute to the study of code switching?

I will first briefly discuss three theories that try to identify constraints on code switching, with special reference to Myers-Scotton’s theory of a matrix language. I will then describe my data and explain why I have chosen to analyse these variables. Then I will discuss my findings and how these findings challenge some theories of code switching. These theories were found lacking, although a modified version of the matrix language theory may help explain the data.

2. **Theories on Code Switching**

In this section I will examine the following theories introduced by linguists to identify structural constraints on code switching:

- The two constraints theory
- The government principle
- The matrix language hypothesis

I will try to explain each theory and the problems associated with it. However, I will especially focus on the matrix language hypothesis, as this is the theory I will adopt in my analysis of the three variables.

2.1 The Two Constraints Theory

This theory is one of the first attempts to identify syntactic constraints on code switching. It was put forward by Sankoff and Poplack (1980). They propose that there are two factors important in
code switching; the free morpheme constraint, and the equivalence constraint. The free morpheme constraint predicts that there cannot be a switch between a bound morpheme and a lexical form, unless the lexical form is phonologically integrated into the language of the morpheme. They depend on their analysis on Spanish/English code switching. This constraint would predict that, *flipeando*, meaning ‘flipping’ is a possible form (1980:565). In this example, -eando is the Spanish progressive suffix. The lexical English form (flip) has been integrated into the phonology of Spanish. But the form *catcheando*, meaning catching, is not possible, because the lexical form catch has not been integrated into the phonology of Spanish. As a result, it cannot take the Spanish progressive suffix-eando.

The equivalence constraint theory states that code switching tends to occur at points where juxtaposition of elements from the two languages does not violate a syntactic rule of either language. It will occur at points were the surface structure of the two languages is the same. Sankoff and Poplack (1980:585), tried to apply this theory to Spanish, and they found that Spanish/English code switching may occur between determiners and nouns but not between nouns and adjectives in the noun phrase. Thus, it would be unacceptable to say:

(1) *his favorito spot
    ‘his favourite spot’

This example cannot occur because it will violate the surface structure of Spanish. In Spanish the adjective must come after the noun, whereas in English it comes before. They admit, however, that sometimes switching does occur, where there is no structural equivalence between the languages. But if this happens it is always accompanied by omissions, repetitions... etc.

2.1.1 Problems with this Theory

The two constraint theory can be applied very neatly to code switching between Spanish and English, because both languages, more or less, share the same word order, and they both have the same categories, i.e. noun, verb, definite article, adjective... etc. However, it would be very difficult to apply the same theory to two languages that do not share the same categories, like Tok Pisin and English (Romaine,
1995:129). It is hard to assign categories to a language like Tok Pisin (verb, article, negative particle... etc). Moreover, Berk-Seligson (1986:328) found that in Hebrew/Spanish code switching, many ungrammatical utterances occur, like the omission of the definite or indefinite article, because the indefinite article does not exist as a grammatical category in Hebrew. Besides, this theory relies on linear order and adjacency (surface structure of the sentence) and not on hierarchical order as the theory of government for example (Romaine, 1995: 129).

Note the following example from my data that cannot be explained by the two constraints theory.

(2) aal lak il maal da S1
‘He told you this money.’

This example shows that the equivalence constraint does not work either, since the surface structure of MSA and ECA is not the same in this example, and still switching occurs. In MSA the demonstrative would come before the noun not after. However, there is still switching between the noun, which is in MSA, and the demonstrative, which is in ECA.

2.2 The Government Principle

Di Sciullo, Muysken and Singh (1986) postulate that code switching is universally constrained by the government relation holding between sentence constituents. Based on their data, they found that it is possible to switch between verbs and subjects but not between verbs and objects. So they tried to identify the constraints that lead to this. They define government in the following way: X governs Y if the first node dominating X also dominates Y, where X is a major category N(noun), V(verb), A(adjective), P(preposition). They argue that “if X has language index q and if it governs Y, Y must have language index q also” (1986:5). Their main claim is that within a maximal projection no switch is allowed, i.e. within a verb phrase or a noun phrase for example, no switching is allowed. According to them, code switching occurs only between elements that are not related by government. This explains why there are no switches between a verb and an object in their data.
2.2.1 Problems with this Theory

The government principle is more promising than the equivalence constraint, because it allows switching between languages with different word order and different distribution of categories. It also assumes that switching depends on hierarchical structure rather than linear structure, and can thus take account of a wider range of languages. However, this theory still has its pitfalls and cannot really be applied universally, at least for the time being.

Bentahali and Davies (1983) found instances of switching between an Arabic pronoun and a French verb accompanied by its clitic; this violates the government and binding constraint. Moreover, according to the government principle, there cannot be a switch within a prepositional phrase, or between a verb phrase and a noun phrase. In the following example quoted from Romaine (1989:130), of Panjabi/English code switching, an English noun is governed by a Panjabi postposition. This is not allowed according to the government principle.

(3) Family de nal.
   ‘In the family.’

In this example, there is a switch within the preposition phrase, which is a maximal projection. DiScuillo et al (1986) would have predicted that there could be no switch in this case, as the noun and the preposition should have the same language index.

Note the following example from my data, which provides a counter argument for the government principle.

(4) il itene il qanaash dool kaanu... L1
   ‘These two counsels were... ’

In this example the speaker switches within the noun phrase, between the demonstrative and the noun, which should not be allowed according to the government principle.

2.3 Myers-Scotton’s Matrix Language Hypothesis

Myers-Scotton (1998a) states that a matrix language (ML) is defined by ‘system morphemes.’ There are two kinds of morphemes, based on the lexical feature of plus or minus ‘thematic roles.’ Content
morphemes assign or receive thematic roles, like agent, experiencer, beneficiary...etc. This category includes nouns, descriptive adjectives and most verb stems. System morphemes, on the other hand, cannot assign or receive thematic roles. This category includes inflections, determiners, possessive adjectives and intensifier adverbs. Thus, an ML supplies system morphemes, which are syntactically relevant. The embedded language supplies only content morphemes. Her main hypothesis is that “languages can sustain structural incursion and remain robust, but the taking in of alien inflections and function words is often a step leading to language attrition and language death” (1998a:289).

According to Myers-Scotton, when two languages/varieties are brought together by a bilingual/monolingual, there is a dominant language at work. Thus, one language should be assigned the status of what she terms ‘a matrix language.’ The matrix language supplies the grammatical frame of constituents, while both languages supply morphemes. The main hypothesis of Myers-Scotton is that there is always an ML in code switched discourse, and there is always only one ML at a time. Thus, to analyse a structure, one must first recognise the ML.

Although this hypothesis still has pitfalls, it is in some ways the most promising because it does not rely on a specific theory of grammar as do the government and binding theories; this means that the ML hypothesis will not be influenced by whatever theory of grammatical structure is adopted. This, in fact, is a significant point in light of the fact that syntacticians are now questioning the government and binding theory and are opting for a more lexically based theory. More importantly, Myers-Scotton does not claim that code switching should be studied with reference to structural constraints only. In fact, she emphasizes the intertwined role of the discourse and structural function of code switching. She tries, in an unprecedented manner, to combine a theory that can explain both the discourse function of code switching and the structural constraints on code switching. Myers-Scotton thinks that by having a matrix language, linguists can then start addressing the question more clearly of why people switch between languages, which is still not entirely clear.

The following example quoted from Myers-Scotton (1997:79) will show how the ML theory works in practice.
Swahili/ English code switching:

(5) (leo) [si-ku-come na books z-angu]
    ‘(Today) I didn’t come with my books’

According to Myers-Scotton, this sentence shows not only intra-sentential code switching, but also intra-word code switching (code switching within a word). I will now explain the example in more detail:

(sì ku-come)                    na books z-angu
one word(which is a verbal phrase) na, preposition
the verbal phrase is:          books, noun, English
an ML +EL constituent          z- angu, possessive pronoun
Si, pronoun, first person singular
Ku, tense marker denoting the past, and denoting negation
Come, verb English

Note that in this example, the system morphemes, (pronouns, tense markers, negative markers and prepositions) are all in Swahili, while the content morphemes (verbs and nouns) are in English. Therefore, the ML is Swahili, and English is the EL.

3. Data Analyzed

My data consists of approximately 9050 words from spoken contemporary Egyptian monologues, four political speeches (P), four mosque sermons (S) and one university lecture (L). These are all contexts where one would expect a predominance of MSA, whereas there are many other code switching occasions in Egypt where this is not the case.

First, I would like to point out that MSA and ECA share a large amount of vocabulary, as well as phonological, syntactic and morphological variables. Hence it is quite difficult at times to make a clear distinction between them. Clyne (1987:754) when working on Dutch, German and English also found it difficult to differentiate between two closely linked languages like Dutch and English. Note the following example given by Clyne:
(6) *Meestal hier at the local shops and in Doncaster.*

"Mostly here at the local shops and in Doncaster."

The preposition ‘in’ is the same in English, German and Dutch and so is the word ‘here’ (hier, here).

To illustrate the point further, it is sometimes difficult to differentiate two languages or two varieties used together in the same stretch of discourse. That is to say, while it is possible for example to distinguish ECA from MSA when they are used separately, it is difficult to differentiate between both if they are used together in the same stretch of discourse, and the speaker mixes between them all the time. Note that Parkinson (1996:99) for example, thinks that ECA and MSA are both “parts of a single expressive system.”

Although one should strive to study the structural differences between utterances, one should still consider those utterances within a larger framework. Every stretch of discourse is unique in itself as well as being a unified whole. Therefore, one has to make a decision for the utterance as a whole as to what language it is in.

I try to classify a portion of discourse as MSA, ECA, a mixture between both more geared towards MSA, or a mixture of both more geared towards ECA depending on the quantity of MSA or ECA features in the text. The nature of the text (whether it is more geared for example towards MSA or ECA) is also relevant when analysing how switching works.

4. **Variable Criteria**

Negation, demonstratives, and the b-prefix are three useful switch-identifying variables because they are different in MSA and ECA. Linguists have tended to study these variables separately (cf. Rammuny (1978) negation, El Hassan (1980) demonstratives, Mitchell & El Hassan (1994) verbal aspect) rather than in concert. These variables will also bring into question some ideas about syntactic constraints on code switching. I will examine these variables in relation to Myers-Scotton’s idea of a matrix language (1997). I categorize them as system morphemes because they do not assign thematic roles.

5. **Variable Combinations and Code Switching**

In my data there are examples of mixed forms consisting of:

- A negative marker in one variety and a verb in another.
• A demonstrative marker in one variety and a noun in another.
• An aspectual marker (the b-prefix) in one variety and a verb in another.

All the mixed forms consist of:
• ECA negative marker + MSA-like verbs
• ECA demonstrative marker + MSA-like nouns
• An ECA aspectual marker (b-prefix) + MSA-like verbs

There are no examples of:
• MSA negative marker + ECA verbs
• MSA demonstrative marker + ECA nouns

According to the matrix language hypothesis, if the three markers I discuss in my data, which are system morphemes, are realised in ECA, then the ML is ECA, and the EL is MSA. The basis of mixing is an ECA syntactic sub-structure, into which MSA lexical elements are inserted. Examples in my data tend to support this view. Note the following examples:

**Negative particles:**
This example comes from sermon one which is geared towards ECA.

(7) il ʔimaan miš kaafi S1

‘That is why believing by itself is not enough’

**System morphemes:**

- il  definite article ECA
- miš  negative particle ECA

**Content morphemes:**

- ʔimaan  noun (belief) MSA
- kaafi  adjective (enough) MSA

In this example, one ML is at work. The negative marker, and the definite article, which are considered by Myers-Scotton to be system morphemes are in ECA, while the content morphemes, which assign thematic roles, like nouns and adjectives, are in MSA. Therefore, the
ML in this example is ECA. ECA provides the sub-structure of the sentence, while the speaker fills in the lexical gaps with MSA items.

**Demonstratives:**
Sermon two is geared towards MSA.

(8) il ʿaqli da madda S2

‘That is to say, this mind is a substance that produces thinking.’

**System morphemes:**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>il</td>
<td>definite article ECA</td>
</tr>
<tr>
<td>da</td>
<td>demonstrative marker ECA</td>
</tr>
</tbody>
</table>

**Content morphemes:**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ʿaql</td>
<td>noun (mind) MSA</td>
</tr>
<tr>
<td>madda</td>
<td>noun (substance) MSA</td>
</tr>
</tbody>
</table>

Again in this example the morphemes that do not assign theta roles, like the demonstrative, and the definite article are in ECA, while the morphemes that assign theta roles like nouns are in MSA. Therefore, the ML in this example is ECA.

**The b-prefix:**
This example is from lecture one which is more geared towards ECA.

(9) illi bitaqaʿi ʿalaa hood il baḥr il mutawassit L1

‘That is to say, all the countries that lie on the Mediterranean Sea.’

**System morphemes:**

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>illi</td>
<td>relative marker ECA</td>
</tr>
<tr>
<td>bi</td>
<td>aspectual marker ECA</td>
</tr>
<tr>
<td>ʿalaa</td>
<td>preposition ECA and MSA</td>
</tr>
<tr>
<td>il</td>
<td>definite article ECA or MSA since it is not clear whether the vowel preceding the article is the case ending of the preceding word, or the ECA realisation of the definite article.</td>
</tr>
<tr>
<td>il</td>
<td>definite article ECA or MSA (same problem as above).</td>
</tr>
</tbody>
</table>
The system morphemes in this example, like the aspectual marker `b-`, occur in ECA rather than MSA, while the content morphemes like nouns, verbs and adjectives are from both varieties ECA and MSA.

In the above examples, the system morphemes are realised in ECA, while the content morphemes are realised in MSA. Apparently, then, one may conclude that the ML in the Egyptian community is ECA, and the EL is MSA. This neat distinction between system morphemes and content morphemes does not always seem to work, however there are a number of examples in my data that pose problems for this hypothesis, and these examples suggest that the situation in the Egyptian community is more complicated. Therefore, I have to abandon the idea of an ML in favour of a more sophisticated framework that can explain more precisely what takes place in the Egyptian community, since so many of the examples are less clear-cut.

6. Problems with the ML Hypothesis

The following examples pose problems for the ML hypothesis; this example is from political speech one part two which is more geared towards MSA.

(10) kaan fii ?ittifaat bitunaffad P 1,2
‘And agreements were being applied’

System morphemes:

- `bi` aspectual marker ECA
- `u` and `a` in the verb `tunaffad`, denote the MSA passive form of the verb (discontinuous passive morpheme).
If one accept the ML-EL hypothesis, then in this example, it is impossible to decide which is the ML (cf. Bassiouney 1998). As the aspectual marker of the verb is in ECA, one might expect the other ML variables to be ECA also, as was the case in the past examples. In this example, however, there is not only one ML at work. In fact, there is one discontinuous system morpheme taken from MSA, and two taken from ECA. The speaker uses the existential *fii*: which is ECA, and also uses the b-prefix, which is an ECA feature. However, he also uses the discontinuous passive morpheme u-a, which is quintessentially MSA. This ‘internal’ form of the passive is not available for him in ECA, in the sense that it cannot be applied to ECA passive verbs, which use the *it*-prefix. To clarify this point further, I will give the ECA and MSA counterparts of this example:

ECA counterpart:

(11) kaan fii ?ittifaqaat bititnaffiz

‘And agreements were being applied’

MSA counterpart:

(12) wa kaanat humaaka ?ittifaqaat tunaffad

‘And agreements were being applied’

Existence is expressed in ECA by a morpheme, which is etymologically a locative adverb ‘*fii*’, which has no morpho-semantic equivalent in MSA. In MSA it is *yuujad* (a passive verb) or ‘*humaaka*’ (an adverb). This means that there is no possibility of mixing between the two systems to express existence; the two systems do not overlap at this point of morpho-syntactic structure. The ECA form has been selected. Similarly, as far as the verb (being applied) is concerned, the passive form is expressed in MSA by structures that have no morpho-semantic equivalent in ECA. Here the MSA form has been selected. If
this is the case, one cannot say, in a sentence such as the one quoted, that the ML is ECA. The speaker obviously knows specific morpho-syntactic forms of both ECA and MSA, and uses them in this example. The example is particularly interesting, because it has intra-word code switching: in the same word we have an aspectual morpheme from ECA, and a passive morpheme from MSA.

Other examples, which cause problems for the ML hypothesis, are as follows:

(13) haaða k kalaam laysa kaafiyan P2
'I think this is not enough'

This example is from political speech two which is a mixture of MSA and ECA.

**System morphemes:**
- haaða: demonstrative MSA
- k: definite article ECA
- laysa: negative marker MSA
- -an: case marker MSA

**Content morphemes:**
- kalaam: noun (talk) ECA or MSA
- kaafii: adjective (enough) MSA

In this example one finds system morphemes from both MSA and ECA in the same sentence, posing a problem for the ML hypothesis.

(14) ya’ni biyib’ā fiīh ḫīmaaya fiī haaðihi l manaaṭiq L1
'that is to say, there is protection in these regions'

This example is from lecture one, which is more geared towards ECA.

**System morphemes:**
- b: prefix ECA aspectual marker
- fi: preposition ECA or MSA
- haaðihi: demonstrative MSA
- l: definite article MSA or ECA
In this example, the content morphemes that occur are from both varieties. This is expected by the ML hypothesis. The problem is that there are also system morphemes from both varieties as well as system morphemes that are shared by both varieties.

(15) wa yurfa‘i ū‘an t takliif da S2
‘And this responsibility is lifted off his shoulders if he is not sane’

This example is from sermon two, which is geared towards MSA.

In this example, the speaker uses only MSA content morphemes and mostly MSA system morphemes. The negative marker is in MSA. He also uses MSA pronouns and case markers. However, in the same clause there is an MSA negative marker and an ECA demonstrative. The speaker thus uses system morphemes from both varieties. This is a good counter example, with all evidence showing that MSA is the ML of the example, but with the embedded form being a system morpheme, which contradicts the hypothesis.
This issue was presented at the Islamic Conference.

This example is from political speech one, part two, which is a mixture of both MSA and ECA.

**System morphemes:**
- **il** definite article ECA or MSA
- **da** demonstrative ECA
- **u-i** discontinuous passive morpheme MSA
- **il** indefinite article MSA or ECA
- **'alaa** preposition (at, on) MSA or ECA
- **l** definite article MSA or ECA
- **il** definite article MSA or ECA

**Content morphemes:**
- **mawdu‘i** noun (issue) ECA or MSA
- **kaan** verb (to be) ECA or MSA
- **surid** verb (to be presented) MSA
- **mu'tamar** noun (conference) ECA or MSA
- **islaamii** adjective (Islamic) ECA or MSA

In this last example, the speaker uses an ECA system morpheme, the demonstrative da, as well as an MSA system morpheme, the discontinuous passive morpheme u-i.

These examples resist interpretation within the framework of diglossia, and challenge ideas about syntactic constraints on code switching. First, as was said before, linguists studying diglossia have not explained cases of mixing of this sort, and have tended to study ECA and MSA variables separately. For example, although Mitchell (1986:9) acknowledges the existence of a mixed variety that he calls “educated spoken Arabic”, he does not explain rules or patterns that govern it.

The two constraints theory (Sankoff and Poplack 1980) and the government principle (Di Sciullo et al. 1986) and the ML hypothesis all fall short of explaining the above examples. For example, the two-constraint theory and the government principle theory would predict that there can be no switching between an aspectual marker like the b-
prefix and a verb, which is not the case. The ML hypothesis predicts that all the system morphemes will be in the ML, but in these examples we see examples of system morphemes from both varieties being used together in the same word, which would indicate that both ECA and MSA are MLs according to Myers-Scotton’s theory. There is an intuitive sense in which the matrix language hypothesis seems to be insightful for some of the data. However, there are other examples for which the hypothesis simply does not seem to account for the data, with system morphemes in single constituents or even single words coming from both MSA and ECA.

The idea of a basic sub-structure (an ML) falls short of explaining the last examples for a number of reasons. First, the ML hypothesis is based on the idea that there is always only one ML at work at a time and that once there are competing MLs, it is an indicator of language death. As was shown in all the examples above, however, sometimes there seems to be more than one ML at work. Although the theory of ML is indeed insightful in certain cases, it does not seem to give a clear-cut picture of what goes on in this diglossic community.

Another problem with applying this theory is the fact that in certain examples it is very difficult to decide whether a certain morpheme belongs to ECA or MSA. The MSA definite article ‘(v)’, where, v stands for vowel, has no vowel of its own, the vowel quality of the article changes according to its position in the sentence (Holes, 1995). Thus, it is sometimes difficult to classify it as MSA or ECA. Moreover, some prepositions (which are system morphemes) are shared by both varieties.

Because ECA and MSA are different varieties sharing a large amount of content and system morphemes, it is almost impossible sometimes to say whether a certain morpheme belongs to ECA or to MSA. This creates problems for an analysis that needs every morpheme clearly identified as belonging to one variety or the other as the ML theory does. The ML theory seems to work much better with very different languages, i.e. languages with different morpheme systems, like Arabic and English, or Turkish and French, rather than German and Dutch or two varieties of Arabic like ECA and MSA.
7. The Composite ML Hypothesis

Myers-Scotton (1998a) has recently, however, proposed another idea to explain some cases of code switching. This idea is the idea of a composite matrix language (1998a:289). She posits that when there is a change in progress in a bilingual community, the ML is a composite, based on structures from both languages, but moving towards the new ML. Usually the change in progress is completed, and a language shift follows, with the new language playing the role of an ML.

Myers-Scotton (1998a:310) supports her hypothesis by giving evidence from other linguistic studies on code switching, like the study of Bolonyai (1996).

Bolonyai gives an example of a young Hungarian girl who lives in the USA with her parents, and how her ML changes from Hungarian, to a composite of Hungarian and English, and finally the new ML becomes English. The following example demonstrates the use of a composite ML.

(17) Mom, en meg-find-t-am a ket quarters-t
    Mom 1sg perf-find-past-1sg det two quarters-acc
    ‘Mom, I found two quarters’ (Bolonyai 1996:12)

The word ‘quarters-t’ is an English word with an English system morpheme (s), and a Hungarian system morpheme (t), which is a Hungarian case marker. In this example, a composite ML is leading to language change.

The idea of a composite ML can help explain what is going on in the Egyptian speech community. The problem with this hypothesis is that it is just not possible to claim that there is change in progress in Egypt, language shift, or language death. It could simply be a stable system that arises from a diglossic language situation with close cognate forms.

All the monologues analyzed have common characteristics: the speakers are all educated, and have thus been exposed to MSA. They are all over 40, that is to say, their exposure to MSA has been over a long period of time. I think there are certain system morphemes from both MSA and ECA, which they have mastered. It is quite difficult to say that these speakers have an ECA sub-structure only (an ECA ML). The ECA sub-structure may be dominant but they still make use of
certain MSA system morphemes. Their ML is not just ECA, their ML is a composite and depending on the text ECA or MSA may dominate.

The fact that their ML is a composite does not necessarily mean that one variety is changing or dying, and it does not mean that one variety will take the place of another. There are two kinds of Arabic in Arab countries, as Ferguson claims (1959), a high form, in our case, MSA and a low form in our case ECA. The two varieties have survived in Egypt for different reasons (Bassiouney, 2000). For example, although MSA is the formal written language rather than the spoken one (Ferguson, 1959), it has strong appeal in Egypt. This is because MSA is considered both the language of religion and the way to understanding the Quran as well as the language of formal literature, classical films, songs and soap operas in Egypt. Singers still sometimes use MSA in their songs. All religious and historical films are in MSA. Children are exposed to MSA in their first year in school. MSA, is part of the average educated Egyptian’s every day existence. ECA however, is different from other vernaculars in the Arab world. It is considered a kind of a ‘semi-standard’ in the Arab world (Holes, 1995:5) due to the fact that it is the language of popular culture and art in the Arab world. Most Egyptian films are shown in other Arab countries, many Egyptian singers, actors and poets are quite famous in other Arab countries. Therefore, the attitude of Egyptians towards both varieties MSA and ECA is positive. This positive attitude makes one doubt the idea that one of these languages, ECA or MSA, is going to replace the other. Even the idea that a new composite of both is going to replace any of them is difficult to prove. However, I think the question of permanent language change is beyond the scope of this work.

8. Conclusion

In summary, I have discussed the conclusions that I reached concerning code switching in light of three grammatical variables. I examined Myers-Scotton’s ML theory with reference to my data, and then discussed the problems that the examples pose for her theory. I found that there is an asymmetry in the way the three system morphemes are used in my data, because it is possible to have an ECA negative marker followed by an MSA verb-form but it is not possible to have an MSA negative marker followed by an ECA verb form. In
addition, while it is possible to have an ECA demonstrative followed by an MSA noun, it is not possible to have an MSA demonstrative followed by an ECA noun. The b-prefix can and does occur with MSA verb-forms. Speakers in Egypt use a composite ML. I think Myers-Scotton’s idea of a composite ML is very useful in explaining what goes on exactly in the Egyptian community. Unlike others (Sankoff & Poplack, 1980; Di Scuillo et al., 1986...etc.), she proposed an explanation that is at least partially consistent with the Egyptian data, although it may need to be further expanded so as to account for the Egyptian community.

For now, the composite ML seems to occur in transitional phases according to Myers-Scotton. Whether other “Arabic-speaking” diglossic communities also use a composite ML is a question worth examining. Studies that try to account for and describe more syntactic variables in actual data in the light of recent code switching theories are needed, as well as studies that compare the attitudes of different diglossic communities towards their vernacular, and how this attitude affects language change.

Meanwhile, the question of why people use both varieties together in the same stretch of discourse has not been addressed. The reasons why people switch between different varieties may help us understand better how this switch occurs (cf. Holes, 1993). Studies that try to account for and describe more syntactic variables in actual data are needed, as well as studies that compare the attitudes of different diglossic communities towards their vernacular, and how this attitude affects language change.

This paper is to my knowledge the first attempt to apply recent theories of code switching to the Egyptian community. In this paper I also try to improve on Myers-Scotton’s theory of an ML and expand it so as to account for the situation in Egypt.
REFERENCES


1. Introduction

This paper analyzes the speech action expressions (SAX) in a part of the sûrat Al-Baqarah, and the problems arising when translating them into structurally completely different languages such as German and English. Specifically, this preliminary study is a close look at two German translations of some speech act verbs like qaala in two verses of the sûrat Al-Baqara from the Qur’an.

1.1 Speech Action Expressions (SAX)

In order to account for the differences between the Quranic original and the English and, especially, German translations, we have to briefly deal with the speech act characteristics of the verbs under consideration. Only the background and theoretical statements pertaining to the present study are introduced here. Traditional speech act theory (SAT) (most recently Allan 1998a, b) deals with utterances which constitute actions by their virtue of being uttered, e.g. to explain, or, to apologize. Its main contention is that human beings usually do not just utter something, but actually “do things with words” (Austin, 1962). Speech act verbs “denote actions usually performed by speaking (to tell, to describe)” (Reinelt, 1996: 924), including report verbs, i.e. “main verbs that report speech, thought and physical perception” (Klamer, 2000: 71. To account for the fact that not all languages have a clear-cut distinction between verbs and other syntactic (and other) categories (e.g. in English: saying that...), we also have to include derived forms and compounds, as long as they are within the same meaning range, all summarily called “speech action expressions” (for recent discussions see Reinelt, 2000; Klamer, 2000; Harras, 2001).
It is now interesting to see how speech action expressions and related phenomena in a fixed text, such as the Qur’an, are translated into English or German, where the resulting text is not fixed. Our aim is to find out which choices are made in the translations of the two verses under consideration, and why this is the case.

1.2 Al-Baqarah

Al-Baqarah, being neither completely dialogic nor only descriptive, contains a sufficient number of occurrences of speech action expressions. This guarantees at least some representativeness. We found well over 160 relevant occurrences of qaala and its forms. In total, more than 100 occurrences of about 50 other SAX altogether, such as shakara, kalama, etc. hint at the wide variety of SAXs in the Baqarah original.

In the following, we will concentrate on qaala translations, because they can serve to demonstrate some of the approaches most clearly. In verses 67 and 68, discussed below, the two most important ones appear close together.

2. SAX in Translations: The Treatment of Qaala Forms

For a cross-linguistic analysis, we compare the SAX originals in verses 67 and 68 with their translations into German and, in part, English. The following is the original.

وَذَٰلِكَ مَا ۖ وَقَالُوا أَوْحَيْنَا إِلَيْهِ الْعَذَابَ فَأَيَّامُ الْعَذَابِ عَظِيمٌ

ادعِ لَنَا رَبَّكَ بِيَوْمِ مُجَابَهَةٍ وَلَا نَضْرَبْنَا قِسْطًا بِنِسْبَةٍ لَّا فَارِضٌ وَلَا يَكْرُرُ عَذَابُهُ بِبِكَارٍ

قالوا فَاعْلَوْا مَا تَوَارَىْنَ

English Transliteration:
(bracketed words discussed below)
While the Quranic original is beyond discussion, translations can vary considerably. As an introduction to the discussion, compare the 6 English translations (see Appendix for a list) of the SAX in 67 and 68 (Arabic original in brackets):

<table>
<thead>
<tr>
<th>Verse</th>
<th>Speech Action Expression</th>
<th>List of English Translations</th>
</tr>
</thead>
<tbody>
<tr>
<td>67</td>
<td>1 qala</td>
<td>said to (4)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>told his folk (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>said unto (1)</td>
</tr>
<tr>
<td></td>
<td>2 ya?murukum (?amara=to order)</td>
<td>commands (6)</td>
</tr>
<tr>
<td></td>
<td>3 qaaloo</td>
<td>they said (6)</td>
</tr>
<tr>
<td></td>
<td>4 ?atattaxi?unaa</td>
<td>makest thou a laughing-stock of us (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>do you take us for a jest (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>do you make fun of us (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>dost thou make game of us (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>do you ridicule us (1)</td>
</tr>
<tr>
<td></td>
<td>5 qala</td>
<td>he said (5)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>he answered (1)</td>
</tr>
<tr>
<td>68</td>
<td>6 qaaluu</td>
<td>they said (6)</td>
</tr>
<tr>
<td></td>
<td>7 ?id'uu (da?a=to ask)</td>
<td>beseech on (1)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>call on (2)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>appeal to (1)</td>
</tr>
</tbody>
</table>
The English translations already show a great variety, which I will here leave to the reader to systematize.

There are, among others, two German translations available, which represent two extremes, one by Mirza for Ahmadiyya (n.d.), and one by Henning for Reclam (1976). We take the case of qaala, for which there are at least 3 words in German (in the following part SAX appear in capital letters): SPRECHEN (to speak?), SAGEN (to say?), and REDEN (to speak informally?), and others (ERKLAREN to explain, GEBOTEN ordered, BITTEN to ask, DEUTLICH MACHEN to explain, ANTWORTEN to answer, to reply) as required by e.g. situational circumstances.

1.1 A Unitary Approach
Reclam (1976) takes a unitary approach by translating all occurrences of qaala with forms from SPRECHEN (to speak with/to):

63. Und als Moses zu seinem Volk SPRACH: “Siehe, Gott gebietet euch eine Kuh zu opfern”, SPRACHEN sie: “Trebst du Spott mit uns?” Er SPRACH:
"Da sei Gott vor, dass ich einer der Toren ware." ... Sie SPRACHEN: "BITTE deinen Herrn fuer uns, uns zu ERKLAEREN, was es fuer eine Kuh sein soll." Er SPRACH: "Siehe, er SPRICHT, es sei eine Kuh, weder alt noch ein Kalb; in mittlerem Alter zwischen beiden; und nun tut, was euch GEBOTEN ist." (Reclam 1966, p. 33)

In this translation, qaala (disregarding the 3rd person, number and stem marking for a second) appears as a grammaticalized quotation marker, whose function it is to simply indicate a change in speaker or spoken parts without hinting at or explicitly indicating any (discoursal or other) function of the part said (Klamer, 2000). A respective German translation would usually be (er) sagte.

On a personal note, this translation is not easy to understand, and SPRECHEN becomes a boring albeit heavy expression (for conveying contents), if all too abundantly used for native speakers of German.

1.2 Differentiating

Mirza (n.d.) takes a different approach. He differentiates qaala translations according to their functions in their respective situations. Differences between the various occurrences can be found on two levels: On the lexical level, differences in meaning realms account for overlapping and differentiation in the translations. On the discourse level, considerations of how the instructions are presented lead to additional differentiations of SAX in the German translations.

Mirza’s translations are as follows:


68. Sie SPRACHEN: “BITTE deinen Herrn fuer uns, dass er uns DEUTLICH MACHE, was sie ist.” Er ANTWORTETE: “Er SPRICHT, es ist eine Kuh, weder alt noch jung, voll erwachsen, zwischen beidem, nun tut, wie euch GEBOTEN.” (Mirza, p. 13)

1.2.1 The Lexical Level

In 68, Mirza uses SPRECHEN for words by Moses, and SAGEN for utterances by others in this situation. While this differentiation is not followed strictly throughout the whole sura, it shows how
overlapping realms of meaning and ranges of words with essentially the same or very similar meanings can be exploited (locally or throughout) to differentiate speakers or the way something is said, resulting in at least three alternatives for qaala: SPRECHEN (to speak) SAGEN (to say) and REDEN (to speak informally).

For the reader, this makes the translation more vivid, and attributes, at least locally, additional features such as whether someone makes the respective utterances considerately or not.

1.2.2 The Discourse Level

In 69, the translation is slightly different. While the SPRECHEN vs. SAGEN distinction is not kept up, the translation ANTWORTE (to answer) for qaala lets the situation appear as dialogic. Of course, answering is quite a different speech action from just saying something (which may even be monologic). By choosing this SAX, the discourse level is introduced (or at least foregrounded). In it, the reader is shown how the contents of Allah’s words in Moses’ words is given as the dialogic reaction of an involved party following a preceding utterance that deserves or at least entices an appropriate linguistic reaction such as an answer.

Not surprisingly, this part resembles real-live situations, and through its vividness enhances the translation considerably. This approach also opens the possibility for further differentiations such as reply, etc. (see the various English translations for da’a in 68).

3. Conclusions

Basically, there are two kinds of translations. Hoque (n.d.) describes them as “academic” and “informative” styles. He states:

Not all translation of the Qur’an is the same. Not all translations match the Arabic Qur’an in rhythm of text, depth of meaning, syntax of sentences, word usage and adherence to the divine words. While some translations are truly loyal to actual Arabic words, some are liberal in bringing out the meaning of the verses. Some translations are truly academic in nature, while others are informative in their objective. Some translators disliked strict loyalty to each and every Arabic word, for fear of obscuring the inherent meaning; while others were vehemently loyal to the revelation, yet succeeded in conveying the

---

1 e.g. REDEN in 170; note that in colloquial German this verb means to speak informally and with little weight on the contents.
meaning. Some translators enjoyed slight liberty with choice of words, while others guarded themselves from what could very well be interpolation of thoughts."

"Aademic" translations, such as Reclam's, in order to transfer the contents correctly, tend to stick to the original phrasing, and for reasons of (lexical) accuracy use uniform translation throughout. In our example, Reclam consistently used SPRECHEN for qaala. In doing this, "academic" translators collapse the multifunctionality of the original word, often leading to boring and difficult-to-understand translations.

Liberal, "informative," translations, however, are better suited for readers unfamiliar with the content. They can, for example, indicate relationships between the actors (M irza 68) and discourse levels (M irza 69) by making use of different meanings of speech act verbs in the target language. Such translations maintain the multifunctionality of the original and, because of their pragmatic accuracy and vivacity, can lead to a better understanding.
REFERENCES

English Translations at Hoque, Zohorul (n.d.) http://www.qurantoday.com/BaqarahIntro.htm
Reinelt, Rudolf 2000. “Sprechakt/handlungsausdruecke (SAX) und Schriftenentwicklung” (“Speech act/action expressions (SAX) and the development of writing”) Memoirs of the Ehime University Faculty of Law and Letters, Department of Humanities 9: 269-85.
APPENDIX

6 Translations of 2:67-68.
Hoque, Zohorul (n.d.) http://www.qurantoday.com/BaqarahIntro.htm

A Comparison of Different Translations

Verse 2:67

Yusuf Ali:
And remember Moses said to his people: “God commands that ye sacrifice a heifer.” They said: “Makest thou a laughing-stock of us?” He said: “God save me from being an ignorant (fool)!”

Zohorul Hoque:
And behold! Musa said to his people, “Allah indeed commands you that you should slaughter a heifer.” They said: “Do you take us for a jest?” He said: “I take refuge with Allah from being of the ignorant.”

T. J. Irving:
When Moses told his folk: “God commands you to sacrifice a cow,” they said: “Do you take us for a laughingstock?” He said: “I seek refuge with God lest I become so ignorant!”

T.U. Hilali-M. Khan:
And (remember) when Musa (Moses) said to his people: “Verily, Allah commands you that you should slaughter a cow.” They said, “Do you make fun of us?” He said, “I take Allah’s Refuge from being among Al-Jadilan (the ignorants or the foolish).”

M. Pickthall:
And when Moses said unto his people: Lo! Allah commandeth you that ye sacrifice a cow, they said: Dost thou make game of us? He answered: Allah forbid that I should be among the foolish!

M.H. Shakir:
And when Musa said to his people: Surely Allah commands you that you should sacrifice a cow; they said: Do you ridicule us? He said: I seek the protection of Allah from being one of the ignorant
Verse 2:68

Yusuf Ali:
They said: “Beseech on our behalf Thy Lord to make plain to us what (heifer) it is!” He said; “He says: The heifer should be neither too old nor too young, but of middling age. Now do what ye are commanded!”

Zohurul Hoque:
They said: “Call on your Rabb, for our sake, to make clear to us what she is.” He replied: “Verily He says, she indeed is a heifer neither too aged nor too young, of a middle age between these. Now, do what you are commanded.”

T. J. Irving:
They said: “Appeal to your Lord for us, to explain to us what she is.” He said: “He says she is neither a worn-out cow, nor a heifer, but of an age in between. Do as you are ordered!”

T.U. Hilali-M. Khan:
They said, “Call upon your Lord for us that He may make plain to us what it is!” He said, “He says, ‘Verily, it is a cow neither too old nor too young, but (it is) between the two conditions’, so do what you are commanded.”

M. Pickthall:
They said: Pray for us unto thy Lord that He make clear to us what (cow) she is. (Moses) answered: Lo! He saith, Verily she is a cow neither with calf nor immature; (she is) between the two conditions; so do that which ye are commanded.

M.H. Shakir:
They said: Call on your Lord for our sake to make it plain to us what she is. Musa said: He says, Surely she is a cow neither advanced in age nor too young, of middle age between that (and this); do therefore what you are commanded.
1. **Introduction: Background, Premises, and Some Observations.**

The copy theory—part of a larger body of “minimalist” proposals by Chomsky 1993, 1995—has provided a compelling account of some well-known reconstruction effects. For example, it captures the fact that elements pronounced in a wh-phrase or topic behave as if they also occupied a position lower down in the structure. Here’s how: suppose that A′-movement creates an instance of a phrase (a copy) in a target position while keeping another in the source position. Source-position “tail” copies are not pronounced, they are gaps. But even so, we know that they exist because they trigger variable-binding and Condition C effects; variable-binding and Condition C (among other tests sensitive to LF structure) essentially can track the history of a moved element. Copy—theoretically “reconstruction”—is just a name for the successful detection of an unpronounced copy.

---

1. I would like to thank Jim McCloskey for his kind, generous and patient help in every aspect of producing this paper. Judith Aissen, Daniel Buring, Sandy Chung, Bill Ladusaw, Jaye Padgett, and an anonymous reviewer provided insightful comments at various stages. Thanks also to audiences at the Fifteenth Annual Symposium on Arabic Linguistics at the University of Utah, and at the 2000 Linguistics at UCSC colloquium. The members of the Fall 2000 Lx 290 Research Seminar at UCSC provided a good deal of advice on the presentation and content of the paper’s material; I am grateful to Chris Potts especially. All the errors in this paper are probably mine.


3. Though see Pesetsky 1998 for another view.

Other research programs, including Cresti (1995), Engdahl (1986), Jacobson (1994, 1999), and Sharvit (1997) (among others) see reconstruction in a different light. These researchers have argued that enriching the principles of semantic interpretation derives certain types of reconstruction without any appeal to syntactic operations. For example, one semantic approach has dealt relatively well with variable-binding reconstruction in certain types of relative clauses (Sharvit, 1997). In fact, reconstruction in relative clauses remains a serious source of empirical discomfort for the copy theory—presuming that nominal phrases and their relative clauses are never related by movement, the copy theory appears to be helpless to explain reconstruction effects between them.\(^5\)

On the other hand, Chomsky (1995), Fox (2000), and Romero (1996) have claimed that a syntactic account of reconstruction like the copy theory has at least one advantage over the semantic account. Though the argumentation is too intricate to summarize briefly, Fox (2000) and Romero (1996) demonstrate that semantic reconstruction comes up short when faced with the fact that scope-sensitive operations and Condition C obligatorily correlate under reconstruction. Because semantic accounts do not refer to syntactic structure, they cannot adequately predict\(^6\) that scope-sensitive elements interpreted under reconstruction cannot escape (they “feed”) Condition C effects, effects presumably determined by LF structure. On the other hand, the correlation between Condition C and scope reconstruction flows quite naturally from a syntactic account like the copy theory. Let us assume these arguments to be persuasive, and take up the copy-theoretic explanation for reconstruction.

The argument maintained below will be this: discomfort notwithstanding, copy-theoretic assumptions provide a fairly reasonable account of two problems raised by reconstruction effects in Syrian Arabic relative clauses.\(^7\) The abstract schema (1) sketches the phenomenon: certain elements E pronounced inside a definite nominal phrase can be detected by the principles of variable-binding and Condition C inside that phrase's relative clause modifier.

---

5 Of course, the same embarrassment holds for other constructions not obviously derived by A'-movement which nonetheless show connectivity effects, clefts and pseudo clefts being prime examples (Higgins 1973, Percus 1997, and many others).

6 (Without resorting to some ad hoc semantic account of Condition C that would correlate with scope)

7 The particular relative clauses under investigation are restrictive relatives.
RELATIVE CLAUSES IN SYRIAN ARABIC: TWO RECONSTRUCTION PROBLEMS

(1) Syrian Reconstruction Generalization:

\[
\begin{array}{c}
\text{DP(definite) \ nominal \ \relclause \ \_ \ \_ \ \_} \\
\end{array}
\]

(1) illustrates the basic scheme, but the situation is complicated. Reconstruction occurs in two types of relative clause: ones with traces and ones with resumptive pronouns.

Resumptive pronouns freely alternate with (DP) traces in the object gap position\(^8\) of the Syrian relative, as in (2) and (3).\(^9\)

(2) Definite nominal:

a. damašq heyya \[lm\(\text{mā\-dina}\) fi-suriyya] [yallī maa-šrifha pro]. Damascus 3fs DEF.city.fs in.DEF.Syria C NEG-saw.1s.3fs “Damascus is the city in Syria that I didn’t see.”

b. damašq heyya \[lm\(\text{mā\-dina}\) fi-suriyya] [yallī maa-šrif \(t\)]. Damascus 3fs DEF.city.fs in.DEF.Syria C NEG-saw.1s “Damascus is the city in Syria that I didn’t see.”

(3) Indefinite nominal:

a. damašq heyya \[m\(\text{dina}\) fi-suriyya] [mā-šrifha pro]. Damascus 3fs city.fs in.DEF.Syria NEG-saw.1s.3fs “Damascus is a city in Syria that I didn’t see.”

b. damašq heyya \[m\(\text{dina}\) fi-suriyya] [mā-šrif \(t\)]. Damascus 3fs city.fs in.DEF.Syria NEG-saw.1s “Damascus is a city in Syria that I didn’t see.”

The paradigm in (2) differs from (3) just in the definiteness of the nominal object: the definiteness prefix \(\text{al-}\) marks \(\text{mdina}\) ‘city’ as definite in (2). In (3) \(\text{mdina}\) is indefinite and thus bare.\(^{10}\)\(^{11}\)

\(^8\) Detailed investigation into Syrian Arabic subject resumptives is made complicated by obligatory pro-drop in neutral contexts and other ill-understood properties of clause-initial subjects. In this domain of subject relative clause gaps is a large topic; too large to address in this particular discussion.

\(^9\) Cowell (1964) claims that Syrian resumptives, while optional in principle, are more common than gaps in speech and writing. The Syrian speakers I have polled so far find both structures perfectly acceptable, though they admit to using resumptive pronouns more. Brustad (2000), on the other hand, claims that resumptives are obligatory in Syrian (90). It seems that resumptive optionality can vary within the dialects covered by the term ‘Syrian.’

\(^{10}\) The definite-marking prefix \(\text{al-}\) undergoes two allophonic phonological processes in Arabic: (1) the initial schwa vowel obligatorily deletes when preceded by a vowel
the gaps of the relatives in (2) and (3) have been boldfaced—the (a) examples include relative clauses with resumptive pronouns, the (b) examples DP-traces.

Note that my representation in the (a) examples supposes that resumptives in Syrian are null. Some recent work in Semitic syntax has made use of the notion that the spec/head relation triggers the realization of agreement clitics e.g. Aoun & Benmamoun (1998), Benmamoun 1999), and others. In line with this work, I will assume that null object resumptive pronouns trigger (and thus can be detected by) agreement on V, or whatever predicate governs them. For instance in (2a) and (3a) the clitic —ha on the verb 2uft ‘I saw’ instantiates 3rd person feminine singular features, in agreement with the resumptive pro. I will postpone discussion of how the resumptive pro obtains the phi-features of the head noun of the relative clause until Section 3.

Let’s now get back to the significance of relative clause resumptive pronouns for reconstruction. If the copy theory provided an account of reconstruction into Syrian relative clauses with trace-type gaps, that alone would be surprising enough. But reconstruction into pro-type gap positions is a very serious challenge for any theory that explicitly ties reconstruction to A’-movement: standard analyses of resumptive pronouns hold that they are not derived by movement at all (e.g. Borer 1984, e.g. when the preceding word ends in a vowel); (2) the [l] phoneme assimilates the [coronal] place of its following consonant.

11 The morphology of the relative clause complementizer appears to be sensitive to the definiteness of the head nominal (just as in Modern Standard Arabic): “definite” complementizers are realized as yalli, illi, or halli (depending on dialect), while the “indefinite” complementizer is phonetically null. See Cowell 1964, Brustad 2000 for a more detailed discussion.

12 Note that this type of agreement only happens with pronominal objects. NP arguments in object position do not trigger agreement. I have no serious explanation for this fact. Aoun & Choueiri 1997 and Aoun & Benamoun 1998 specifically identify the site of clitic agreement with pro as a CliticP above VP but below IP. See footnote 30 and Section 4 for a different analysis. Jim McCloskey (p.c.) points out that this same pattern of agreement exists in the Celtic languages.

13 Another explanation for the existence of clitics in the place of independent object pronouns might be that the clitics are themselves pronouns: “weak” pronouns that must incorporate onto their governing predicate. If such an incorporation [i.e. head-movement] proposal turned out to be valid it might have some solid consequences for the theory below—a theory that makes crucial use of object pros in D—depending on whether the trace of head-movement can be treated as a variable (particularly as a pronominal variable) or not.
RELATIVE CLAUSES IN SYRIAN ARABIC: TWO RECONSTRUCTION PROBLEMS

McCloskey 1990). The copy theory faces two reconstruction problems in the Syrian relative clause domain.

(4) Two Reconstruction Problems:
1. How can the reconstruction effects in (1) come about at all?
2. How can reconstruction effects be tied to resumptive pro positions?

I submit that the copy theory achieves reasonable answers to both these challenges by compelling us to adopt a ‘head-raising’ analysis of Syrian relative clauses. This type of analysis sits squarely within the famous proposal phylum of Shachter (1973), Vergnaud (1974, 1982), Munn (1994), Kayne (1994) and Bianchi (2000). Supposing that the head nominal phrase in (1) raises by A'-movement out of the gap position of the relative clause modifier, a tail copy of the head nominal will remain to be detected by Condition C and variable-binding. This is the basic analysis to be developed below.14

Aoun & Choueiri (1997) implicitly address the two problems above as they apply to variable-binding reconstruction in Lebanese Arabic. The next section and part of Section 3 outline this proposal and show it to be only partially satisfactory for Syrian. The bulk of Section 3 discusses Problem #1, establishing the empirical generalizations of Condition C and variable-binding reconstruction for Syrian Arabic and detailing the head-raising analysis. Section 4 discusses Problem #2, attempting to cash out the claim that the tail copy of a phrase can involve a resumptive pronoun. Section 5 includes some initial and tentative suggestions for a semantic theory of Syrian Arabic relative clauses, and the conclusion briefly addresses some larger-scale implications and problems.

2. A Prior Account of the Problems in Question

Aoun & Choueiri (1997) concern themselves with Lebanese Arabic examples like (4), cases of variable-binding under reconstruction between definite nominal phrases and restrictive relatives.

14 Note that the reconstruction effects in (1) appear to hold only for definite nominal phrases. I will not address the issue of indefinite nominals much more in this paper.
(5) \[\text{tolmiizun}_1 \text{ wala}_1 \text{wala}_1 \text{all}_1 \text{DEF.teachers}_2 \text{prefer}_2 \text{3mp}_2 \text{3ms}_2 \text{students}_2 \text{3mp}_2 \text{3ms}_2 \text{.}
\]

... rah_yirba_yeezze.

FUT win prize

[Lebanese: A&C 20a]

“Their\textsubscript{1}\text{ good student}_1, that all the teachers\textsubscript{2} prefer him\textsubscript{2} is going to win a prize.”\textsuperscript{15}

In (5), the definite head nominal phrase \[\text{tolmiizun}_1 \text{ wala}_1 \text{all}_1 \text{DEF.teachers}_2 \text{prefer}_2 \text{3mp}_2 \text{3ms}_2 \text{students}_2 \text{3mp}_2 \text{3ms}_2 \text{.}
\]

An explanation for this kind of interpretation must resolve both reconstruction problems given in (4): how does \(-u\textsubscript{1}\) come to be bound, and (how) does this binding involve the resumptive clitic \(-u\textsubscript{2}\)? Aoun & Choueiri (1997) claim that an indirect relation between the nominal head phrase and the relative clause’s resumptive pronoun induces variable-binding. First, a pro element within the gap position DP raises into the relative Spec, CP headed by \text{yalli} to check \{[+definite], [Case], and [f]-features. Next, the tail pro in the A’-chain \{pro ... pro\} is bound at LF by a c-commanding quantifier in the relative clause: \text{kall lm\textsuperscript{ll}allmiit}_1 ‘all the teachers’ in (5). The bound-variable interpretation is transferred to the upper pro (or else both pros are instances of the same element), and the overt pronoun in the head nominal phrase receives the interpretation of the upper pro via a ‘predication relation’ between the relative clause and the nominal phrase.\textsuperscript{16}

\textsuperscript{15}Glosses adapted from Aoun & Choueiri 1997. I am not sure whether the Lebanese construction \{\text{kall} + definite plural N\} is a true universal quantification. Baker (1996) and others have argued that determiners like all cross-linguistically do not have the same quantifier properties as every. In Modern Standard Arabic and Syrian, there exists an alternative construction \{\text{kall} + indefinite singular N\} usually translated “each/every N.” For this section, I will just assume (with Aoun & Choueiri 1997) that Lebanese \text{kall} (+ definite plural N) is a quantifier that can bind pronominal variables. In discussion below I will try to sidestep this issue by using the Syrian quantifier wala ‘no’, which has the added advantage of forestalling e-type readings.

\textsuperscript{16}In the terminology of Aoun & Choueiri (1997), “the relativized DP, the relative clause, and the pro in COMP are all coindexed. . . . [T]he pro and its trace form a chain generated by movement [and since] reconstruction occurs. . . . with chains generated by movement, reconstruction will be available [between the pro and its trace]. . . .” (16).
As for the question of how reconstruction may be related to a resumptive position (Problem #2), Aoun & Choueiri (1997) propose a complex internal syntax for Lebanese resumptive DPs—the pronounced, obligatory clitic occupies the D head and the pro DP occupies the specifier of the resumptive clitic DP. This pro is the element that rises to Spec, CP, transferring its interpretation to the nominal phrase by predication.\footnote{\textit{This is actually a simplification of the claims of Aoun & Choueiri 1997. The full claim is that the resumptive clitic undergoes head-movement to a clitic projection between VP and IP; the resumptive pro in Spec, DP would thus undergo \(\lambda\)-movement to Spec, CliticP and then \(\lambda\)'-movement to Spec, CP. Reconstruction, strictly speaking, would take place into the CliticP, not the resumptive DP. See fn 26 for some brief discussion of this proposal for Syrian.}}

In sum, the derivation for (5) would proceed as in (6).

\begin{equation}
\text{(6)}
\end{equation}

There are some additional advantages of the pro-raising analysis. Aoun & Choueiri (1997) remark that adjuncts may not be relativized in Lebanese. DPs like (7a) or (7b) are unacceptable.

\begin{equation}
\text{(7) a. *ssabab yalli raht.o . . .}
\end{equation}
DEF.reason(ms) C left.1s.3ms
“The reason that/why I left..”
[Leb: A&C 41]

b. *sabab yallı ṭoḥ.ta...reason(ms) C left.1s.3ms
“A reason that/why I left.”
[Leb: A&C 86]

Such facts appear mysterious but are entirely expected in the pro-raising perspective—non-argument positions cannot license pro. Moreover, Aoun & Choueiri (1997), having amassed some arguments that pro is strictly definite, claim that pro raises to Spec, CP only to check a [+definite] feature on C. When C is [-definite] (as determined by an indefinite nominal phrase head) pro will not raise and thus reconstruction effects do not occur, as evidenced by the absence of variable-binding in (8):

(8) *šūft [taʃmiiz min taʃmiizun1], [kall lmʃallmiiṯ1, bifaddlu₂ ]
saw.1s:student(ms) of student(mp).3mp all DEF. teachers prefer.3mp.3ms
“I saw [a student of theirs1], that all the teachers, prefer him.”
[Lebanese: A&C 29a]

pro-raising appears to meet the challenge posed by the two reconstruction problems in (4): variable-binding in a nominal head occurs because the variable of Lebanese relative clauses is always a resumptive pronoun pro, not t. This pro can be bound, and can transfer its bound interpretation to a pronoun in the nominal phrase. Aoun & Choueiri (1997) grant that a head-raising analysis might suffice to predict variable-binding in cases like (5), but explicitly reject such an account for Lebanese. They note, for instance, that the ban on adjunct relativization would be unexplained under the head-raising account. Leaving that objection aside for now (we will return to it in the conclusion), in the following section we will investigate reconstruction in Syrian Arabic as a test to decide whether pro-raising or head-raising accounts for the facts of Syrian Arabic.

This section is bipartite: the first establishes the variable-binding and Condition C reconstruction generalizations in Syrian Arabic (the facts sketched in 1). The second discusses the significance of those generalizations for the pro-raising hypothesis and for the copy theory, attempting to address Problem #1: how can the reconstruction effects come about at all?

3.1 Two Generalizations

Let us first address variable-binding in Syrian. As might be expected (Heim & Kratzer, 1998), semantic variable-binding of pronouns appears to require c-command. The paradigm in (9) demonstrates this requirement.

(9) a. [wala ʿtaālib], habb ʿattah/ii/ ʿaleyu pro.
   No student(m) liked(3ms) DEF.investigation of.3ms
   “No student, liked the investigation of him.”

b. *huwwa/pro habb ʿattah/ii/ ʿaleyu [wala ʿtaālib].
   3ms liked(3ms) DEF.investigation of no student(m)
   “He liked the investigation of no student.”

c. *al/ustaāda yallii [wala ʿtaālib], maa-habbha ʿhabbat.
   DEF.teacher(f) C no student(m) NEG-liked(3ms).3fs liked.3fs
       ... ʿattah/ii/ ʿaleyu pro.
       ... DEF.investigation of.3ms
   “The teacher, that no student, liked, liked the investigation of him.”

The sentences in (9a) and (9b) were read aloud to informants, who were asked whether they accurately described a situation where every student in school has been investigated by the teachers and deeply represents the process. The sentence (9b) where the pronominal variable asymmetrically c-commands the quantifier was quite clearly rejected in this scenario. Of course, this type of example does not necessarily demonstrate that variable-binding requires c-command since it also instantiates a Condition C violation. The sentence in (9c), though, where the QNP is embedded in a subject relative clause was also judged to be unacceptable. This example clearly cannot be ruled out by Condition C (at
any level of representation): it and (9a) suggest that variable-binding requires c-command.

The speakers polled for the sentences in (10) were asked whether the sentences might describe a likely scenario in a situation where every student in school is extremely modest. Confronted with the DP in (11), the speakers were asked if the phrase could refer to an investigation of students that the students all hated (11a) or that did not particularly affect them (11b). Given that variable-binding requires c-command, the bound-variable interpretation of the pronominal pro1 in sentences (10a) and (11a) is puzzling: the QNP wala taalib1 does not (visibly) c-command the position of the pronominal.18

(10) a. \textbf{pro} \ldots QP
\textit{?aqṭirahū \ ['liqṭirah yīnnu pro zekīl]}, \ldots suggested1s DEF.suggestion C .3ms smart .\ldots

b. [\textit{yallī wala taalib}, ūbīl fiīh \textit{pro}]
\textit{c} no \textit{student} \textit{accepted} in.3ms
“I made the suggestion that he₁ was smart that no student accepted.”

c. *\textbf{QP} \ldots pro
*\textit{?aqṭirahū \ ['liqṭirah yīnno wala taalib], zekīl]}, \ldots suggested1s DEF.suggestion C no \textit{student} \textit{smart}

\ldots [\textit{yallī pro} /huwwe ūbīl fiīh \textit{pro}].
\textit{c} 3ms \textit{accepted}(3ms) in.3ms
“I made the suggestion that no student was smart that he₁ accepted.”

18 Jim McCloskey (p.c.) points out that (11) might be used to refer to an investigation of a person who no student liked. This ambiguity might account for the fact that two speakers did not find there to be an acceptability difference between (11a) and (11b). The grammaticality judgments indicated in (11) were shared by five other speakers. Examples similar to (11) using the construction \textit{kall} (‘all, every’) + indefinite singular were accepted by one speaker. I attribute this fact to an e-type reading available for pronouns in relation to \textit{kall} + indefinite singular. Such a reading is generally unavailable for pronouns in relation to quantifiers like no, thus I will stick to the Syrian quantifier wala for this discussion.
(11) a. [wala taalib],
    [illī [wala taalib],
    ma’a-habbu pronom pro 1
    DEF. investigation of.3ms C no student
    "The investigation of him 1 that no student 1 liked."
    NEG-liked.3ms relativized object

b. * [wala taalib],
    [illī [za’aal [wala taalib],]
    pronom pro 1
    DEF. investigation of.3ms C upset no student relativized subject
    "The investigation of him 1 that upset no student."

The examples in (10a) and (11a) are structurally similar; they each contain a (boldfaced) null pronominal pro 1 inside a phrasal complement to the definite head noun of the relative clause. The null pronouns are variables bound by the (boldfaced) QNP wala taalib ‘no student.’ In each case, the gap position of the relative clause is an object pro detectable by object agreement on the verb. Note that the QNP subject c-commands the gap in the (a) cases.

Suppose that LF ‘reconstruction’ of pronom pro 1 in the (a) examples (however implemented20) targets the relative gap position; this would explain how variable-binding comes about, since pro 1 would be c-commanded by the QNP subject wala taalib ‘no student’ at LF. (10b) and (11b) provide possible reasons to think that this idea is on the right track. (10b) seems uninteresting: it switches the positions of the pronoun and the QNP from those in (10a). The unavailability of variable-binding is expected in (10b) no matter if the QNP is interpreted in its pronounced position or if it has ‘reconstructed’ to a position below the pronominal subject of the relative clause. In neither position would it c-command the pronoun it is attempting to bind. However, (10b) does show that the QNP in the head nominal phrase cannot reconstruct to a position c-commanding the relative clause subject. This can be taken as weak evidence that the gap position of the relative is the locus of reconstruction. (11b) provides stronger evidence: variable-binding of the pronominal pro 1 is unavailable with a relative clause containing a gap subject and a QNP object. We can conclude that in (11a) the QNP c-commands the

20 The term ‘reconstruction’ in this discussion refers merely to the phenomenon where an element is interpreted or has various effects in a position different from where it is pronounced. It does not refer to any particular analysis of this phenomenon.
position of reconstruction while in (11b) it does not. Since the gap is c-commanded in (11a) but not (11b), it makes sense to identify the gap with the position of reconstruction. We have (perhaps\textsuperscript{21}) established the generalization in (12).

(12) Syrian variable-binding reconstruction generalization:

“A pronoun within a definite nominal phrase can 'reconstruct' to the gap position of that phrase's relative clause modifier.”

\[ \text{e.g. } [\text{DP-definito} [\ldots \text{pronoun} \ldots]]_2 [\text{C QNP} \ldots \text{pro}_1] \]

With this generalization in hand, let us now see whether Condition C effects hold under reconstruction in Syrian. Condition C forbids names bound by pronouns or other names\textsuperscript{22} (Chomsky 1986, Reinhart 1983). This restriction holds in Syrian, as demonstrated by (13) and (14) below.

(13) a. \textit{ahmed\textsubscript{1} Paal} \textit{pinn layla haabatu\textsubscript{2}}.
    \textit{said C} \textit{liked\textsubscript{3fs}}
    “Ahmed\textsubscript{1} said that Layla liked him\textsubscript{2}.”

b. *\textit{huwwa/pro\textsubscript{1} Paal} \textit{pinn layla haabat ahmed\textsubscript{1}}
    \textit{3ms} \textit{said C} \textit{liked\textsubscript{3fs}}
    “He\textsubscript{1} said that Layla liked Ahmed\textsubscript{1}.”

\textsuperscript{21} Strictly speaking, all that has been shown is that reconstruction occurs into a position below the relative clause subject (10a), (11a) and above the relative clause object (11b). Conceivably reconstruction could target some fixed position between the subject and the object of the relative clause every time: it might have nothing to do with the gap. Jim McCloskey (p.c.) points out that this position might be identified with the left edge of \textit{vP}. Section 4 presents some direct evidence that it is in fact the gap position that is relevant to relative clause reconstruction.

\textsuperscript{22} For the purposes of this discussion, assume that names are base-generated with an index that gets assigned a value by some assignment function, just like pronouns. Whether or not the binding of names by pronouns (or pronouns by names) is true “semantic binding” or simply accidental coindexation is irrelevant here.
(14) a. *huwwa\pro_2 qajib ahmed_1.
    likes
    “He likes Ahmed.”

b. *ahmed_2 qajib ahmed_1.
    likes
    “Ahmed likes Ahmed.”

In (13) it appears that the binding configuration name_1 > pronoun_1 is allowed across a clause-boundary (a), but that pronoun_1 > name_1 is disallowed (b). Clause-internally, the pronoun_1 > name_1 and name_1 > name_1 configurations are not permitted, as (14) demonstrates. The Condition C requirement that names not be bound appears to hold.

Given that Condition C entails c-command (part of binding), examine the paradigms below. (15a) and (16a) were judged unacceptable: they contain a name in a CP complement to N, and a pronominal subject of the relative clause modifier of that N. Notice that the object position has been relativized and contains a resumptive pro (detectable by agreement on the verb). These sentences become acceptable when the name and the pronoun are switched, as in the (b) cases. The speakers consulted for the (b) cases require the name to be mentioned in the preceding discourse to accept the sentences (due to the strangeness of having a pronoun precede its coreferent name), but otherwise found the sentences to be directly contrastive to the (a) cases.

* subject pronoun, relativized object:
(15) a. *qaqtirahu [liqitirahu ?innu ahmed habb layla_1]
suggests DEF.suggestion C loved

*Which fan of Madonna does she like best?
Which fan of Madonna does she like best?
Because simply switching the pronoun and the name in (15) and (16) determines the grammaticality of the sentences, it is reasonable to believe that the source of unacceptability in the (a) cases is a Condition C effect under reconstruction. Specifically, it appears that a name embedded in a CP complement to N can be bound at LF by a subject pronoun in the relative clause modifier of that N.
The following examples establish some legitimacy for this claim. The sentences in (17) are virtually equivalent to (15a) and (16a) except that the nominal phrase containing a name has been related to a subject gap, a gap that c-commands the pronoun coindexed with the name.

\[(17) \ a\ \text{suggest}^{\text{(PERF).1cs}}\text{DEF.suggestion}\ C \text{loved}\]

\[
\begin{align*}
\text{a.} & \ yall_{i_2} \text{za'aaalha} \text{pro}_i \ldots \text{kahtiiran} \\
& \text{C bothered.3fs very much} \\
& \text{“I made the suggestion that Ahmed loved Layla that bothered her very much.”}
\end{align*}
\]

\[
\begin{align*}
\text{b.} & \ \text{al-}'aadiiia} \text{ 'akaset} \text{DEF.judge} \text{f reversed.3fs DEF.decision}\ C \text{ was innocent}\]

\[
\begin{align*}
& \ yall_{i_2} \text{maa-kaan} \ldots \text{za'aaalu} \text{pro}_i \text{kahtiiran} \\
& \text{C NEG-was bothered.3ms very much} \\
& \text{“The judge reversed the decision that Ahmed was innocent that had not bothered him very much.”}
\end{align*}
\]

The grammaticality of (17a) and (17b) indicates the relevance of the gap position in determining reconstruction effects in relative clauses. Supposing that reconstruction of the name inside the head nominal phrase occurs into the gap position, we predict (17a) and (17b) to be acceptable since at no time would the name be bound by the (object) pronoun.

The Condition C generalization for Syrian appears to be that a definite nominal phrase containing an embedded name coindexed with a relative clause subject cannot then be related to the relative object gap, as shown by (15a) and (16a). In other words, a name in a head nominal phrase ‘reconstructs’ to the gap position of the relative. As expected, ungrammaticality disappears when the pronoun and the name are reversed (15b) (16b) and when the definite nominal phrase containing the name is related to a gap that c-commands the pronoun (17). In sum, we have the generalization in (18).
Syrian Condition C reconstruction generalization:

“A name within a definite nominal phrase ‘reconstructs’ to the gap position of that phrase’s relative clause modifier.”

e.g. *[DP(definite) [ . . . name1 . . . ]]2 [ C pronoun1 . . . pro2]]

With the variable-binding generalization (12) and the Condition C generalization (18) established, let’s see what significance they might have for the pro-raising analysis and for the assumptions of the copy theory.

1.2 The Copy Theory and the Head-Raising Analysis

Recall that the Aoun & Choueiri (1997) pro-raising analysis predicts Condition C to be absent under reconstruction in relative clauses (since at no time in the derivation would a name in the nominal phrase ever be inside the relative clause). I submit that pro-raising is an incomplete account of Syrian relative clauses in light of the Condition C generalization (18).

What does the copy theory have to say about Condition C? Suppose that the principles governing Condition C operate on LF representations (Chomsky, 1995; Fox, 2000; Reinhart, 1983). Since Condition C is sensitive to structure at LF, specifically whether or not names are bound, it stands to reason that Condition C effects under reconstruction provide positive evidence for the existence of structure in the reconstruction site at LF. Put into copy theory terms, a Condition C effect under reconstruction locates an unpronounced copy of a moved phrase that contains a name.

Applying this reasoning to Syrian, the Condition C generalization (18) becomes an excellent case for the actual structure of a head nominal phrase existing in the relative clause’s gap position at LF. The same reasoning applies to the variable-binding generalization in (12): since variable-binding requires c-command at LF, (12) indicates that a copy of the head nominal phrase exists in the gap position at LF, allowing binding of the variable inside the nominal phrase. We have assumed that copies represent instances of a single phrase that has undergone (A’-) movement; this assumption virtually compels us to say that the head nominal phrase moves from the gap position of its relative clause, leaving a tail
copy that induces Condition C effects (18) and variable-binding effects (12).

As mentioned earlier, this type of ‘head-raising’ analysis of relative clauses originated with work by Shachter (1973) and Vergnaud (1974, 1982). Researchers within the antisymmetry framework of Kayne (1994) have been the most prominent advocates of the theory in recent years, Bianchi (2000) being the most comprehensive discussion. Here, it only matters that the head nominal undergoes some type of A'-movement out of the gap position (presumably to Spec, CP). It does not matter whether this is the final position of the nominal (Kayne, 1994; Bianchi, 2000) or whether the head nominal moves on to another, ultimately more standard A'-position (Vergnaud, 1982).

Having adopted a copy-theoretic head-raising perspective on the generalizations in (12) and (18), let us be more clear about what the structure of relative clauses looks like in Syrian. The variable-binding reconstruction case (10a) is repeated below as (19) with its corresponding structure. 25

(19) ḥaqṭiraḥtu [liqṭirahā ḫinnuu [pro kān zeky] yalli suggested.1cs DEF.suggestion C. 3ms was smart C

wala ṭāalībī talkīf fiḥ [pro].
no student accepted to.3ms

“I made the suggestion that he was smart that no student accepted.”

25 Given this discussion’s agnosticism with regard to the ultimate landing site of the head nominal phrase of the relative clause, this structure may well be a “pre-LF” representation. The point is merely to show that at LF, copies of the moved element exist in the gap position and in Spec, CP of the relative clause.
In the structure above, the head nominal phrase \([_{NP \ liq\text{tir\text{a}ah } \ Pinnuu \ pro_1 \ ka\text{an} \ zeky\text{y}^2}]\), originates in the gap position (here, complement position to the resumptive pro\(^2\)) and moves through Spec, DP into Spec, CP. The copies have been indicated with oval outlines. The pronoun pro\(_1\) in the tail copy becomes bound by the QNP \(\text{wala taalib}\) at LF.

As for cases that violate Condition C under reconstruction, let us articulate the structure for (15a), repeated as (20) below.

---

\(^2\)This structure and the one that follows makes a claim about the internal structure of the gap position (that contains a resumptive pronoun) that will not be defended until the next section. Specifically, it claims that resumptive pros are derived by NP-movement through Spec, DP and into Spec, CP, leaving an NP copy in the gap position.
(20) *?aqithatu [liqtirah ʔannu ahmed ḥabb layla1 yalli pro1 suggested.1cs DEF.suggestion C loved C ʕaaraādetu pro2 . . . . . . tamaāman]).

objecēt.3fs.3ms completely

“I made the suggestion that Ahmed loved Layla, that she objected to completely.”

As in (19), the reconstruction effect is derived by the head nominal leaving a tail copy in the gap position. The NP [ liqtirah ʔannu ahmed ḥabb layla1 ], ‘the suggestion that Ahmed loved Layla’ moves through Spec, DP into Spec, CP; its tail copy (or perhaps the one in Spec, DP) triggers the Condition C effect at LF since the name Layla1 is bound by the relative subject pro1.

In sum, we have seen that Condition C and variable-binding indicate the presence of a copy of a nominal head within the Syrian relative clause, motivating a head-raising analysis. Such an analysis reasonably addresses Problem #1 in (4) by claiming that reconstruction effects arise
due to movement of the head nominal from the gap position. At this point, we make the clear prediction that the gap position particularly, the one involving the resumptive pro, gives rise to reconstruction effects. How can we test whether this position and not another triggers Condition C and variable-binding? Someone who finds head-raising distasteful might claim that reconstruction consistently targets a fixed position between the relative clause subject and the relative clause object (Spec vP, for example, or Spec, CliticP). Such a hypothesis arguably deduces the facts in section 3.1. We need an answer to Problem #2 to firmly establish head-raising from the gap position as a valid theory: is reconstruction really related to the resumptive gap position and if so, how?

4. Problem #2: Resumptive DPs and Reconstruction.

In preparing an answer to the question of how (and even whether) reconstruction effects are tied to resumptive pronoun positions, an issue of their external syntax, let’s first try to address their internal syntax. Recall from (2) and (3) that resumptive pronouns—as detected by agreement clitics on the main verb—alternate with trace gaps. Koopman (1999) has proposed two types of movement to account for various options in the morphosyntax of pronouns cross-linguistically: DP-movement and NP-movement.27 DP-movement targets the entire nominal DP, and NP-movement moves through Spec, DP to strand the D head.

Suppose that relative clauses derive by movement of a nominal phrase out of a gap position: if the head nominal is a DP that can undergo either type of movement, then we have an account of the optionality between resumptives and traces. When the entire DP raises, a DP copy is left behind (i.e. a DP-trace gap of the familiar type). When the NP inside the head nominal raises, though, it must move through Spec, DP, triggering agreement on D (following Koopman’s analysis). This agreement in features is realized as pro, which we have assumed triggers the pronominal agreement clitic on the main verb.28

27 Koopman 1999 actually refers to the latter type as “NumP” movement. For this discussion the existence of NumP is not relevant; let’s simply assume “NP movement” to refer to movement of some phrase below DP.

28 Notice that we now predict that DP-movement should display the standard effects of DP-traces but that NP-movement might not. This prediction appears to be borne out in Modern Standard Arabic, which displays a similar type of ambiguity of resumption in relative clauses, where resumptive clitics (i.e. NP-movement stranding a
Let's assume for sake of argument that the optionality between DP- and NP-movement out of relative clauses is motivated by a feature [\textit{wh}], which can be generated either on D or on N of the head nominal phrase (see Börer, 1999 for proposals that [definite] behaves this way in Semitic). The trees structures in (19) and (20) indicate what NP-movement of the nominal head looks like for cases of variable-binding and Condition C reconstruction. As it turns out, we make a surprising prediction about these structures.

Heycock (1995) and Huang (1993) have suggested that A'-movement of a predicate results in obligatory reconstruction of that predicate. Whatever the ultimate explanation for this generalization, it is relevant to the structures in (19) and (20): the NPs that move out of the gap position through Spec, DP to Spec, CP are predicates. This predicate, whatever its denotation (\textit{<<, <e,t>>}-type function, ‘property’, etc.), cannot be incorporated into the argument structure of the verb by itself. Longobardi 1994 has argued that D performs this function; D transforms its complement NP into an argument. Essentially, DP encodes ‘referentiality’ but NP does not. Since the moved NPs in (19) and (20) are predicates, we predict them to reconstruct obligatorily. In copy theory terms, this means that their tail copy must not be deleted at LF. The same prediction does not hold of moved DPs, being ‘referential’ and being arguments they do not fall under the Heycock/Huang generalization. The tail copies of DP movement are free to delete (at least partially) at LF.

What does all this tell us about (19) and (20)? The Heycock/Huang generalization provides a test to determine whether reconstruction targets the gap position specifically. If the gap position truly induces the Condition C effects in (20), we expect Condition C to correlate only with NP-movement (i.e. with a resumptive pro and NP copy in gap position). Why? Because in (20), the lowest NP copy must not delete; if it did it would violate the Heycock/Huang generalization. Since the NP copy contains a name coreferent with a c-commanding pronoun, Condition C effects arise. But if the DP had moved instead, leaving no resumptive pro and no pronominal agreement on the verb, then the NP restriction of the tail copy of the moved DP should be able to delete, alleviating the Condition C effects. The acceptability of (21) bears out this prediction:

\textit{pro} and an NP-copy) cannot license parasitic gaps but DP-copies can (Darrow 1999).
The acceptability of (21) stands in sharp contrast to (20). Here, the entire DP of the nominal head phrase raises out of the gap position of the relative clause. Assuming that deletion applies to LF representations (either optionally or to effect a certain semantic interpretation, it does not matter which), the derivation as shown above is possible, where the NP complement of the DP tail copy has deleted.²⁹ Presumably, the D

²⁹ Safir 1999 develops an argument against precisely this kind of reasoning. He claims that the absence of Condition C effects under reconstruction is no argument for the absence (or partial deletion) of a copy, since the name inside the tail copy could simply have undergone “vehicle change”, transforming it into a pronoun. It is not obvious to me what predictive power vehicle change might have for the facts in Syrian: why should vehicle change be disallowed in relative clauses with resumptive pronouns but allowed for trace-type relative clauses? Safir explicitly distinguishes between ‘overt resumption’ and ‘covert introduction of a pronominal variable’ (618), but while this distinction arguably explains why trace-type (DP-movement) copies do not show Condition C effects, it does not capture why Condition C effects arise in the presence of resumptive pronouns.
trace in the tail copy cannot be deleted: if it were there would be no variable in the relative clause. Note that another derivation for (21) exists where none of the tail copy is deleted: this derivation will be ruled out by Condition C just like (20).

The Condition C reconstruction contrasts between DP-movement and NP-movement are further illustrated in (22).

*pro subject, relativized object with clitic:
(22) a. *aqtiraH [liqtiraaH ?’in nu ahmed1 kaan yby]2 [yalli pro1
suggested 1s DEF. suggestion C was C

Saaradu pro2 . . . tamaaman].
objected (3ms)3ms completely
“I made the suggestion that Ahmed1 was stupid that he2 objected to completely.”

*bpro subject, relativized object w/o clitic:
(22) b. aqtiraH [liqtiraaH ?’in nu ahmed1 kaan yby]2 [yalli pro1
suggested 1s DEF. suggestion C was C

Saarad t2 . . . tamaaman].
object (3ms) completely
“I made the suggestion that Ahmed1 was stupid that he2 objected to completely.”

The NP/DP-movement theory makes a different set of predictions about variable-binding reconstruction. Since the tail copy of NP movement may not delete due to the Heycock/Huang generalization, we predict a pronoun in the tail NP copy to be able to get a bound-variable reading by a c-commanding quantifier. Of course, we have already seen instances of this type of reconstruction in, for example, sentence (19)/(10a). And since the tail copy of DP movement need not undergo deletion, we predict variable-binding in DP movement cases, too. That is, we do not predict an asymmetry between reconstruction effects and clitic presence like we did for Condition C, since the criteria inducing variable-binding are positive, not negative. This prediction is borne out by the examples in (23): in each case, variable-binding is available.
*QP subject, relativized object with clitic:

(23) a. [[stuha'ii? 'aleyu  pro1 ]] [yalli [wala  taalib] maa-habbu pro2 ]
  DEF.investigation of.3ms C no student NEG-loved.3ms
  “The investigation of him1 that no student2 loved...”

*QP subject, relativized object w/o clitic:

b. [[stuha'ii? 'aleyu  pro1 ]] [yalli [wala  taalib] maa-habbu t1 ]
  DEF.investigation of.3ms C no student NEG-loved
  “The investigation of him1 that no student1 loved...”

We thus appear to have arrived at a solution to Problem #2. The head nominal of a relative clause in Syrian originates in the gap position, where it can undergo NP-movement to derive a resumptive pro variable or DP-movement to derive a trace variable. Condition C reconstruction effects arise when the tail copy of NP movement induces Condition C violations, but the NP restriction in the tail copy of DP movement can delete, escaping Condition C. Variable-binding reconstruction effects arise under both kinds of movement because in both DP and NP movement a derivation will exist with the NP restriction on D present. And since the presence or absence of the resumptive variable pro determines the presence or absence of Condition C reconstruction, we can safely say that it is the object gap position particularly that triggers reconstruction effects, and thus that the head nominal phrase of the relative clause originates in that gap position.30 The theory of NP and DP movement and the head-raising analysis thus provides a plausible answer

---

30 One might think to identify the ultimate site of reconstruction in relative clauses to a CliticP between VP and IP, as proposed by Aoun & Benmamoun 1998 for Lebanese Clitic Left-Dislocation (see fn 12). But it is unclear to me whether CliticP can—alone—deduce that Condition C and the agreement clitic on V obligatorily correlate. That is, suppose that the Syrian resumptive pronoun originated in Spec, CliticP: why do Condition C effects exist only when the clitic is present? Moreover, it is unclear why it is possible to “turn off” agreement in CliticP in Syrian: how can relative clauses be generated without agreement on V? Incorporating CliticP into the head-raising theory is possible in principle: the head nominal phrase could be base-generated in Spec, CliticP and the verb could raise to the Clitic head position to agree. However, it is still unclear how agreement can be turned off. The main empirical motivation for CliticP in Aoun & Benmamoun 1998 was a variable-binding reconstruction asymmetry between preverbal and postverbal subjects. The consultants I have interviewed so far have been extremely resistant to postverbal subjects in the highest CP in a relative clause, so it has not yet been possible to test whether the same asymmetry holds for Syrian Arabic.
to the problem of how reconstruction can be tied to resumptive pronoun positions. The next section offers a few suggestions for interpreting the LF structures that result from NP and DP movement.

5. One and a Half Proposals for a Semantics

With respect to the analysis developed so far, two types of structure must be provided by semantics. First, we have relative clauses derived by NP movement, where the NP restriction on the pro variable must be present (by the Heycock/Huang generalization). Second, we have relative clauses derived by DP movement, where the NP complement to the D head containing the t variable can be deleted optionally.

It might be possible to identify an interpretation procedure for the DP movement case with the more familiar relative clause domain of English. English relative clauses only allow traces as gaps. Suppose that a DP movement head-raising analysis were appropriate for English relative clauses. We would then predict for English exactly the reconstruction effects found in Syrian relatives derived by DP movement. In fact, it has been argued by a number of researchers (e.g. Merchant, 2000; Munn, 1994; Safir, 1999) that Condition C reconstruction effects are not found in English relative clauses. The examples in (24) appear to support this claim:

(24) a. I have a report on Bob1’s division that he1 won’t like.
    [Merchant 2000, fn. 5]

b. I read every report on Bob1’s division that he1 ever submitted.
    [Merchant 2000, fn. 5]

c. I read the report on Bob1’s division that he3 submitted yesterday. It was aw-
   ful!

The absence of Condition C effects in (24) correlates with the absence of Condition C in Syrian relatives. As for variable-binding reconstruction, Jim McCloskey (p.c.) has pointed out to me that these effects can be found in English relative clauses, as in the following cases:
These are the kinds of intrusions into his private life that [no president] should have to endure.

*The kinds of intrusions into his private life that [no president] should have to endure would surely sap his morale.

Feinstein and Boxer both ran strong campaigns...

However, Murdoch's papers published allegations about her private life that [neither candidate] could effectively counter.

*However, allegations about her private life that [neither candidate] could counter damaged her campaign.

The grammaticality of the (a) cases above but not the (b) ones patterns with the variable-binding reconstruction found in the Syrian examples (11a) and (11b). In sum, variable-binding and Condition C reconstruction effects in English parallel that of DP-movement relatives in Syrian. At this point, let us be content with reducing the issue of a semantics for Syrian Arabic DP movement-derived relative clauses to the issue of a semantics for English relative clauses.

As for Syrian relatives derived by NP movement, consider the Syrian Condition C violation (16a), repeated as (27) below:

,*αlPaadiya 'akaset [[ālhekem ? innu ahmed, kaan DEF.judge. f reversed.3fs DEF.decision C was

bāri?l2 [yall pro2 kaan habbu pro2]],

innocent C was liked.3ms

"The judge reversed the decision that Ahmed was innocent that he had liked."

The trick to interpreting a head-internal relative clause is to interpret the tail copy of the head, the copy inside the relative. That LF tail copy of the sentence in (27) has pro2 as its head D; both copies are spelled out below:
(28) Partial LF representation of (27):

\[ ([\text{\textipa{\textae}}k\text{\textipa{\textem}}m\text{\textipa{\textinnu}} \text{\textipa{\textahmed}}, \text{\textipa{\textkaan barii?}}],_2 \text{\textipa{\textyall}} \text{\textipa{\textpro}}, \text{\textipa{\textkaan \texthaabbu}} \text{\textipa{\textdecision}}, \text{\textipa{\textc}} \text{\textipa{\textw\texta\texti\textnt\texte\textcient}}, \text{\textipa{\textc}} \text{\textipa{\textw\texta\texti\textnt\texte\textliked}}, 3\text{\textms})

\[ ([\text{\textipa{\textpro}}, \text{\textipa{\text\textae}}k\text{\textipa{\textk\texthekm}}, \text{\textipa{\text\textinnu}} \text{\textipa{\textahmed}}, \text{\textipa{\text\textkaan barii?}}], \ldots \text{\textipa{\textdecision}}, \text{\textipa{\textc}} \text{\textipa{\textw\texta\texti\textnt\texte\textcient})}

In the spirit of work on head-internal interpretation by Grosu & Landman (1998) suppose that \textipa{\textpro} induces NP-movement of the head NP \textipa{\texthekm}, \textipa{\text\textinnu}, \textipa{\textahmed}, \textipa{\textkaan barii}? ‘decision that Ahmed was innocent’ because \textipa{\textpro} is lexically encoded as a function from \textipa{\text<e, t>} type functions to \textipa{\text<e, t>} type functions. It takes the denotation of its sister NP, the tail copy of the head of the relative clause, and yields that same denotation. Since this type of D is unavailable in English, NP-movement is unavailable in English head-internal relative clauses; we predict resumptive pronouns to be unavailable in English relative clauses. Once the D containing the \textipa{\textpro} variable in (27) and (28) denotes an \textipa{\text<e, t>, <e, t>} function, the interpretation procedure of the entire relative clause can proceed straightforwardly: the \textipa{\textpro} variable can function as the variable of the relative clause, and the upper copy of the relative clause head as the relative clause operator. These suggestions for a semantics are few and vague. They are intended merely to show that a plausible semantics can be given in principle for the LF structures posited in this paper.

6. Conclusions

Surprisingly enough, the copy theory can provide a fairly satisfactory account of Condition C and variable-binding reconstruction in Syrian Arabic relative clauses. We have addressed both the problem of how

\textsuperscript{31} One way to make sense of this claim is to say that there are two ways to extract a relative-clause-internal head in Syrian: either the D or the N of the DP head can, in principle, be generated with a [wh] feature. Borer 1999 suggests this type of base-generation option for the syntactic feature [definite]. When D is generated with a [wh] feature, a DP-movement derivation results (just like in English). When N is generated with a [wh] feature, the D head that takes that NP as its sister is an \textipa{\text<e, t>, <e, t>} function.

\textsuperscript{32} We can assume that the semantic function of the upper NP copy of the moved relative clause head is to initiate Predicate Abstraction over the \textipa{\textpro} variable (e.g. \textipa{\textpro} in 27 and 28). This is not so outlandish a proposal: it parallels Heim & Kratzer’s 1998 synctagmecomatic triggering of Predicate Abstraction by relative pronouns and by indices in English.
reconstruction effects come about at all and also the problem of how reconstruction can be tied to resumptive pronoun positions. Accepting the premises of the copy theory and a head-raising analysis of relative clauses for Syrian derives the Condition C and variable-binding reconstruction generalizations, the optionality between resumptives and traces in relative clauses, and the correlation between Condition C and resumptivity (given some assumptions about DP movement and predicate reconstruction).

Two points remain to be addressed concerning the pro-raising proposal of Aoun & Choueiri (1997). Recall the unacceptable Lebanese relative clauses involving extraction of adjuncts in (7), repeated here as (30):

(29) a. *ssababo yallioorªHt.oo.o.o.
   DEF.reason(ms)  C  left.1s.3ms
   “The reason that/why I left . . .” [Leb: A&C 41]

b. *sabab yalli yraht.o . .
   reason(ms)  C  left.1s.3ms
   “A reason that/why I left . . .” [Leb: A&C 86]

Aoun & Choueiri (1997) claim that (29) demonstrates an absolute ban on relativization of adjuncts. Such a ban clearly does not hold for Syrian Arabic.33 Witness the examples in (30):

(30) a. ssababo yallia darabet ahmed . .
   DEF.reason  C  hit.3fs
   “The reason that Layla hit Ahmed . . .”

b. sabab Layla Darabet ahmed . .
   reason  hit.3fs
   “A reason that Layla hit Ahmed . . .”

---

33 I will not try to answer why Lebanese might have such a restriction on the relativization of adjuncts. It does seem likely to me that a plausible answer might link up with the fact that Lebanese resumptives are obligatory whereas Syrian ones are optional. Note that if the NP-movement-type head-raising analysis is on the right track for Lebanese, we predict Condition C effects to hold under reconstruction in that dialect.
Aoun & Choueiri (1997) claim also that the pro-raising analysis deduces that relative clauses headed by indefinite nominals will not show reconstruction, since pro raises to check (among other features) [+definite]. Under Aoun & Choueiri’s (1997) analysis, indefinite nominals do not involve movement of pro, just base-generation. A similar analysis might, in principle, be extendable to NP- and DP-copies, but I will leave this issue open.34

Several important issues remain for the head-raising analysis. For one, we have given up the standard base-generation analysis of relative clause resumptives (exemplified by McCloskey, 1990), instead proposing that NP-movement derives the resumptive pro variables. And so, it is now entirely unclear how one might explain the fact that resumptive clitics appear inside islands in Syrian Arabic. Aoun & Choueiri (1997), who derive resumptive pronouns by pro-raising, claim that island-internal resumptives are of a different kind altogether: they are base-generated. That is one possible approach: another take might be that NP-movement is not island-sensitive (though why this might be is a total mystery).

Sandy Chung has pointed out to me that if the DP/NP-movement version of the head-raising analysis is correct for Syrian Arabic, we might predict many more types of reconstruction than just Condition C and variable-binding. For instance, an indefinite embedded inside the head nominal phrase should be able to bear narrow scope with respect to an element inside the relative clause. This element could be negation, a universal quantifier, or perhaps an intensional verb. Such predictions remain open at this time, awaiting further fieldwork.

---

34 It seems to me that the issue of what types of determiner can host a reconstructed nominal phrase should coincide with a theory about the ultimate landing site of the head nominal phrase. Kayne 1994 and Bianchi 2000 propose a structure for relative clauses where the relative CP is a sister to a relative determiner; the head nominal phrase under this theory sits in Spec, CP of the relative clause. It is quite difficult to find direct and obvious positive evidence in support of this theory, especially in languages without rich morphophonological material at the level of the relative CP. Nonetheless, future research into the head-raising analysis for Arabic will have to focus on the legitimacy of the Kayne/Bianchi relative clause-structure w/r/t particular aspects of Arabic syntax. It is my hope that investigation into the landing site of the head nominal phrase will shed some light on why indefinite nominals do not reconstruct.
http://www.usc.edu/dept/LAS/linguistics/semitic.
____________ & Elabbas Benmamoun. 1998. "Minimality, Reconstruction, and PF 
Linguistic Inquiry 30:4.621-642.
Berlin: Mouton de Gruyter.
Borer, Hagit. 1984. “Restrictive Relatives in Modern Hebrew”. Natural Languages 
and Linguistic Theory 2.219-260.
____________. 1999. “Deconstructing the Construct”. Beyond Principles and Pa-
rameters ed. by K. Johnson & I. Roberts, 43-89.
town Univ. Press.
from Building 20 ed. by Kenneth Hale & Samuel Jay Keyser, Cambridge: MIT 
Press.
____________. 1998. “Minimalist Inquiries, the Framework”. Ms., MIT. To ap-
pear in Step by step: Essays on Minimalist Syntax in honor of Howard Lasnik 
ed. by Roger Martin, David Michaels, & Juan Uriagereka, Cambridge: MIT 
Press.
Georgetown Univ. Press.
3.79-122.
Harvard University.
Engdahl, Elisabet. 1986. Constituent Questions: The syntax and semantics of ques-
Heim, Irene & Angelika Kratzer. 1998. Semantics in Generative Grammar. Cam-
26.4.547-570.
of SALT IV ed. by M. Harvey & L. Santelmann , 161-78. Cornell University.
Philosophy 22:2.117-84.


1. Introduction to the Problem


(1) jaa?-a al-awlaad-u
    came-3msg the-boys-nom

The subset of cases with which we will be concerned is that in which the plural subject is a conjoined NP, such as (2) and (3), from Mohammad (1990:97):

(2) a. qara?-a al-walad-u wa al-bint-u kitaab-a-n
    read-3msg.pft the-boy-nom and the-girl-nom book-acc-indef
    “The boy and the girl read a book.”

b. *qara?-aa al-walad-u wa al-bint-u kitaab-a-n
    read-dual.pft the-boy-nom and the-girl-nom book-acc-indef
    “The boy and the girl read [dual] a book.”

(3) a. qara?-at al-bint-u wa al-walad-u kitaab-a-n
    read-3fsg.pft the-girl-nom and the-boy-nom book-acc-indef
    “The girl and the boy read a book.”

b. *qara?-aa al-bint-u wa al-walad-u kitaab-a-n
    read-dual.pft the-girl-nom and the-boy-nom book-acc-indef
(2a) and (3a) exhibit the impoverished agreement characteristic of VSO sentences in SA (Cantarino 1975:96): the singular verb agrees with the first conjunct of the compound subject NP in gender but does not agree with the entire conjoined NP in number, since the latter is dual (as SVO order proves: \textit{al-waladu} \textit{wa al-bintu} \textit{xarajaa} “the boy and the girl left” exhibits full number agreement). If the verb is inflected for the dual to agree with the number of the entire conjunction, as in (2b) and (3b), the sentence becomes ungrammatical.

In representing the agreement relation in (2a, 3a), the null hypothesis would be that it holds between the verb and a conjunction of NPs, as in (4), omitting details:

\begin{equation}
V \ [NP \ C \ NP]
\end{equation}

Aoun, Benmamoun, and Sportiche (1994–hereafter ABS) challenge this intuitively plausible analysis. They observe that if a pre- or post-verbal conjoined NP and a plural verb, exhibiting full agreement, occur with an expression (such as a quantifier or a reciprocal), which requires a semantically plural associate or antecedent, the sentence is grammatical. However, if the NP is post-verbal and the verb singular—an impoverished agreement structure—in the context of such an expression, the sentence becomes ungrammatical (211). The sentences in (5), from Lebanese Arabic (LA), illustrate the contrast with the reciprocal anaphor \textit{ba\'dun} “each other” (214):

\begin{equation}
(5)\quad\begin{array}{ll}
a. & \text{\textit{kariim} w marwaan bi\'hjb-o ba\'dun} \\
   & \text{Kariim and Marwaan love-3pl some} \\
   & \text{“Kariim and Marwaan love each other.”} \\
\hline
b. & \text{bi\'hjb-o kariim w marwaan ba\'dun} \\
   & \text{love-3pl kariim and marwaan some} \\
   & \text{“Kariim and Marwaan love each other.”} \\
\hline
c. & \text{*bi\'hjb kariim w marwaan ba\'dun} \\
   & \text{love kariim and marwaan some} \\
   & \text{“Kariim and Marwaan love each other.”} \\
\end{array}
\end{equation}

In (5a, b), \textit{kariim} \textit{wa marwaan} antecedes \textit{ba\'dun}, indicating that it is a conjunction of NPs; if it were not, it would not be semantically
plural and so would be ineligible to antecede the reciprocal. If *kariim wa marwaan* were likewise a conjunction of NPs in (5c), it should also be able to antecede *ba’dun*. The reason it cannot, ABS argue, is that it is not a conjunction of NPs but of clauses (Inflection Phrases, or IPs), (the NPs of) which do not form a semantic plurality. Schematically, the structure of such a clausal conjunction would be as in (6):

\[(6) \ V_i \left[ [\text{IP} \ [\text{NP}] \ldots t_i] \ C \ [\text{IP} \ [\text{NP}] \ldots e_i] \right] \]

$V_i$ has been raised out of the first conjunct; $e_i$ is a null verb anaphoric on $V_i$. ABS (1994:217) offer (7), from Moroccan Arabic, as a concrete example of (6):

\[(7) \ n’as [\text{IP} \ kariim \ldots t_j \ldots t_i \ldots] \ w [\text{IP} \ marwaan \ldots e_j \ldots t_i \ldots] \ [fi-l-bayt], \]

"Kariim and Marwaan slept in the bedroom."

The PP *fi-l-bayt* is assumed to have been right node raised out of the first IP to adjoin to the root IP dominating both conjuncts.

The diagnostic ABS offer for NP conjunction-co-occurrence with an expression that requires a semantically plural antecedent-predicts that sentences like (5c) will be ungrammatical-sentences, that is, with [1] VSO word order, [2] impoverished number agreement, and [3] a conjoined NP anteceding an anaphor. The prediction is apparently incorrect for Form VI verbs in SA. Form VI verbs that derive from Form III and have plural subjects are understood reciprocally (Cuvaly-Haak, 1997:104), as in these examples:

\[(8) \ a. \ tua-kataab-naa \]

recip-write-1pl.pft

"We corresponded with each other."

b. tua-qaabal-uu

recip-meet-3pl.pft

"They met each other."

c. tua-qaatal-tunna

recip-fight-2fpl.pft

"You [fem pl] fought each other."

I will argue below that the prefix *tua-* is a reciprocal complement to Form VI verbs. Assuming this to be the case, we expect the analog to
(5c) with a Form VI verb to be ungrammatical. But the expectation is not borne out; (9) is well formed in SA:

(9) ta-qaabal-a kariim-u-n wa marwaan-u-n
    recip-meet-3msg.pft Kariim-nom-indef and Marwaan-nom-indef

"Kariim and Marwaan met [each other]."

If kariimun wa marwaanun antecedes a reciprocal, then it must be a semantic plurality, as it is in (5a, b). Nonetheless, the post-verbal NP enters into impoverished agreement with the masculine singular taqaabala, contrary to what the ungrammaticality of its analog (5c) predicts.

The grammaticality of (9) raises a second problem. ABS (1994) cite (restrictive) relative clauses as another line of evidence against the phrasal analysis of conjoined singular post-verbal NPs in simple sentences. This argument is also based on the requirement that NP conjunctions be a semantic plurality. A compound NP that is post-modified by a relative clause must be a single constituent; therefore, it ought to exclude agreement between only the verb external to the relative clause and the first conjunct in VSO order. This expectation is borne out in LA, as the following data, with both full and partial agreement, show (ABS, 1994:213):

(10) a. raah-o l-walad w l-m'allim yalliuulta;-o 'ala l-mdrase
    went-3pl the-boy and the-teacher who met-3pl on the-school
    "The boy and the teacher who met went to school."

b. *raah l-walad w l-m'allim yalliuulta;-o 'ala l-mdrase
    went the-boy and the-teacher who met-3pl on the-school
    "The boy and the teacher who met went to school."

The ungrammaticality of (10b) is consistent with that of (5c); in both cases, there is an element following the compound subject that requires a phrasal subject, and this requirement, ABS (1994) argue, is inconsistent with partial agreement. In the light of the grammaticality of (9), we therefore expect that a relative clause with a Form VI verb will be grammatical as well. This prediction is not straightforwardly confirmed, however; judgments on the following sentence were divided:
The teacher (f) and the student who met [each other] left.

Four of my consultants (and an anonymous reviewer) found (11) grammatical, while three found it ungrammatical. By parity of reasoning, if (5c) and (10b) are both ungrammatical in LA, we would predict that (9) and (11) would be grammatical in SA, but that is not clearly so. An account of the variation in judgments of (11) will be offered in 5.2.

The analysis to follow will be set in the context of two global assumptions. First, I will argue for a phrasal rather than a clausal account of NP conjunction in verbal sentences. Two reasons for pursuing this course are implicit in what has been said already. Since (9) contrasts with (5c) in being grammatical, the clausal analysis that explains the grammaticality of the latter is inappropriate for the former. Evidently, karīmun wa marwaanun in (9) is a semantic plurality, since it antecedes a reciprocal, but it is nonetheless able to enter into impoverished agreement. This conclusion points to an analysis that is methodologically parallel to ABS’s. They postulate a conjunction of clauses underlying NP conjunction so that (the subject NP of) the first conjunct can enter into an agreement relation with the verb (210). The phrasal analog to this would be for the first conjunct of a compound NP to agree with the verb preceding it. This is essentially the tack that I will follow (subject to the qualification in 3.2 concerning null expletives).

The second reason for pursuing a phrasal analysis of NP conjunction follows from the existence of both full and impoverished agreement in verbal sentences in LA. The contrast between (5b, c) plausibly reduces to a parameterized morphological property (Chomsky, 1993:31) within LA, namely “rich” VSO agreement in (5b) and “poor” VSO agreement in (5c). (Cf. Bolotin, 1995:22, who contrasts SA and Berber along this dimension.) This analysis does not mean that poor agreement does not exist in LA; alongside of the contrast between (5b, c), we find the parallelism in (12):
(12) a. raah kariim w marwaan
    left-3msg kariim and marwaan
    “Kariim and Marwaan left.”

b. raah-o kariim w marwaan
    left-3msg kariim and marwaan
    “Kariim and Marwaan left.”

(12) underlines the fact that the clausal analysis of NP conjunction depends on two facts about LA: [1] the ungrammaticality of poor VSO agreement holds only with expressions requiring a plural antecedent like ba’sidun; [2] rich agreement exists in the same context. But neither condition holds of SA, assuming the reciprocal complement analysis for Form VI: poor VSO agreement with a conjoined subject antecedent to a reciprocal is grammatical, as in (9), and rich VSO is unavailable (for lexical NPs); the following counterpart to (1), for example, is ungrammatical:

(13) *jaa-uu al-awlaad-u
    came-3mpl the-boys-nom

These facts call the tenability of the clausal analysis into question for SA, and favor the phrasal analysis as the null hypothesis. I will offer two further arguments for a phrasal analysis in Section 3.4.

The other global assumption underlying what follows is that the grammaticality patterns in (9) and (11) can be explained by appeal to checking theory in the minimalist program. Specifically, the concepts of feature strength and feature interpretability will figure prominently in the discussion, along with the internal structure of nominal and pronominal NPs. The paper will therefore be organized as follows. Section 2 will lay out the general framework of assumptions, and Section 3 will explain and justify background assumptions regarding the Arabic data. Section 4 will explicate the relevant aspects of checking and feature theory, and these will be applied to the Arabic data in Section 5.
2. **General Assumptions**

In this section, I briefly set out the assumptions that I make about clause structure, NP structure, and the interpretation of indices in compound NPs.

2.1 Clause structure

I make three assumptions about clause structure in the analysis. First, I assume a fairly standard X-bar analysis of the clause as a Tense Phrase (TP), with T[ense]₀ as its head, the element in subject position as its specifier, and VP as its complement. I employ the familiar bar levels and category symbols instead of a bare phrase structure analysis (Chomsky 1994, 1995) for clarity of exposition. TP and its constituents are formed by iteration of the operation Merge, which combines words into larger units (Chomsky, 1995:226). Second, I provisionally accept Chomsky’s (1995) proposal that the agreement heads AgrS and AgrO and their projections should be excised from the inventory of functional categories (349-55), contra, e.g., Simpson & Wu (2002). In place of Agr as a mechanism for checking agreement relations is the concept of a sub label, to be explained below. Third, I adopt the VP-internal hypothesis for subjects (Speas, 1990; Shlonksy, 1997 for the Semitic languages in general, and Plunkett, 1993 for Arabic in particular). On this assumption, subjects originating in the specifier position of the verb phrase, [Spec, VP], may raise to subject position—that is, to the specifier of Tense Phrase, [Spec, TP]. This assumption underlies certain of the movement analyses below.

2.2 NP structure

I make two assumptions concerning phrase structure. The first is to adopt the DP analysis of NPs (Radford, 1993; Ritter, 1991 for Hebrew, and, for Arabic, Benmamoun, 1993; Kaplan, 1993; LeTourneau, 1995; Hasan, 1995), under which nominal phrases are headed by D[eterminer] (which may be phonetically null) taking an NP complement.

Second, I adopt the anti-symmetry condition of Kayne (1994) as it applies to coordinate structures. Kayne’s basic intuition is that the linear ordering of terminal symbols must match the ordering of non-terminal elements imposed by asymmetric c-command (which holds for
two non terminals X and Y if X c-commands Y but not conversely). This requirement is stated as the Linear Correspondence Axiom (LCA):

\[ \text{Linear Correspondence Axiom (1994:6)} \]
\[ d(A) \text{ is a linear ordering of } T. \]

\[ T \text{ is a set of terminal symbols in a tree; } A \text{ is the set of ordered pairs of non terminal symbols } \{X_i, Y_i\} \text{ such that } X_i \text{ asymmetrically c-commands } Y_i; \text{ and } d(A) \text{ is the set of dominance relations between the non terminal elements in } A \text{ and the terminals they dominate. The LCA states, in effect, that the ordered pairs of non-terminals in } A \text{ impose a linear order of the terminals in } T, \text{ so that the latter matches the former. One consequence of the LCA which Kayne deduces is that coordinate structures containing a coordinating conjunction are headed by the conjunction, with the first conjunct asymmetrically c-commanding both the conjunction and the second conjunct (1994:11-12), as proposed in Munn (1992) and seconded by ABS (1994:207). If they were not so headed, a flat structure like (4) above, in which the conjuncts symmetrically c-command each other, would be permitted, in violation of the LCA. I will adopt a hierarchical analysis of DPs below. The LCA will also bear on the derivation of VSO word order.} \]

2.3 Fusion of anaphoric indices

Fiengo & May (1994:38-39) develop a theory of indexing for coreferential and non-coreferential expressions that includes the proposal that a conjunction of NPs has two sets of indices: one for each conjunct and one for the entire conjunction. The latter is a fusion of the indices of each conjunct, and this fused index is also assigned to a pronoun that takes the conjunction as its antecedent. Thus, Jack and Jill think that they should carry each other up the hill would have the indexical representation (15):

\[ \text{(15) } [\text{Jack}_1 \text{ and } \text{Jill}_2]_{*1*2} \text{ think that they}_{*1*2} \text{ should carry each other}_{*1*2} \text{ up the hill} \]

I will make use of this mechanism in discussing the LF representation of sentences with Form VI verbs in 3.4.2.
3. **Specific Assumptions concerning Standard Arabic**

In this section, I will explain and, where necessary, justify the assumptions I make about VSO word order, impoverished subject-verb agreement, and reciprocal arguments in Form VI in SA.

3.1 VSO word order

A continuing controversy in Arabic syntactic research has been the choice of a basic word order (see Mohammad, 1990 for a review of earlier proposals and his own arguments for SVO, and Fassi Fehri, 1993 for counterarguments). The adoption of the VP-internal hypothesis and the LCA as sketched in 2.1 and 2.2 derives SVO as the basic word order for SA. A corollary of the LCA is that specifiers universally precede their heads (Kayne, 1994:33-36). If subjects originate in [Spec, VP], as the VP-internal hypothesis asserts, it follows that subjects precede verbs in initial syntactic structures.

Underlying SVO order is generated by Merge; VSO word order arises from applying the operation Move. Two types of movement are licensed: adjunction of an X^0 to a head and substitution of an XP (a maximal projection) for a specifier. VSO order is an instance of adjunction: V^0 adjoins to T^0. Derived SVO order is the result of substitution: the DP in [Spec, VP] raises to [Spec, TP]. Move is in both cases “morphologically driven”: elements raise to check their inflectional features in the domain of a functional head—for our purposes, T^0. Feature movement, introduced in Section 4, is a special case of adjunction; features may not undergo substitution, since specifiers must be maximal projections (Chomsky 1995:270).

3.2 Impoverished agreement in VSO

Accounting for VSO word order by verb raising leaves the mechanism of partial agreement unexplained. Two proposals are relevant to what follows. Bolotin (1995) develops a suggestion of Chomsky (1993:31) to account for SVO and VSO orders in Arabic in terms of strong and weak agreement features. Since I am assuming that there are no Agr projections (2.1), the specifics of Bolotin’s proposal will not figure in the analysis, but the core idea of feature strength in her article is central to it.

The other relevant proposal is the influential null expletive pronoun analysis of Mohammad (1990, 1998), seconded by van Gelderen (1996,
In essence, their proposal is that there is in SA VSO structures an expletive pro subject in [Spec, TP] with a singular number feature that agrees with the verb, which has adjoined to T. Schematically, the structure would be as in (16):

(16) \[ TP \{ Spec \{ pro \} \{ T' \{ T \{ \{ V \{ \{ t \{ (NP) \} \} \} \} \} \} \} \{ V.P \{ Spec \{ DP \} \{ V' \{ t \{ (NP) \} \} \} \} \} \] \]

Note that the subject remains in situ in [Spec, VP]. It cannot raise to [Spec, TP] to check its inflectional features, since that position is occupied by the expletive. We will resolve this matter in 5.1.1. Because the subject is in [Spec, VP], it is not in an agreement relation with the verb after the verb adjoins to Tense. The configuration for subject-verb agreement is a specifier-head relation (Chomsky, 1986a:24, 1993:8), with the subject in the specifier and the verb either being the head of the same minimal maximal projection or else adjoined to it. Spec-head agreement so conceived holds in SVO structures, with the subject DP in [Spec, VP] before verb raising or [Spec, TP] after. Spec-head agreement between the verb and the subject DP is not possible if the subject DP remains in situ and the verb raises. It is possible, however, with the expletive pro in [Spec, TP], which, because it is singular, yields impoverished number agreement with the verb. As van Gelderen (1998) notes, the analysis does not account for gender agreement; we will also address that in 5.1.1.

3.3 Reciprocal arguments in Form VI

In this section, I review evidence for the claim that Form VI verbs take a reciprocal anaphor as complement and then discuss objections to this claim.

3.3.1 Evidence for \textit{ta-} as a reciprocal argument

The first piece of evidence in favor of the reciprocal argument analysis is the interpretation of Form VI verbs in comparison with their Form III sources. The glosses for the sentences in (8) suggest that the verbs in these representative examples have a reciprocal argument, realized prefixally as \textit{ta-}. That the prefix is the morphological realization of a featurally stipulated argument is evident from the fact that Form III verbs without the prefix do not permit, let alone require, a reciprocal interpretation. The complements of Form III verbs must
simply satisfy selectional restrictions consonant with the lexical semantics of the verb in question, as in (17):

(17) a. kaatab-naa shadiq-a-naa
     write-1pl.pft friend-acc-1pl
     “We corresponded (with) our friend.”

  b. qaabal-uu umm-a-ka
     meet-3pl.pft mother-acc-2msg
     “They met your mother.”

  c. qaatal-tunna al-jaysh-a al-masriyy-a
     fight-2fpl.pft the-army-acc the-Egyptian-acc
     “You fought the Egyptian army.”

That the prefix is a reciprocal argument is confirmed not only by the glosses in (8) but also by the fact that ta- realizes a reflexive argument in several derived verbal measures: verbs like taqaddama “present oneself”, tafallata “extricate oneself”, and tahabbaba “ingratiate oneself” in Form V (Wehr-Cowan, 1979:876, 841, 179); middle verbs like iyatasala “wash (oneself)” and irtaqaa “raise (oneself)” in Form VIII; causative-reflexive verbs like istahamma “bathe (oneself)” in Form X; the second measure of quadriliteral verbs like tapamraka “to present oneself as an American” or tafaalsa “to present oneself as a philosopher”; and occasionally even in Form VI verbs, as in taraa?aa, one meaning of which is “show oneself, appear” [Wehr-Cowan, 1979:369]). Insofar as reflexives and reciprocals are both anaphors with the feature specification [-pronominal, +anaphor], (Chomsky 1982:78-79), it is not surprising that a single affix realizes both.

A second piece of evidence to support the contention that the prefix realizes a reciprocal argument in the syntax comes from operator movement in Form III verbs and the Form VI verbs derived from them. Compare (18) and (19):

(18) a. kaatab-naa shadiq-a-naa
     write-1pl.pft friend-acc-1pl
     “We corresponded (with) our friend.”

  b. man kaatab-naa?
     who write-1pl.pft
     “Whom did we correspond (with)?”
(19)  a.  ta-kaatab-naa
    recip-write-1pl.pft
    “We corresponded (with) [each other].”

b.  *man  ta-kaatab-naa?
    who  recip-write-1pl.pft
    “Whom did we correspond (with) [each other]?”

(18b) is grammatical because *kaatab- does not select a reciprocal argument as its complement, so man “who” is free to undergo first Merge (forming a head-complement construction with *kaatab-naa under V’) and then Move (to reach [Spec, CP] (the specifier of Complementizer Phrase). (19b) is ungrammatical, evidently because *takaatab- does select a complement, so that Merge of the verb with man is either blocked (say, under binary branching, which prohibits a head from taking two complements) or else yields a non-convergent derivation because merger results in a q-discharge violation, with one q-role being assigned to two arguments. As a secondary effect, Move is also blocked. As in (19a), the argument blocking man is understood reciprocally. Additional arguments in support of a reciprocal argument as a complement to Form VI verbs can be found in Letourneau (1998:93-99).

Before we proceed, a categorical clarification is in order. I have spoken of ta- as realizing in the morphology an argument in the syntax. Extending this analysis, we might take not just the feature complex realized by ta- but ta- itself as a syntactic argument, specifically a clitic, which would be a DP in its q-position but a D 0 after raising to T (Chomsky, 1995:249). This is an attractive option in that it can presumably be assimilated to incorporation as X 0 movement (Baker 1988). The D 0 would subsequently be realized as an affix in the morphology. The clitic analysis aids in clarity of exposition, and I will adopt it here. (9) would therefore have the abbreviated representation (20) after Merge and before Move, with fused indices as per 2.3.


3.3.2 Counterarguments concerning ta as a reciprocal argument

The claim that a phrasal analysis of DP conjunction with Form VI verbs is correct depends crucially on the premise that there is, in fact, a
reciprocal complement to these verbs. I now discuss and answer objections that may be raised against this claim.

The first objection pertains to sentences like (11), where the conjoined DP enters into an overt number agreement relation with the pronoun subject and predicate of the relative clause that postmodifies it. In these circumstances, it may be asked whether the semantic plurality of the conjoined DP depends on its full agreement with the following dual verb (*taqaabala*) rather than its antecedence of a reciprocal object to the verb. If so, the motivation for postulating a reciprocal argument is diminished if not eliminated.

This argument assumes that one type of agreement relation (subject-verb) may obviate the need for another (pronoun-antecedent), an assumption that in turn presupposes that the latter agreement relation is dependent on the former. That presupposition is evidently incorrect: subject-verb number agreement and pronoun-antecedent agreement are independent. Full number agreement may exist independently of antecedence of a pronoun, as in *al-rijaalu wasaluu ʿamsi* “the men arrived yesterday.” Conversely, antecedence of a pronoun is independent of number agreement with a verb, as in *al-baytu, istaraahu zaydun* “The house, Zayd bought it” (from Al-Shorafat, 1999:11), in which subject-verb agreement between *ištaraa* and *zaydun* is independent of pronoun-antecedent agreement (including number agreement) between the topic DP *al-baytu* and the resumptive pronoun *-hu*. As a corollary of this, number agreement in a relative clause does not in fact exclude the inadmissibility of pronoun-antecedent agreement between a DP and a pronominal. Consider (21), discussed in LeTourneau (1998:98):

(21) a. *haaḥulaaʔi hum al-rijaal-u, ʿalla-δ-iina, pro,
   these.m.pl thay the-men-NOM who.m.pl
   ya-ta-qaabaluuna
   m-recip-meet-pl.ind
   “These are the men who meet [each other].”

b. *haaḥulaaʔi hum al-rijaal-u, ʿalla-δ-iina, pro,
   these.m.pl thay the-men-nom who.m.pl
   ya-ta-qaabal-uuna-humi
   m-recip-meet-pl.ind-them
   “These are the men who meet [each other] them.”
Both sentences in (21) exhibit number agreement among \((al-rijaalu, Palladina, pro, and yataqaabaluuna)\). If subject-verb agreement obviated the need for pronoun-antecedent agreement, then both sentences should be grammatical, because subject-verb agreement obtains in both. That (21b) is ungrammatical shows that no such dependency between the two agreement relations holds. In that sentence, pronoun-antecedent agreement between \(al-rijaalu\) and the object suffix \(-hum\) is barred due to a Condition B violation of the binding theory: \(-hum\) cannot be bound by \(pro\), since they are in the same minimal binding domain (the relative clause) and so cannot be coindexed. Only if the direct object of the verb is the reciprocal prefix \(ta-\), as in (21a), does pronoun-antecedent agreement between \(al-rijaalu\) and the direct object as well as number agreement between \(pro\) and \(yataqaabaluuna\) hold, because the latter configuration satisfies Condition A (requiring an anaphor, here \(ta-\), to be bound in its minimal binding domain—here, by \(pro\)). Whether or not agreement between a pronoun and a potential antecedent is licit depends on the binding theory, not on subject-verb agreement. The two operate independently, so the latter does not obviate the former.

A second objection to treating \(ta-\) as a syntactic argument is that at least some Form VI verbs, for example, \(tagaatal-\) “fight”, may be understood intransitively, on analogy with LA \(lta-\) “meet”, and so cannot have a complement. But \(lta-\) is understood reciprocally (cf. [10]), so it is not strictly correct to speak of it as intransitive, a term better reserved for unaccusative and unergative verbs that take exactly one argument (Marantz, 1984:35; Haegeman, 1994:329). The reciprocal interpretation of \(lta-\) suggests that it too takes a reciprocal argument, a conclusion supported by the following SA contrast:

\[
\begin{align*}
(22) \quad &a. \quad laqi-a \ kariim-un \ wa \ marwaan-u-n \\
&\quad \text{met.3sg Kariim-nom-indef and M arwaan-nom-indef} \\
&\quad \text{mudiira-hum} \\
&\quad \text{director-acc-their} \\
&\quad \text{“Kariim and M arwaan met their director.”} \\
&b. \quad man laqi-a \ kariim-u-n \ wa \ marwaan-u-n? \\
&\quad \text{who met.3sg Kariim-nom-indef and M arwaan-nom-indef} \\
&\quad \text{“W hom did Kariim and M arwaan meet?”}
\end{align*}
\]
(23) a. itlaqaa kariim-u-n wa marwaan-u-n
met.recip.3sg Kariim-nom-indef and Marwaan-nom-indef
"Kariim and Marwaan met [each other]."

b. *man itlaqaa kariim-u-n wa marwaan-u-n?
who met.recip.3sg Kariim-nom-indef and Marwaan-nom-indef
"Whom did Kariim and Marwaan meet [each other]?"

The form laqiya is the simple transitive Form I from which Form VIII itlaqaa, evidently cognate with LA *ita?, is derived. The grammaticality patterns in (22) and (23) exactly parallel those in (18) and (19) for Form VI takaatab- derived from Form III kaatab-. This parallelism refutes the counterargument that certain Form VI verbs are intransitive and so cannot take a reciprocal complement.

The third objection to *ta- as a syntactic argument has, at first sight, more weight. If the conjoined DP subject of a Form VI verb is a semantic plurality, then it should be able to be modified by the kinds of quantifiers that ABS cite as diagnostic for semantic plurality in DP conjunctions. This is, however, not generally possible. In the examples that follow, nunation is omitted, and sentences without asterisks are those that were accepted by one speaker but rejected by another:

(24) a. ta-jaa9ab-a kariim wa salwa fii nafs-i al-waqt-i
recip-attract-3msg Kariim and Salwa in same-gen the-time-gen
"Kariim and Salwa attracted [each other] at the same time."

b. ta-qaarab-a kariim wa marwaan wa ?ahmad kull-u-hum
recip-approach-3msg Kariim and Marwaan and Ahmad all-nom-them
"Kariim and Marwaan and Ahmad all approached [each other]."

c. *ta-ta-maa?al-u kariim-a wa marwaan al-iqnaani
f-recip-resemble-3sg.impf Kariim-f and Marwaan the-two
"Kariima and Marwaan both resemble [each other]."

d. *ta-raa?aa kariim wa marwaan kull-u waahid-i-n
recip-see.3msg.pft Kariim and Marwaan each-gen one-gen-indef
"Kariim and Marwaan each saw [each other]."

e. *ta-qastal-a kariim wa marwaan sawaa?-a-n
recip-kill-3msg Kariim and Marwaan together-acc-indef
"Kariim and Marwaan fought [each other] together."
Modifiers of plurality such as kulluhum “all (of) them”, aliqaani “both”, kullu wahidin “each one”, and sawaan “together” are disallowed either in a majority of the examples or in all of them, depending on the speaker consulted. This might be taken to imply that the conjoined NP is not semantically plural and, consequently, that it does not antecede a null reciprocal. But this does not seem to be the most natural interpretation of the facts, even if we ignore the divided judgments about (24a, b). A more plausible explanation for the unacceptability of (24c-e) seems to be that the modifiers are redundant because of the reciprocal, as they are in the English glosses, which are also (if more marginally) unacceptable, presumably for the same reason. No such redundancy arises in the examples that ABS (1994) provide as diagnostics for clausal conjunction. This is the expected result in that the examples in ABS with indicators of plurality are not understood reciprocally. (It may be significant that the only one that is so understood, llta:o “they met [each other],” does not include an expression that demands semantic plurality of the subject DP.) Redundancy in indicating plurality can then only occur in the context of a reciprocal. Thus, the patterns in (24) confirm rather than disconfirm the hypothesis that such reciprocals are present and syntactically active.

In summary, none of the counterarguments against postulating a reciprocal complement to Form VI verbs afford sufficient reason to reject the hypothesis. The objection that full number agreement between a conjoined DP and a following verb obviates the need to assume pronoun-antecedent agreement between the DP conjunction and ta-fails because the two agreement patterns operate independently. The objection that certain Form VI verbs are in fact intransitives that cannot take complements is falsified by the parallelism between takatab- and iltaqaa, since the latter is not truly intransitive but rather reciprocal. Finally, the objection that conjoined DPs are not semantically plural (and so do not antecede reciprocals) because they cannot take the range of plurality markers ABS present as diagnostic for phrasal rather than clausal conjunction is also faulty. Unlike the underived verbs that ABS examine, such associates of plural arguments are redundant with Form VI verbs. Such redundancy is possible only if those verbs do take reciprocal arguments. I conclude, then, that such reciprocals do in fact exist, as represented in (20).
3.4 Phrasal vs. clausal conjunction

In Section 1, I presented two arguments for a phrasal vs. a clausal analysis of VSO sentences with Form VI verbs. In this section, I offer two additional arguments.

3.4.1 Plural subject pronouns in VSO

A third reason for preferring a phrasal analysis is based on the distribution of plural pronouns in VSO sentences. Sentences with VSO order cannot take pronoun subjects that differ from the verb in number; contrast (25a) (from Fassi Fehri 1993:107) and (25b) (modeled on ABS 1994:205, citing Bahloul and Harbert 1992):

(25) a. *naam-at hunna
    sleep-fsg.pft they.f
    “They slept.”
 b. nim-na
    sleep-fpl.pft we.pl
    “We slept.”

This fact is compatible with a clausal analysis of DP conjunction in that a pronoun cannot replace a conjunction of clauses; that is, the following substitution pair is impossible:

    came-3msg the-boy-nom and the-girl-nom
    “The boy and the girl came.”
 b. *jaai?-a hum
    came-3msg they.m
    “They came.”

However, this compatibility does not of itself confirm a clausal analysis, because pairs like (27), in which conjunction plays no role, are also impossible:

(27) a. jaa?-a al-awlaad-u
    came-3msg the-boys-nom
    “The boys came.”
 b. *jaai?-a hum
    came-3ms they.m
    “They came.”
(27) shows that the ungrammaticality of *jaa?a hum need not be attributed to the fact that it illicitly substitutes for a conjunction of clauses in (26). In view of (26) and (27), there is no obstacle to positing a substitution pair like (28), with the pronoun replacing a conjoined DP rather than conjoined clauses:

(28) a. jaa?-at umm-u-n wa ?ibnat-u-haa
    came-3fsg mother-nom-indef and daughter-nom-her
    “A mother and her daughter came.”

    b. *jaa?-at hunna
    came-3fsg they.f
    “They came.”

The fact that subject personal pronouns are barred irrespective of whether they substitute for conjoined or nonconjoined DPs, on the one hand, or for phrases or clauses, on the other, allows for the possibility that DP conjunction is in fact phrasal rather than clausal.

3.4.2 Indexing and binding with Form VI

The fourth reason for preferring a phrasal to a clausal analysis of VSO with Form VI verbs concerns the binding relation between the reciprocal and its conjoined DP antecedent at LF. Under the phrasal analysis, the LF representation for (9), after Move has applied, is (29):

(29) \[TP \{\text{pro} \} \{T qaabal-ak\} \{V P \{\text{DP kariim(un)i wa marwaan(un)j}i* \} \}

Move adjoins the verbal root qaabal- to T⁰; in this position, it enters into spec-head agreement with the expletive subject in [Spec, TP]. The conjoined DP subject remains in situ in [Spec, VP], where it serves as the antecedent of the reciprocal affix ta-, which is the complement of qaabal- before Move raises the verb (cf. [20]). The compound DP and the reciprocal bear the same fused index under the indexing fusion convention introduced in Section 2.3. Thus, the entire DP conjunction antecedes the reciprocal pronoun, and the sentence bears the required reciprocal interpretation. The indexing configuration satisfies Principle A of the binding theory: the reciprocal is bound in the minimal domain D containing both the reciprocal and its antecedent.
Suppose now that the LF representation for (9) is instead (30), assuming the clausal analysis:

\[
(30) \text{ qaabal-} a, [\text{TP kariim(un), ta-}] \text{ \ qaridim (un)i \ met-3msg kariim \ recip and marwan (un)j \ met-3msg marwan \ recip }
\text{“Kariim met himself and Marwaan met himself.”}
\]

In this representation, the anaphoric ta- must be construed reflexively rather than reciprocally, because it occurs in each clause and so takes either kariimun or marwaanun as its antecedent. (If we insist that ta- be construed reciprocally because it is the argument of Form VI verb, the interpretation is gibberish: “Kariim met each other and Marwaan met each other.”) Since the antecedents are in conjoined clauses rather than a conjoined DP, the disjoint indexes cannot be fused; therefore, the two tokens of the anaphor bear distinct indexes, and Principle A of the binding theory is satisfied for each one individually, under a reflexive interpretation, but not for a single reciprocal argument with a fused index. The resulting interpretation, while indexically and perhaps even semantically coherent, does not accord with the meaning of taqaabala. We cannot avoid this result, as ABS do for filbayt in (7), by assuming a right node raising structure in which ta- has scope over the two TP conjuncts. If that were the case, the anaphor would not be bound in its minimal domain (if that notion can even be defined here). I conclude that a phrasal analysis for Form VI verbs in SA is preferable to a clausal one.

4. Feature Theory: Strength, Checking, and Interpretability

The substitution pairs in (26)-(28) show that there is a general prohibition against dual or plural subject pronouns following singular verbs, in contrast to dual or plural DPs, which may do so. Since both pronouns and lexical DPs may be either dual or plural, this distributional difference between them cannot reside in their semantic plurality. What other property might account for this difference? I will argue that the distributional differences in the (a) and (b) sentences of (26)-(28) and the grammaticality contrast between (9) and (11) can be accounted for on the basis of [1] feature strength, [2] feature interpretability, and [3] different internal structure in lexical and pronominal DPs. Before proceeding with the analysis, I offer in this section a sketch of feature strength and interpretability, in the context
of checking theory. I then consider a simple sentence like (9) and its analog with a plural subject pronoun in 5.1 and a relative clause like (11) in 5.2.

Lexical and non-lexical (functional) categories enter a derivation through Merge. These categories consist of sets of formal features such as category, tense, and number, known as f-features, some of which must be checked in order for the derivation to converge. Features are checked by raising to a certain domain of a head (called its checking domain) under Move. The category to which a feature raises is called the target, adapting earlier usage.

A feature may raise either to check a feature of the target or to be checked itself (Chomsky 1995:266). Checking a feature normally entails deleting it, meaning that it remains accessible to further computation but invisible at LF (Chomsky 1995:280). Features of targets are either strong or weak. Strong features force overt movement of a full category (that is, a word rather than only one of its features) in violation of the economy condition Procrastinate, which demands that movement be delayed until LF and so be covert rather than overt if possible. Weak features satisfy Procrastinate, licensing movement covertly at LF instead of in the overt syntax, leaving the word of which a moved feature is a part in situ (Chomsky, 1993:30; 1995:266, 272). Feature strength is defined as in (31):

(31) If $F$ is a strong feature, then $F$ is a feature of a nonsubstantive category and $F$ is checked by a categorical feature.

“Non-substantive category” is a synonym for ‘functional category’ (cf. Chomsky, 1995:342), particularly T[ense], in contrast to lexical categories such as nouns and verbs (Chomsky, 1995:232). A strong feature of a category is what I will call a non-categorical feature, in the sense that it is a feature other than the one that defines the category to which the non-substantive belongs. For example, T has a non-categorical D-feature that may be strong or weak; if it is strong, it is checked and eliminated by the categorical D-feature of a nominal or pronominal subject DP, which raises along with the word containing it. The categorical feature of T, as of targets generally, does not enter into checking relations (Chomsky, 1995:278), for a reason to be given.
below. In effect, then, categorical features and strong features are in (near) complementary distribution: a strong feature cannot be categorical, and a categorical feature cannot be strong. One implication of this complementarity is that nominal or pronominal DPs cannot be differentiated in terms of the feature strength of their categorical D-feature. The reason is that while categorical D, like T, is a non-substantive category, it cannot be checked by the (strong) non-categorical D-feature of T; under the second consequent of (31), strong features must be checked by categorical features. Nor can it be checked by the categorical feature T of T\textsuperscript{0}, since such features of targets do not enter into checking relations. This exclusion of feature strength from categorical D narrows the range of permissible analyses, a theoretical desideratum. It also forestalls a potential redundancy: pronominal and nominal DPs are distinguished structurally by the presence or absence of an NP complement, rendering a difference in feature strength unnecessary and so presumably illicit on minimalist assumptions. Finally, the restriction of feature strength to targets such as T or C does not favor the allowing of non-substantive categories that are not targets to vary in strength.

D is not only a categorical feature but also a [+Interpretable] feature, that is, a feature that contributes to the interpretation of a sentence at LF. D is Interpretable by virtue of being a categorical feature (Chomsky, 1995:277-78) and because of its associations with referentiality and definiteness (Chomsky, 1995:240, 342). Features that are [+Interpretable] contrast with those that are [-Interpretable], such as Case and agreement, which play no role in LF interpretation, “assuming that they are distinguished from the semantic properties that they closely reflect” (Chomsky, 1995:278). Because Interpretable features must be present at LF for semantic interpretation, they cannot be deleted when checked; because non-Interpretable features do not contribute to LF interpretation, they must be deleted when checked for a derivation to converge (Chomsky, 1995:277). If a non-Interpretable feature is left unchecked at LF, the derivation crashes. Only non-Interpretable features of targets enter into checking relations; Interpretable features do not (Chomsky 1995:282-83).

As an illustration of these concepts, consider the wh- question (32):
Al-Shorafat (1999:10) derives (32b) from (32a) as follows. In (32a), the functional head C (of CP) has a strong interrogative feature Q that must be checked by the categorical feature of a (pro)nominal category, that is, by (a variant of) D. The categorical feature that checks Q, F₀, is the interrogative (wh-) feature of the wh- operator maaḍḍaa, a variant of D; maaḍḍaa is attracted to [Spec, CP], the checking domain of C, by the strong Q feature of the latter and substitutes for its specifier. Because maaḍḍaa checks a strong feature, it raises overtly, in violation of Procrastinate, yielding the word order in (32b). This movement obeys two computational principles, which specify what are legitimate steps in a syntactic derivation (Kitahara, 1997:13-17), Last Resort in (33) and the Minimal Link Condition (MLC) in (34):

(33) Last Resort: K attracts F if F is the closest feature that can enter into a checking relation with a sublabel of K. (Chomsky 1995:297)

(34) Minimal Link Condition: K attracts a [a category or feature] only if there is no b, b closer to K than a, such that K attracts b. (Chomsky 1995:311).

"The closest feature that can enter into a checking relation" in (33) is one that matches a feature of K; a closer feature that does not match cannot be attracted. A sublabel of K "is a feature of the zero-level projection of the head H(K) of K" (Chomsky, 1995:268). A zero-level projection of a head may be minimal or maximal. A minimal zero-level projection is the head itself; a maximal zero-level projection, designated H⁰max, is an X⁰ category to which one or more features have been adjoined (Chomsky, 1995:245, 299, 311). In (32b), the categorical V-feature of the verb kataba raises overtly to check the strong V-feature of T; adjunction of the categorical V-feature to T yields T⁰max. H(K) = C must then attract a =maaḍḍaa to [Spec, CP], where its feature F₀ may enter into a checking relation with Q = the sublabel of C. However,
there is a category $b = V$ which is closer to $C$ than is $maa\delta aa$ under the definition of the relation ‘closer’ in (35):

(35) $b$ is closer to $HP$ [the maximal projection of a head $H$] if $b$ c-commands $a$ and is not in the minimal domain of $CH$ [a chain $(g, t)$ formed by adjunction of $g$ to $H$].

The ‘minimal domain of a chain’ is defined as in (36)-(38):

(36) $Max(a)$ is the smallest maximal projection including $a$.

(37) The domain of a chain $d(CH)$ is the set of categories included in $Max(a)$ that are distinct from and do not contain $a$ or $t$.

(38) The minimal domain of a chain $Min(d(CH))$ is the smallest subset $K$ of $d(CH)$ such that for any $g$ an element of $d(CH)$, some $b$ an element of $K$ reflexively dominates $g$.

To see how these definitions work, consider the tree (39), after the raising of $katab$- and prior to the raising of $maa\delta aa$ in (32b):
Adjuction of \( V \) to \( T \) in (39) forms \( \text{CH} (V,t) \). Let \( maa\d^\d\alpha = a \) and \( katab- = b \) in (35). This configuration satisfies (35): \( katab- \) c-commands \( maa\d^\d\alpha \) and is not in \( \text{Min}(d(CH)) \). It is not in \( \text{Min}(d(CH)) \) because it is not in \( d(CH) \); it is not in \( d(CH) \) because that excludes \( a = V \) by (37). \( Katab- \) is therefore closer to \( C \) than \( maa\d^\d\alpha \). Why does it not raise to \( C \) to check \( Q \) instead of the latter raising to [Spec, CP] to do so? The reason is that while \( V \) movement satisfies the MLC, it does not satisfy Last Resort, because the \( V \)-feature and \( Q \) do not match. Therefore, \( H(K) \) attracts \( maa\d^\d\alpha \) instead.

Movement of \( Zayd \) to [Spec, CP] instead of \( maa\d^\d\alpha \) is excluded on similar grounds. Formation of \( \text{CH} (V,t) \) enlarges \( d(CH) \) from \( VP \) to these categories in TP: \( d(CH) = \text{Spec}_2 ([\text{Spec, TP}]), \text{Spec}_3 ([\text{Spec, VP}] = \text{DP}_1 = Zayd), \) and \( \text{DP}_2 (maa\d^\d\alpha) \). This set of categories is also \( \text{Min}(d(CH)) \). It is customary for domination to be a reflexive relation. For \( b \) to reflexively dominate \( g \) is equivalent to saying \( b = g \) or that \( g \) dominates itself (cf. Partee, ter Meulen, and Wall 1990:440). In the light of (37), I understand “reflexively dominates” in (38) to mean ‘domination of categories that nether dominate nor are members of CH.’ In this case, and apparently generally, \( d(CH) = \text{Min}(d(CH)) \) (Chomsky, 1993:13-14; 1995:298-99). Since the two specifiers are in the same minimal domain, they are equidistant from \( maa\d^\d\alpha \), permitting it to cross the lower specifier when raised. Now if \( a = \text{DP}_2 = maa\d^\d\alpha \) and \( b = \text{Spec}_3 = \text{DP}_1 = Zayd \), then (35) is only partly satisfied: \( b \) c-commands \( a \), but both are in \( \text{Min}(d(CH)) \). Therefore, \( Zayd \) is not closer to [Spec, CP] than \( maa\d^\d\alpha \). Last Resort and the MLC are then satisfied for \( maa\d^\d\alpha \): it has an interrogative feature \( F_Q \) to check the Q feature of \( C \), and there is no closer category to \( C \) with that feature. Consequently, it raises to substitute for [Spec, CP]. To license this substitution, \( V \) raises again to adjoin to \( C \), so that \( \text{Spec}_1 \) and \( \text{Spec}_2 \) are in the same minimal domain and so equidistant from \( maa\d^\d\alpha \). \( Zayd \) raises covertly to [Spec, TP] to check the weak D-feature of \( T \), and \( maa\d^\d\alpha \) substitutes for [Spec, CP], crossing \( Zayd \) to yield (32a). The interrogative feature of \( maa\d^\d\alpha \) is Interpretable (Chomsky, 1995:289), so while it is checked, it is not deleted (or, a fortiori, erased). However, the strength feature of \( Q \), as distinct from \( Q \) itself, is non-Interpretable (Chomsky, 1995b:278) and so deletes and erases at LF.
One final aspect of this derivation deserves comment, as it will figure in later discussion. The feature of a target entering into a checking relation with a raised feature must be non-Interpretable, unlike the categorical feature that checks it. If it were Interpretable, then the target would not be able to delete, so neither the category undergoing checking nor its checker would be altered by the operation, making it ‘locally superfluous’, a property inconsistent with minimalist language design by virtue of the proliferation of computational possibilities that it entails (Chomsky, 1995:282). This is why the categorical feature of a target cannot enter into checking relations; being categorical, it is Interpretable. Now both Q and F₀ are Interpretable; consequently, the latter need not be checked and so need not undergo raising under Last Resort. It raises in this instance not to check a non-Interpretable feature but to check a strong feature, namely Q (Chomsky, 1995:290).

5. Impoverished Agreement with Form VI Verbs

5.1 Impoverished agreement in uniclausal sentences

Let us now consider how these minimalist ideas account for agreement in a sentence like (9), repeated here as (40), with the first conjunct changed to feminine to make gender agreement between it and the verb more prominent:

(40) ta qaabal-at kariim-at-u wa marwaan-u-n
    recip-meet-3fsg.pft Kariim-f-nom and Marwaan-nom-indef

“Kariima and Marwaan met [each other].”

Assuming, as per 2.1-2.3, the VP-internal subject hypothesis, the DP analysis of noun phrases (with D null, as represented by e, since each of its NP complements is a proper noun), the elimination of Agr from the set of functional projections, antisymmetry in conjoined DPs, and fusion of indices in conjunctions, the post-Merger structure for (40) is (41):
In this and the following analyses, the V-feature of T is strong and its D-feature weak. I assume "SVO" order—that is, that the VP-internal subject in [Spec, VP] precedes the V head, which it must do as a corollary of Kayne's antisymmetry hypothesis, as noted in 3.1. In order to derive VSO order, the V \( qaabal- \) must raise overtly to T. By (31), the strong V-feature of T, a non-substantive category, attracts the categorical V-feature \( F_V \) of V. \( F_V \) checks the strong V-feature, eliminating the non-Interpretable feature [strength], with the categorical (and so Interpretable) V-feature itself left intact. Because strong features trigger overt movement, the whole word \( qaabal- \) raises along with its categorical feature to adjoin to T, contravening Procrastinate. Adjunction of V to T yields the intermediate structure (42), with the verb c-commanding the subject DP as a condition for (or a corollary of) impoverished number agreement (Kayne, 1994:51):
We now turn to the categorical D-feature and the f-features of the subject DP. The D-feature of *kariimatuuwa marwaanun* is weak, since it is a categorical feature and so cannot be strong as a corollary of (31). Since the D-feature of T is, ex hypothesi, also weak, it does not force overt movement of the conjoined DP; hence, the formal features of the latter raise covertly under Procrastinate. The movement of formal features is governed by the economy condition (43) and the amplification of the definition of Move (44) (Chomsky, 1995:262, 265):

(43) *F* carries along just enough material for convergence.
(44) *Move F* automatically "carries along" FF(LI) (the formal features of a lexical item).

(44) defines “just enough material for convergence” in (43) as (at least) the entire set of formal features of a lexical item, not just the one being checked. Let us examine three of these—D, gender, and number—in the light of these two economy conditions.

D is the most straightforward: it adjoins to T—technically, to [T V - T]—to check the weak D-feature of the latter, perhaps the norm for covert raising in VSO structures (Chomsky, 1995:272). The categorical
D-feature of the DP, being Interpretable, cannot delete; the strength feature of the non-categorical D-feature of T, which is non-Interpretable, deletes and erases. In a coordinate structure like kariimatuu wa marwaanun, it is unclear where the categorical feature originates; in principle, it could be from the D⁰ head of either conjunct. However, if FF(LI) raise together as per (43)-(44), it must be the D-feature of the first conjunct that raises along with (at least, perhaps at most) the gender feature of that conjunct.

Gender is an f-feature of nouns, and, as such, Interpretable (Chomsky, 1995:282). This conclusion is plausible in the case of natural gender but counterintuitive in the case of grammatical gender, as found in SA. It is not clear how the latter, being independent of the natural gender of the referent of a noun, is an Interpretable feature. In fact, interpretability is most visible when natural gender overrides grammatical gender. This is quite rare, but Haywood and Nahmad (1976:98-99) cite the following example:

\[(45)\] qatal-a               al-xaliif-at-u al-waziir-a
    kill-3msg.pft the-caliph-f-nom the-minister-acc

"The caliph killed the minister."

qatala “he killed” occurs with al-xaliifatu “the caliph” despite the feminine suffix -at- of the latter, implying that natural (lexical) gender determines gender agreement in this case. Nonetheless, grammatical gender is restricted in Arabic (in contrast, say, to German) to nouns with inanimate referents, for example, maktab “office” vs. maktaba “desk.” With animates, especially higher animals, grammatical and natural gender regularly coincide: rajul- “man” vs. ?imraa?a(t)- “woman.” Therefore, I tentatively adopt Chomsky’s conclusion that gender is Interpretable for SA. Assuming so, it raises along with D by (44) as a ‘free rider’ which can enter into a checking relation with a sublabel of H(K) of projection K as defined above, with T = H(K) to TP = K. Sublabels evidently allow for checking agreement relations in the absence of Agr projections (Chomsky, 1995:351). A plausible candidate for the sublabel would be the (non-Interpretable) gender feature of V; if so, FF(LI) attach to a maximal zero-level projection (T⁰max) rather than a minimal one (T⁰). Hence, because it is only the gender feature of the first conjunct that raises, the verb agrees in gender
with only that conjunct, as in (2a), (3a). Feature movement of D and \( FF_{FG} \) (the formal feature feminine gender) yields (46):

\[
(46) \quad TP
\]

\[
\begin{array}{c}
\text{Spec} \\
T \\
V \\
\text{Spec} \\
Cj' \\
D \\
\hline
___qaabalk \ -at \ [e \ kariima, \ wa e \ marwaanj\]i*j \\
t_k \ ta-i*j \\
\end{array}
\]

Covert movement of the gender feature of the first conjunct along with categorical D as part of \( FF(L1) \) at first appears to provide a model for the checking of another of its f-features, number. Since I am taking DP conjunction to be phrasal, singular number in verbal sentences cannot be accounted for by agreement with the subject DP of the first conjunct of a conjunction of clauses. An ostensible alternative would be as follows. The number feature of the first conjunct in (46), \( kariima(tu) \), is singular. If it raises covertly with D as a free rider to adjoin to T by (44), it can then enter into singular number agreement with \( qaabalat \) just as its gender feature enters into gender agreement with the verb. However, this analysis holds only for DP conjunctions with singular conjuncts; it cannot be extended either to plural DPs such as \( al-muhaasibuuna \) “the lawyers” or, a fortiori, to a conjunction of plural conjuncts such as \( al-banaatu wa al-awardu \) “the girls and the boys.” This weakness does not validate a clausal analysis of DP conjunction, because such an analysis suffers from the same defect: If either
conjunct is plural, then it must agree fully with the following verb; thus, [al-banaatu ji?na] wa [al-awlaadu e] will yield full agreement in VSO after verb raising as *ji?na [al-banaatu t] wa [al-awlaadu e], contrary to fact. Can impoverished agreement be uniformly treated under the assumption of phrasal conjunction for DPs?

In order to answer this, we must look at an additional agreement pattern. ABS (1994:210) observe that the following pattern obtains for postverbal conjoined singular pronominal subjects in LA:

(47) a. naam-o huwwe w hiyye bi-l-bayt
   slept:3pl he and she in-the-house
   “They [he and she] slept in the house.”

b. naam huwwe w hiyye bi-l-bayt
   slept:3msg he and she in-the-house
   “They [he and she] slept in the house.”

c. *naam-it huwwe w hiyye bi-l-bayt
   slept-3fsg he and she in-the-house
   “They slept in the house.”

(47b, c) show that, as usual, only gender agreement with the first conjunct holds with VSO word order. ABS note that this pattern holds in SA as well, for which the paradigm would then be this:

(48) a. naam-uu huwa wa hiya fi-l-bayt-i
   slept:3mpl he and she in-the-house-gen
   “They [he and she] slept in the house.”

b. naam-a huwa wa hiya fi-l-bayt-i
   slept-3msg he and she in-the-house-gen
   “They [he and she] slept in the house.”

c. *naam-at huwa wa hiya fi-l-bayt-i
   slept-3fsg he and she in-the-house-gen
   “They [he and she] slept in the house.”

If we substitute Form VI taqaabal- “met each other” for naam-“slept”, we obtain the following paradigm:

(49) a. ta-qaabal-uu huwa wa hiya fi-l-bayt-i
   recip-meet-3pl he and she in-the-house-gen
   “They [he and she] met each other in the house.”
b. ta-qaabal-a huwa wa hiya fi-l-bayt-i
   recip-meet-3pl he and she in-the-house-gen
   “They [he and she] met each other in the house.”

   c. *ta-qaabal-at huwa wa hiya fi-l-bayt-i
      recip-meet-3fsg he and she in-the-house-gen
      “They [he and she] met each other in the house.”

Adding the patterns in (48) (and [49]) to the agreement patterns discussed previously yields the following options for VSO order in SA, with “+” indicating that the combination is grammatical and “*” that it is not:

\begin{equation}
\begin{array}{cccc}
   & naam-a & naam-uu \\
a. & alawlaadu & + & * \\
b. & alwaladu wa albintu & + & * \\
c. & alawlaadu wa albanaatu & + & * \\
d. & hum & * & + \\
e. & huwa wa hiya & + & + \\
\end{array}
\end{equation}

If these grammaticality patterns divide along a single property, there are two options: [1] the nominal DPs pattern against the pronominal DPs or [2] the compound DPs pattern against the non-compound DPs. Neither option is viable. The nominal DPs \textit{alawlaadu} “the boys”, \textit{alwaladu wa albintu} “the boy and the girl”, and \textit{alawlaadu wa albanaatu} “the boys and the girls” in (50a-c) pattern together against the pronominal DP \textit{hum} in (50d), which exhibits the converse pattern. If these four were the only possibilities, it would be natural to suppose that the presence or absence of an NP complement to the D head is responsible for the attested patterns. The nominal DPs (50a-c) consist of a null D with and NP complement: [DP [D e] [NP]]. The pronominal DPs (50d, e) consist of a bare D dominating the pronoun: [DP D]. (50a-c) allow for impoverished but not full VSO agreement, while (50d) reverses the pattern. But this explanation fails for (50e), which contrasts with (50a-d) in permitting both full and impoverished agreement. Pairing compound and non-compound DPs also fails: the non-compounds (50a, d) are mirror images with respect to agreement, while the compounds (50b, c) share impoverished but not full agreement, with (50e) again allowing both. In view of these facts, there
is evidently no way that all four patterns can be accommodated by a single analysis. I will therefore offer two complementary analyses: one to account for nominal DPs and a second to account for pronominal DPs.

5.1.1 Impoverished agreement with conjoined nominal DPs

The analysis I offer in this section is an adaptation of the null expletive proposal of Mohammad (1990, 1998) and van Gelderen (1996, 1998) previously outlined in 3.2, that in SA VSO structures an expletive pro subject in [Spec, TP] with a singular number feature agrees with the verb. Recall that categorical features as a set are Interpretable (Chomsky, 1995:278), hence, available for computation throughout the derivation. This entails that the categorical feature of V, after adjunction to T, can participate in further operations. Suppose that V has targeted T and raised to adjoin to it and that Merge has only formed T' from T and VP. T must project to TP in order to present a maximal projection at LF, intermediate projections presumably being ineligible for LF interpretation (Chomsky, 1995:242). Therefore, T' must combine with [Spec, TP], either by Merge or Move. Overt movement is inadmissible here; such movement takes place to satisfy the Extended Projection Principle (requiring that clauses have subjects, whether null or non-null) by checking a strong D-feature of T. But the D-feature of T is, ex hypothesi, weak. Covert feature movement is also inadmissible. Such movement is always adjunction and never substitution; unless it carries along additional lexical material for convergence under (44), a feature cannot target [Spec, TP] for raising (Chomsky, 1995:270). If it did, it would violate the uniformity condition (51), “where the phrase structure status of an element is its (relational) property of being maximal, minimal, or neither” (Chomsky, 1995:253):

\[(51)\] A chain [formed by Move] is uniform with regard to phrase structure status.

A category that substitutes for a specifier must be a maximal projection, a condition features obviously do not satisfy. Since neither overt nor covert movement is possible, [Spec, TP] must be targeted by Merge. This in turn entails that an element must be available with
which T' can merge to form TP. That element must meet a number of conditions: [1] it must be an XP to satisfy (51); [2] in particular, it must be a DP in order to enter into number agreement with the verb; [3] it must be singular in number to account for impoverished number agreement; [4] it must not have a gender feature, since gender is checked against the DP in [Spec, VP]; [5] it must have no feature for Case, specifically, nominative Case--otherwise, either the Case feature of this DP or that of *kariimatu wa marwaanun* will not be checked (since the Case feature of T will delete and erase once it has checked either DP), causing the derivation to crash because a [-Interpretable] feature survives to LF; [6] it must be phonetically null. An empty expletive that has a singular number feature but no gender feature satisfies these requirements (van Gelderen, 1996:759; 1998:20). The absence of (nominative) Case (along with f-features) is a property of overt expletives like there (Chomsky, 1995:287–but cf. Groat, 1999 for a contrary view) and so presumably holds, a fortiori, of their covert counterparts.

Merger of an expletive with T' becomes subject to checking theory if the definition of closeness in (35) is extended to apply to discrete syntactic objects a and K, where a = DP (the expletive pro) and K = T'. The categorical feature CF of the head of a is D. T' can attract D under the MLC (34) if there is no b closer to T' than a = D such that T' attracts b instead. By the definition of ‘closer’ (35), b is closer to T' than D if b c-commands D and is not in the minimal domain of some chain CH. Let *qaabal-* = b. *qaabal-* heads its chain (V, t) and so is not in the minimal domain of that chain: since d(CH) consists of categories distinct from and not containing the members of the chain, the same holds a fortiori for Min(d(CH)). *qaabal-* therefore satisfies the second condition of (35). However, it does not satisfy the first condition: it does not c-command the expletive DP, either before Merge applies, since the DP is not yet part of the syntactic configuration, or after, at which point DP c-commands V. There is therefore no b closer to T' than D by (35). Consequently, the Minimal Link Condition is satisfied, permitting T' to attract the DP. CF = D enters into a checking relation with a sublabel of the head of K, that is, with a feature of \( H^{d_{\text{max}}} = T^{d_{\text{max}}} \) (\( T^0 \) becomes \( T^{d_{\text{max}}} \) due to adjunction of the V-feature of V.) A plausible candidate for the sublabel is the N-feature of the subject DP, as I will argue next. Since the N-feature is Interpretable, it is available for
further computation and so can check or be checked by CF, allowing the expletive DP to become [Spec, TP]. To exclude unwanted checking relations (e.g., between the DP in [Spec, VP] and the verb), Merge in such cases is restricted to non-arguments (Chomsky, 1995:311), including, of course, expletives.

The sub label of T with which expletive D establishes a checking relation is N. The reasoning supporting this conclusion is as follows. Unlike the categorical feature D of pronominal and lexical DPs, the D-feature of an expletive is [-Interpretable], because the expletive lacks semantic features that Interpretable D corresponds to (Chomsky, 1995:287, 384). Being non-Interpretable, D must be checked and deleted. How is the expletive checked? Since it is inserted into [Spec, TP] by Merge, it cannot be checked by being raised to that position; consequently, it must be checked by the adjunction of some feature F to it. If there is a null expletive in [Spec, TP], then it must have a DP associate, that is, the subject DP. The features of the associate raise covertly for checking by T (Chomsky, 1995:287). Chomsky proposes that the categorical feature N of the NP complement of D in lexical DP subjects adjoins to the D of an expletive to check and delete it. In order for N to adjoin to the expletive, it must first adjoin to T as a free rider along with the categorical feature of D in the manner described above (assuming feature percolation from NP to D, as suggested in n. 6).

Supposing this analysis to hold for null as well as overt expletives, the derivation of (46) continues after merger of pro and T' with the DP associate of the null expletive, *kariimatu wa marwaanun*, presenting the N feature of the NP *kariimatu* for raising as a free rider of the D-feature of the D₀ head to adjoin to T. If there is a sublabel of T that enters into a checking relation with N (which seems unlikely), it is checked by the latter, N being a categorical feature. Because N is also Interpretable, it is still available for computation. It then adjoins to D in [Spec, TP] and checks and deletes its [-Interpretable] feature, permitting convergence. On this analysis, the verb agrees not with the subject DP in [Spec, VP] but rather with the expletive pronoun in [Spec, TP]. The number feature of V is non-Interpretable, unlike the number feature of nouns (Chomsky, 1995:278); therefore, it must be checked and eliminated for convergence. With V adjoined to T, it is in a checking configuration with the expletive in [Spec, TP], which checks its singular number feature. These computations leave the formal feature number of the
subject DP unchecked; however, since it is Interpretable and is not attracted by a strong feature, it can remain unchecked without blocking convergence, like the D-feature of the DP in [Spec VP] in (16) (recall the problem noted in Section 3.2, which this analysis resolves).

This analysis accounts for the invariant singular number of the verb in verbal sentences irrespective of the internal structure of the post-verbal lexical DP. The segregation of number and gender also explains why gender agreement invariably holds between a verb and its post-verbal subject but number agreement appears to hold only in the case of a conjunction of singular DPs. Finally, treating VSO structures as expletive-associate constructions accords with the homomorphism Chomsky notes between the two (1995:361).

If this is on the right track, then the structure supporting number agreement between qaabal- and pro is (52):

```
(52)                          TP
     Spec                 T'
     |                      |
     DP                    T     VP
     |                      |
     D         V          T     Spec
     |                      |
     V    FF D/G        CjP    V    DP
     |                      |
     DP                     Cj'    D
     |                      |
     D          NP   Cj   DP
     |            |
     N         D          NP

pro qaabal, -at [e kariima, wa e marwaan]i*j tk ta-i*j
```

5.1.2 Impoverished agreement with pronominal DPs

We now turn to the case of impoverished agreement with pronominal subjects, both compound and non-compound, as in (53) and (54).
While both conjoined and non-conjoined pronominal DPs allow full agreement in VSO (pace Fassi Fehri’s dissent concerning nimna hunna), only the conjoined singular subject pronouns enter into impoverished agreement in VSO. What difference between them accounts for this? A reasonable answer turns on the fact that the first conjunct of a lexical DP conjunction can raise overtly over the verb (Ali Farghali, pc):

(55) kariim-at-u ta-qaabal-at wa marwaan-un
    kariima-f-nom recip-meet-3fsg.pft and marwaan-nom
    “Kariima and Marwaan met [each other].”

Agreement between kariimatu and taqaabalat is presumably licensed in a spec-head relation; assuming adjunction of V to T, this implies that kariimatu is in [Spec, TP]. This being the case, it further follows that [Spec, TP] must be empty prior to raising, ruling out a null expletive in that position. Recall from 5.1.1 that expletive pro is inserted into [Spec, TP] by Merge and that this is permitted only with non-arguments to exclude unwanted checking relations. Since kariimatu is an argument, it cannot be inserted into [Spec, TP] by Merge. The only alternative is Move.

Movement of only the first conjunct in the conjunction must be licensed by the economy conditions and computational principles. The raising of a feature to a specifier is disallowed by the uniformity condition (51), which entails that only maximal projections may substitute for specifiers. This requires an XP, a condition that the DP kariimatu satisfies. Raising the first conjunct of the conjunction satisfies the economy condition (43), which requires that movement of a feature F must carry along the minimal amount of material necessary for convergence, namely, one DP conjunct rather than the whole DP.
conjunction. Movement of the entire conjunction is barred: it would yield number feature mismatch between the subject DP and the singular verb, in violation of Last Resort, canceling the derivation. Movement of the minimal DP is also licensed by the computational principles (33), Last Resort, and (34), the MLC. So far we have assumed that the V-feature of T is strong (to license overt verb raising) and that its D-feature is weak (to license covert feature movement of D). However, in order to force overt movement of the DP to [Spec, TP], in violation of Procrastinate, the non-categorical D-feature must also be strong. Assuming so, this strong D-feature of T attracts the categorical D-feature of kariimatu to check and delete the strong feature. Since the D-features of T and the DP match, (33) is satisfied, and attraction is possible. By (34), the D-feature and its DP can be attracted only if there is no D-feature that is closer to T; there is none. The only plausible candidate would be the root DP of the conjunction, and this DP satisfies neither of the criteria for the relation ‘closer’ in (35). Adjunction of V to T yields the CH (V, t_v), and Min(d(CH)) = {[Spec, TP], [Spec, VP], DP}. The root DP of the conjunction = [Spec, VP] is in the minimal domain of CH, contrary to the second condition of (35). In addition, the root DP does not c-command the first DP conjunct, since it dominates it. There is then no D-feature closer to [Spec, TP], so kariimatu raises overtly. Under (43)-(44), the raising of the masculine gender feature (FF_{MG}) as a free rider with categorical D entails that it is the first conjunct that moves. Because FF(LI) must raise when a feature is moved, the singular number feature FF_{NSG} of the conjunct also raises as a free rider to D (along with FF_{MG}) and checks FF_{NSG} of the verb. The number feature of the DP must raise, even though it is [+Interpretable] and need not be checked, in order to check the [-Interpretable] number feature of the verb in the absence of expletive pro. The derived representation for (55) would then be (56), omitting details:

\[(56) \{TP \{kariimatu\} [[T \{qabal{-\eta} at\} [VP \{ta wa Marwaanun\} [V' \{t, t_v\} \{dp ta{-\eta}\}]]]]\}\]

I now would like to suggest that a case of impoverished agreement like (49b) is the covert analog of the discontinuous coordination in (56). That is, it is an instance of covert phrasal movement. Pesetsky (2000:1) defends a tripartite typology of movement: feature movement,
overt phrasal movement, and covert phrasal movement. The cases of covert phrasal movement that Pesetsky considers are instances of operator (wh-) movement to [Spec, CP]; the present case is one of argument (DP) movement to [Spec, TP]. Since movement is covert, the noncategorical D-feature of T is weak, in contrast to the derivation for (55). Assuming so, the structure generated from (49b) by Move is (57), the covert analog to (56):

(57) [TP [Spec huwa] [[T$_1$. qaabal-a] [VP [Spec t, wa hiyaa] [V’ [t, t$_2$] [DP ta-i*]]]]

Covert movement of *huwa to [Spec, TP] is licensed by the same principles of economy and computation as govern overt movement in (56), permitting the subject pronoun to agree with the verb in masculine gender and singular number.

A technical difficulty with both (56) and (57) is that the trace left by movement of the first conjunct of the compound DP is NP-trace (DP-trace). NP-trace is an anaphor that must be bound in its minimal domain under Principle A of the binding theory (Chomsky 1981, 1982, 1993). Because the trace is in the specifier of VP, that projection cannot be its minimal domain, since the specifier cannot be c-commanded by another category within the VP. If the minimal domain for binding must be a maximal projection—the normal case—then TP is the next largest maximal projection, and since huwa is in [Spec, TP], the trace is properly bound. If the domain does not need to be maximal, that is, if it is T’, then it may be that the suffix -a is the binder by virtue of its third person value for the person feature.

We turn now to the case of a plural subject pronoun in (53), *taqaabalat hunna “they (f.) met each other.” The post-Merger structure underlying (53) before projection of T’ to TP is (58):
As in (57), the D-feature of $T^0$ is weak but its V-feature strong. I also assume, as per Section 4, that the D-feature of the subject pronoun *hunna* is weak, since it is a categorical feature. *qaabal-* has no D-feature; its categorical feature is V, and D, as a categorical feature, is not one of the f-features of V. Consequently, *hunna* cannot check its D-feature within the VP; even though the pronoun is in a checking configuration, it is not in a checking relation with the verb (Chomsky 1995:310). To check its D-feature, *hunna* must therefore either target [Spec, TP] to raise overtly or adjoin its D-feature to $T$ covertly. Since the D-feature of $T$ is weak, overt movement of the subject is disallowed (though *qaabal-* adjoins overtly to $T$ to check and delete the strong V-feature of $T$). Raising must therefore be covert. In view of the discussion of impoverished agreement with lexical DPs in 5.1.1, there are two derivations to consider: one with expletive *pro* and another without it.

Suppose first that expletive *pro* is not selected. Then there will be no means to project $T$ to TP in the absence of a strong D-feature for $T$. The derivation for (58) will, after overt V movement to $T$, enter LF without a maximal projection and so crash. Nor is covert phrasal movement of *hunna* to [Spec, TP] an alternative. This movement would satisfy (43), because the pronoun is a minimal DP, unlike a conjoined DP. However, this analysis is barred for the same reason that overt movement is unavailable for an entire conjoined DP: feature mismatch between the plural pronoun and the singular verb would violate Last Resort (33), causing the derivation to cancel.

Suppose now that expletive *pro* is selected. Merger of the expletive targeting [Spec, TP] with $T'$ will form TP, satisfying the maximal
projection requirement for LF interpretation. The derivation then proceeds as for a lexical DP: the singular number feature of V is checked by the number feature $FF_{Nsg}$ of the expletive $pro$ in [Spec, TP]; the categorical D-feature of *hunna* adjoins covertly to T to check the weak D-feature of T, with its gender feature $FF_{FG}$ raising as a free rider; and the Interpretable gender feature of *hunna* checks the non-Interpretable gender feature of the verb, which latter feature erases and deletes. The derivation parallels the one for lexical DPs, yet (40) is grammatical but (53) ungrammatical. Why?

The answer may lie in the different internal structure of nominal and pronominal DPs. Recall that the expletive analysis of impoverished agreement requires treating the postverbal DP as the associate of the null expletive. Among the free riders to the categorical D-feature when it raises is categorical N of the NP complement of D in nominal DPs. After adjunction of D to T, the N-feature is in a position from which it can raise to adjoin to the expletive. This N-raising option is unavailable for pronominal DPs, because, consisting of a bare D head, they lack an NP complement. There is therefore no N to adjoin to the D of the expletive; as a result, the derivation underlying (53) enters LF with an unchecked [-Interpretable] feature—that is, the D-feature of the expletive—and so crashes.

5.2 Impoverished agreement in relative clauses

We turn next to the derivation of the (un)grammatical (11), repeated here as (59):

(59) (*xaraj-at al-mudarris-at-u wa al-Taalib-u ; alla-9-aani
     left-3fsg the-teacher-f-nom and the-student-nom who-m-dual recip-
     ta-qaabal-aa
     meet dual.pft
     “The teacher and the student who met [each other] left.”

Under the assumptions for (40) and (53), the post-Merge structure underlying (59) is (60), in which “[. . .]” represents the relative clause subtree (61), with selected fused indices omitted:
al-mudarrisatu wa al-taalibu is in [Spec, DP], since it is part of the relative clause headed by the dual relative pronoun Ḵalladaani (as pronoun substitution in SVO order confirms: in Ḵumaa xarajaa, Ḵumaa replaces al-mudarrisatu wa al-taalibu Ḵalladaani tqaabalalaa and not only Ḵalladaani tqaabalalaa). As such, the compound DP is in a specifier-head relation to the relative pronoun as a token of D0, and so in a checking configuration with it. In contrast to the case of the subject pronoun in [Spec, VP] in (58), however, the CjP in [Spec, DP] and D0 also are in a checking relation, since they each have a categorical D-feature, the former by percolation from the first DP conjunct. (If so, Merge establishes a checking relation between an argument, rather than
a nonargument, and a head; see Chomsky (1995:311-12 for ways of accommodating this possibility.) Since each one's D-feature is categorical, both are Interpretable, but neither is strong, since neither is a D-feature of T. In this case, as with F₂ in (32), the D-feature of al-mudarrisatu wa al-taalibu need not be checked. However, unlike that case, there is no strength feature to eliminate; checking would be a 'locally superfluous' operation that would leave both features unaffected, so it does not apply (Chomsky 1995:291). The pro subject and the verb qaabal- of VP₂ raise to [Spec, TP₂] and T₀₂ -aa (by substitution and adjunction, respectively). These computations yield the derived structure (62) for the relative clause, with the internal structure of the DP in [Spec, DP] omitted:

(62)

\[
\begin{array}{c}
\text{Spec}_2 \\
\text{DP} \\
\text{DP} \\
\text{D} \\
\text{D'} \\
\text{TP}_2 \\
\text{Spec}_3 \\
\text{T} \\
\text{V} \\
\text{T} \\
\text{Spec}_4 \\
\text{V'} \\
\text{V} \\
\text{DP} \\
\text{D'} \\
\end{array}
\]

Turning to the matrix clause in (60), the V xaraj- is attracted to the strong V-feature of T to adjoin overtly and check its f-features. As in (52), the number feature of V is checked against the expletive in [Spec, TP]. After these features have been checked, the D-feature of T remains to be checked by categorical D. There are two candidates: the D-feature of the DP in [Spec, DP] and the D feature of the D head allaabani. The categorical feature of the entire relative clause DP belongs to the latter, not to the DP in its specifier. Therefore, the D-feature of the head is the one eligible for raising and checking, provided that the conjoined DP in
the specifier is not closer to T; otherwise, the Minimal Link Condition (34) will not be satisfied, and the derivation will be canceled. To verify this, we apply the definition of ‘closer’ in (35) to the two D-features.

Let b in (35) = the D-feature of al-mudarrisatu, a = the D-feature of allađaαani, and HP = TP. Because the first DP of the conjunction is in [Spec, CjP], it fails to c-command the D⁰ head of the relative clause DP, there being a maximal projection CjP which dominates the category containing b but not the one containing a. b therefore does not c-command a, contrary to the first condition of (35). Adjoining xaraj- to T in (60) forms the CH (V, t). Then by (37) d(CH) is Spec₁ ([Spec, TP] = expletive pro DP) and Spec₂ ([Spec, VP] = the compound DP al-mudarrisatu wa al-τaalibu); there is no object DP because xaraj- is intransitive. Therefore, Min(d(CH)) = {{Spec, TP, [Spec, VP]}}, it follows that both a and b are in Min(d(CH)), contrary to the second condition of (35). Consequently, neither stipulation of (35) is satisfied, and b is not closer to T than a. It follows that a is not closer to T than b, and that neither a nor b is closer to T than the other. They are equidistant from T.

The MLC does not stipulate which of two equidistant features must move. Given the silence of (34) on this matter, it is reasonable to assume that the choice is forced only when one or the other is closer to some HP and is free—that is, subject to ideoloccal variation—otherwise if other licensing conditions have been satisfied. To illustrate: the root DP of (56) cannot raise even though it is equidistant with the first conjunct from T because to do so would violate Last Resort. In the present case, by contrast, Last Resort is satisfied by either movement. We can then exploit this construal of (34) to account for the divided judgments concerning (11). On the one hand, speakers who find (11) grammatical select the categorical D-feature of the first conjunct of the DP in [Spec, VP] to raise to T to covertly check the weak noncategorical D-feature of the latter. This DP contains an NP complement to D, al-mudarrisatu; hence, an N-feature is available to check the N-feature of the expletive in [Spec, TP], as in uniclausal sentences like (40), and the derivation converges. On the other hand, speakers who find (11) ungrammatical select the categorical D-feature of the relative pronoun to check the D-feature of T. However, the D⁰ of the relative pronoun, unlike that of either conjunct of the DP conjunction, does not stand in a head-complement relation to an NP but in a head-postmodifier (adjunct)
relation to a TP. Therefore, the D-feature of _allaḍaani_ presents no N-feature as a free rider when it raises. This means that the non-Interpretable D-feature of the expletive in [Spec, TP] cannot be checked. As a result, the derivation enters LF with an unchecked [-Interpretable] feature and crashes.

6. **Conclusion**

I asserted in Section 1 that the reciprocal object argument for Form VI verbs in SA offers apparent counterevidence to the clausal analysis of DP conjunction in Arabic VSO sentences put forward in ABS (1994). Under their analysis, DP conjunctions must be clausal in order to account for the fact that they do not co-occur with expressions that signal semantic plurality. ABS's argument has strong empirical support in the case of underived (Form I) verbs with such expressions. However, Form VI verbs prohibit their DP subjects from being the residue of conjoined clauses, due to the impossibility of distributing a reciprocal argument across two clauses (cf. [30]). In view of this, I have proposed that the traditional phrasal analysis of DP conjunction is correct, at least for Form VI verbs. The phrasal analysis of DP conjunction and impoverished agreement in VSO accounts for [1] the grammaticality of uniclausal sentences like (9), with lexical DPs, and (49b), with conjoined pronominal DPs; [2] the ungrammaticality of non-conjoined plural postverbal pronouns as in (53); [3] the divided judgments of sentences containing relative clauses like (11)/(59).

These data are accounted for by the minimalist concepts of feature strength and feature interpretability and the differing internal structure of nominal and pronominal DPs. Specifically, I have assumed that the V-feature of T is strong but the (non-categorical) D-feature weak. These assumptions explain the observed VSO word order: the verb raises overtly to check the strong V-feature, whereas the gender feature of the first conjunct in the conjoined DP raises covertly. The singular number feature is introduced by merger of expletive pro with T' to form TP with the expletive in its specifier in order to explain impoverished number agreement in plural lexical DPs (and, a fortiori, conjunctions of plural DPs), as well as conjunctions of singular DPs. The grammaticality of singular conjoined pronominal DPs as in (49b) is a consequence of the possibility of covert phrasal movement of the first conjunct to [Spec, TP], where it enters into agreement with the verb in
a spec-head relation as in (57). The ungrammaticality of non-conjoined pronoun subjects in VSO order with impoverished agreement as in (53) is a consequence of the inability of a pronominal DP to check the [-Interpretable] D of the expletive because the former lacks an N-feature. And the (un)grammaticality of conjoined DP specifiers in relative clauses like (59) in VSO order is a consequence of selecting either a D-feature that is able to check the non-Interpretable D-feature of the expletive pro in [Spec, TP] or one that is unable to do so.

REFERENCES


Quirk, Randolph, & C. L. Wrenn. 1958. An Old English Grammar. London:


III

CORPUS LINGUISTICS
1. Introduction

In this paper we present an empirical investigation of horizontal regional variation in Modern Standard Arabic of radio news bulletins based on a corpus-linguistic approach. For that purpose we have compiled a text corpus of radio news bulletins from linguistically very different countries, e.g., Algeria, Egypt and Saudi Arabia. The completely transcribed corpus of about 320,000 words has been tagged primarily on word level, using the traditional, mixed grammar as reference. Further on we have developed several computer applications to explore the corpus. The emphasis of this study lies in the synchronic description of the use of complementary particles with reference to contemporary classical Arabic grammar. In general, the investigation confirms a great uniformity between the three countries on the syntactic level, but we will also demonstrate that in the case of particles with a similar function an important shift in function occurs, compared with the syntactic descriptions of these particles in the classical grammar.

1.1 Aim of the investigation

The main aim of our investigation was to find out to what degree Modern Standard Arabic (hereafter MSA), such as it is used nowadays in spoken media, is uniform on a syntactic level. Primarily we focused on horizontal variation, e.g. variation within one speech level in a larger area, which has been analyzed synchronically, that is to say within the same period of time. On the other hand, we also wanted to
contrast our findings to the contemporary traditional Arabic grammars. Indeed, in the Arab world it is assumed that language use is only correct when it conforms to the rules as they have been formulated in traditional writings centuries ago (El-Ezabi, 1967; Ad-Dahdah, 1990).

We based our investigation on standard language, which we describe as the language variety which is used nowadays in most writings and on the spoken level in formal situations, but which was codified a long time ago (cf. Joseph, 1987). In spite of the fact that this language variety underwent certain developments during the last century, in particular on the lexicographic level, grammatical rules, in general, are considered not to have changed at all compared to the rules of Classical Arabic (Shorrab, 1984). MSA can also be considered a standard language because it is the language that is spread through education and formally used in the media. It is clear, however, that although MSA is to be considered an independent language variety it is not surprising that language users of this variety from time to time might be influenced by their local dialect, Classical Arabic and modern European languages.

2. Methodological problems

The Arabic language situation has been described in terms of diglossia, triglossia, quadriglossia and even multiglossia (Marçais, 1930; Ferguson, 1960; Nofal, 1980; Morsly, 1990; Meiseles, 1975 and Hary, 1996). The fact that the Arabic language can be subdivided into different levels means that prudence is necessary for the scholar who wants to conduct synchronic investigations. Indeed in choosing language samples of different language levels the results of the investigation might be biased. If we were, for instance, to define the language sample simply as the language of the media the odds are that different language levels or varieties would be compared. It is requisite for synchronic investigations that a corpus of similar language samples of different countries be compiled.

2.1 Corpus linguistic approach

In order to make a statistical analysis of language, the compilation of a representative corpus becomes a basic necessity. This approach is followed in corpus linguistics in order to test the relevance of linguistic hypotheses on representative corpora. One of the interesting features of
computerized corpus linguistics is that it is possible to check whether certain grammatical rules are in fact applied in current language use. Indeed, grammatical work is often equal to the work of muqalliduun or copyists. Grammatical rules, such as they have been determined by Arab or western grammarians, are often considered axioms copied by other grammarians.

2.2 Definition of the language layer
In order to make a sound comparison of standard language in different Arab countries, we chose to make a comparison of monologues. These monologues encompass the formal language that is used by journalists on the radio to inform listeners of certain developments and ongoing facts in the world. Until now, no detailed investigation into Arabic monologues has been carried out.

In my view language use in news broadcasts is very interesting material for investigation into variation, because different speakers use MSA in identical circumstances. Moreover, the language use in news broadcasts shows three interesting characteristics. A first feature - and one not without importance - is that there is great diversity in the language situation that serves as a basis for news reports. Local news, for example, will be based on an Arabic language context, whereas international news has to be translated from a foreign language into Arabic. The second characteristic is that the information has to be processed very quickly. This means that we can obtain a more accurate image of how people daily treat MSA in a more or less spontaneous way. A third important characteristic of news broadcasts is that a lot of persons are involved, although that they all belong to one category, viz. the journalists.

2.3 External structure of the corpus
As far as the external structure of the corpus is concerned questions are to be answered about the size of the corpus, its representativeness, the method of compilation, the execution of the compilation and the representativeness of the corpus for the whole language stratum or universe that is to be investigated.
2.4 Definition of the language area

In principle, fifteen Arab countries were eligible for our investigation. It is, however, very important to choose countries in such a way that they manifest the greatest possible difference on a linguistic level, not only now, but also in the past. For this reason we have chosen the following three countries: Algeria, Egypt and Saudi Arabia. No doubt no Arab country has been so thoroughly influenced by a European language as Algeria. On the other hand, it is clear that Egypt has always played a leading role both in the field of the Arabic language and in formal education. Finally, we chose Saudi Arabia as the third country because of its isolated character. Because of the religious doctrine of the Wahhabites, Saudi Arabia was more isolated from the rest of the world than most other Arab countries. From these three countries we took a randomly chosen language sample of news broadcasts. For methodological reasons these were further subdivided in core news, reports, politic commentaries and sports news.

2.5 Size of the corpus

In order to define the size of the corpus we postulated the principle of saturation. Geeraerts (1989) uses this concept in lexicography to indicate the point where in a certain language layer no new noticeable number of words is found. At that moment lexicographical investigation reaches a point of saturation. We postulate this principle for the definition of the size of the corpus. We assume that once no new noticeable amount of words appear in a certain language layer, there will also no longer occur a significant number of new grammatical elements. Computer analysis revealed that the optimal size for the corpus is 80,000 words per country. The total size for the three countries was approximately 320,000 words.

2.6 Internal structure of the corpus

As far as the internal structure of the corpus is concerned, questions have to be answered about the definition of words, the categorization of words in grammatical categories, the encoding or tagging of the corpus, the transcription used, etc.
2.7 The tagging of the corpus

For our investigation we have chosen a tagged corpus. We opted for an orthographic transcription because compared to a phonemic one it has the advantage that every word can be written in the same way even when it is pronounced in a different way. In tagging we also used a special encoding for dialectal interference. Because words in news bulletins are often pronounced at high speed, it was not always possible to recognize words clearly. When transcribing the corpus we used a safety factor. When doubts arose about the identification of a word, it was marked. Sentences in which unclear words occurred were excluded from the analysis.

For the encoding of texts we have to keep in mind that the tagged text ought to be accessible for as many linguistic levels as possible (Huber, 1988). This means that it is of primordial importance to encode the corpus in such a way that the encoding remains open enough to cover as many levels of investigation as possible. On the other hand it is extremely important to work as much as possible in a structured and consistent way. We decided to implement the tagging on a word level, which meant that we had to attribute to every word the corresponding grammatical category.

2.8 Specific problems in tagging Arabic texts

At that level the investigator is confronted with three problems which all have to do with the specific characteristics of the Arabic language. The first problem involves the definition of a word and the delimitation of it in a text. Because of the agglutinative character of Arabic words, they cannot be defined as a string of characters between two blanks, which for the research of other languages is less problematic. Therefore we developed a tagger, which makes it possible for the computer to discriminate prefixes and suffixes from the words. This made it possible to attribute a grammatical tag to words that were isolated that way.

The second problem concerns the attribution of grammatical categories. It has to be decided which grammatical system will form the background for the tagging. In the case of Arabic there are at least two choices, the Arab grammatical approach and the different western approaches. A lot of contemporary grammatical descriptions use both western and Arabic terminology, so that we might speak of a traditional
mixed grammar. We also chose this traditional mixed grammar for the classification of the grammatical categories because it has the advantage that not only synchronic comparisons can be made, but also diachronic ones. The choice of a linguistic model was, in our case, also influenced by computer technical possibilities. In our view a ‘word based grammar’ seemed to be the most practical approach. However, in developing the tag-system we took both the western traditional grammar and the traditional Arabic grammar into account. This means that the user of the database will be able to conduct searches both on Arabic and western traditional grammatical categories.

And last, but not least, there is the problem of ambiguity, which is an essential problem for all languages, but which takes on an extra dimension in Arabic because of the consonantal writing system. Arabic tagging can be based both on a paradigmatic approach and on a syntagmatic one. Paradigmatic tagging for Arabic only gives a partial solution. It is impossible to tag Arabic words solely on a paradigmatic basis. In Arabic the disambiguation does not only have to take place between adjectives and nouns. It is much more complex because of the consonantal writing. To a homomorph word such as, for instance, the particle *maa*, as many as nine grammatical categories can be attributed, according to the division of Hans Wehr (1979). Tagging the word *maa* therefore means that we must be able to make a difference in the tagging between, for example *maa* as a negation particle, as a relative pronoun or as an interrogative one. This level of discrimination also requires a syntagmatic approach.

2.9 Practical approach

In order to apply the encoding as soundly as possible all the texts were translated. Every encoding was attributed on the basis of a complete translated text and on basis of the syntactical value of the word in context.

Further on we developed different computer programs to explore the corpus. We made a spelling and tag-checker, also a program to define sentences in texts, a KWIC-index for Arabic based on sentences, a program to count words and a program to transform tagged texts into normal Arabic texts.
2.10 Investigation of complementary particles

We wanted to limit our investigation into syntactical variation between different Arab countries to the use of particles. We also wanted to investigate the relationship between the functions of certain particles, such as how they were used in Classical Arabic and internal language evolution. We assumed that in every language there is institutionalized variation. In this institutionalized variation we can distinguish between constructions and grammatical elements that are equivalent, constructions and grammatical elements that are similar, and constructions and grammatical elements that are unique. In the first case constructions are involved which are mutually switchable without an occurrence of change in meaning. Pure equivalent constructions are probably rare in a language. Nevertheless, some authors do point out that certain constructions are mutually exchangeable (Schulz, 1981; Fischer, 1986). This does not alter the fact that stylistically there still may be differences.

With similar constructions and grammatical elements we mean constructions and elements that have a comparable, but not identical function, which means that certain grammatical elements can hold a similar function. This means that they do hold another function, but the difference between the functions of those is so small that mistakes and also confusions may easily occur. For instance, the two particles of the future, viz. *sa* and *sawfa*, have a similar function. In principle, however, they are not arbitrarily switchable.

In the last case constructions are involved which hold a unique function. This means that the function of those elements is so clearly defined that a shift in use is very unlikely. In other words, we may expect that these constructions are much more stable and are less subject to shifts in use. The difficulty, however, is that in some domains particles hold a similar function, while in other domains they hold a unique function. Since some particles hold a complementary function, as time goes by the language user may confuse the two functions, which might lead to internal language evolution or shifts in language use. We assume that internal language evolution and shifts in language use most easily occur in particles with a similar function. We suggest calling them complementary particles.
On the basis of a thorough study of the grammars of Cantarino (1974) and Krah & Reuschel (1980) we distinguish the following complementary particles: wa and fa, Ram and Raw, Riṣaa and Rīn, the particles of the future sa and sawfa, the negation particles maa and lam, the negation particles lān and the collocation sawfa lāa, and the interrogative particles har and sa. For the investigation we established the following procedure: in the first place, we checked how contemporary Arab writers, such as Ya’qub (1988), Al-Umari (1993), Aḥmad-Dhāddāḥ (1990), Ni’mā (1973), and Shalaby (1985) describe the functions of these complementary particles in order to define very precisely which of those functions are similar, equivalent and unique. It was clear that those writers depart from an indivisible fuṣḥā. This fuṣḥā is essentially based on the Koran and some typical sample sentences.

Further on, we investigated how contemporary western descriptions evaluate the syntax of these particles basing ourselves on the works of Cantarino (1974), Fischer (1982, 1986), Reuschel (1988), Kouloughli (1994), and Holes (1995) and a number of specialized articles in order to examine whether there are clearly demonstrable data indicating a shift in the use of such particles.

Next we compared these data to a qualitative and quantitative analysis of the corpus. In order to define significantly statistical differences between the usage of particles in the three countries we used the chi-square test. The results of this quantitative and qualitative analysis are reported in the following section.

3. Synchronic results

In the first place, regional differences in standard language seem to be limited to differences in style. We discovered, for instance, that word order differs a lot in Algeria compared to Egypt and Saudi Arabia. The reason is the much more lively way in which Algerian news is presented on the radio, which influences not only the structure of the news broadcast itself, but also the structure of the sentences. In general, we can claim that, as far as standard Arabic in news broadcasts is concerned, there is, for the time being, no reason to assume that different standards are developing in the Arab world.
We could not determine any regional variation on a syntactic level for those particles. On many points the analogy in the use of particles is striking.

As an example we give the distribution of the negative particles *maal* and *lam* with the following results:

<table>
<thead>
<tr>
<th>Negative particles</th>
<th>Algeria</th>
<th>Egypt</th>
<th>Saudi Arabia</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N %</td>
<td>N %</td>
<td>N %</td>
</tr>
<tr>
<td>Total</td>
<td>120 100</td>
<td>111 100</td>
<td>86 100</td>
</tr>
<tr>
<td>lam</td>
<td>101 84</td>
<td>83 75</td>
<td>65 76</td>
</tr>
<tr>
<td>maa</td>
<td>19 16</td>
<td>28 25</td>
<td>21 24</td>
</tr>
</tbody>
</table>

Starting from the zero hypothesis and assuming that there is no regional difference in distribution of the two particles for the different countries, we can check this by applying the $x^2$-test. This test yields for this table a value of 3.64, whereas the table-value with a certainty of 99.5% is 10.60. This means that we cannot determine a statistically significant difference in the distribution of the two particles in the different countries. We cannot find regional variation in the use of the negative particles *maal* and *lam*. What strikes one most, however, is that its use in the three countries contrasts in the same way with ‘classical Arabic grammar’. The particle *maal* certainly keeps its specific function of referring to the present, but this function seems to be almost exclusively limited to the collocation with the verb *zaala*. The particle *lam* has become generally common for the negation of the past tense. The different negation possibilities of the verb *zaala* however show some regional variation.

Table 2 gives a survey of the collocation between the negation particle *maal* and the verb *zaala* (stay). The verb *zaala* shows a great variation both as regards the particle, which can precede the verb, and the tense of the verb *zaala*. In the following table a survey is given of all the possibilities.

Application of the $x^2$-test on Table 2 gives us an $x^2$-value of 18.547 with a 99.5% certainty, whereas the $x^2$-table value is 21.32. In other words here we did find a statistically significant difference between the three countries in the various possibilities of negation for the verb *zaala*. 
The Egyptian and Saudi corpora showed a marked preference for the ‘proper’ forms, with *maa* going with past tense and *lā* with present tense a majority of the time, while the Algerian corpus shows no such preference at all. When we run the countries against each other separately, we find that, as far as the collocation ‘negation particle -zaala’ is concerned, Egypt and Algeria are wildly different, Saudi and Algeria somewhat less so, but still statistically significantly different, whereas Egypt and Saudi Arabia are not statistically different from each other.

4. **Diachronic results**

On the other hand, we discovered important differences in the current use of complementary particles in MSA-RN and the rules about their use such as they are discussed by various contemporary Arab grammars and, strikingly, also by most western grammarians.

We found empirically that in every set of two complementary particles one of them loses importance as far as frequency is concerned. The particles *fa, pin, maa, sawfa* and *sam* occur less frequently than their complementary counterparts *wa, ḫa, sa* and *law*. Moreover, we discovered that this is the case in the three countries under investigation.

Qualitative analysis revealed that the low frequency of certain particles is due to the fact that their function has been taken over by the other complementary particle. This confirms the thesis postulated that particles with similar functions can evolve more easily in function because of the fact that the difference in nuance is very small. The particle *wa* has taken over the function of the particle *fa* in a logical or sequential succession of events. The same goes for the conditional function which has almost completely been taken over by the particle...
which originally had a purely temporal function. We also discovered that the particle ٍذٍ has taken over functions which have been reserved by Arab grammarians for the particle ٍذٍ. On the other hand we discovered that the complementary particle which loses importance, is often reduced to certain fixed expressions or semantic relations. The use of the particle ٍذٍ, for instance, seems to be limited to expressions that are semantically limited to the meaning of speaking. The use of the particle ٍذٍ too is limited to a few fixed expressions, such as, for instance ٍذٍ ٍذٍ ٍذٍ. The same goes for the particle of negation ٍذٍ, which is almost exclusively used in collocation with the verb ٍذٍ, leaving its counterpart ٍذٍ for all the other occasions where a negative particle is required in the past tense.

Why the use of a particle is limited while the other one takes over certain functions is not clear. Some authors postulate the thesis of conduct avoidance. (Cantarino, 1974; Kouloughli, 1988). People use the negation particle ٍذٍ much more instead of the particle ٍذٍ because the latter is considered to be dialectal. Even if this is not the case from an objective point of view we can understand that Arabs have the subjective impression of ٍذٍ having a rather dialectal character. Our investigations, however, demonstrate that this thesis is difficult to sustain. The particles ٍذٍ and ٍذٍ and to a lesser degree ٍذٍ which undoubtedly are characterized as ٍذٍ also become less important under the influence of particles such as ٍذٍ, ٍذٍ and ٍذٍ which are used frequently in certain dialects, unless we assume that from a subjective point of view these particles are considered to be less dialectal.

5. Conclusion

Our investigations have demonstrated that MSA, such as it is used in different countries in news bulletins on the radio, follows new regularities that are not found in the descriptions of contemporary Arab grammarians. On the other hand, neither do these regularities correspond to the descriptions made by western grammarians on the written language.

The question arises whether the shifts in language use observed also do occur in written MSA. There are some indications that this is,
indeed, the case. Indications found in literature are, however, not always based on a systematic and thorough analysis of empirical data. In order to gain a clear insight in the evolution of MSA, it is useful to compile larger samples of representative corpora of the Arabic language, which, of course, ought to be tagged in a consistent way. In particular when compiling corpora of the Classical language as well, we will obtain more information on the structure of Classical Arabic and its evolution.

REFERENCES


Marçais, William. 1930. “La diglossie arabe”. In L’Enseignement Public, XCVII.
1. Introduction

There is a long history of research on the linguistic characterization of speech and writing. Although a variety of approaches have been adopted, the shared goal of most previous studies has been to identify specific linguistic features that distinguish between the two modes. Many studies also offer an overall linguistic characterization of speech and writing. In general, writing is claimed to be:

- More structurally complex and elaborate than speech, indicated by features such as longer sentences and greater use of subordination (O’Donnell, 1974; Kroll, 1977; Chafe, 1982; Akinnaso, 1982; Tannen, 1982; 1985)

- More explicit than speech, in that it has a complete idea united with all assumptions and logical relations encoded in the text (Devito, 1966; Olson, 1977)

- More decontextualized or autonomous than speech, so that it is less dependent on shared situation or background knowledge (Kay, 1977; Olson, 1977)

---

1 This work was done during a 2-year period in which the author stayed in the Netherlands under the supervision of the Egyptian government. The research was initially made possible by: the teaching, help and support of Dr. Everhard Ditters, Department of Languages and Cultures of the Middle East (TCNMO), Faculty of Arts, Nijmegen University; and by the discussion and guidance of Prof. Dr. Seham El-kareh, Phonetics department, Alexandria University. Her support has aided in improving the description presented in this paper.
• Less personally involved than speech and more detached and abstract than speech (Blankenship, 1974; Chafe, 1982; Chafe and Danielewicz, 1986)

• Characterized by a higher concentration of new information than speech (Stubbs, 1980; Kroch and Hindle, 1982; Brown and Yule, 1983)

• More deliberately organized and planned than speech (Ochs 1979; Rubin 1980; Akinnaso 1982; Brown and Yule 1983)

The generalization that writing is decontextualized, while speech is contextualized is based on the perception that speech depends on a shared situation and background while writing does not depend on such a shared context. Tannen (1982, 1985) notes that while this characterization is true for the linguistic differences between conversation and expository prose, the two genres most frequently used to represent speech and writing, it is not true for speech in general. Spoken genres such as academic lectures do not show a high dependency on a shared context, while written genres such as personal letters do show such a dependency.

The claims that writing is more complex, elaborate, and explicit than speech are the most widely accepted of the above characterizations. This greater complexity is generally attributed to two distinctive characteristics of writing: the lack of strict time constraints during production and the need to establish cohesion strictly through the lexical-syntactic channel. Chafe (1982:37) notes that in writing one has time to mold a succession of ideas into a more complex, coherent, and integrated unity while speech, because it is produced online, is more fragmented. Tannen (1982:3) notes that ‘cohesion’ is established in spoken discourse through psycholinguistic and non-verbal channels such as intonation, gestures and eye-gaze, while ‘cohesion’ is established in writing through lexicalization and complex syntactic structures that make connectives explicit.

1.1 Previous quantitative studies

To support generalizations concerning speech and writing, researchers have looked at the distribution of specific features in
spoken and written texts. For example, structural complexity and elaboration have been measured by the frequency of different subordinate constructions, prepositional phrase series, adjectives etc. Complex texts are claimed to make frequent use of these features. Explicitness has been measured by features such as word length and type/token ratio, (the ratio of different words to the total number of words). Detachment and decontactualization have been measured by frequency of personal pronouns, questions, exclamations and similar features.

Disagreements are common among studies. For example, Blankenship (1962, 1974) found sentence length in speech and writing to be nearly the same, while other researchers, e.g., O’Donnell (1974); Poole and Field (1976), have found the mean length of sentences and clauses to be considerably longer in writing. Also, Horowitz and Newman (1964) describe speech as containing more ideas and subordinate ideas, and as being more elaborated than writing, in contrast to the generally accepted view that writing is more elaborated than speech.

Most studies have found that writing has a much higher degree of subordination than speech, reflecting its greater structural complexity (O’Donnell, 1974; Kroll, 1977; Café, 1982). Other studies do not support these results, finding little difference in the overall number of subordinate clauses between speech and writing (Blass and Siegman, 1975). In fact, some studies have found the opposite. Poole (1979) found a higher index of embedding in speech, Price and Graves (1980) found a higher ratio of dependent clauses in speech, and Halliday (1979) found more complex structures in speech.

Beaman (1984) is one of the few researchers who used a quantitative approach to address these widespread contradictory findings. She notes that the failure to control for differences in register, purpose, degree of formality and planning contributed to the confusing picture which emerged from previous quantitative studies. Other researchers have abandoned quantitative approaches for these reasons. In the next section, some non-quantitative studies of speech and writing are examined.
1.2 Non-quantitative approaches

Akinnaso (1982) claims that quantitative studies have not addressed the important issues relating to speech and writing. Thus he questions the validity of previous generalizations regarding the overall relationships between speech and writing:

Comparative studies of spoken and written language have emphasized general, rather than specific, consequences of writing on language structure, the working assumption being that written language is generally more complex than the spoken language. Yet contradictions occur when different studies are compared on specific findings. (1982:110)

For example, he mentioned that a central problem of previous studies is:

their quantitative orientation, each researcher deciding on what and how to count. It is common-place in social science that statistical counts usually capture only ‘etic’ rather than ‘emic’ categories, thereby ignoring the underlying logic behind surface behavior (1982:110).

In contrast to the quantitative approach of earlier studies, Akinnaso (1982) proposes to study speech and writing from the perspective of thematic cohesion, an analysis of the ways in which surface structure elements in a text are connected to make their unified function in developing a common theme (Halliday and Hassan, 1976). By investigating the ways in which thematic cohesion is achieved in each mode, this researcher (Akinnaso) attempts to uncover the underlying difference between speech and writing.

Several other researchers have also opted for a non-quantitative approach. Tannen (1982) analyzes two texts in detail with respect to linguistic features of integration and involvement. Rader (1982) analyzes a written narrative in detail to show that writing as well as speech can rely heavily on context. These and other researchers have chosen a non-quantitative approach because of the greater detail and depth of analysis possible when the linguistic characteristics of a text are directly interpreted in terms of their functions in the communicative interaction.

In fact, both quantitative and non-quantitative analyses have complementary strengths and weaknesses. Quantitative analysis gives
solid empirical foundation to the findings; non-quantitative analyses are required for interpretation. Either type of analysis in isolation gives an incomplete description.

In the present study, a new approach for comparison is intended. The approach depends on writing a formal grammar for the automatic syntactic analysis of two corpora of spoken and written Modern Standard Arabic (MSA). This approach will help in giving concrete information in terms of frequency and usage about the two modes of the language (in our case the Arabic language). Once the grammar is felt to be adequate for the description, large amounts of texts can be automatically analyzed, giving more and more global information about each mode.

2. The Proposed Linguistic Approach (Model) for Describing The NP in MSA

The purpose of this study is to compare two corpora of MSA, namely written and spoken, taking into consideration the NP level (see the next section). We explained the framework in which the NP category occurred. In describing the NP, we dealt with Immediate Constituents and with smaller units into which these NPs can be analyzed. Constituent analysis is associated with alternating function and category layers until the terminal lexical entries are reached. In this respect, a distinction is made between the functions that these units can fulfill and the categories in which these functions can be realized. The same distinction in terms of functions and categories is used in the description of the elements in which a phrasal constituent can be analyzed. Some structures consist of a single obligatory central element with optional expansions giving way to a description in terms of head and modifiers. Other structures, for example a prepositional phrase embedded within the NP structure, are rather described in terms of ‘header’ and ‘complement’. In this context the relations between the elements of a constituent are dealt with. For more details about this linguistic approach cf. Ditters (1992, 1999, 2000) and Owens (1988).

2 Since the present paper does not mainly deal with the parser itself, the linguistic approach is surveyed in summary.
This discussion of the NP is accompanied by a detailed subclassification of the Arabic class of ‘noun’ as far as the terminal entries are concerned (see section 3.2). In addition, a number of semantic features are used to complete the description of the lexical entries.

It is important at this point to emphasize that the adoption of a given linguistic approach does not imply the rejection of other approaches since other descriptive models may offer, in many cases, an equally adequate description of language facts. It was felt that our approach was the most reliably handled by an automatic parser.

Elaborated in this way, we believe that our descriptive model can be used for an automated comparison of NPs in different corpora.

2.1 Parts-of-speech needed for the study

Arabic words are traditionally classified into three main classes: nouns, verbs and particles. Verbs are generated from triliteral or tetraliteral roots. Nouns are classified into derivational nouns or non-derivational nouns. Derivational nouns are those derived from Arabic verbs, thus they are originally derived from Arabic roots. For example “maktab” ‘office’ is derived from the verb “kataba” ‘wrote’. Derivational nouns are semantically related to the root and their measures (morphological patterns) for each category are fixed. In contrast, non-derivational nouns are not derived from verbs. For example “rajul” ‘man’ does not have a verb from which it is derived. Particles form a closed set. Their number is relatively small and their features are well established.

El-Kareh and Al-Ansary (2000) have presented a detailed description of the Arabic parts of speech relying on the nature of Arabic described above and on the nature of computational needs. Since, in the present study, we are interested in NPs, we are going to describe those parts of speech related to the sub classification of the main class ‘noun’ as well as considering the subclass ‘prepositions’ from the main class ‘particles’. Both of these are needed to describe the notion of NP.

2.1.1 Subclasses adopted

Arabic is a highly derivational root and pattern language. For example, the root _k-t-b_ can be used to form the nouns _kaatib_ ‘writer’,

Each word takes a specific morphological pattern to express a different meaning. It is supposed by us that there is a close relationship between the nature of the morphological patterns, or the subclass, and the slot that can be filled in a specific syntactic structure. The sub classification process is thus intended to code as much information as possible to allow for study of this issue in a more reliable way. It could prove to be of great importance to impose restrictions on the syntactic description to eliminate undesired parses. Thus, subclasses were selected to describe this morphological nature. Each derivation takes a part of speech and describes its main class and its subclass. This, in fact, will be a compromise between the new computational approach, in which more subclasses for parts of speech are needed, and the traditional Arabic approach, which insists on the existence of only 3 parts of speech. For example, the words in the following table are all nouns and, thus their main class will be ‘N’, as follows:

<table>
<thead>
<tr>
<th>Word</th>
<th>Main Class</th>
</tr>
</thead>
<tbody>
<tr>
<td>kaatib ‘author’</td>
<td>N</td>
</tr>
<tr>
<td>maktuub ‘written’</td>
<td>N</td>
</tr>
<tr>
<td>maktab ‘desk’</td>
<td>N</td>
</tr>
<tr>
<td>kattaab ‘author’</td>
<td>N</td>
</tr>
<tr>
<td>kitaaba ‘writing’</td>
<td>N</td>
</tr>
</tbody>
</table>

However, even though it may be adequate for the traditional classification of parts of speech in Arabic, it is clear that ‘N’ is not sufficient to describe these words for computational purposes; it is linguistically very broad. Since the shape of each word gives information about the type of noun, it is this information that has been recorded as a sub classification of the noun in the following table:
On the basis described above, Arabic nouns are classified into forty-six subclasses (e.g., Active participle, Passive participle, Exaggeration pattern, Adjectival noun, A dverbial noun, Infinitive noun, Common noun, Pronoun, Quantifier, etc.), while particles are classified into twenty-three subclasses (e.g., Additional, Resumption, Indefinite, Conditional, Confirmational, Prohibition, Imperative, Optative, Reasonal, Dubious, etc.).

2.2 The proposed level of syntactic description

As has already been mentioned, the main purpose of the present study is to compare spoken MSA with its written counterpart through corpora of both varieties; it is necessary at this point to make two important issues clear. First, the NP selected for the study is not the smallest building block of the sentence, as far as nouns are concerned. Rather, it is a unit that has a single function in the sentence. For example, at certain levels, the NPs in (2) are NPs in a sentence like (1). But the main scope of our description includes an NP like the one in (3) as well since it has one function in the sentence, the subject. So, our formal description is supposed to account for all those NPs, in (2) and (3), as separate entries to the grammar.

(1) ۲۰ادارـة الـقاہیرـة ۱۹امـسی ۲۰ابـعو ۱۹ابـی الـدکتـور ۱۹ـقیـس ۱۹ـسـمید ۱۹وزـیر ۱۹نـکـذـیـلی ۱۹ـکـیـلـی ۱۹ـبـنـالـدـکـتـور ۱۹ـقیـس ۱۹ـسـمید ۱۹ـبـنـالـدـکـتـور ۱۹ـقیـس ۱۹ـسـمید ۱۹ـبـنـالـدـکـتـور ۱۹ـقیـس ۱۹ـسـمیд ۱۹ـبـنـالـدـکـتـور ۱۹ـقیـس ۱۹ـسـمید ۱۹ـبـنـالـدـکـتـور ۱۹ـقیـس ۱۹ـسـمید ۱۹ـبـنـالـدـکـتـور ۱۹ـقیـس ۱۹ـسـمید ۱۹ـبـنـالـدـک~

(2) al-ضاـماـل ـ الـدکـتـور ـ وزـیر

(3) alـدـکتـور ـ ۲۰سـامـعو ۱۹ـسـمید ۱۹ـبـنـالـدـکـتـور ۱۹ـقیـس ۱۹ـسـمید ۱۹ـبـنـالـدـک~

Each word is tagged to identify both its main class, represented by the first character in the tag, and its subclass, represented by the rest of the tag.

<table>
<thead>
<tr>
<th>Word</th>
<th>Tag</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>kaātib ‘author’</td>
<td>NDS</td>
<td>Noun, active participle.</td>
</tr>
<tr>
<td>maktuub ‘written’</td>
<td>NDO</td>
<td>Noun, passive participle.</td>
</tr>
<tr>
<td>maktab ‘desk’</td>
<td>NDP</td>
<td>Noun, derivational noun of place.</td>
</tr>
<tr>
<td>kattaab ‘author’</td>
<td>NDE</td>
<td>Noun, exaggeration pattern.</td>
</tr>
<tr>
<td>kitaaba ‘writing’</td>
<td>NIS</td>
<td>Noun, standard infinitive.</td>
</tr>
</tbody>
</table>
Second, the scope of the present study was limited to the types of NPs that appeared in the corpus being studied.

2.3 The corpus used
The written MSA texts were collected from the Al-Ahram Newspaper, an Egyptian Arabic newspaper considered a primary source of written MSA. It is Egypt’s official newspaper and is widely read. For spoken texts we wanted to find spontaneous speech. We chose to collect spontaneous interviews found in TV news bulletins. The 9 o’clock news bulletin is considered to be a primary source of data of this type. Such interviews generally use MSA as a means of oral communication, and highly educated persons are invited to participate. The present study deals with an 1120 NP corpus, with 560 NPs from each variety of MSA.

3. Research Methods

3.1 The affix grammars over finite lattices (AGFL formalism)
Affix grammars are a family of two-level grammars, where the first or lower level consists of context-free syntax rules, rewriting nonterminals to terminals or to other nonterminals e.g., SENTENCE: SUBJECT, VERB. The second level is a level where affix domains are defined by rewriting affix-nonterminals e.g., number: PLURAL; SINGULAR. The two levels can be combined by using affixes as parameters to nonterminals in the syntax rules. Thus, the one-level grammar can be extended with affixes as parameters e.g., SENTENCE: SUBJECT (number), VERB (number). Note that all occurrences of the same affix within one rule have the same value. This implies unification of affix values. The affix mechanism is a powerful tool to make the grammar more compact, while at the same time allowing for linguistically motivated description.

Affix Grammars over Finite Lattices (AGFL) are a particularly simple form of Affix Grammars. In AGFL, the second level consists of restricted context-free affix-rules (or meta rules). Each affix-rule defines the direct production of a non-terminal affix. Such a direct production can be either a nonterminal affix or a terminal affix. The meta rules are restricted to have a finite domain; this means that
recursion is prohibited. This enables reliable consistency checks and the generation of efficient parsers.  

3 The parser

The parser developed according to the linguistic approach described above achieved a rate of accuracy exceeding 97% of the total NPs tested. The parser successfully analyzed very complicated NPs and dealt with coordination at different linguistic levels (see appendix I for samples of the parse trees using the AGFL formalism). Implementing the parser depended on three main points, namely:

- Conversion of the detailed description of the noun in MSA into a hierarchical affix structure,
- The introduction of a level for special kinds of filters to eliminate non-relevant word subclass in relation to the structure,
- The ability to classify and group affixes (those responsible for the description of the noun subclasses) in a flexible way.

The grammar is converted into an executable parser by the AGFL formalism described in the last section.

3.2.1 Limitations of the parser

Since the parser is limited to NP level (constituent level), some limitations are necessary. First, the ‘case’ of the main head is not instantiated since it depends on the syntactic context, i.e., the position of the constituent being analyzed in the sentence. Also, the case of all heads whose case depends on that of the main head is not instantiated, e.g., ADJHEADs. Second, some alternatives could not be eliminated, e.g., cases where the NP includes a PP at a lower level. For instance almost all Arabic prepositions have more than one semantic value, and this function differs according to the usage or to the context. In such cases, the parser gives more than one parse for the NP. All alternatives given are the same but associated with a different semantic function of the preposition each time. Finally, in some of the NPs analyzed, the grammar is sometimes unable to link the PP with a head, either with the

---

main head or with the nearest head. This kind of ambiguity caused some alternatives to appear. The same general structure is assigned to each alternative, but they differ in giving two possibilities for the attachment of the PP, either to the nearest head or to the main head. It is preferred to leave the first limitation open until shifting the syntactic level to a higher linguistic level—the sentence level. To disambiguate alternatives related to the last two limitations described, deeper linguistic analysis on more abstract linguistic levels will be needed. However, this is beyond our ability at the present stage of the work.

3.3 Parsed corpus / linguistic database conversion

The parsed corpus is converted into a searchable relational linguistic database using the AGFL-Parse-Tree Reader, an interface programmed in-house using Visual Basic v. 5.00. The interface is capable of storing all information obtained from the parser including structure, relations, number of levels, affix values etc. for each NP, in addition to the alternative number being executed to produce each layer in the parse tree of each NP. The last piece of information is extremely important in cases like tracing the output or when the database is interrogated for information about the most commonly used rule in the analysis. We kept a record of each piece of information even if we did not use it at the time, as it may be important for future use of the database. Figure 1 is a screen shot of the interface. It is an example of a query associated with the written corpus.4

4 We will discuss types of queries in section 7.1.
4. Comparing Spoken and Written MSA

In this section, the analyzed spoken and written NPs will be compared in a formal way to reflect the nature of the NPs in both varieties. Using the interface described in the last section, the comparison will concentrate on three axes. The first axis will deal with the variation in NP types i.e., the overall functional structure of the NPs. Variation in NP type could be dealt with as a factor for horizontal complexity. The second axis will deal with the structure of each functional element, at the next lower linguistic level in the overall structure. The third axis will continue to study noun subclasses, considering the number of subclasses appearing in each variety, accompanied by the frequency of occurrence of each subclass and the distribution of each subclass within different functions.

4.1 Overall functional structure

In this subsection, the sequences of the functional slots inside the NP will be dealt with, leading to a description of co-occurrence and accumulation of functions. By the overall functional structure we mean
the functional structure assigned to the NP at the highest linguistic level of the description (the constituent level). The overall functional structure has a direct relation to the NP type. Before going into details, it is worth looking at a concrete example. Consider the following NPs in (4):

(4)  a. al-kitaabu ‘book’
     b. al-kitaabu al-kabiiru ‘the big book’
     c. kitaabu al-taalibi ‘the Student’s book’
     d. kitaabu al-taalibi al-mujtahidi ‘the book of the industrious student’
     e. kitaabu hu al-jamiilu ‘his nice book’
     f. kitaabun mina l-kutubi ‘a book of the books’

We note that each of these NPs consists of a specific functional structure on the highest level as in (5) below:

(5)  a. PREDET NHEAD
     b. PREDET NHEAD ADJPOM
     c. NHEAD POD
     d. NHEAD POD
     e. NHEAD POD ADJPOM
     f. NHEAD PPOM

Accordingly, we can define the ‘overall functional structure’ of the NP as the general structure considering the highest functional level of description. We are considering ‘type’ as directly related to the overall functional structure. The same overall functional structure indicates the same type. A difference in overall functional structure implies different types. If we have two NPs with the same overall structure but one of them has more elaboration at a lower linguistic level they will be considered the same type. (4c) and (4d) with their corresponding

---

5 The following acronyms are used:
PREDET = Predeterminer
POD = Postdeterminer
POM = Postmodifier
PPOM = Prepositional phrases
ADJPOM = Adjective phrases
NPOM = Noun phrases
ADV POM = Adverb phrases
overall structures (5c) and (5d) are good examples. Both of them have the same structure at the highest functional level (NHEAD, POD) but the POD in (4d) has an ADJPOM at a lower level connected to the NHEAD of the POD (____) not to the main NHEAD of the main NP (____). This means that the difference between these two NPs lies in the POD function not in the overall functional structure of the whole NP (we will discuss this type of difference in the next subsection). We can conclude that the NPs in (4) have four types, each with one token, and one type with two tokens (for a total of five types with six tokens). It should be clear that the ‘overall functional structure’ is the syntagmatic axis that concentrates on the possibility of the accumulation of sequences of functions inside the NP.

According to this criterion described above the NPs of spoken and written MSA were scanned for type/token relationship considering the overall functional structure. Table 1 below represents types of WMSA NPs associated with the frequency of occurrence of each type (tokens).

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEAD</td>
<td>156</td>
</tr>
<tr>
<td>NHEAD, ADJPOM</td>
<td>21</td>
</tr>
<tr>
<td>NHEAD, ADJPOM, ADJPOM</td>
<td>5</td>
</tr>
<tr>
<td>NHEAD, ADJPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, NPOM</td>
<td>2</td>
</tr>
<tr>
<td>NHEAD, POD</td>
<td>140</td>
</tr>
<tr>
<td>NHEAD, POD, ADJPOM</td>
<td>11</td>
</tr>
<tr>
<td>NHEAD, POD, ADJPOM, ADVPOM</td>
<td>2</td>
</tr>
<tr>
<td>NHEAD, POD, COMPLEMENT</td>
<td>3</td>
</tr>
<tr>
<td>NHEAD, POD, NPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, PPOM</td>
<td>24</td>
</tr>
<tr>
<td>NHEAD, POD, PPOM, PPOM</td>
<td>3</td>
</tr>
<tr>
<td>NHEAD, PPOM</td>
<td>16</td>
</tr>
<tr>
<td>NHEAD, PPOM, PPOM</td>
<td>2</td>
</tr>
<tr>
<td>PREDET, NHEAD</td>
<td>73</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM</td>
<td>12</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, ADJPOM</td>
<td>4</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, ADJPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, CONFIRMATION</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, PPOM</td>
<td>2</td>
</tr>
</tbody>
</table>
Now we will shift the attention to the nature of the SMSA NPs. Table 2 below represents types of SMSA NPs associated with frequency of occurrence of each type (tokens).

Table 2: Type/token relationship of the spoken NPs

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEAD</td>
<td>161</td>
</tr>
<tr>
<td>NHEAD, ADJPOM</td>
<td>26</td>
</tr>
<tr>
<td>NHEAD, ADJPOM, ADJPOM</td>
<td>3</td>
</tr>
<tr>
<td>NHEAD, ADJPOM, PPOM</td>
<td>2</td>
</tr>
<tr>
<td>NHEAD, ADJPOM, PPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, ADVPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, NPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD</td>
<td>120</td>
</tr>
<tr>
<td>NHEAD, POD, ADJPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, ADJPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, ADVPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, CONFIRMATION</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, NPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, PPOM</td>
<td>10</td>
</tr>
<tr>
<td>NHEAD, POD, PPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, PPOM</td>
<td>18</td>
</tr>
<tr>
<td>NHEAD, PPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>PREADDRESS, NHEAD</td>
<td>6</td>
</tr>
<tr>
<td>PREADDRESS, NHEAD, NPOM</td>
<td>2</td>
</tr>
<tr>
<td>PREADDRESS, PREDET, NHEAD</td>
<td>9</td>
</tr>
</tbody>
</table>
Table 2 reflects some facts about the structure of the spoken NPs. Some contrasts with the data from WMSA NPs may be noticed:

- The 560 NPs studied represent 43 types.
- The POM is the most likely function to follow the NHEAD.
- If the NP is preceded by a PREDET, it tends to be followed by an ADJPOM. If not, the presence of ADJPOM is less likely. This means that definite NPs are more frequent than indefinite NPs. Consequently, definite ADJPs will be expected to be more frequent than indefinite ADJPs, as is shown in the table above.
- Regarding different types of postmodifiers, the ADJP is the most frequent modifier of the NHEAD followed by the PP then NP and finally ADVP.
• Flexibility is detected for the NP in postmodifying the NHEAD; the NPOM does not seem to have any specific structure connected to it. It can follow the NHEAD, the POD or even the ADJPOM. It may be more frequent when a POD precedes it.

• The functional sequence ‘ADJPOM + ADJPOM’ has the highest rate of succession of the same function.

• Though not frequent, some functions can follow the PPOM, e.g., number complement.

By comparing table 1 and 2, one may notice very clear differences concerning the presence and absence of certain types in each variety. Table 3 represents types that exist in the SMSA corpus but not in the WMSA one.

Table 3 lists 18 type-differences that could be considered characteristic types of Spoken Modern Standard Arabic. A number of new functions and types appeared. The new functions are thought to be highly restricted to the spoken variety and are expected not to appear in the written variety even at a later stage of the work. REFORMULATION and PHRASAL INSERTION are good examples, as the speaker has neither time nor space to think about changing what is being said. Thus he resorts to these functions to add or change components from the sentence.

As for the frequency of types that appear in table 3, it is clear that some functional combinations appear to be associated more strongly with the spoken variety of MSA than others, since they occur in high frequency in comparison with the other types in this table, e.g., to have a PREADRESS followed by predetermined NHEAD (type number 6).

---

6 Considering that, at the present stage, the work is limited to the corpus being studied.
### Table 3: Types unique to the SMSA corpus

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEAD, ADJPOM, PPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, ADVPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, ADJPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, ADVPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, CONFIRMATION</td>
<td>1</td>
</tr>
<tr>
<td>PREADDRESS, PREDET, NHEAD</td>
<td>9</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, ADJPOM, ADJPOM</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, NPOM</td>
<td>2</td>
</tr>
<tr>
<td>PREDET, NHEAD, CONFIRMATION, REFORMULATION</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, NPOM</td>
<td>3</td>
</tr>
<tr>
<td>PREDET, NHEAD, NUMBER COMPLEMENT</td>
<td>2</td>
</tr>
<tr>
<td>PREDET, NHEAD, PHRASAL INSERTION, ADJPOM</td>
<td>1</td>
</tr>
<tr>
<td>PRE EXTENSION, PREDET, NHEAD, ADJPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>CHEAD, POD</td>
<td>4</td>
</tr>
<tr>
<td>CHEAD, POD, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, CHEAD, PPOM, NUMBER COMPLEMENT, ADJPOM</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, CHEAD, PPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>VOCATIVE INTRODUCTION, NHEAD</td>
<td>2</td>
</tr>
</tbody>
</table>

In contrast, Table 4 represents types that exist in the written MSA corpus, but not the spoken. This table represents 8 type-differences, which are characterized by more complicated structures and by more reliance on grammatical devices that express certain relations between components of the NP. This means more time and concentration are needed for the good-preparation of such structures; consequently, such structures are suitable mainly for the written medium. The seventh type in Table (9) is a good example of the situation where the NHEAD is postdetermined, and then followed by a complement with two successive extensions, each of which expresses a given relation to the NHEAD.7

---

7 To follow the complexity of this type you may need to examine the following NP: ؛أكبر دول العالم في 1991 "The biggest weapon exporting country in the world in 1991".
Table 4: Types unique to the WMSA corpus.

<table>
<thead>
<tr>
<th>Type</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEAD, POD, ADJPOM, ADVPOM</td>
<td>2</td>
</tr>
<tr>
<td>NHEAD, POD, COMPLEMENT</td>
<td>3</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, ADJPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, CONFIRMATION</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, SUBSTITUENT</td>
<td>2</td>
</tr>
<tr>
<td>SUPERLATIVE NHEAD, POD, SUPERLATIVE COMPLEMENT</td>
<td>5</td>
</tr>
<tr>
<td>SUPERLATIVE NHEAD, POD, SUPERLATIVE COMPLEMENT, SUPERLATIVE EXTENSION</td>
<td>2</td>
</tr>
<tr>
<td>ENUMERATIVE HEAD, ENUMERATIVE POM</td>
<td>2</td>
</tr>
</tbody>
</table>

In spite of these contrasts, tables 1 and 2 indicate that there are many shared types between each variety. Table 5 represents those core types.

Table 5: Types shared by both the WMSA and the SMSA corpora.

<table>
<thead>
<tr>
<th>Type</th>
<th>Written Tokens</th>
<th>Spoken tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEAD</td>
<td>156</td>
<td>161</td>
</tr>
<tr>
<td>NHEAD, ADJPOM</td>
<td>21</td>
<td>26</td>
</tr>
<tr>
<td>NHEAD, ADJPOM, ADJPOM</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>NHEAD, ADJPOM, PPOM</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>NHEAD, NPOM</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD</td>
<td>140</td>
<td>120</td>
</tr>
<tr>
<td>NHEAD, POD, ADJPOM</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, NPOM</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>NHEAD, POD, PPOM</td>
<td>24</td>
<td>10</td>
</tr>
<tr>
<td>NHEAD, POD, PPOM, PPOM</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, PPOM</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>NHEAD, PPOM, PPOM</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>PREADRESS, NHEAD</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>PREADRESS, NHEAD, NPOM</td>
<td>9</td>
<td>2</td>
</tr>
<tr>
<td>PREDET, NHEAD</td>
<td>73</td>
<td>77</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM</td>
<td>12</td>
<td>40</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, ADJPOM</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, PPOM</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>PREDET, NHEAD, CONFIRMATION</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, PPOM</td>
<td>9</td>
<td>7</td>
</tr>
<tr>
<td>PRE EXTENSION, PREDET, NHEAD</td>
<td>19</td>
<td>16</td>
</tr>
<tr>
<td>PRE EXTENSION, PREDET, NHEAD,</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 5 represents 25 types representing what we will call the core types. Though shared, the list provides some more information about how frequent each type is used in comparison with each variety. Table (10) emphasizes the fact that as spoken structures become more complicated, they become less frequent. In almost all structures where the NP is simple, i.e., nothing follows the NHEAD, the structure is very frequent, exceeding the frequency of the same structure in the written variety. Though these types of NPs are more frequent in spoken than in written, they still occur in high frequency in written because they are the basic NP forms that are the building blocks for any more elaborated structures. For example, let us have a look at more elaborated NP structures. Consider the following selected types:

<table>
<thead>
<tr>
<th>Type</th>
<th>Written</th>
<th>Spoken</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEAD, ADJPOM</td>
<td>21</td>
<td>26</td>
</tr>
<tr>
<td>NHEAD, ADJPOM, ADJPOM</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>NHEAD, PPOM</td>
<td>16</td>
<td>18</td>
</tr>
<tr>
<td>NHEAD, PPOM, PPOM</td>
<td>2</td>
<td>1</td>
</tr>
</tbody>
</table>

It is clear that a sharp decrease in the frequency of the elaborated type in general is detected once another functional element is added. Note that the decrease in the spoken corpus is sharper than that which occurred in the written corpus. Thus, the written corpus has more occurrences than the spoken corpus. Another revealing example of the same phenomena is shown in the next table, in which the more complicated structure tends heavily toward the written and the less complicated structure tends heavily toward the spoken.

<table>
<thead>
<tr>
<th>Type</th>
<th>Written</th>
<th>Spoken</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREADDRESS, NHEAD</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>PREADDRESS, NHEAD, NPOM</td>
<td>9</td>
<td>2</td>
</tr>
</tbody>
</table>

Even when an element is inserted rather than open-ended we see a similar pattern.
While the first type attains a small frequency difference between the spoken and the written, the second type shows a sharp difference. Clearly, syntactic elaboration is a much stronger feature of the written corpus than of the spoken.

4.2 Functional elements

According to the linguistic approach described in section 3, each functional element in the overall structure is realized in the form of a category. This category in turn is described in terms of functions in a lower level. In this section, we will examine three functions in a lower level, namely the HEAD, the POD and the ADJPOM, to see how different the same function is in written and spoken varieties of MSA.

4.2.1 The head function

In analyzing the NP in MSA, three head functions are distinguished, namely, the nominal head (NHEAD), the adjectival head (ADJHEAD), and the adverbial head (ADVHEAD). The NHEAD is the main head of the NP, the ADJHEAD is the head of an adjectival phrase that may postmodify the head of the NP, and the ADVHEAD is the head of an adverbial phrase that may also postmodify the head of the NP in a certain position in the structure. The ADJHEAD and the ADVHEAD are the heads of the categories that realize the ADJPOM and ADVPOM functions respectively. In this subsection we will examine in detail the possible subclasses (subcategories) that can realize the broadest head function, the NHEAD (some information about the ADJHEAD function will be given in section 4.3).

4.2.2 The nominal head

A wide range of subclasses can realize the function NHEAD in MSA. The likelihood of each subclass realizing the NHEAD function depends on the degree of its priority. The priority in turn depends on how frequent each subclass is in realizing this functional slot. Table 6 below lists the frequencies with which some subclasses realize the NHEAD in spoken and written MSA.

<table>
<thead>
<tr>
<th>Type</th>
<th>Written</th>
<th>Spoken</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEAD, ADJPOM</td>
<td>21</td>
<td>26</td>
</tr>
<tr>
<td>NHEAD, POD, ADJPOM</td>
<td>8</td>
<td>1</td>
</tr>
<tr>
<td>Subclass</td>
<td>Tag</td>
<td>Spoken</td>
</tr>
<tr>
<td>--------------------------</td>
<td>-----</td>
<td>--------</td>
</tr>
<tr>
<td>Common noun</td>
<td>NNC</td>
<td>348</td>
</tr>
<tr>
<td>Standard infinitive</td>
<td>NIS</td>
<td>167</td>
</tr>
<tr>
<td>Pronoun</td>
<td>NPRO</td>
<td>145</td>
</tr>
<tr>
<td>Place proper noun</td>
<td>NPP</td>
<td>23</td>
</tr>
<tr>
<td>Passive participle</td>
<td>NDO</td>
<td>20</td>
</tr>
<tr>
<td>Preceding noun of title</td>
<td>NOP</td>
<td>18</td>
</tr>
<tr>
<td>Human proper noun</td>
<td>NPH</td>
<td>15</td>
</tr>
<tr>
<td>Active participle</td>
<td>NDS</td>
<td>14</td>
</tr>
<tr>
<td>Derivational n. of place</td>
<td>NDP</td>
<td>9</td>
</tr>
<tr>
<td>Adjectival noun</td>
<td>NDJ</td>
<td>6</td>
</tr>
<tr>
<td>Following noun of title</td>
<td>NOF</td>
<td>5</td>
</tr>
<tr>
<td>Relational noun (Nesba)</td>
<td>NW</td>
<td>-</td>
</tr>
<tr>
<td>Exaggeration pattern</td>
<td>NDE</td>
<td>-</td>
</tr>
</tbody>
</table>

Examination of Table 6 highlights some important facts about the nature of the category that can realize the NHEAD in MSA generally, and in spoken and written varieties particularly. Three points are clear from the MSA data: first, NNC, NIS, NPRO and their derivatives\(^8\) are the main subcategories that realize the NHEAD function. Second, the NW is rarely used to realize an NHEAD, since according to the grammar written, it is filtered to fill the slot of ADJHEAD not that of the NHEAD\(^9\). Third, NDJ and NDE are also rarely used as forms of common nouns (nominalized adjectives).

In addition to the differences in frequency of realization of subcategories, spoken and written varieties display two other features. First, NOF is clearly a spoken property. Second, the spoken variety displayed 28 subclasses with 946 tokens while the written variety displayed 26 subclasses with 1142 tokens.\(^{10}\) More information will be given in section 4.3 where the distribution of different subclasses within different functions will be dealt with.

\(^{8}\) Derivatives are those subclasses that have a ‘D’ associated with their tags.

\(^{9}\) When the NW realizes a NHEAD it will be considered as a form of common nouns.

\(^{10}\) Because the NHEAD could be realized by a broad list of categories, not all subclasses are listed in table (6). The 28 and 26 subclasses appeared in the SMSA and WMSA respectively include those listed and others.
4.2.3 The POD function

The examination of the structure of the functional elements is very important. Two NPs may have the same structure on the highest linguistic level, but this does not imply that they share the same structure in a lower level. For Example, (12) below has three NPs with the same overall functional structure (NHEAD, POD):

(6) a. qi'la't al-?a'mali
   the works sector
b. qi'la't al-tammeyyati al-?idaariyyati
   the sector of administrative development
c. qi'la'u al-?a'mali wa-al-tammeyyati ?al?idaarriyyati
   the sector of works and administrative development

Though each of these NPs has a POD function that follows the NHEAD, each of the PODs has a completely different structure, PREDET+NHEAD, PREDET+NHEAD+ADJPOM and COORDINATION respectively. Table 7 represents the functional structure of the NP realized in a POD function in the spoken corpus.

Table 7: The functional structure of the NP realized in the POD function in SM SA

<table>
<thead>
<tr>
<th>Spoken POD Type</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEAD</td>
<td>86</td>
</tr>
<tr>
<td>PREDET, NHEAD</td>
<td>54</td>
</tr>
<tr>
<td>NHEAD, POD</td>
<td>17</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM</td>
<td>12</td>
</tr>
<tr>
<td>PREADDRESS, NHEAD</td>
<td>4</td>
</tr>
<tr>
<td>COORDINATION</td>
<td>4</td>
</tr>
<tr>
<td>NHEAD, PPOM</td>
<td>4</td>
</tr>
<tr>
<td>NHEAD, ADJPOM</td>
<td>4</td>
</tr>
<tr>
<td>PRE EXTENSION, PREDET, NHEAD</td>
<td>3</td>
</tr>
<tr>
<td>CARDINAL NUMBER</td>
<td>2</td>
</tr>
<tr>
<td>PREADDRESS, PREDET, NHEAD</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, ADJPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, ADJPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>PRE EXTENSION, PREDET, NHEAD, ADJPOM</td>
<td>1</td>
</tr>
<tr>
<td>ADJPOM</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, ADJPOM</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, NHEAD, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>CHEAD, NUMBER COMPLEMENT</td>
<td>1</td>
</tr>
</tbody>
</table>
Table 7 represents 21 POD-Types sorted according to frequency of occurrence. The concentration of occurrences shows that the first 10 POD-types are the most frequent. The rest, though infrequent, show other types that could occur in the POD function. These types could be verified for frequency by examining more data. In most cases, the frequency of each POD-type goes down as more than one extra functional element follows the NHEAD. This addition of functions implies lower levels for the description that could be interpreted in terms of complexity.

Table 8 represents the functional structure of the NP realized in a POD function in the written corpus.

<table>
<thead>
<tr>
<th>Written POD type</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>NHEAD</td>
<td>113</td>
</tr>
<tr>
<td>PREDET, NHEAD</td>
<td>51</td>
</tr>
<tr>
<td>NHEAD, POD</td>
<td>46</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM</td>
<td>13</td>
</tr>
<tr>
<td>PRE EXTENSION, PREDET, NHEAD</td>
<td>12</td>
</tr>
<tr>
<td>COORDINATION</td>
<td>9</td>
</tr>
<tr>
<td>NHEAD, POD, PPOM</td>
<td>6</td>
</tr>
<tr>
<td>NHEAD, ADJPOM</td>
<td>6</td>
</tr>
<tr>
<td>3 COORDINATED NPs</td>
<td>5</td>
</tr>
<tr>
<td>CARDINAL NUMBER</td>
<td>4</td>
</tr>
<tr>
<td>PRE EXTENSION, PREDET, NHEAD, ADJPOM</td>
<td>4</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM</td>
<td>3</td>
</tr>
<tr>
<td>CHEAD, NUMBER COMPLEMENT</td>
<td>2</td>
</tr>
<tr>
<td>PREDET, NHEAD, CONFIRMATION</td>
<td>2</td>
</tr>
<tr>
<td>PREDET, NHEAD, ADJPOM, PPOM</td>
<td>1</td>
</tr>
<tr>
<td>4 COORDINATED NPs</td>
<td>1</td>
</tr>
<tr>
<td>NHEAD, POD, ADVPOM</td>
<td>1</td>
</tr>
</tbody>
</table>

The table lists 17 POD-types, again sorted by frequency. The concentration of frequency shows that the lowest frequency is reached at the 15th type. This might explain why the occurrences of the POD
function are more frequent in written than in spoken MSA. Still, more tokens for more elaborated types occur within the most frequent types. The types began to be less frequent only in three cases: when the ADJPOM is followed by a PPOM or when coordination includes four NPs or when an ADVPOM modifies postdetermined NHEAD.

Comparing tables 7 and 8 leads us to a confirmation of the results, related to the POD that were reached by comparing tables 6 and 7. The spoken POD-types are greater in frequency than the written POD-types but the total number of POD-tokens are greater in frequency in the written than in the spoken (201 for spoken and 279 for written). This result means that the written POD structures are more frequent but occur in more fixed patterns while the spoken POD structures are less frequent but occur in a wider range of patterns. The concentration of frequency in both tables shows that in written MSA, there is a wider range of types through which the POD function could be realized in high frequency than in spoken MSA (14 types out of 17 for written and 10 types out of 21 for spoken). This might explain why the occurrences of the POD function are more frequent in written than in spoken MSA.

4.2.4 The ADJPOM function

Applying the same strategy as followed in the study of the POD function, the ADJPOM is examined. The resulting types for each variety are listed in tables 9 and 10 below.

The ADJPOM tables reveal that in both spoken and written varieties, the basic ADJPOM structure that consists of an ADJHEAD or PERDET + ADJHEAD has the highest frequency. Some elaborated types are connected to each variety showing the possibility of different functional elements being connected to the ADJHEAD. The spoken variety is higher in both number of types and total number of tokens. A limited number of ADJPOM types are detected within the written variety. Although the difference in total number of tokens between the spoken and written ADJPOMs is not very high, there is a sharp difference in the number of types. There are some types in which the spoken variety displays more complexity than the written e.g., the permission of coordination inside the ADJPOM, (the coordination reaches three coordinated adjectival phrases), and the possibility of an ADJPOM to modify the ADJHEAD (this could be considered as a form
of recursion because the whole type is an ADJPOM that modifies another ADJHEAD in a higher or preceding linguistic level).

Table 9: The functional structure of the ADJP realized in the ADJPOM function in SMSA

<table>
<thead>
<tr>
<th>Spoken ADJPOM type</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDET, ADJHEAD</td>
<td>82</td>
</tr>
<tr>
<td>ADJHEAD</td>
<td>37</td>
</tr>
<tr>
<td>COORDINATION</td>
<td>6</td>
</tr>
<tr>
<td>PREDET, ADJHEAD, ADJPOM</td>
<td>4</td>
</tr>
<tr>
<td>ADJHEAD, PPOM</td>
<td>2</td>
</tr>
<tr>
<td>ADJHEAD, OD</td>
<td>1</td>
</tr>
<tr>
<td>PREDET, ADJHEAD, POM</td>
<td>1</td>
</tr>
<tr>
<td>3 COORDINATED ADJPs</td>
<td>1</td>
</tr>
<tr>
<td>PRE MODIFIER, ADJHEAD</td>
<td>1</td>
</tr>
<tr>
<td>ADJHEAD, ADJPOM</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 10: The functional structure of the ADJP realized in the ADJPOM function in WMSA

<table>
<thead>
<tr>
<th>Written ADJPOM type</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>PREDET, ADJHEAD</td>
<td>85</td>
</tr>
<tr>
<td>ADJHEAD</td>
<td>29</td>
</tr>
<tr>
<td>ADJHEAD, PPOM</td>
<td>8</td>
</tr>
<tr>
<td>PREDET, ADJHEAD, PPOM</td>
<td>2</td>
</tr>
<tr>
<td>COORDINATION</td>
<td>2</td>
</tr>
</tbody>
</table>

4.3 Subclasses distribution
Section 4.2.1 gives an idea about the possibility of the subclasses listed in table 6 to realize the function NHEAD. Note that the statistics given in table 6 do not mean that those frequencies are the total frequencies of those subclasses appearing in the corpus being studied. Still the same subclasses may realize some other functions. In this section we will study the distribution of some subclasses inside different functions found inside the NP. For the time being we will concentrate on the distribution of some subclasses within these functions namely the NHEAD, the ADJHEAD and the
The distribution will learn how subclasses are used, and if the subclass can realize more than one function, it will then learn how frequent such subclasses are in realizing different functions. Such information will help in obtaining the best information possible concerning the usage of lexical entries in the varieties being studied. Table 11 below represents 10 examples of this kind of investigation.

Table 11: Distribution of subclasses within different functions.

<table>
<thead>
<tr>
<th>Subclass</th>
<th>Tag</th>
<th>Frequency</th>
<th>NHEAD S.</th>
<th>NHEAD W.</th>
<th>ADJHEAD S.</th>
<th>ADJHEAD W.</th>
<th>PRE EXT. S.</th>
<th>PRE EXT. W.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Common Noun</td>
<td>NNC</td>
<td>348</td>
<td>440</td>
<td>348</td>
<td>438</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Standard Infinitive</td>
<td>NIS</td>
<td>167</td>
<td>226</td>
<td>167</td>
<td>226</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Relational Noun</td>
<td>NW</td>
<td>78</td>
<td>45</td>
<td>-</td>
<td>77</td>
<td>42</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Demonstrative</td>
<td>NM</td>
<td>52</td>
<td>66</td>
<td>22</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adjectival Noun</td>
<td>NDJ</td>
<td>51</td>
<td>59</td>
<td>6</td>
<td>6</td>
<td>45</td>
<td>53</td>
<td></td>
</tr>
<tr>
<td>Passive Participle</td>
<td>NDO</td>
<td>33</td>
<td>20</td>
<td>20</td>
<td>12</td>
<td>13</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>Active Participle</td>
<td>NDS</td>
<td>31</td>
<td>52</td>
<td>14</td>
<td>18</td>
<td>17</td>
<td>34</td>
<td></td>
</tr>
<tr>
<td>Cardinal Noun</td>
<td>NE</td>
<td>24</td>
<td>14</td>
<td>23</td>
<td>13</td>
<td>1</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Ordinal Noun</td>
<td>NF</td>
<td>6</td>
<td>3</td>
<td>4</td>
<td>-</td>
<td>2</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>Exaggeration</td>
<td>NDE</td>
<td>3</td>
<td>10</td>
<td>-</td>
<td>1</td>
<td>3</td>
<td>9</td>
<td></td>
</tr>
</tbody>
</table>

The table conveys interesting information about how the subclasses listed are used in both varieties of MSA. The following remarks can be made:

- **Common nouns and infinitives** are the most frequent subclasses so far. They display high stability in instantiating one and only one function, the NHEAD function.

- **NW** has the highest priority in realizing the ADJHEAD.

- **NW** is the expectable subclass in realizing the ADJHEAD in spoken MSA while NDJ is the expectable subclass in written MSA.

---

**NP STRUCTURE TYPES IN SPOKEN AND WRITTEN MSA CORPORA**

**PREDETERMINER EXTENSION** is a function given only to demonstratives. Normally, according to our approach for parsing MSA, demonstratives, in certain structures, could be used as a NHEAD or as an extension to the predeterminer.
NW is rarely used as a common noun; only two instances appeared in the written MSA.

The usage of NM differs significantly between the spoken and written varieties of MSA. In the former, it tends to be used as a NHEAD while in the latter it tends to be used as a predeterminer extension.

Though having the highest frequencies in realizing the ADJHEAD, NW and NDJ do not behave the same in realizing other functions. NDJ exceeds the NW in realizing the NHEAD.

Participles are very common in realizing both the NHEAD and the ADHEA D. However, NDS is commonly used as an ADJHEAD in written, while NDO is commonly used as NHEAD in spoken.

Cardinal and ordinal nouns are rarely used to fill the slot of the ADJHEAD.

NDE fills the ADJHEAD slot in written MSA more frequently than in spoken MSA. No instances are detected for this subclass in realizing NHEAD in spoken MSA. The reason might be because NDE is more or less a highly classical derivative; thus it is more a written property than a spoken one.

As it appears, adjectival heads display a high stability regarding the number of subclasses that can realize this function. Only 7 subcategories realize the ADJHEAD function. This number of subcategories is the same in spoken and written MSA corpora. There is no significant difference between the total numbers of tokens for each ADJHEAD type in both corpora (158 for spoken and 150 for written).

5. Conclusion

This paper has compared the spoken and the written varieties of Modern Standard Arabic, focusing on the NP. We have relied on building a tool for the automatic formal analysis of two corpora of spoken and written Modern Standard Arabic. The grammar written was compiled using AGFL formalism. The output (parse trees) is converted into a searchable relational linguistic database by an interface.
programmed in Visual Basic 5.0. The study has focused on three areas: first, the overall structure of the NP in spoken and written MSA, second, the structure of each functional element in a lower linguistic level, and third, the distribution of noun subclasses within different functions. The frequency of occurrence is included in each of these points. The following conclusions could be made:

- Detailed description of the Parts-of-Speech is very important. It helps in filtering the terminal categories according to the realized function. This in turn will be applicable in: automatic elimination of undesired parses and in developing automatic rule based POS taggers that can assign the word a tag according to the syntactic context

- The general structure of the NPs makes clear that there is a core grammar of MSA shared between the two varieties of Arabic. 45% of the SMSA types are deviating from the core while 24% of the WMSA types are deviating from the core.

- There are a small number of written types that occurs frequently.

- There are a larger number of spoken structures, each appearing less frequently.

- Grammars and tools implemented for written language applications cannot be used with spoken language applications, and vice versa. Using tools of one variety with the other will hinder the different systems in one of the following ways:
  - The system may assign more alternatives to the string being analyzed leading to complications in choosing the correct one;
  - The system may assign many parse trees, and still not assign the correct one.
  - The system may fail to give an analysis;
  - It may slow down the system.
• Not only are results of the proposed comparison important for automated language processing applications but also they are important for various fields associated with teaching Modern Standard Arabic.

The results will be used to enrich the parsing tool used in this study with the statistical information required for the application of statistical techniques in case of high number of alternatives, improving the parser performance towards a new set of data.

REFERENCES


---------- & Ruqaiya Hassan. 1976. Cohesion in English. London:


APPENDIX

Samples of the parse trees obtained by the grammar implemented (in AGFL format).

Parsing sample 1

zu'amaa'u zabhati al-pinqaadi

NP(DEF BY POD, NNP, BPLU, MASC, THIRD, case)
  NHEAD(NNP, BPLU, MASC, THIRD, EMPTY, HUMAN)
  NOUN(NNP, BPLU, MASC, THIRD, EMPTY, HUMAN)
"zu'amaa'u"
  POD(DEF BY POD, SING, FEM, THIRD, GEN)
  POD-NP-FILTER(DEF BY POD, SING, FEM, THIRD, GEN)
  NHEAD(NNC, SING, FEM, THIRD, EMPTY, NONHUMAN)
  NOUN(NNC, SING, FEM, THIRD, EMPTY, NONHUMAN)
"zabhati"  
  POD(DEF BY PREDET, SING, MASC, THIRD, GEN)
  POD-NP-FILTER(DEF BY PREDET, SING, MASC, THIRD, GEN)
  PREDET
  PART
  "al-

NHEAD(NIS, SING, MASC, THIRD, MONOTRANS, NONHUMAN)

  NOUN(NIS, SING, MASC, THIRD, MONOTRANS, NONHUMAN)
  "pinqaadi"

1. INTRODUCTION

Arabic has millions of valid words. The words are built from base roots and morphemes for number, gender, noun case, verb tense, etc. There are also a number of less rigid morphemes for semantic properties like verb-transitivity and intensity called ‘measures’ in El-Sadany and Hashish (1989:600-612) and ‘melody’ in McCarthy (1981:376-418). Some morphemes are determined by affixes, but many by internal vowel changes or infixed letters.

There are also hundreds of near-morpheme patterns which may be applied to a root, with systematic but overlapping and inconsistent semantic meanings. The definite article /al-/ and the conjunctions /fa-/ and /wa-/ as well as the prepositions /bi-/, /ka-/, and /li-/ are written without spaces to separate them from the following word. Object and possessive pronouns are attached as suffixes. Thus a single word may sometimes convey the sense of an entire English sentence as in /faHalaqathu/ (then she shaved him.)

All of these possibilities result in a large vocabulary. Kenneth R. Beesely (1998) claims to distinguish 72 million distinct word-forms, even though his lexser is based on a dictionary and discards unattested possibilities. If we were to include syntactically well-formed and semantically intelligible words without regard for prior usage we could generate many more.

Arabic is conventionally written without vowels, so words that would be distinct if voweled may appear identical. For example, the
unvoiced word /raj/ could be the verb /rajala/ (he walked) or the adjective /rajil/ (afloat) or the nouns /rajul/ (man) or /rijl/ (leg) in any of their cases.

Some of the affix meanings overlap, for example, the third-person masculine singular perfect verb has the same suffix as the accusative definite noun, and the second-person masculine singular imperfect verb has the same suffix as the third-person feminine.

Thus, although a single word may convey a great deal of information, the written form of the word may be open to several interpretations, which must be resolved in context. If we were able to attach probabilities to each interpretation, we would have made a giant step toward resolving the ambiguity.

Clearly this is an invitation to examine a large corpus. However, it might have to be very large: If words and their forms were independently identically distributed, we’d expect that in a corpus of 72 million words, 26 million would not occur even once\(^1\) and in a corpus of 144 million words, the chances are still \(1/e^2\) that any given word will not occur. The fact that some words are actually much more common, and others less popular makes the size of sample needed for statistical confidence on less common word-forms even greater.

In our work, we instead measure frequencies of individual morphemes, including roots, affixes, and semantic forms, and compute a figure of merit for each analysis based on its components. The merit rating is not the same as the probability, because the various morphemes are not independent, but may be used in the same way.

2. METHODOLOGY

The simplest lexical analyzer is a simple dictionary lookup. This works well for English, which has little structure, but there are too many possible word-forms in Arabic. It is certainly possible to store 72 million words, but we would need a systematic generation procedure to create the dictionary; it is too large to create by hand.

A large fraction of traditional Arabic grammars (Thatcher, 1982; Wightwick and Gaafar, 1998) is devoted to discussing verb forms. These are a small list of patterns from which a verb root may be extended in meaning. Western grammarians give them numbers, but in

\(^1\) \((1-1/n)^n \approx 1/e = .3679\)
Arabic grammar the patterns are described by their effects on the verb stem /f'l/ (he did.) Form IV, for example is /'af'al/, and traditional grammars provide more rules rationalizing which of the vowels can change and which may not.

Most print dictionaries group words by root, not surface spelling, (e.g. Wehr, 1994) because the variety of related meanings cast light on each other.

We have set out to build a lexical analyzer for Arabic words that builds on this traditional model, and extends it to incorporate tense, number, etc. Our approach aims to optimize machine recognition rather than human learning.

Our target final lexer consists of a list of several thousand patterns and interpretations, as in Fig 2. Each pattern combines both measures and affixes. Such a list can be compiled into a finite state machine that examines each character once and performs all the actions for successfully matched strings. The format is inspired by lex, but we
use a small program (treeflex.pl in Fig. 1) that combines the
patterns in a trie, with interpretations at the leaves; where patterns are
ambiguous, there may be several different interpretations at a single
leaf. Search of the trie proceeds character by character, with each node
of the trie corresponding to at least one character of the input word. The
word is scanned only once, and the trie is descended only once. When
a leaf of the trie is reached, corresponding to recognition of a pattern,
the root characters are taken from their fixed positions in the pattern,
and the root is sought from a separate database. If the word is
ambiguous, several different roots might be sought.

Figure 2: Some patterns and outcomes for a lexical analyzer.

<table>
<thead>
<tr>
<th>Words</th>
<th>Possible interpretations</th>
</tr>
</thead>
<tbody>
<tr>
<td>/na?aa&quot;,ilu/</td>
<td>.0001:verb,9,1st,impf,either,pl</td>
</tr>
<tr>
<td>/na?asta&quot;,ilu/</td>
<td>.0001:verb,10,1st,impf,either,pl</td>
</tr>
<tr>
<td>/ta&quot;,ilu/</td>
<td>.0004:verb,1,2nd,impf,fem,sg</td>
</tr>
<tr>
<td>/tu&quot;,aliu/</td>
<td>.0003:verb,2,2nd,impf,fem,sg</td>
</tr>
<tr>
<td>/tu?aa&quot;,iliu/</td>
<td>.0003:verb,3,2nd,impf,fem,sg</td>
</tr>
<tr>
<td>...</td>
<td></td>
</tr>
</tbody>
</table>

We can prepare the pattern-interpretation lexer source in several
ways:

1. By hand-editing. However, the large number of possible patterns
   makes this daunting.
2. Under program control from a yacc-like grammar that describes
   the combinations of affixes and stem-patterns. The program can
   read a data file to compute combined frequencies from individual
   production frequencies.
3. By mining patterns from the corpus. Fig. 1 shows how each of
   these possibilities is realized.

The high-level description, sr2.y1, describes how morphemes are
composed. It doesn’t explicitly mention roots, but includes patterns
built on consonants. As in yacc, the only supported combining
operation is concatenation, but adding context-sensitive concatenation
operations may eventually be necessary. (For example, Arabic
grammars note that the pattern for verb form VIII, /\'i\'ifta\'al/ becomes
/Pita\text{\texttextst{\texttextst{a}}}\text{\texttextst{\texttextst{al}}}/ when the first consonant of the root is a /\text{\texttextst{t}}/ or /\text{\texttextst{d}}/. This can be easily dealt with as an extra pattern, or by adding a special class of consonants used everywhere and distinguished only here, but might be more transparent using a special concatenation operator.)

`y1comp.pl` is a program which builds `sr2.y1out`, a list of regular expressions and consequences (see Fig. 2) from the productions of `sr1.y1` and the frequencies in `prog.stat`.

Our instrumented lexer `wrds.pl`, which finds the frequencies is a perl program which performs a recursive descent parse. It is prepared from the same database of productions as the target lexer, but manually. Its process of pattern matching is on exactly the naive boundaries a human would use referring to a traditional grammar. So one stage considers whether the word form might have a leading conjunction, and if it appears to have one, whether the appearance is correct, or the apparent conjunction is actually part of the word. (See Table 5.) Another stage considers the verb endings, and whether e.g. a form with an apparent feminine plural ending actually parses successfully or not. A third stage considers the various verb forms. In the fourth stage, a word form that has passed all previous stages may have a non-existent root. Finally, a completely correct analysis may be wrong in context. It writes an output file containing zero or more possible parses for each word. If the word fails to match any of the roots or word-forms in its database, it can give parses, which include speculative roots or word-forms. As we review the speculative output, we can add correct speculations to the parsing database from which both programs are built. For subsequent runs, these interpretations are no longer speculative. This feedback is called hand edits in the figure.

To avoid some false parses, to keep speculations from being too general, and to improve the reliability of our statistics, we gather our statistics using a voweled corpus, an Arabic New Testament (Intl. Bible Society). Using a voweled corpus does not restrict later use of the data gathered from it; it is easy to selectively ignore the vowel information when examining a voweled or semi-voweled text, especially since we are already prepared to deal with ambiguity.

3. RELATED WORK

A number of lexical analyzers for Arabic words have been reported in the literature. Sadany and Hashish (1989:600-612) give a good
breakdown of the morphological analysis problem for Arabic. They wrote a Prolog program for carrying out the analysis. Beesley (1998) has a morphological analyzer based on a good dictionary, which is a finite state machine. Since the states in the parser are related closely to the dictionary, they do not produce speculative analyses for nonce words. They may (correctly) give dozens of alternative possibilities for a word-form.

Kiraz (2000) presents a notation and a framework for "Multi-tiered" finite state machines, which are essentially several finite state machines running in parallel. Thus one machine discovers voweling, one machine discovers roots, one machine discovers verb forms, and resulting final states are a cross-product of the individual machines. Since the vowel-search machine would not change state on seeing a second consonant in the input, there is a notion of an empty input transition. The resulting machines are quite simple in comparison to the very large FSA generated by Beesley’s method. However, since the parallel machines don’t interact, the technique can find un-attested words, unlike Beesley’s dictionary-driven method. Also, the treatment of prefixes is unclear; they seem to be handled by a separate dictionary component.

Beesley (1998) and Kiraz (2000) have different techniques to combine a search for the root with a scan of the word. Instead, we scan for a pattern, and then do a dictionary lookup for the root, which appears in a fixed position in each different pattern. In case of ambiguity, we might look up several different roots.

Our technique of scanning for a word-form and then looking to see whether the root is valid is similar to Maloney and Niv (1981:376-418), but their dictionary consists of previously verified stems, rather than roots, and is thus larger, while they have far fewer patterns. Maloney and Niv (1981:376-418) are concerned mostly with recognizing names in Arabic text. Their parser is satisfied to assign a single part-of-speech without considering alternative possibilities. Al-Shalabi and Evens (1998) give a discussion of their algorithm for extracting tri-lateral and quadrilateral radicals from Arabic words. They report success with a simple deterministic algorithm that removes affixes; this is surprising, since in addition to part-of-speech, gender, and case ambiguity we sometimes encounter ambiguity of the root.
The work of Beesley (1998) and Kiraz (2000) show that Arabic words can be described by a regular language. This is interesting because the regular languages can be recognized or transduced very efficiently, in time proportional to the number of characters in the input string. Some of the notations in Beesley (1998) appear to require substantial processing in order to convert them for easy recognition. Our work attempts to build on this idea of using finite-state-machine operations to build a lexical analyzer. We translate our grammar into the intermediate form of regular expressions, for which there are many well-known recognition algorithms. We also include a probability component, extracted from the corpus.

4. Future Work

Our focus has been on verbs. We expect to extend our work to noun and adjective forms.

Although it appears to be straightforward to simply remove vowels from the pattern lists, we haven’t yet attempted to do so, but a plausible lexcer for Arabic must work on unvoweled text.

Even with a voweled corpus there are still ambiguous analyses; the most frequent error is mistaking some other part of speech for a third-person-perfect-masculine verb, the simplest verb declension. For the moment we have left these ambiguous parses in the frequency database. However, the statistics would be more reliable if they reflected the actual usage, and we will eventually tag the corpus so that some of the correct-but-inappropriate analyses can be detected and pruned. In fact, for resolving ambiguity, statistics on how often a morpheme occurs in an inappropriate analysis are probably more interesting than how often it occurs. Since an unvoweled word has more ambiguity than a voweled one, these statistics should be collected on unvoweled text. The hand-corrected tagged corpus would be valuable in other contexts, as well.

We have not yet tested how useful the lexical probabilities are. This may require using the lexical analyzer in some larger system, perhaps the tagging described in the previous paragraph.

Our technique of transforming regular expressions offers a systematic opportunity for “two-level” lexical analysis, in which surface forms may differ from underlying lexical forms. At present we are not making use of this technique.
5. CONCLUSION

We are building a lexical analyzer for Arabic words that builds on the traditional pattern model, in which words are built up by inserting roots into patterns.

To help deal with ambiguity, we attach a numerical value to each lexical analysis of a word; the numerical values are computed from constituent frequencies. Thus based on values in tables 1 through 4, a first-person imperfect plural verb might have a numeric descriptor of .0003. Of course, as the system is extended to other parts of speech, the numeric descriptor value will drop.

Table 5 shows frequency of conjunction and preposition prefixes for some verbs in the corpus. We see from the table that an initial letter of /w/ is part of the root word only 9% of the time; otherwise it is a conjunction prefix. Similarly, an initial /f/ turns out to be the conjunction prefix 88% of the time.

<table>
<thead>
<tr>
<th>Table 1: Frequencies of person</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>First</td>
</tr>
<tr>
<td>Second</td>
</tr>
<tr>
<td>Third</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 2: Frequencies of tense</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tense</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>Imperfect</td>
</tr>
<tr>
<td>Perfect</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 3: Frequencies of gender</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
</tr>
<tr>
<td>--------------</td>
</tr>
<tr>
<td>Either</td>
</tr>
<tr>
<td>Feminine</td>
</tr>
<tr>
<td>Masculine</td>
</tr>
</tbody>
</table>
Table 4: Frequencies of number

<table>
<thead>
<tr>
<th>Number</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dual</td>
<td>86/1295</td>
</tr>
<tr>
<td>Plural</td>
<td>311/1295</td>
</tr>
<tr>
<td>Sing</td>
<td>898/1295</td>
</tr>
</tbody>
</table>

Table 5: Parses with leading conjunctions and prepositions

<table>
<thead>
<tr>
<th>Initial</th>
<th>/b/</th>
<th>/f/</th>
<th>/k/</th>
<th>/l/</th>
<th>/w/</th>
</tr>
</thead>
<tbody>
<tr>
<td>Conjugation/Preposition</td>
<td>1</td>
<td>162</td>
<td>16</td>
<td>0</td>
<td>390</td>
</tr>
<tr>
<td>Root</td>
<td>1</td>
<td>22</td>
<td>12</td>
<td>4</td>
<td>39</td>
</tr>
</tbody>
</table>

REFERENCES


1. Introduction

Anyone who works with large Arabic corpora and investigates forms that can be considered alternatives of each other will encounter an enormous amount of data that look much like the data in Table 1.

Table 1: Relative frequencies of ِْٓىٍىاٍ, ِ١٥٢, ِ٢٥ in the Ahram and Hayat Newspapers

<table>
<thead>
<tr>
<th></th>
<th>Ahram</th>
<th>Hayat</th>
</tr>
</thead>
<tbody>
<tr>
<td>ِْٓىٍىاٍ</td>
<td>20896 (67.6%)</td>
<td>23733 (67.9%)</td>
</tr>
<tr>
<td>ِ١٥١</td>
<td>6975 (22.6%)</td>
<td>8143 (23.3%)</td>
</tr>
<tr>
<td>ِ٢٥</td>
<td>3031 (9.8%)</td>
<td>3066 (8.8%)</td>
</tr>
</tbody>
</table>

These data are taken from a corpus of an entire year of both the Ahram and the Hayat newspapers, with over 35 million words (35951294). The Ahram is, of course, Egyptian, while the Hayat is Saudi owned, but published in London with an overwhelmingly Lebanese staff. The table compares the three common particles that mean ‘if’ in Arabic, ِْٓىٍىاٍ, ِ١٥١, ِ٢٥. It shows an almost amazing consistency of relative rates of usage of these three forms in the two newspapers over the course of a year. None of the percentages vary by more than a single percentage point across the two papers, and most are much less than that. The table indicates that when writers for these two papers come to use an if-form, they choose ِْٓىٍىاٍ almost exactly two-thirds of the time, ِ١٥١ a little less than one-fourth of the time, and ِ٢٥ a little less than one-tenth of the time. If you divide the years in half, or look at a subsequent year, the results stay amazingly consistent. If you
divide by genre of article, you find more variability, but again an amazing consistency within any particular type of article.

It is important to stress that this is just one example of many. Many other similar examples could be given of “variable” parts of Arabic grammar or word choice where speakers clearly use more than one form, but do so in a statistically consistent manner, rather than “randomly.” Data such as these need two kinds of explanations: 1) how to explain the variability between the forms in a single newspaper (i.e. why does \(i\ddot{a}\dot{a}\) consistently get two-thirds and \(law\) less than one-fourth, etc.), and 2) how to explain the variability or lack thereof between the two papers, comparing individual forms.

For the first kind of explanation we typically rely on differences in meaning, even if subtle, between the forms. In this case, we might say that \(law\) is a contrary to fact conditional, while \(i\ddot{a}\dot{a}\) is not, and that it is possible simply that contrary to fact situations come up a lot less than regular conditional situations. Specifying the difference in meaning between \(i\ddot{a}\dot{a}\) and \(in\) is somewhat harder, but still possible, but it is certainly much less clear why this minor difference should lead to such a dramatic difference in their rates of usage. In other cases, it is simply not possible to find a difference in meaning, and we are left with other less satisfying explanations that have to do with stylistic differences, and personal, regional, and other preferences.

Explanations for the second kind of variability (or in this case consistency) have to do with the underlying unity of the language, and the existence of regional preferences. In this case we might say that the data for the conditional particles indicate that the mainly Lebanese writers at the Hayat newspaper seem demonstrably to be using the same linguistic system as the mainly Egyptian writers at the Ahram, giving evidence that the language of the two papers is in fact the same. If differences, such as the Hayat’s slight preference for the other two forms over \(in\) compared to the Ahram, prove to be significant, then we would try to explain them by invoking changes in progress, regional variation in the meaning of the forms, regional preferences, and the like.

It is against a background of data like these, data that demonstrate the essential unity of newspaper Arabic across many countries, that the data I will discuss in this paper must be understood. It is important not to overreact to any single analysis, or to make sweeping generalizations.
based on only a few points of view. On the other hand, I believe that it is both interesting and instructive to make these kinds of comparisons and try to understand them from as many sides as possible.

2. Technical Details and Discussion

This paper will present an analysis of Arabic future particles based on a set of data gathered from a number of newspapers from across the Arab World, specifically a Moroccan, an Egyptian, a Kuwaiti, and a Lebanese paper. Mark van Mol (2000) has made the point that dealing with a raw rather than a grammatically coded corpus is difficult and likely to lead to a large number of ‘false hits’ and other problems which compromise the results. While I agree in general with this statement, one of the purposes of this paper was to develop techniques that with at least some phenomena and some forms would produce reliable results from a much larger corpus than is typically possible with a grammatically coded one.

The case of sawfa and sa- is instructive here. It is quite easy to design a search algorithm for sentences containing sawfa that produces no false hits at all, since sawfa stands as a separate word, it does not resemble other words, and it has only a limited number of possible prefixes (like wa- and fa-). Doing the same thing for sa-, however, proves to be annoyingly difficult. No matter how clever the algorithm, huge numbers of false hits are produced, which must either be eliminated by hand (a prohibitively time-consuming process), or simply ignored. For this paper, I tried two different techniques, both based on strictly limiting the search to a particular kind of form that I could control, and which I was confident would lead to few or no false hits.

For both techniques I generated the 1000 most commonly used verbs from one of the newspaper corpora, and edited this list by hand, getting rid of forms that were clearly going to generate many false hits when combined with sa-. This left 931 verbs in the list. I then did a search first for only the masculine singular present tense forms, without pronoun endings. A fairly careful spot-check of the results indicated that this technique produced a negligible number of false hits. Encouraged, I then tried a search for all present tense forms, with or without pronoun endings. While this led to a few more false hits than the previous method, the number is still much too small to affect the overall percentages in a significant way. Because the overall N is
large, and the obviously false hits vanishingly few, I concluded that these techniques did provide a way to mine a large, raw, corpus, with a fair degree of confidence.

3. The Future Particles
   The future particles include sawfa, an independent word that occurs immediately before an imperfect indicative verb form, and the prefix sa-, which is prefixed to an imperfect indicative verb form. These forms appear to be interchangeable with each other (some might call it ‘free variation’), and the choice between them to some extent random. For Classical Arabic, Wrights grammar makes no distinction between them, although van Mol (2000), referring to Ya’qub and Al-Umari, states that in Classical Arabic sawfa refers to the remote future and sa-to the near future. Whether or not this was the case for Old Arabic, the distinction appears to have been forgotten for Modern Standard and/or Newspaper Arabic. Modern teaching grammars make no distinction between them, Al-Khawalda (2000) specifically stating, “it is difficult to identify any syntactic or semantic reason why one would be chosen over the other.”

What at least some modern works do refer to is relative frequency of use and formality of style. The form sa- is considered to be an abbreviation or shortened form of sawfa, and as such it is considered less formal. Some works declare it also to be more common (“sawfa is usually shortened to sa-“). The fact that sa- is considered an abbreviation is clearly related to its being felt as less formal, and probably to its being more common as well. Evidence for all of these features will be searched for in this study.

4. The Corpus
   The newspapers examined for this study include the Ahram (Egypt), the entire year 1999, the Hayat (Lebanon-Saudi-International), the entire year 1997, Al-Tajdid (Morocco), half of 2002, and Al-Watan, Kuwait, half of 2002. The number of words of this corpus totals over 45 million.

5. Results
   As a point of comparison, the two future particles were also counted in the Quran. The data are summarized below. Table 2a
shows the results for the count based on the third person singular masculine forms of my list of 931 verbs with no pronoun endings, while Table 2b shows the results for the count based on all forms of the 931 verbs, including pronoun endings. This latter should be compared with Table 2c, taken from Al-Khawalda and based on a hand count, of the actual total number of these two forms in the Quran. (Given the relatively small numbers here, and that the 931 verbs in the list were based on a modern rather than medieval corpus, these results give a certain amount of confidence that the counts below, which do not have the benefit of a hand count for comparison, do reflect the actual situation.)

Table 2a: Future Particles in the Quran, Count Technique A*

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>sawfa</td>
<td>5</td>
<td>25.0%</td>
</tr>
<tr>
<td>sa-</td>
<td>15</td>
<td>75.0%</td>
</tr>
</tbody>
</table>

*MasSing forms only of 931 Most Common Verbs, no pronoun suffixes

Table 2b: Future Particles in the Quran, Count Technique B*

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>sawfa</td>
<td>31</td>
<td>29.0%</td>
</tr>
<tr>
<td>sa-</td>
<td>96</td>
<td>71.0%</td>
</tr>
</tbody>
</table>

*All forms of 931 Most Common Verbs, including pronoun suffixes

Table 2c: Future Particles in the Quran, Al-Khawaldi’s Hand Count

<table>
<thead>
<tr>
<th></th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>sawfa</td>
<td>42</td>
<td>26.9%</td>
</tr>
<tr>
<td>sa-</td>
<td>114</td>
<td>73.1%</td>
</tr>
</tbody>
</table>

Clearly, approximately one-forth of the future particles in the Quran are the full form sawfa, while three-fourths are the shortened form sa-. These data are not presented to imply anything about the extent to which the present situation ‘descended’ from the situation in the Quran. Rather, they are given simply as background in terms of which the tables to be presented will be better understood.

The basic data on sawfa and sa- for the newspapers investigated are presented in Table 3.
Table 3a: Future Particles in Four Arabic Newspapers, Count Technique A*

<table>
<thead>
<tr>
<th></th>
<th>Ahram</th>
<th>Hayat</th>
<th>Tajdid</th>
<th>Watan</th>
</tr>
</thead>
<tbody>
<tr>
<td>sawfa</td>
<td>5767 (19.7%)</td>
<td>723 (2.5%)</td>
<td>169 (5.5%)</td>
<td>1132 (9.0%)</td>
</tr>
<tr>
<td>sa-</td>
<td>23464 (80.3%)</td>
<td>28604 (97.5%)</td>
<td>2877 (94.5%)</td>
<td>11492 (91.0%)</td>
</tr>
</tbody>
</table>

*3MasSing forms only of 931 Most Common Verbs, no pronoun suffixes

Table 3b: Future Particles in Four Arabic Newspapers, Count Technique B*

<table>
<thead>
<tr>
<th></th>
<th>Ahram</th>
<th>Hayat</th>
<th>Tajdid</th>
<th>Watan</th>
</tr>
</thead>
<tbody>
<tr>
<td>sawfa</td>
<td>12366 (20.4%)</td>
<td>1671 (2.5%)</td>
<td>410 (5.2%)</td>
<td>2586 (8.7%)</td>
</tr>
<tr>
<td>sa-</td>
<td>48314 (79.6%)</td>
<td>64041 (97.5%)</td>
<td>7449 (94.8%)</td>
<td>27164 (91.3%)</td>
</tr>
</tbody>
</table>

*All forms of 931 Most Common Verbs, including pronoun suffixes

There are several things to notice about these data. The first and most obvious is that it is universally true that sa- is more common, actually much more common, than sawfa. Even without other information, this would normally imply that use of sawfa is somehow 'marked' or special. It appears to be true that sa- is somehow the normal or regular way to form a marked future. This does not mean that sawfa necessarily means something different from sa-, but it is where we should look for something extra.

Another thing to notice is that the difference between the two techniques is fairly minimal, less here with much larger N than in the Quran. It is probably safe to assume that the true percentage of sawfa forms in each case is somewhere between that listed for method B, and that listed for method A.

The most striking thing about the data in Table 3 is the enormous difference in the percentage of sawfa between the various newspapers (and thus, we assume, between the usage in the various countries). In the Ahram, about 20% of all marked futures are sawfa, making it relatively common, and not terribly different from the percentage in the Quran. In the Hayat, on the other hand, the percentage of sawfa is about 2%, meaning that only one in fifty marked futures uses sawfa. This difference is really quite remarkable, since usage of sawfa in the Hayat is almost 10 times lower than it is in the Ahram. This difference in percentage of use is consistent over various divisions of the data. I divided the newspaper data into two parts and ran them separately, and got consistent data for both parts. I had another year of the Hayat, and it also produces a very similar percentage. In short, these patterns are
very stable, highly significant, rather surprising, and just a little confounding. What would account for such an extreme, consistent difference between future particle use in these two papers/countries? Why is sawfa so much more highly marked in the Hayat than it is in the Ahram?

The data for the other two papers is also interesting, with the percentages of both lying in between the extremes of the Ahram and the Hayat. The Moroccan paper, Al-Tajdid, is closer to the low percentage of the Hayat (even though its usage of sawfa is over twice as high), and the Kuwaiti paper is closer to the higher percentage of the Ahram.

Another way of looking at this matter is to look at the average occurrence of sawfa and sa- per 100,000 words of running text for each paper. These data are summarized in Table 4.

<table>
<thead>
<tr>
<th></th>
<th>Ahram</th>
<th>Hayat</th>
<th>Tajdid</th>
<th>Watan</th>
</tr>
</thead>
<tbody>
<tr>
<td>sawfa</td>
<td>35.0 (19.7%)</td>
<td>3.7 (2.5%)</td>
<td>5.8 (5.6%)</td>
<td>17.5 (9.0%)</td>
</tr>
<tr>
<td>sa-</td>
<td>142.4 (80.3%)</td>
<td>146.9 (97.5%)</td>
<td>98.5 (95.4%)</td>
<td>178.0 (91.0%)</td>
</tr>
<tr>
<td>Total</td>
<td>177.4</td>
<td>150.6</td>
<td>104.3</td>
<td>195.5</td>
</tr>
</tbody>
</table>

*MasSing forms only of 931 Most Common Verbs, no pronoun suffixes

100,000 words represent about two days or two issues of one of these papers. If you read through two issues of the Ahram, according to this table, you would encounter about 177 marked futures, and about 35 of them would be marked with sawfa. On the other hand, if you read through two issues of the Hayat, you would encounter about 150 marked futures, but only 3 or 4 of them would be marked with sawfa, and so on for the other two papers. Looking at the data in this way allows us to see differences in the overall use of marked futures. Since it is quite possible in Arabic to use the plain imperfect to refer to the future, it is usually not actually required to use one of these particles. We can see from the totals in Table 4 that the Moroccan paper used marked plurals in general a lot less often than any of the other papers, and that the Kuwaiti paper used them a lot more often. Again, these appear to be stable, replicable trends in the data.
The percentage data in Table 4 is almost identical to the percentages in Table 3 again showing the enormous difference between the Ahram and the Hayat in sawfa usage.

As a further check on the data, and a different way of looking at it, I divided the Hayat corpus into signed and unsigned articles, and was able to determine the nationality of many of the authors. This left me with a quite large corpus of articles by Lebanese writers, a somewhat smaller one of Egyptian writers who wrote articles for the Hayat newspaper, and quite small corpora for a number of other Arabic speaking countries. Because the size of the corpora is so much smaller than what we have been looking at, these data should be expected to vary more widely and more erratically. They are summarized in Table 5.

Table 5: Future Particles in Country specific groups of Hayat Authors, Count

<table>
<thead>
<tr>
<th>Technique A*</th>
<th>Egypt</th>
<th>Lebanon</th>
<th>Morocco</th>
<th>Kuwait/Saudi</th>
<th>Iraq</th>
</tr>
</thead>
<tbody>
<tr>
<td>sawfa</td>
<td>19 (1.2%)</td>
<td>170 (4.4%)</td>
<td>0 (0%)</td>
<td>15 (9.7%)</td>
<td>25 (3.0%)</td>
</tr>
<tr>
<td>sa-</td>
<td>1575 (98.8%)</td>
<td>3657 (95.6%)</td>
<td>420 (100%)</td>
<td>140 (90.3%)</td>
<td>798 (7.0%)</td>
</tr>
</tbody>
</table>

*3M asSing forms only of 931 Most Common Verbs, no pronoun suffixes

These data are at first glance not particularly comforting. The Kuwaiti data matches the Watan newspaper data from Kuwait quite closely, but the data for each other country is troubling in one respect or another. The Moroccan data contains no sawfa at all. One explanation for this is that within any country, particular writers have preferences, and it is only in the aggregate that a country will match its usual average. Since most of the Moroccan data in the Hayat is from a very few writers, it is not unlikely that these few disfavor sawfa even though Moroccans writers in general use it sometimes. The Lebanese authors used sawfa a little more frequently than the overall paper does (although still not very frequently). Since the majority of articles in the Hayat are unsigned, and we can assume that most of the unsigned articles are also by Lebanese authors, what we can assume from this is that sawfa is a little more likely to appear in signed than in unsigned articles. We will give possible reasons later in the paper for why this might be the case. The Egyptian numbers, however, are truly confounding. They match the overall Lebanese rather than the
expected Egyptian use of *sawfa*. One might propose that writers for the Hayat, no matter what their nationality, try to match a known Hayat style, but since the choice between *sa*- and *sawfa* is very un-salient and authors are likely to be unaware of it, this seems a rather unlikely and counter-intuitive explanation. I believe that I do have an explanation for this anomaly, but to understand it we will have to cover some other material first, so it too will wait till later in the paper.

5.1 Genre Studies

The articles in the newspaper corpus are divided by the papers themselves into types. Each paper chooses its own categories, so they are not directly comparable, but we might be able to find some general trends by looking at the data divided into these ‘genres’ of articles. A look at an initial tabulation of data from the Ahram in Table 6 clearly shows an effect of article type, with some types showing an extremely high use of *sawfa*, and others a much lower use.

<table>
<thead>
<tr>
<th>Technique</th>
<th>sa-</th>
<th>sawfa</th>
<th>%sawfa</th>
</tr>
</thead>
<tbody>
<tr>
<td>A MOD</td>
<td>1105</td>
<td>1528</td>
<td>42.0%</td>
</tr>
<tr>
<td>FILE</td>
<td>1233</td>
<td>2149</td>
<td>36.5%</td>
</tr>
<tr>
<td>POST</td>
<td>210</td>
<td>458</td>
<td>31.4%</td>
</tr>
<tr>
<td>INVE</td>
<td>961</td>
<td>2380</td>
<td>28.8%</td>
</tr>
<tr>
<td>WRIT</td>
<td>748</td>
<td>1977</td>
<td>27.4%</td>
</tr>
<tr>
<td>OPIN</td>
<td>1448</td>
<td>3895</td>
<td>27.1%</td>
</tr>
<tr>
<td>SPOR</td>
<td>1911</td>
<td>8338</td>
<td>18.6%</td>
</tr>
<tr>
<td>FRON</td>
<td>1601</td>
<td>7485</td>
<td>17.6%</td>
</tr>
<tr>
<td>EGYP</td>
<td>591</td>
<td>3009</td>
<td>16.4%</td>
</tr>
<tr>
<td>REPO</td>
<td>715</td>
<td>3756</td>
<td>16.0%</td>
</tr>
<tr>
<td>ECON</td>
<td>723</td>
<td>3947</td>
<td>15.5%</td>
</tr>
<tr>
<td>ARTS</td>
<td>243</td>
<td>1478</td>
<td>14.1%</td>
</tr>
<tr>
<td>WORL</td>
<td>568</td>
<td>4056</td>
<td>12.3%</td>
</tr>
<tr>
<td>ARAB</td>
<td>309</td>
<td>3858</td>
<td>7.4%</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td>12366</td>
<td>48314</td>
<td>20.4%</td>
</tr>
</tbody>
</table>

*All forms of 931 Most Common Verbs, including pronoun suffixes*

Notice that the top six categories have high percentages of *sawfa*, over one-fourth, while the bottom eight categories descend from about one-fifth to almost one-twentieth. These differences are so striking that
one cannot help wondering if there is something here that might help us tease out the social or rhetorical ‘meaning’ of sawfa. The four letter categories provided by the Ahram for these articles refer to, in order: Columns, Ahram Files (longer interpretive articles), Letters to the editor, Investigations (another type of longer, interpretive article), Writers (i.e. well-known columnists), Opinion articles, Sports news, Front Page news, Egypt news, Reports from Ahram correspondents (usually stationed abroad), Economy, Arts news, World news, and Arab World news. Clearly, the results for these categories lump together in a way that makes sense. All the categories with more than 25% sawfa are clearly non-news items: letters to the editor, editorials, and longer pieces that delve into a subject. The categories with less than 20% sawfa are news oriented, the only exceptions being Arts, which would include much news and possibly some commentary. The basic pattern, therefore, seems to be that a relatively low percentage of sawfa is used in straightforward news articles, while a relatively higher percentage is used in commentary, opinion, and longer articles. It is almost as if the ‘contracted’ nature of the sa-prefix is reflected in its higher usage in ‘contracted’ settings where expansiveness is not expected, and that the percentage of sawfa grows as authors feel they have more room to write extensively about their subject.

The data for the Hayat categories looks quite different because the Hayat has fewer categories, as can be seen in Table 7.

<table>
<thead>
<tr>
<th>Category</th>
<th>sa-</th>
<th>sawfa</th>
<th>% sawfa</th>
</tr>
</thead>
<tbody>
<tr>
<td>GEN</td>
<td>1268</td>
<td>17971</td>
<td>6.6%</td>
</tr>
<tr>
<td>COM</td>
<td>59</td>
<td>841</td>
<td>6.6%</td>
</tr>
<tr>
<td>SPO</td>
<td>68</td>
<td>5241</td>
<td>1.3%</td>
</tr>
<tr>
<td>SCI</td>
<td>4</td>
<td>329</td>
<td>1.2%</td>
</tr>
<tr>
<td>NEW</td>
<td>189</td>
<td>26379</td>
<td>0.7%</td>
</tr>
<tr>
<td>BUS</td>
<td>83</td>
<td>12834</td>
<td>0.6%</td>
</tr>
<tr>
<td>CAR</td>
<td>0</td>
<td>446</td>
<td>0.0%</td>
</tr>
<tr>
<td></td>
<td>1671</td>
<td>64041</td>
<td>2.5%</td>
</tr>
</tbody>
</table>

*All forms of 931 Most Common Verbs, including pronoun suffixes

GEN is a very large category that seems to include all opinion pieces, commentary, and longer, investigative articles. COM
(computers), CAR (automobile), and SCI (science) are very small categories with articles and news about their respective fields. SPO (sports), BUS (business news) and NEW (news) cover the expected areas. Ignoring the three small categories, we see the same pattern here that we saw in the Ahram. The opinion and commentary pieces have significantly higher use of sawfa, while the news articles have a very low use of sawfa, close to one in a hundred. In both papers, interestingly, sports news patterns with news, but has a higher percentage of sawfa than other news categories. Notice, of course, that the Hayat percentages as a group are much lower than the Ahram percentages. The Ahram percentages varied between 42% and 7%, while the Hayat ones vary between 7% and 0%. But the overall pattern, and the general correspondence between the two, is strikingly clear: within their range, non-news articles had a much higher percentage of sawfa use than news articles.

The other two papers have similar patterns, although again their categories differ so much that there are not directly comparable. The Watan includes its editorials, opinion pieces and longer articles in the same categories as the news articles on the same subjects, so the differences between the categories are smaller. Still, front page news, foreign news, business news and Gulf news have lower than average use of sawfa, while last page articles, and articles on local topics have higher than average use.

The Moroccan newspaper has a much less clear pattern, and with more data we might discover that it really is different than the other papers. However, since the categories are more or less completely non-equivalent to those in the other papers, no such claim will be made here. But it is interesting to note the the ‘world’ category has a higher than average use of sawfa, unlike all the other papers, while several non-news categories like ‘thought’ and ‘last page’ have low uses of sawfa. Also in contrast to the other papers, the ‘sports’ category has the lowest use of all. So it is possible that the rhetorical ‘meaning’ of sawfa is different in Morocco than it is in the eastern parts of the Arab World. I have data from studies of other types of markers showing that Maghrib-area papers pattern differently from all the others, so this would not be the only result of this nature.

These genre differences lead us to a convincing explanation as to why the signed Lebanese articles should show a higher percentage of
sawfa usage than the paper as a whole. We have seen that sawfa is much more likely to be used in a non-news article, and signed articles are much more likely to be non-news as well. Over 60% of the signed Lebanese articles in the Hayat are in the GEN category, while a majority of the unsigned articles are in the NEW category. It thus seems reasonable that the corpus of signed Lebanese articles would show a higher percentage of sawfa than the newspaper as a whole.

5.2 Individual Preferences

It was stated above in relation to the Moroccan signed articles in the Hayat that individuals might vary considerably in their usage from the norm in their country, and that the stability of these percentages is established only with large numbers of writers who ‘even each other out.’ This became evident when I went looking for some explanation as to why the Egyptian signed articles in the Hayat demonstrated such a low percentage of sawfa compared to the Ahram. Two reasons became apparent. The first is that, for whatever reason, 85% of the signed articles in the Egyptian authors Hayat corpus are in the NEW, BUS and SPO categories, with only 15% in the GEN category. This prejudices the corpus against sawfa from the beginning. Even in the GEN category, however, the percentage is lower than expected. I therefore chose two of the authors who wrote regular GEN columns for the Hayat and analyzed them individually. Their results are interesting. The first is Ahmad Bahgat, a well know columnist who also writes for the Ahram. His topics are wide-ranging, but often more human interest than political. The other is Wahid Abdelmajid. He is the assistant director of the Ahram Center for Political Studies, and thus works for the Ahram newspaper’s parent organization, but he also writes a regular political column for the Hayat. Their results are vastly different as can be seen in Table 8.

<table>
<thead>
<tr>
<th></th>
<th>sa- sawfa</th>
<th>%sawfa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ahmad Bahgat</td>
<td>41</td>
<td>17</td>
</tr>
<tr>
<td>Wahid Abdelmajid</td>
<td>171</td>
<td>4</td>
</tr>
</tbody>
</table>

*All forms of 931 Most Common Verbs, including pronoun suffixes
Ahmad Bahgat gives a high percentage of sawfa, higher than the average for the Ahram as a whole, but in line with non-news items in the Ahram. Wahid Abdelmajid, on the other hand, gives a very low percentage of sawfa, more in line with the Hayat as a whole, and even below what Lebanese writers in the Hayat are giving in non-news articles. There is a sense, then, in which the latter is an individual anomaly. Because there are only a few Egyptian authors in the corpus of Egyptian Hayat authors, his results overwhelm the others. We can assume that there are similar authors in Egypt as a whole who avoid sawfa, but there are few enough of them that when considered as part of a large corpus, they have little effect on the results.

It is fairly clear, therefore, that the anomalous Egyptian author data from the Hayat can be explained by a combination of the fact that most of the signed Egyptian articles in the Hayat are news-oriented, and that even of the few non-news oriented writers, one very prolific writer personally has a preference for sa-.

5.3 Individual Words

Another way of looking at these data is to consider the patterning of individual words. There is neither time nor space to do much of that for this paper, but to get a basic idea I took some of the most frequent verbs and looked at their percentage of use with sawfa and sa-. It turns out that many of the verbs that seem to prefer sawfa are content verbs (ṣaṣara, iktaṣaṭa, inḍamma, istamaṭa ihtaaja, istamarra), while many of the verbs that preferred sa- were generic verbs like kaana, tamma, ittāxaḍa, takawwana, etc. There were, of course, many exceptions. For illustrative purposes I include here the data for two common verbs, one with high sawfa usage and one with low. Although there are many unanswered questions about these patterns, the patterns themselves are seductive. Table 9 shows the basic patterns for the two verbs across the Ahram and Hayat newspapers.

Table 9: Future Particles with istamarra and tamma in the Ahram and Hayat, Count Technique B*

<table>
<thead>
<tr>
<th>istamarra</th>
<th>Ahram</th>
<th>Hayat</th>
</tr>
</thead>
<tbody>
<tr>
<td>sawfa</td>
<td>317</td>
<td>48</td>
</tr>
<tr>
<td></td>
<td>43.1%</td>
<td>6.6%</td>
</tr>
<tr>
<td>sa-</td>
<td>418</td>
<td>681</td>
</tr>
<tr>
<td></td>
<td>56.9%</td>
<td>93.4%</td>
</tr>
<tr>
<td>total</td>
<td>735</td>
<td>729</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
DILWORTH B. PARKINSON

A n examination of Table 9 seems to clearly show both a 'word-effect' and a 'country-effect,' with tamma having a much lower percentage of sawfa in both papers than istamarra, but within a context of relatively high sawfa use for the Ahram, and relatively low sawfa use for the Hayat. Table 10, while harder to figure out, shows just as clearly at least a semi-independent genre effect.

Table 10A: Future Particles with istamarra and tamma in the Ahram and Hayat across article genres, Count Technique B

<table>
<thead>
<tr>
<th></th>
<th>Ahram</th>
<th>Hayat</th>
</tr>
</thead>
<tbody>
<tr>
<td>sawfa</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is-</td>
<td>503</td>
<td>25</td>
</tr>
<tr>
<td>sawfa %</td>
<td>11.0%</td>
<td>1.2%</td>
</tr>
<tr>
<td>total</td>
<td>4568</td>
<td>2121</td>
</tr>
<tr>
<td></td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

*All forms of 931 Most Common Verbs, including pronoun suffixes

A n examination of Table 9 seems to clearly show both a ‘word-effect’ and a ‘country-effect,’ with tamma having a much lower percentage of sawfa in both papers than istamarra, but within a context of relatively high sawfa use for the Ahram, and relatively low sawfa use for the Hayat. Table 10, while harder to figure out, shows just as clearly at least a semi-independent genre effect.

Table 10A: Future Particles with istamarra and tamma in the Ahram and Hayat across article genres, Count Technique B

<table>
<thead>
<tr>
<th></th>
<th>Ahram</th>
<th>Hayat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sa-</td>
<td>sawfa</td>
</tr>
<tr>
<td>FILE</td>
<td>15</td>
<td>43</td>
</tr>
<tr>
<td>AMOD</td>
<td>10</td>
<td>20</td>
</tr>
<tr>
<td>INVE</td>
<td>9</td>
<td>11</td>
</tr>
<tr>
<td>OPIN</td>
<td>27</td>
<td>33</td>
</tr>
<tr>
<td>POST</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>REPO</td>
<td>32</td>
<td>29</td>
</tr>
<tr>
<td>ARTS</td>
<td>10</td>
<td>8</td>
</tr>
<tr>
<td>SPOR</td>
<td>36</td>
<td>26</td>
</tr>
<tr>
<td>ECON</td>
<td>29</td>
<td>18</td>
</tr>
<tr>
<td>FRON</td>
<td>108</td>
<td>66</td>
</tr>
<tr>
<td>EYP</td>
<td>30</td>
<td>18</td>
</tr>
<tr>
<td>WRT</td>
<td>12</td>
<td>6</td>
</tr>
<tr>
<td>WRL</td>
<td>56</td>
<td>23</td>
</tr>
<tr>
<td>ARAB</td>
<td>42</td>
<td>14</td>
</tr>
</tbody>
</table>
tamma

<table>
<thead>
<tr>
<th></th>
<th>Ahram</th>
<th>Hayat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sa-</td>
<td>sawfa</td>
</tr>
<tr>
<td>AMOD</td>
<td>33</td>
<td>18</td>
</tr>
<tr>
<td>FILE</td>
<td>47</td>
<td>25</td>
</tr>
<tr>
<td>WRT</td>
<td>22</td>
<td>7</td>
</tr>
<tr>
<td>OPIN</td>
<td>75</td>
<td>22</td>
</tr>
<tr>
<td>INVE</td>
<td>330</td>
<td>92</td>
</tr>
<tr>
<td>POST</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>SPOR</td>
<td>441</td>
<td>97</td>
</tr>
<tr>
<td>REPO</td>
<td>107</td>
<td>12</td>
</tr>
<tr>
<td>FRON</td>
<td>671</td>
<td>69</td>
</tr>
<tr>
<td>WORL</td>
<td>220</td>
<td>18</td>
</tr>
<tr>
<td>ECON</td>
<td>787</td>
<td>59</td>
</tr>
<tr>
<td>EGYP</td>
<td>947</td>
<td>69</td>
</tr>
<tr>
<td>ARAB</td>
<td>272</td>
<td>10</td>
</tr>
<tr>
<td>ARTS</td>
<td>109</td>
<td>4</td>
</tr>
</tbody>
</table>

Table 10B
Summary of sawfa % of Table 10A

<table>
<thead>
<tr>
<th></th>
<th>Ahram</th>
<th>Hayat</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sawfa</td>
<td></td>
</tr>
<tr>
<td>istamarra</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A hram</td>
<td>Hayat</td>
<td></td>
</tr>
<tr>
<td>non-news</td>
<td>63.4%</td>
<td>16.9%</td>
</tr>
<tr>
<td>News</td>
<td>36.9%</td>
<td>3.2%</td>
</tr>
<tr>
<td>tamma</td>
<td></td>
<td></td>
</tr>
<tr>
<td>A hram</td>
<td>Hayat</td>
<td></td>
</tr>
<tr>
<td>non-news</td>
<td>24.4%</td>
<td>7.1%</td>
</tr>
<tr>
<td>News</td>
<td>8.7%</td>
<td>0.3%</td>
</tr>
</tbody>
</table>

Although there may be something to the notion that the ‘word-effect’ comes from the fact that particular words are more likely to be used in a particular genre, and is thus derivative of the ‘genre-effect’, that does not appear to explain these particular data, since istamarra, the word with high sawfa percentages, is actually used more in news than in non-news articles. On the other hand, tamma, the verb with low overall sawfa percentages, is used quite a bit more frequently in news
articles indicating that for it ‘word-effect’ and ‘genre-effect’ are working together. It is still relatively common in non-news articles, however. The most striking thing for me is how clear the ‘genre-effect’ is, even for individual verbs and taking the ‘word-effect’ into account. Within their respective ranges of variation set by the paper they appear in, both words clearly have much more sawfa in the non-news than in the news articles, but this is more than twice as high with the ‘pro-sawfa’ word than it is with the ‘anti-sawfa’ word. These patterns are so obviously significant that I don’t think we can afford to ignore them in our quest to understand the meaning of sawfa-sa variation.

Until a large number of these word studies are done, however, it would be a mistake to read too much into this. We simply don’t know the effect of word size (istamarra is a ‘big’ word, and ‘tamma’ is a small, little word), person, and a host of other variables that could potentially affect the choice.

One would hazard a prediction, however, that a computer model could be devised that, given the paper, country, type of column and the identity of the words, could do a pretty good job of predicting sawfa-sa choices in particular instances, and even better at predicting their relative percentages overall.

5.4 Non-marked futures

It is well known, of course, that many verbs that refer to a future time in Arabic are not marked with either future particle, but are, rather, simply plain imperfect. An accurate count of which imperfect verbs actually refer to a future time and which do not would require a tagged corpus. To give a ball-park figure of simply how many imperfect verbs that COULD have been marked with either sawfa or sa- actually were, I did a machine count of all the bare imperfect verbs and subtracted from that number all those preceded by sawfa and by any other particle that would preclude sawfa, such as tan or kaii. This total was then compared with the totals for sawfa and sa-. This procedure was carried out for two separate years of the Hayat, and for two separate halves of the Ahram year. It turns out that a fairly stable percentage of all verbs that can be, are marked with one of the future particles, between 8% and 9% generally. This means that all that constrained variability that has been presented in the rest of this paper is going on in a subset of less than 10% of the imperfect verbs that could accept these markers.
In other words, we have been concentrating on a set of patterns that are certainly embedded into a broader set of patterns that would need to be investigated to understand the whole system.

5.5 Near and Far Future

I searched for the occurrence of sawfa and sa- in conjunction with a number of time expressions that might code for near and far future, and found no relationship whatsoever. For example, of all the occurrences of the word al-yawm ‘today’ in the Ahram sawfa-sa corpus, 23% occurred with sawfa and the rest with sa-. This is, clearly, close to the basic percentages of sawfa and sa- in the corpus in the first place, and indicates no preference of one over the other in the presence of al-yawm. Several similar sentences were found with both sawfa and sa-; for example, sentences declaring that a particular person’s funeral will be held later today occurred with both particles. For modern, newspaper usage, therefore, I think we can safely discard the notion that near and far future have anything to do with the choice between these particles.

5.6 Possible Explanations

What have we learned, then, about the social or rhetorical meaning of choosing sawfa as opposed to sa-? Clearly we need some way of either explaining or characterizing the kinds of strong statistical effects that we have seen in the data.

5.7 The Country Effect

The ‘country-effect’ appears to be very strong and stable. Although it interacts with individual preferences, and with individual genre and word effects, it clearly has a strong influence of its own. I don’t believe there is any evidence that anyone is aware of this effect, so it seems unlikely that anyone would consciously or subconsciously make their choices specifically to mark themselves as, say, Egyptian or Lebanese. Rather, we would need to say that in the overall community of Standard Arabic readers and writers, there are strong sub-communities which individuals belong to, and which define most of the literary interactions of most individuals. In other words, Egyptians read and write for, mainly, other Egyptians, and Lebanese the same. Since the majority of models for these writers are from a single
community, it makes sense that the unconscious patterns these writers acquire would be from their own community. What kind of mental machinery allows individuals to acquire and produce statistically significant, patterned data seemingly outside the grammatical system of the language is a question beyond the scope of this paper; I will say, however, that the machinery proposed by Skousen’s Analogical Modeling of Language Theory seems to provide the necessary mechanism without proposing counterintuitive features, or trying to cover every outcome with complicated rules.

One thing the country-effect implies is that choosing sawfa really has a different meaning in different countries. In all countries sawfa is the less common, and thus marked, form. But in Egypt, it is relatively common overall (chosen about one-fifth of the time), and in sawfa-favoring contexts (non-news contexts, with particular words), the likelihood of sawfa being chosen can rise to well over 50%. In the Lebanese context, sawfa is quite rare overall (chosen only about one in 50 times), and even in sawfa-favoring contexts the percentage of sawfa rarely rises above 7-15%. A text that is clearly marked in the Lebanese context as having high-sawfa would be read in an Egyptian context as having low-sawfa. Thus, whatever the rhetorical force of choosing sawfa is, it is only interpretable within a single sub-community, and then not in terms of a single choice, but only in terms of the overall pattern of choices in a text. The country effect tells the writer and the reader what the ‘expected’ range of usage is, i.e. what the normal usage is, thus allowing writers to manipulate and readers to properly interpret deviations from normal on either the high or low side. In other words, for example, nothing is keeping a Lebanese writer from using 40% sawfa in a news-oriented article, but doing so would give an extremely marked, non-newsy feel to the article. The strong norm-setting of this effect, therefore, allows for a kind of rhetorical or social meaning that would not be possible without it.

Actually explaining the country effect, however, is quite difficult. Why is it that Egyptians use so much more sawfa than others? There is so much we don’t know that any proposed explanation would be speculation. If we assume that the 25% sawfa usage of the Quran is the base from which all country usages developed, then we would need to find out when things started to change in places like Lebanon, how fast they changed, whether they changed only in the newspaper genre, or
elsewhere as well, and a host of other factors. We would need to look at other papers in Egypt, ones with a less ‘stuffy’ reputation, as well as at other genres, to see if perhaps a similar change is happening on the fringes in Egypt. Possible speculations about the ‘meaning’ of this difference include the idea that sawfa is somehow more old-fashioned than sa-, and that Lebanese writers have no interest in identifying with old-fashionedness, while Egyptian writers have a stronger tie to the old heritage, and in some instances want to identify with it. Another possibility is that sawfa is more formal than sa- and Lebanese writers as a whole avoid marking their texts as formal, while Egyptians don’t mind doing so. There are other possibilities as well. None strike me as terribly convincing or seductive, although they are certainly possible.

5.8 The Genre Effect
The ‘genre-effect’ is also strong, stable and clear. What is not clear is whether we have characterized it adequately (as a news-oriented/non-news-oriented distinction); in fact, it is almost certain that we have not, given the specific nature of the data on which it is based. We have characterized it here as a news-oriented and non-news-oriented distinction, but this is almost certainly an artifact of the fact that we have had to rely on the newspapers own characterizations of their articles for our categories. The important question is: what (hopefully more intuitive) distinction underlies the category effects that the data show. Formal vs. informal? Abbreviated vs. long-winded? Fact-oriented vs. opinion oriented? I am somewhat uncomfortable simply characterizing it as formal vs. informal, since news articles, particularly in highly respected newspapers like those considered here, could hardly be characterized as informal, and the pieces included in the non-news-oriented category would include both highly formal ‘thought’ pieces, and quite informal, and often humorous, essays and other pieces. Further, there is nothing about the expression of opinion, per se, that would lead one to assume a particular style. There is probably something to the idea that since sa- is considered a shortened form of sawfa, use of sawfa iconically represents taking one’s time, speaking expansively about a topic, while use of sa- iconically implies getting the most content in the least amount of verbiage. This distinction might be characterized as thoughtful vs. efficient, in terms of
information packaging and marking a text in the way you want it to be taken.

The interesting thing about the genre effect is that it appears clearly to be at work in all countries in more or less the same way, but within the percentage constraints imposed by the overarching country effect. In other words, everyone greatly raises their percentage of sawfa in non-news articles, and dramatically lowers it in news articles, but the range in which they do so is vastly different for the different countries. Thus, the meaning of a particular usage, or of the pattern of usage of a particular article, can only be interpreted in its own country context, since a low usage in the Ahram could be read as a high usage in the Hayat. Both Egypt and Lebanon share the pattern and meaning of the genre effect, but readers from either place who are not intimately acquainted with writing from the other place will almost certainly misinterpret the ‘feel’ of articles from the other place, perhaps making news articles written by Egyptians and read by Lebanese feel not direct and straightforward enough, and perhaps making non-news articles written by Lebanese and read by Egyptians feel less thoughtful than expected, or perhaps too curt.

5.9 The Word Effect

The word effect investigation here was very preliminary, but the results for the two verbs investigated were very clear. Certain verbs strongly favor sawfa, while others favor sa-, over all countries and genres. Much more would need to be done before deciding whether this is because (as implied above) that ‘content’ verbs tend to be more sawfa oriented, and generic verbs more sa- oriented, or because of some other feature of the semantics, context, or morphology of the individual verbs. The effect in the two verbs investigated, however, is so strong and striking that this effect, like the other two, cries out for further investigation and convincing explanation.

6. Conclusion

Linguistic theory has tended to fairly narrowly define the aspects of language that are interesting and worth studying. Things that might have an effect on proposed grammatical, morphological and phonological rules are given primacy, and much other data is simply dismissed as irrelevant and uninteresting. Recent statistical approaches
to language continue to demonstrate the poverty of such an approach. We know all kinds of things, when we ‘know’ a language, beyond what is normally modeled as linguistic knowledge in a set of rules. Items in supposed ‘free variation’ with each other often turn out, upon investigation, to be highly constrained in interesting ways. We have demonstrated in this paper that the choice between sawfa and sa- is one such area. Its use varies dramatically from one country to another, from one genre to another, and even from one individual verb to another. I believe these results are clear and solid. What is not so solid at the present time is the appropriate way of characterizing these three kinds of variability, and giving an adequate explanation for them, and putting them into a broader context of a whole variable system. In other words, much remains to be done to understand what I have presented here. I hope, however, that the data has been compelling enough that it will impel both me and others to continue to search for that more cogent explanation, characterization and broader context, not only of sawfa and sa-, but of the host of other variables with which Arabic is so rife.

REFERENCES


Van Mol, Mark. 2000. “Exploring Annotated Arabic Corpora, Preliminary Results”. International Conference on Artificial and Computational Intelligence For Decision, Control and Automation In Engineering and Industrial Applications (ACIDCA), Monastir, Tunisia, March, 2000.
<table>
<thead>
<tr>
<th>Subject</th>
<th>Page Numbers</th>
</tr>
</thead>
<tbody>
<tr>
<td>affix grammars</td>
<td>157</td>
</tr>
<tr>
<td>Al-Ahram newspaper</td>
<td>157, 191, 192, 198-199</td>
</tr>
<tr>
<td>Al-Hayat newspaper</td>
<td>191, 192, 198-199</td>
</tr>
<tr>
<td>Arabic</td>
<td></td>
</tr>
<tr>
<td>future particles</td>
<td>194, 207</td>
</tr>
<tr>
<td>particles</td>
<td>191</td>
</tr>
<tr>
<td>roots and patterns</td>
<td>154-155, 183</td>
</tr>
<tr>
<td>syntax</td>
<td>46, 56, 73, 104</td>
</tr>
<tr>
<td>word classification</td>
<td>154</td>
</tr>
<tr>
<td>word order</td>
<td>93</td>
</tr>
<tr>
<td>clause structure</td>
<td>91</td>
</tr>
<tr>
<td>code switching</td>
<td>19</td>
</tr>
<tr>
<td>code switching theories</td>
<td></td>
</tr>
<tr>
<td>composite matrix language</td>
<td>35</td>
</tr>
<tr>
<td>government principle</td>
<td>22</td>
</tr>
<tr>
<td>matrix language</td>
<td>19-20, 23-24</td>
</tr>
<tr>
<td>two constraints</td>
<td>20</td>
</tr>
<tr>
<td>complementary particles</td>
<td>142</td>
</tr>
<tr>
<td>condition C</td>
<td>64, 67, 68, 77</td>
</tr>
<tr>
<td>conduct avoidance</td>
<td>145</td>
</tr>
<tr>
<td>contextualization</td>
<td>150</td>
</tr>
<tr>
<td>copy theory</td>
<td>53, 56, 57</td>
</tr>
<tr>
<td>coreferential</td>
<td>92</td>
</tr>
<tr>
<td>corpus linguistics</td>
<td>136-137</td>
</tr>
<tr>
<td>Arabic text tagging</td>
<td>136, 139-140, 193</td>
</tr>
<tr>
<td>external structure</td>
<td>137</td>
</tr>
<tr>
<td>internal structure</td>
<td>138-139</td>
</tr>
<tr>
<td>speaking/writing</td>
<td>149</td>
</tr>
<tr>
<td>voweled corpora</td>
<td>187</td>
</tr>
<tr>
<td>diglossia</td>
<td>19, 136</td>
</tr>
<tr>
<td>discourse particles</td>
<td>4</td>
</tr>
<tr>
<td>Dutch</td>
<td>26</td>
</tr>
<tr>
<td>existential fii</td>
<td>33</td>
</tr>
<tr>
<td>feature strength</td>
<td>103-104</td>
</tr>
<tr>
<td>gender</td>
<td>112</td>
</tr>
<tr>
<td>head-raising</td>
<td>57, 68-69</td>
</tr>
<tr>
<td>Hebrew</td>
<td>22</td>
</tr>
<tr>
<td>horizontal syntactical variation</td>
<td>135</td>
</tr>
<tr>
<td>Hungarian</td>
<td>35</td>
</tr>
<tr>
<td>impoverished number agreement</td>
<td>85</td>
</tr>
<tr>
<td>phrasal vs. clausal analysis</td>
<td>89, 101, 128</td>
</tr>
<tr>
<td>lexical analyzers</td>
<td></td>
</tr>
<tr>
<td>morphemes</td>
<td>23-24, 181</td>
</tr>
<tr>
<td>negation</td>
<td>6</td>
</tr>
<tr>
<td>noun phrase</td>
<td>91-92, 161</td>
</tr>
</tbody>
</table>
oath phrases., 3-4

Panjabi, 23
parser, 158-159
politeness, 10
power, 12, 15

Quran, 47, 194-195

reconstruction, 53-54
restrictive relative clauses, 88
resumptive pronouns, 72

saturation principle, 138
Semitic syntax, 56
Skousen, 208
Spanish, 21
speech act expressions, 41
Suurat al Baqarah, 42-45
Swahili, 25

ta- as a reciprocal complement, 87,
94, 96-99
Tok Pisin, 21

uniformity condition, 116

variable binding, 61, 68
vowelling, 181-182

Wahhabites, 138
Current Issues in Linguistic Theory (CILT) is a theory-oriented series which welcomes contributions from scholars who have significant proposals to make towards the advancement of our understanding of language, its structure, functioning and development. CILT has been established in order to provide a forum for the presentation and discussion of linguistic opinions of scholars who do not necessarily accept the prevailing mode of thought in linguistic science. It offers an outlet for meaningful contributions to the current linguistic debate, and furnishes the diversity of opinion which a healthy discipline must have.

A complete list of titles in this series can be found on the publishers’ website, www.benjamins.com

303 TORCK, Danièle and W. Leo WETZELS (eds.): Romance Languages and Linguistic Theory 2006. Selected papers from ‘Going Romance’, Amsterdam, 7–9 December 2006. ca. 300 pp. Forthcoming
302 FERRARESI, Gisella and Maria GOLDBACH (eds.): Principles of Syntactic Recontruction. xvii, 216 pp. + index. Expected November 2008
288 CAMACHO, José, Nydia FLORES-FERRÁN, Liliana SÁNCHEZ, Viviane DÉPREZ and María José CABRERA (eds.): Romance Linguistics 2006. Selected papers from the 36th Linguistic Symposium on Romance Languages (LSRL), New Brunswick, March-April 2006. 2007. viii, 340 pp.
287 WEIJER, Jeroen van de and Erik Jan van der TORRE (eds.): Voicing in Dutch. (De)voicing – phonology, phonetics, and psycholinguistics. 2007. x, 186 pp.


SATTERFIELD, Teresa, Christina TORTORA and Diana CRESTI (eds.): Current Issues in Romance Languages. Selected papers from the 29th Linguistic Symposium on Romance Languages (LSRL), Ann Arbor, 8–11 April 1999. 2002. viii, 412 pp.


ELŠIK, Viktor and Yaron MATRAS (eds.): Grammatical Relations in Romani. The Noun Phrase. with a Foreword by Frans Plank (Universität Konstanz). 2000. x, 244 pp.


<table>
<thead>
<tr>
<th>Title</th>
<th>Editors/Authors</th>
<th>Year</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clitics, Pronouns and Movement.</td>
<td>BLACK, James R. and Virginia MOTAPANYE (eds.)</td>
<td>1997</td>
<td>375 pp</td>
</tr>
<tr>
<td>Microparametric Syntax and Dialect Variation.</td>
<td>BLACK, James R. and Virginia MOTAPANYE (eds.)</td>
<td>1996</td>
<td>xviii, 269 pp</td>
</tr>
<tr>
<td>Theoretical Linguistics and Grammatical Description. Papers in honour of Hans-Heinrich Lieb.</td>
<td>SACKMANN, Robin and Monika BUDDE (eds.)</td>
<td>1996</td>
<td>x, 375 pp</td>
</tr>
<tr>
<td>Germanic Linguistics. Syntactic and diachronic.</td>
<td>LIPPI-GREEN, Rosina L. and Joseph C. SALMONS (eds.)</td>
<td>1996</td>
<td>viii, 192 pp</td>
</tr>
<tr>
<td>Recent Advances in Natural Language Processing. Selected Papers from RANLP '95.</td>
<td>MITKOV, Ruslan and Nicolas NICOLEV (eds.)</td>
<td>1997</td>
<td>472 pp</td>
</tr>
<tr>
<td>Guidelines and Romance Languages. Selected papers from the 55th Linguistic Symposium on Romance Languages (LSRL XXV) Seattle, 2–4 March 1995.</td>
<td>ZAGONA, Karen (ed.)</td>
<td>1995</td>
<td>vi, 330 pp</td>
</tr>
<tr>
<td>Case Suspension and Binary Complement Structure in French.</td>
<td>HERSCHENSOHN, Julia:</td>
<td>1996</td>
<td>xi, 200 pp</td>
</tr>
<tr>
<td>Romani in Contact. The history, structure and sociology of a language.</td>
<td>MATRAS, Yaron (ed.): Romani in Contact. The history, structure and sociology of a language. 1995.</td>
<td>1995</td>
<td>viii, 208 pp</td>
</tr>
<tr>
<td>On Subject and Theme. A discourse functional perspective.</td>
<td>HASAN, Ruqaiya and Peter H. FRIES (eds.): On Subject and Theme. A discourse functional perspective. 1995. xii, 414 pp.</td>
<td>1995</td>
<td>xii, 414 pp</td>
</tr>
<tr>
<td>Lexical Knowledge in the Organization of Language.</td>
<td>EGLI, Urs, Peter E. PAUSE, Christoph SCHWARZE, Arnim von STECHOW and Götz WIENOLD (eds.): Lexical Knowledge in the Organization of Language. 1995. xiv, 167 pp.</td>
<td>1995</td>
<td>xiv, 167 pp</td>
</tr>
</tbody>
</table>
98 KIRSCHNER, Carl and Janet Ann DECESARIS (eds.): Studies in Romance Linguistics. Selected papers from the XVII Linguistic Symposium on Romance Languages. 1989. xii, 496 pp.
105 SANKOFF, David: Diversity and Diachrony. 1986. xii, 450 pp.
<table>
<thead>
<tr>
<th></th>
<th>Title</th>
<th>Editors/Compilers</th>
<th>Year</th>
<th>Pages</th>
</tr>
</thead>
</table>