Report on
the Introduction of Arabic to ParGram

Mohammed A. Attia, UMIST

To be presented to the ParGram Fall Meeting 2004
(from August 31 to September 4), at the National
Centre for Language Technology, School of Computing,
Dublin City University, Dublin, Ireland.

Abstract

The purpose of this paper is to shed light on some linguistic phenomena in Arabic which, I think, are not dealt with within the Lexical Functional Grammar and cannot be accommodated by the current formalisms of the Xerox Linguistic Environment. Among these challenges are the Arabic nominal sentence structure, the issue of external governors, and the pro-drop in verbal sentences.

1. ParGram

ParGram or Parallel Grammar (King 2004, Butt et al. 2002, Butt et al. 1998) is a project that aims at providing full syntactic representation for many languages (currently, English, French, German, Japanese, Malagasy, Norwegian, Urdu, and Welsh) within the framework of Lexical Functional Grammar (Bresnan 2001). The project uses the Xerox Linguistic Environment (XLE) as a platform for writing grammar rules and lexical entries. After providing enough rules and lexical entries, the system is expected to parse sentences and provide both the c(onstituent)-structure and f(unctional)-structure representation for each one. While c-structure accounts for language-specific lexical idiosyncrasies and syntactic particular differences, the f-structure is supposed to represent a level of abstraction higher enough to capture parallelism among different languages and reduce cross-linguistic syntactic differences.

For decades, many research centres across the world have been working on the computational analysis of different languages. Each group has been working within a different theoretical framework and sometimes even without a tangible theoretical framework at all. And each group has been employing different methodologies in dealing with different or even similar linguistic phenomena. What is both intriguing and ambitious about the ParGram project is that it has researchers and grammar writers in different languages working within the same theoretical framework and using the same formalisms and set of features and terminology. The ParGram has actually become a testing ground for the LFG where hypotheses are applied and continuously contested.

With every new language incorporated into the ParGram project, there is a new challenge as well as an added benefit. The challenge is to accommodate the language specific structures and the benefit is to introduce new ideas from that language. ParGram is not a set of rigid moulds in which each language must fit in, but rather a

flexible tool that can change to provide linguistically-motivated explanations and analyses of the different structures of any specific language, whether at the c-structure level or even at the deeper f-structure level. It is a real challenge to provide flexibility and at the same time maintain the consistency needed so that all grammar writers working in different languages can still understand each other.

2. Arabic

The version of Arabic I'm taking in my study is Modern Standard Arabic (MSA). When I mention Arabic throughout this paper I primary mean MSA as opposed to classical Arabic, the language of formal writing until roughly the first half of the 20th century. It was also the spoken language fairly before the medieval times. MSA is also opposed to colloquial Arabic, which is the dialects currently spoken in different parts if the Arab world. MSA is the language of the modern writing and the language of the news. It is the language unanimously understood by all Arabic speakers and the language taught in Arabic classes.

Arabic exhibits many complexities (Daimi 2001, Fehri 1993, Chalabi 2000) which pose no little challenge to theoretical as well as computational linguistics. This is a list some of the major issues involved in Arabic:

1. Arabic typology is different than the Latin alphabet.
2. Arabic writing direction is from right to left.
3. Arabic has a relatively free word order.
4. Beside the regular sentence structure of verb, subject and object, Arabic has a nominal sentence structure of a subject phrase and a predicate phrase, with no verb or copula.
5. Arabic is a highly inflectional language, the matter that makes Arabic morphological analysis complicated. Arabic words are built from roots rather than stems.
6. Arabic writing involves diacritization, which is largely ignored in modern texts, the matter that makes morphological analysis yet more difficult. Ali (2003) gives a good example that can make an English speaker grasp the complexity caused by dropping Arabic diacritization. Suppose that vowels are dropped from an English word and the result is 'sm'. The possibilities of the original word are: some, same, sum, and semi. Chalabi (2000) even claims that the absence of diacritization in Arabic poses a computation complexity "one order of magnitude bigger than handling Latin-based language counterparts".
7. Arabic is a clitic language. Clitics are (Crystal 1980) the morphemes that have the syntactic characteristics of a word but are morphologically bound to other words. In Arabic, many coordinating conjunctions, the definite article, many prepositions and particles, and a class of pronouns are all clitics that attach themselves either to the start or end of words. So complete sentences can be composed of what seems to be a single word. For example:
   wa'a`taitumuuniiha
   divided as:

wa	'a`taitum	uu	nii	ha
and	gave.pl	you.pl	me	it

   (and you gave it to me)
8. The inconsistent and irregular use of punctuation marks. Punctuation marks have been introduced rather recently into the Arabic writing system, yet it is

not as essential to meaning or closely observed as is the case with English. Arabic writers shift between ideas using conjugating conjunctions instead of punctuation marks. In MSA, however, due to the influence if translation which, to some extent, transfers punctuation marks from the target languages, and due to the tendency of modern writers to use punctuation marks more consistently, Arabic has come to see more punctuation. Yet, even in modern writing it is rather impossible to rely on the period, for example, as a demarcation of the sentence boundary.

9. Arabic is a pro-drop language. The subject can be omitted leaving any syntactic parser with the challenge to decide (Chalabi 2004), first, whether or not there is an omitted pronoun in the subject position and, second, what the antecedent of the omitted pronoun is.

3. Arabic Sentence Structure

Transformational-generative grammarians (Anshen 1968, Fehri 1993) have had an acrimonious argument about whether the original word order in Arabic is VSO or SVO. However, within Lexical Functional Grammar we do not have to concern ourselves with this issue. But we have to provide an adequate description for the s-structure and f-structure of all possible sentences that can arise because of the free word order in Arabic.

The traditional classification of Arabic sentences is: nominal for verbless sentences, and verbal for sentences which contain a verb. More explanation is provided in the next two sections.

3.1 Verbal Sentences

For the three elements of subject, verb and object, all different word orders of SVO, VSO, VOS, and OVS are possible. The only combinations that do not occur in Arabic are OSV and SOV.

An example of SVO:
al-waladu akala al-tuffahata
the-boy.nom ate the-apple.acc
(The boy ate the apple)

An example of VSO:
akala al-waladu al-tuffahata
ate the-boy.nom the-apple.acc
(The boy ate the apple)

An example of VOS:
akala al-tuffahata al-waladu
ate the-apple.acc the-boy.nom
(The boy ate the apple)

An example of OVS:
al-tuffahata akala al-waladu
the-apple.acc ate the-boy.nom
(The boy ate the apple)

For the above four sentences we will have as many different s-structures and different parse trees as there are different word orders. However these differences melt away in the f-structure, where the Arabic sentence analysis is no different from an English or a French one, see Figure 1. However, in the first sentence, where the subject comes first, there are two different analyses available. The first is already mentioned, and the second is to consider the subject as the subject phrase and the rest of the sentence as the predicate phrase in which case the subject of the verb is an elliptic pronoun that refers back to the subject. And this why when the subject comes initially the verb must agree in number, a condition not allowed when the subject follows the verb.

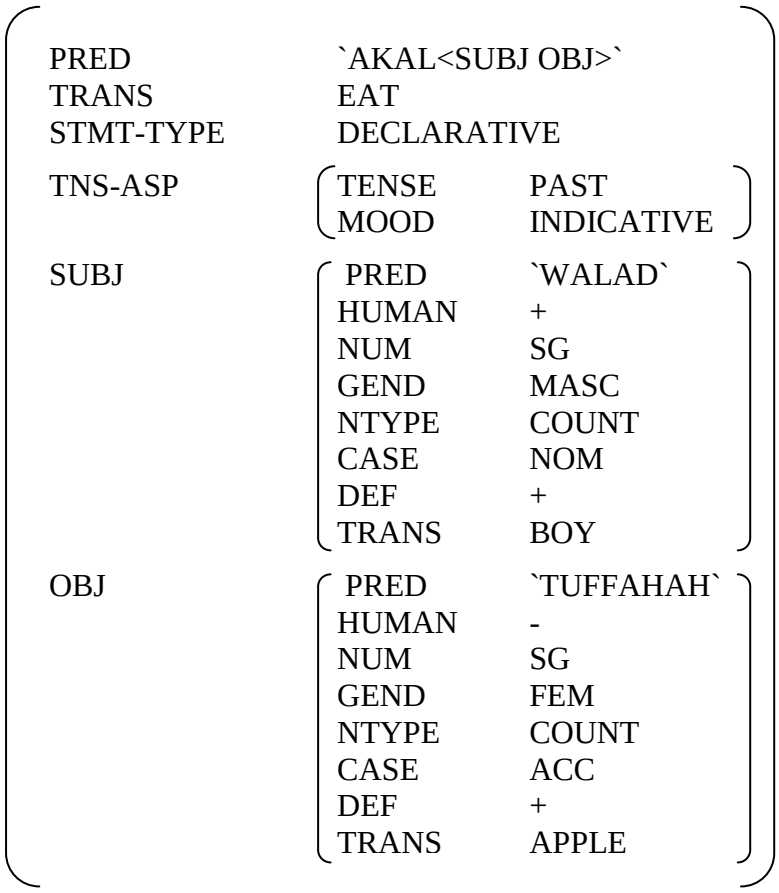

Fig. 1: F-structure of a verbal Arabic sentence

3.2 Nominal Sentences

Nominal sentences are the class of Arabic sentences that contain no explicit verb. They are composed of a subject phrase and a predicate phrase.

S --> NP {AP | NP | PP}

An example of nominal sentence of an NP followed by an AP:
al-shamsu mushriqatun
the-sun shining
(The sun is shining)

An example of nominal sentence of an NP followed by an NP:
Hada rajulun tayyibun
This a-man good
(The is a good man)

An example of nominal sentence of an NP followed by a PP:
Ar-rajulu fi ad-dari
the-man in the-house
(The man is in the house)

Moreover, the predicate phrase does not always have to follow the subject phrase. There are many (constrained) instances where the predicate phrase can be fronted, such as the following example.

fi ad-dari rajulun
in the-house a-man
(A man is in the house)

Japanese has a structure similar to the Arabic nominal sentences. Within ParGram (see Butt 2002), the Japanese sentences which are composed of a noun phrase and an adjective, the adjective is taken to be the main predicate of the sentence. If we copy the Japanese sentence analysis to Arabic, we will get an f-structure analysis as shown in Figure 2.

PRED	`MUSRIQATUN<SUBJ >`	
TRANS	SHINING	
GEND	FEM	
CASE	NOM	
STMT-TYPE	DECLARATIVE	
TNS-ASP	TENSE	PRES
	MOOD	INDICATIVE
SUBJ	PRED	`SHAMS`
	HUMAN	-
	NUM	SG
	GEND	FEM
	NTYPE	COUNT
	CASE	NOM
	DEF	+
	TRANS	SUN

Fig. 2: F-structure of a nominal Arabic sentence

However, I feel that this analysis is not linguistically motivated. There no evidence to support the idea that the adjective is either the main predicate or that it subcategorizes for a subject. Moreover, external governors, as will be shown later, can precede the whole structure and assign new cases to the subject and the predicate. If an external governor can assign case to the subject, this means that the adjective cannot be a main predicate or a case assigner.

Fehri (1993) argues that Arabic “verbless sentences, like verbal ones, are also headed by (abstract) T and AGR”. This means that the sentence is headed by an implied verb that carries the tense and defines the agreement features. This implicit verb must be explicit when the tense is changed either to the past or future. Moreover, nominal sentences in Hebrew, a Semitic language with a structure very similar to that of Arabic, are analysed as mixed category which are categorially nominal and functionally verbal (see Falk 2004). This makes Arabic nominal sentences eligible for an f-structure analysis similar to the English sentences of copula, subject and PredLink, as shown in Figure 3.

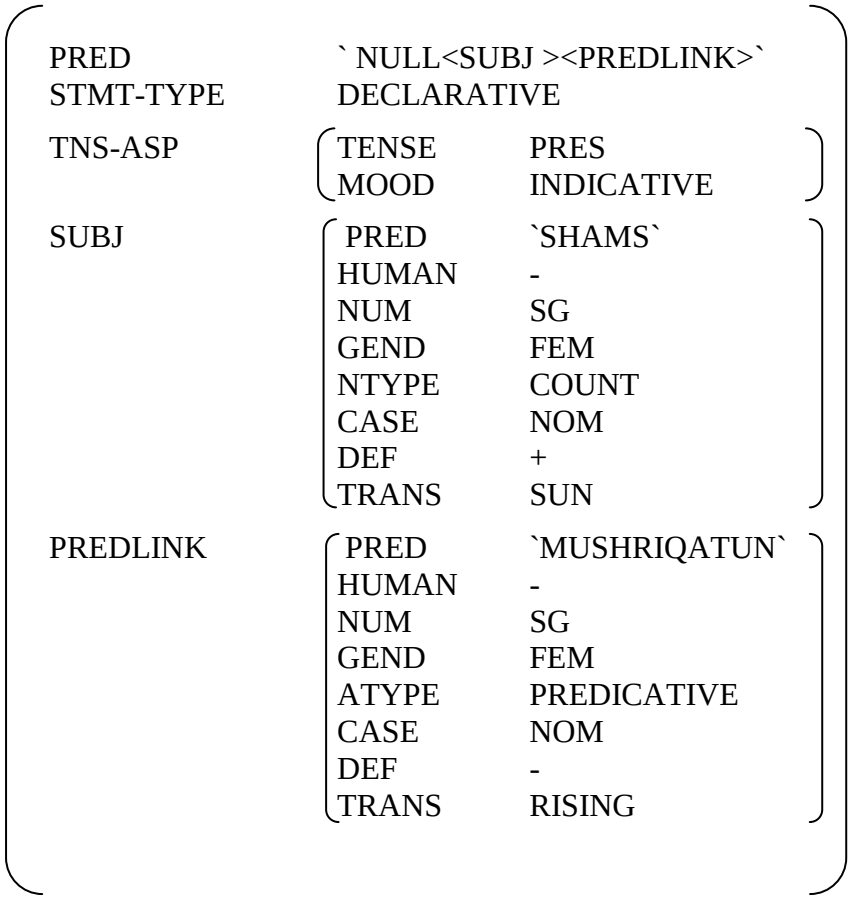

Fig. 3: Proposed f-structure of a nominal Arabic sentence

4. External governors

External governors (Fehri 1993) are a group of lexical items that precede a sentence and change the default case of the constituents. External governors can be verbs or particles. The most common among verb external governors are:

kana (was)

kanat al-shamsu mushriqatan
was the-sun.nom rising.acc
(The sun was rising)

asbaha (became)

asbahati al-shamsu mushriqatan
became the-sun.nom rising.acc
(The sun became rising)

laisa (is-not)
laisati al-shamsu mushriqatan
is-not the-sun.nom rising.acc
(The sun is not rising)

This group of verbs are like modal verbs in that they are not fully inflected. They precede nominal sentences and assign new cases to its two constituents. As seen in the examples, the predicate, which normally takes the nominal case, is now assigned the accusative case. According to traditional grammarians these verbs are case assigners in that they assign the nominative case to the subject and the accusative case to the predicate.

The most common among particle external governors are:

ma (not)
ma al-shamsu mushriqatan
not the-sun.nom rising.acc
(The sun is not rising)

la (not)
la ahadun qa'iman
no person.nom standing.acc
(no person is standing)

inna (affirmation, indeed)
inna al-shamsa mushriqatun
indeed the-sun.acc rising.nom
([Indeed] the sun is rising)

lakinna (but)
lakinna al-shamsa mushriqatun
but the-sun.acc rising.nom
(But the sun is rising)

ka'anna (as if)
ka'anna al-shamsa mushriqatun
as-if the-sun.acc rising.nom
(As if the sun is rising)

la`alla (might)
la`alla al-shamsa mushriqatun
might the-sun.acc rising.nom
(The sun might be rising)

The first two particles assign the nominative case to the subject and the accusative case to the predicate, while the rest assign the accusative case to the subject and the nominative case to the predicate.

Not only do external governors occur with nominal sentences, but they also precede verbal sentences when the subject is fronted or the object is topicalized. Whenever the subject and object occur after the verb, they are governed only by the verb and protected by it from external governors, but when they occur before it, they may be assigned cases different than their default cases.

‘inna at-taliba yahtarimu ustazahu
indeed the student.acc respect his-teacher.acc
(Indeed the student respects his teacher)

‘inna at-taliba yahtarimuhu ustazuhu
indeed the student.acc respect-him his-teacher.nom
([Indeed] the student, his teacher respects him)

In the first example *at-taliba*, the subject, which normally receives the nominative case as a default is now receiving the accusative case because it is preceded by the particle *‘inna*. In the second example *at-taliba*, the topicalized object, which normally receives the nominative because of topicalization which makes it like the subject of a nominal sentence, is now receiving the accusative case because it is preceded again by the particle *‘inna*. We need to note that only topicalized objects, not fronted objects, can come in this context. The difference is that the sentence following a topicazied object must contain a pronoun that takes the object as it antecedent.

The above external governors govern the entire sentence structure, yet there are some situations where only the subject or object is governed in a certain context and assigned a case different from the default case.

kullu ut-talibi yahrimuna ustazahu
all.nom the-students.gen respect teacher.their
(all students respect their teacher)

Here the subject is assigned the genitive case by the specifier *kullu* where they constitute together a possessive construction. In this sentence the specifier takes the nominative case. Let’s look at a similar example but with the object:

ra’aitu thalathata rijalin
saw.1pers.sg three.acc men.gen
(I saw three men)

Here the object is assigned the genitive case by the cardinal *thalathata* where they constitute together a possessive construction. In this sentence the specifier takes the accusative case. Here the specifier in the NP receives the case, not the head of the NP.

External governors can even influence verbs (but not constituents governed by verbs) and change their morphological forms.

ta'kuluuna al-tuffah
eat.pl.2pers the-apples
(You eat the apple)

la ta'kuluu al-tuffah
not eat.pl.2pers the-apples
(Do not eat the apple)

lan ta'kuluu al-tuffah
not eat.pl.2pers the-apples
(You will not eat the apple)

In the above examples when the negation particle *la* or *lan* precede the verb, it is morphologically changed.

5. Pro-Drop

Arabic is a pro-drop language. The pro-drop theory (Baptista 1995 and Chomsky 1981) stipulates that a null category is allowed in the subject position of a finite clause if the agreement features on the verb are rich enough to enable its content to be recovered.

According to Chalabi (2004) there are two challenges that follow the pro-drop in Arabic. The first challenge is to decide whether there is a pro-drop or not. Let's look at the following example:

akalat al-dajajah
ate.fem the-chicken

In the above example we are not sure whether the NP following the verb is the subject (in this case the meaning is "the chicken ate") or the object and the subject is an elliptic pronoun means *she* and understood by the feminine mark on the verb (in which case the meaning will be "she ate the chicken").

The second challenge, after deciding that there is a null pronoun in the subject position, is to resolve the pronoun reference. Let's look at the following examples.

dhahaba 'ila al-hadiqati
went.sg.masc to the-garden
(He/it went to the garden)

dhahabat 'ila al-hadiqati
went.sg.fem to the-garden
(She/it went to the garden)

Dhahabaa 'ila al-hadiqati
went.dual.masc to the-garden
(They went to the garden)

dhahabataa 'ila al-hadiqati

went.dual.fem to the-garden
(They went to the garden)

dhahabuu ‘ila al-hadiqati
went.pl.masc to the-garden
(They went to the garden)

dhahabana ‘ila al-hadiqati
went.pl.fem to the-garden
(They went to the garden)

As noticed from the first two examples pronominal reference ambiguity needs to be resolved. The ambiguity results from the fact that the pronoun system in Arabic distinguishes largely between only two features of gender: masculine and feminine. So the ambiguity cause by a possible reference to a non-human or inanimate object must be resolved. Yet in the rest of the examples, ambiguity can be preserved in English, which has only one pronoun in the plural, but if the target language is not English this ambiguity may also need to be resolved.

6. Corpus

I collected my corpus from articles published on the Al-Jazeera website[1] in different areas (news, science, sports, health, economics, etc.) during 10 months from September 2003 to July 2004. It includes 21,384 articles, containing 11,394,351 words, of them there is a list of 29,592 unique words (i.e., after ignoring the repetition for each word).

I collected this corpus, after reviewing the Copyright and Terms of Use document[2], by searching for five common prepositions expecting that any article is to contain an occurrence of at least one of them. The five words are (*min* ‘from’ *ila* ‘to’ *`an* ‘about’ *`ala* ‘on’ *fi* ‘in’)

My reason for choosing the corpus from Al-Jazeera website is that Al-Jazeera has become the most popular and most influential media channel in the Arab world. Feuilherade (2004), the BBC reporter, states that Al-Jazeera station “enjoys an audience of over 35 million viewers in the Middle East and is probably the only institution of its kind able to reach so many Arab hearts and minds.” Al-Jazeera employs presenters and reporters from across the spectrum of the Arabic-speaking countries.

With data from the corpus I hope to find evidence to prove that some sentence structures are no longer used in modern writing (such as the OVS word order), and so I will not have to accommodate them in my grammar. As a result the grammar will be more simplified and the parse time will be reduced.

[1] www.al-jazeera.net
[2] See Al-Jazeera website:
http://english.aljazeera.net/NR/exeres/769D18A2-298F-4183-B882-C41A5D9BA996.htm?dialogboxmode=1

7. Morphological Analyser

In my analysis of Arabic within the XLE, I'll rely on the Xerox finite-state Arabic morphological analyser[3], which is supposed to be compatible with XLE.

8. Lexicon

I expect to build a lexicon of between 10,000 and 15,000 Arabic words, all extracted from the corpus and used as base forms after removing affixes. The subcategorization frame for each lexical item will be specified and a translation for each word will be provided.

Part of speech, word sense, and subcategorization frames for each lexical item will be limited only to the data provided by the corpus. In this way I will avoid word senses that are no longer used in modern writing.

Bibliography

Ali, Nabil. 2003. The Second Wave of Arabic Natural Language Processing. A paper presented to the Expert Group Meeting on the Promotion of Digital Arabic Content, Beirut, 3-5 June 2003 http://www.escwa.org.lb/wsis/meetings/3-5june/docs/18.pdf

Anshen, Frank and Peter A. Schreiber. 1968. A Focus Transformation of Modern Standard Arabic. In *Language* 44-4: 792-797

Baptista, Marlyse. 1995. On the Nature of Pro-drop in Capeverdean Creole. Harvard Working Papers in Linguistics. Volume 5.

Bresnan, Joan. 2001. Lexical-Functional Syntax. Oxford: Blackwell.

Butt, Miriam, Helge Dyvik, Tracy Holloway King, Hiroshi Masuichi, and Christian Rohrer. 2002. The Parallel Grammar Project. In Proceedings of COLING-2002 Workshop on Grammar Engineering and Evaluation. pp. 1-7.

Butt, Miriam, Tracy Holloway King, Maria-Eugenia Nino, and Frederique Segond. 1998. *A Grammar Writer's Cookbook*. Stanford, CA: CSLI Publications

Chalabi, Achraf. 2000. MT-Based Transparent Arabization of the Internet TARJIM.COM. In White, J.S. (Ed.). *AMTA 2000, LNAI 1934*. Springer: Verlag Berlin Heidelberg, pp. 189-191.

Chalabi, Achraf. 2004. Elliptic Personal Pronoun and MT in Arabic. In JEP-2004-TALN 2004 Special Session on Arabic Language Processing-Text and Speech. http://www.lpl.univ-aix.fr/jep-taln04/proceed/actes/arabe2004/TAAC17.pdf

[3] See Xerox website:
http://www.xrce.xerox.com/competencies/content-analysis/arabic/info/info.html

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Foris. Dordrecht.

Crystal, David. 1980. *A First Dictionary of Linguistics and Phonetics*. Westview Press

Daimi, Kevin. 2001. Identifying Syntactic Ambiguities in Single-Parse Arabic Sentence. In *Computers and Humanities* 35:333-349.

Falk, Yehuda N. 2004. The Hebrew Present-Tense Copula as a Mixed Category. A paper presented to LFG Conference 2004. http://www-lfg.stanford.edu/lfg/lfg2004/abstracts/lfg04-abs-falk.pdf

Fehri, Abdelkader Fassi. 1993. *Issues in the Structure of Arabic Clauses and Words*. Kluwer Academic Publishers, Dordrecht, Holland.

Feuilherade, Peter. 2004. Al-Jazeera debates its future. http://news.bbc.co.uk/1/hi/world/middle_east/3889551.stm

King, Tracy Holloway. 2004. Parallel Grammar Project. http://www2.parc.com/istl/groups/nltt/pargram/