Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Humanities

2008

Mohammed A. Attia

School of Languages, Linguistics and Cultures

Contents

Word Count: 73458

Abstract

Handling Arabic Morphological and Syntactic Ambiguity within the LFG Framework with a View to Machine Translation

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy in the Faculty of Humanities

2008

Mohammed Abdel-Aal Attia
School of Languages, Linguistics and Cultures

This research investigates different methodologies to manage the problem of morphological and syntactic ambiguities in Arabic. We build an Arabic parser using XLE (Xerox Linguistics Environment) which allows writing grammar rules and notations that follow the LFG formalisms. We also formulate a description of main syntactic structures in Arabic within the LFG framework. When tested on short sentences randomly selected from a corpus of news articles, our parser achieved 92% coverage after applying robustness techniques.

Morphological ambiguity in Arabic is a notorious problem due to the richness and complexity of Arabic morphology. We show how an ambiguity-controlled morphological analyzer is built in a rule-based system that takes the stem as the base form using finite state technology. We point out sources of genuine and spurious morphological ambiguities in Arabic and show how ambiguity in our system is reduced without compromising precision. We conduct an evaluation experiment that shows that our morphology outperforms both Buckwalter's and Xerox morphologies with regard to precision and avoidance of spurious ambiguities.

Syntactic ambiguity is also a major problem for large-scale computational grammars which cover a realistic and representative portion of a natural language. We identify sources of syntactic ambiguities in Arabic, focusing on four ambiguity-generating areas which have the greatest impact. These are the pro-drop nature of the language, word order flexibility, lack of diacritics, and the multifunctionality of Arabic nouns. We deal with ambiguity not as one big problem, but rather as a number of divisible problems spreading over all levels of the analysis: pre-parsing, parsing and post-parsing stages. The pre-parsing stage contains all the processes that feed into the parser such as tokenization, morphological analysis or POS tagging. The parsing phase covers the topics of granularity of phrase structure rules, lexical specifications, application of syntactic constraints, and domain specific adaptation. The post-parsing stage controls the selection and ranking of these solutions. We show how applying these techniques results in reducing parse time and keeping ambiguities within a manageable boundary.

XLE includes a parser, transfer and generator components, which makes it suitable for Machine Translation. We demonstrate the MT component in the ParGram project by applying simple transfer rules, and point out what needs to be done in order to produce a fully-fledged MT system.

Declaration

I hereby declare that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright Statement

Acknowledgments

In a few weeks time I will be back to Egypt, where I come from. Yet my stay in the UK was a thought-provoking, eye-opening experience. It has been a chance to learn and ponder, not only through books and lectures, but also through the revealing encounters with the society and the people. It has been a transformational journey for me, my wife and my, now bilingual, children: Mena, Nada and Youssef. As T.S. Eliot puts it “but even if we do end up back where we started, the journey is all.”

I would never have the chance to study in the UK without the financial aid of my sponsor, the government of the Arab Republic of Egypt. I feel deeply indebted to the people of Egypt for everything I learned in the UK. I hope to be able to repay part of the debt by transferring the knowledge and promoting research in Egypt.

I owe an enormous debt of gratitude to my supervisor, Harold Somers, for many reasons. First he put me in the right track since Day One saving me a lot of time and effort exploring uncountable paths. Second he kept motivating and supporting me along the way. He encouraged me to write papers, and attend meetings and conferences. He also provided the financial funds needed for travelling to these research events. Third, at times when I was about to go off rail or lose confidence, his advice helped put me back in the right track.

I’m greatly indebted for the members of my research panel: Allan Ramsay and Paul Bennett. Their remarks were invaluable in orienting and solidifying my research.

I consider myself lucky to have been associated with the members of the ParGram group who are the major contributors to the LFG theory, among them are: Ron Kaplan, Tracy Holloway King, Louisa Sadler, Miriam Butt, Mary Dalrymple, Josef von Genabeth, Victoria Rosén, Paul Meurer, Ingo Mittindof and Martin Forst. I owe them a great debt for giving me an invaluable help and feedback on various aspects of my research project in the ParGram meetings in Dublin, Ireland, 2004; Gotemba, Japan, 2005; and Oxford, UK, 2006. I’d like also to extend my thanks to Lauri Karttunen for his great help and advice on using Finite State Technology.

Author Statement

I would like here to acknowledge that Chapter Two “Managing Morphological Ambiguities” is an enhanced, modified and updated version of my paper titled “An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks” (Attia, 2006a).

Chapter Three “Tokenization” is an updated version of my paper titled “Arabic Tokenization System” (Attia, 2007).

Chapter Four “Handling Multiword Expressions” is an updated version of my paper titled “Accommodating Multiword Expressions in an Arabic LFG Grammar” (Attia, 2006b).

Transliteration Table[1]

Name of letter	Transliteration Symbol	Arabic letter shape
hamzah	ʾ	أ
bāʾ	b	ب
tāʾ	t	ت
ṯāʾ	ṯ	ث
ǧīm	ǧ	ج
ḥaʾ	ḥ	ح
ḫaʾ	ḫ	خ
dāl	d	د
ḏāl	ḏ	ذ
rāʾ	r	ر
zāy	z	ز
sīn	s	س
šīn	š	ش
ṣād	ṣ	ص
ḍād	ḍ	ض
ṭāʾ	ṭ	ط
ẓāʾ	ẓ	ظ
ʿaīn	ʿ	ع
ġaīn	ġ	غ
fāʾ	f	ف
qāf	q	ق
kāf	k	ك
lām	l	ل
mīm	m	م
nūn	n	ن
hāʾ	h	هـ
wāw	w	و
yāʾ	y	ي

Short Vowels

fatḥah	a	ـَ
kasrah	i	ـِ
ḍammah	u	ـُ

Long Vowels

ā
ī
ū

Compound Vowels

aw
ai

[1] We follow the DIN 31635 standard for the transliteration of the Arabic alphabet.

List of Abbreviations and Acronyms

Abbreviation	**Full Form**
1	first person
2	second person
3	third person
A	adjective
acc	accusative
ADJ	adjective
ADJP	adjectival phrase
ADV	adverb
ADVP	adverbial phrase
AGR	agreement
AP	adjectival phrase
ART	article
ATYPE	adjective type
card	cardinal (number)
CL	Computational Linguistics
COMP	complementizer
CONJ	conjunction
COORD	coordination
CP	complement phrase
c-structure	constituent-structure
D	determiner
DAT	dative
decl	declarative
def	definite
DET	determiner
dist	distal
dl	dual
ERG	ergative
fem	feminine
f-structure	functional-structure
fut	future
gen	genitive
GEND	gender
imp	imperative
indef	indefinite
INF	infinitive
int	interrogative
LFG	Lexical Functional Grammar
masc	masculine
MOD	modifier
MSA	Modern Standard Arabic
MT	Machine Translation
MWE	Multiword expressions
N	noun
neg	negative

NLP	Natural Language Processing
nom	nominative
NP	noun phrase
NSEM	noun semantics
NSYN	noun syntax
NTYPE	noun type
NUM	number
OBJ	object
OBL	oblique
ord	ordinal (number)
OVS	object–verb–subject
ParGram	Parallel Grammars
PART	particle
pass	passive
PCFG	Probabilistic Context Free Grammar
PERS	person
pl	plural
POS	part-of-speech
poss	possessive
PP	prepositional phrase
PRED	predicate
PredP	predicate phrase
PREP	preposition
pres	present
PRON	pronoun
prox	proximal
PS	Phrase Structure
quant	quantitive
rel	relative (pronoun)
S	sentence
sg	singular
SL	source language
SPEC	specifier
STMT-TYPE	statement-type
SUBJ	subject
SVO	subject–verb–object
TL	target language
TNS-ASP	tense-aspect
V	verb
VCop	copula verb
VOS	verb-object-subject
VSO	verb–subject–object
VTYPE	verb type
XADJUNCT	open adjunct
XCOMP	open complement
XLE	Xerox Linguistics Environment
XTE	Xerox Translation Environment

1 Introduction

This research investigates different methodologies to manage the problem of morphological and syntactic ambiguities in Arabic. When a computational grammar becomes mature and complex enough to deal with naturally occurring texts ambiguity becomes a natural consequence. When the grammar starts to deal with real data there is an explosion in the number of possible solutions for a given sentence. The number of solutions is usually factored by the number of words in a sentence and the use of certain ambiguity-prone constructions, such as coordination and prepositional phrases. The task of disambiguation requires that ambiguity is controlled at each level of the analysis and that plausible solutions surface as an output while implausible ones are discarded.

Morphological ambiguity in Arabic is a notorious problem due to the richness and complexity of Arabic morphology. We show how an ambiguity-controlled morphological analyzer is built in a rule-based system that takes the stem as the base form using finite state technology. We point out sources of genuine and spurious morphological ambiguities in Arabic and show how ambiguity in our system is reduced without compromising precision.

Syntactic ambiguity is also a major problem for large-scale computational grammars which cover a realistic and representative portion of a natural language. We identify sources of syntactic ambiguities in Arabic, focusing on four ambiguity-generating areas which have the greatest impact. These are the pro-drop nature of the language, word order flexibility, lack of diacritics, and the multifunctionality of Arabic nouns. We deal with ambiguity not as one big problem, but rather as a number of divisible problems spreading over all levels of the analysis: pre-parsing, parsing and post-parsing stages. The pre-parsing stage contains all the processes that feed into the parser whether by splitting a running text into manageable components (tokenization), analyzing words (morphological analyzer) or tagging the text. These processes are at the bottom of the parsing system and the effect of ambiguity in this stage is tremendous as it propagates exponentially into the higher levels. The parsing stage is the process when the syntactic rules and constraints are applied to a text, and the subcategorization frames are specified. The post-parsing stage has no effect on the number of solutions already produced by the parser, but this stage only controls the selection and ranking of these solutions.

Before we could deal with the ambiguity problem we had to develop an Arabic parser. Building the NLP system has not been a straightforward process due to the nature of Arabic. Arabic is well-known for its rich and complex morphology and syntactic flexibility. The syntactic parser for Arabic is developed within the framework of LFG (Lexical Functional Grammar) (Bresnan, 2001, Kaplan and Bresnan, 1982).

In this parser a cascade of finite state transducers are used to cover the pre-processing phases such as normalization, tokenization, morphological transduction and transduction of multiword expressions (MWEs). Beside core transducers there are backup transducers to provide estimates when exact analyses are not possible. Tools for analysing the corpus by breaking a running text into sentences and for providing frequency statistics on lexical entries are developed in Visual Basic. Arabic grammar rules and notations are written using XLE (Xerox Linguistics Environment), (Butt et al., 1999b, Dipper, 2003), which is a platform created by Palo Alto Research Center (PARC) for developing large-scale grammars using LFG formalisms. It includes a parser, transfer and generator components, which makes it suitable for building a Machine Translation (MT) system. Building the NLP system has not been a straightforward process due to the difficult nature of Arabic. Arabic is well-known for its rich and complex morphology and syntactic flexibility.

We also had to formulate a description of the syntactic constructions in Arabic within the framework of LFG. Arabic has intricate, complex and multi-faceted syntactic structures which led researchers to propose differing representations. There is a wide gap between Arab and Western grammarians in their attempts to describe the Arabic syntactic structures, with each applying a different set of criteria to characterize the same phenomena. The challenge is that a complete formal description of Arabic is not available yet (Daimi, 2001), let alone in the domain of LFG. Many aspects of Arabic are not investigated satisfactorily, such as topicalization, agreement, and long-distance dependencies. There is even no agreement among researchers on the basic sentence structures in Arabic. Therefore, in some instances we provide solutions, while in other instances we pose open questions that need further research and investigation.

In this Introduction we explain the theoretical framework on which this thesis is built, what platform is used in the development and what variety of Arabic is the target of analysis and processing. We also review the literature on Arabic parsers and explain the architecture of our parser.

Chapter Two details issues related to the Arabic morphological analyser. We discuss the underspecification of POS classification in Arabic. We also illustrate the sources of ambiguity in Arabic morphology and the techniques that can be followed to manage this ambiguity. We compare our morphological analyser to two of the best known Arabic morphological analysers in the research community and conduct an evaluation experiment to explore the extent to which ambiguity is controlled by the three analysers.

Chapter Three introduces the Arabic tokenizer component. Tokenization in Arabic is a non-trivial task due to the complex nature of the language. Arabic has a group of clitics that encompass a wide range of syntactic categories, such as conjunctions, prepositions, particles and pronouns. These clitics are attached to words and can be concatenated one after the other. The challenge a tokenizer faces is to separate these clitics from words and from each other. The tokenizer is also responsible for identifying MWEs and marking them as units, not as individual words. In the sequence of processing, the tokenizer comes as the initial step of processing. However, our discussion of tokenization occupies a belated position in the order of the thesis as it draws on information from the morphological analyzer, and builds on concepts and ideas from Chapter Two.

Chapter Four explains the MWE transducer. MWEs have high frequency in texts and when they are identified and analyzed correctly they add a sense of certitude to the analysis and reduce ambiguity. However, when MWEs are analyzed compositionally, they lose their meaning and put unnecessary load on the parser. We start by defining and classifying MWEs, and then proceed to show how they can be accommodated at each level of the analysis.

Chapter Five pinpoints grammatical issues in Arabic that usually constitute a source of perplexity when building a parser. We describe the main clausal architecture and sentence types in Arabic, and how they can be accounted for in LFG. We also investigate agreement in Arabic, and show how Arabic is a language with alternate agreement and how agreement is best accounted for within the phrase structure rules. Then we explore functional control and long-distance dependencies in Arabic, and show how agreement and resumptive pronouns are used to mark the relation between the position of the filler and the position of the gap. We end the chapter with a detailed investigation of the approaches to analysing the copula constructions in LFG and argue for the need for a unified representation of what we conceive as a universal predicational construction.

In Chapter Six we investigate the tools and methods for syntactic disambiguation available within the framework. We first identify sources of syntactic ambiguities in Arabic. The problem of ambiguity in Arabic language has not received enough attention by researchers. Although most aspects of the ambiguity problem are shared among human languages, it is still worthwhile to show how the special characteristics of a certain language contribute towards increasing or reducing ambiguities. We focus specifically on four ambiguity-generating areas in Arabic which, in our estimation, have the greatest impact. These are the pro-drop nature of the language, word order flexibility, lack of diacritics, and the multifunctionality of Arabic nouns.

We then move on to explore the full range of tools and mechanisms implemented in the XLE/LFG framework for ambiguity management, showing how they were applied to our Arabic grammar. Handling the ambiguity problem is divided into three stages: The pre-parsing stage contains all the processes that feed into the parser whether by splitting a running text into manageable components (tokenizer), analyzing word categories and morpho-syntactic features (morphological analyzer) or tagging the text (POS tagger). These processes are at the bottom of the parsing system and their effect is tremendous as they directly influence the number of solutions a parser can produces. The parsing stage is the process when the syntactic rules and constraints are applied to a text, and the subcategorization frames are specified. The parsing phase covers the topics of granularity of phrase structure rules, lexical specifications, application of syntactic constraints, and domain specific adaptation. The post-parsing stage has no effect on the number of solutions already produced by the parser. This stage only controls the selection and ranking of the solutions.

Chapter Seven is about grammar development, testing and evaluation. We start by showing that the development of a hand-crafted rule-based grammar is not usually a fast process, but it usually takes years of building and investigation. We then explain the stages of Arabic grammar development and the tools used for processing the corpus for the purpose of testing and developing the grammar. We then conduct an evaluation experiment on unseen set of data to show how much coverage the grammar has achieved at the current stage. We also apply the set of robustness tools (guessers and fragment grammar) and show how these utilities are effective in increasing the coverage.

Chapter Eight concludes the thesis by recapitulating the prospect of MT within the ParGram project. We first define what is meant by and what could be expected from MT. We give short explanation of the rule-based transfer approach. We then demonstrate the MT component in the ParGram project. We apply simple transfer rules to translate a small sentence from Arabic into English, and point out what needs to be done in order to produce a fully-fledged MT system. We also show what possible extensions can be implemented in the system, as a whole, in the future.

1.1 Background

The version of Arabic we are concerned with in this study is Modern Standard Arabic (MSA). When we mention Arabic throughout this research we primary mean MSA as opposed to classical Arabic, the language of formal writing until roughly the first half of the 20th century. Classical Arabic was also the spoken language before the medieval times. MSA also contrasts with colloquial Arabic, which is any of the various dialects currently spoken in different parts of the Arab world. MSA, the subject of our research, is the language of modern writing and the language of the news. It is the language universally understood by Arabic speakers and the language taught in Arabic classes.

Our work is part of the ParGram (Parallel Grammar) project (Butt et al., 1999b, Dipper, 2003). ParGram is a project that aims at providing full syntactic representation for a range of languages (currently, English, French, German, Japanese, Norwegian, Urdu, Welsh, Arabic, Chinese, Hungarian, Vietnamese and Malagasy) within the framework of LFG (Bresnan, 2001, Falk, 2001, Kaplan and Bresnan, 1982, Sells, 1985). There is an essential assumption among the LFG community that while the c-structure representation accounts for language-specific lexical idiosyncrasies and syntactic particular differences, the f-structure represents a level of abstraction high enough to capture parallelism among different languages and bypass cross-linguistic syntactic differences. Our aim is to write a core grammar for Arabic that covers major constructions of MSA with emphasis on ambiguity resolution for the purpose of MT.

Arabic grammar rules and notations are written using the XLE platform (Butt et al., 1999b, Dipper, 2003) created at PARC for developing large-scale grammars using LFG notations. It includes a parser, transfer and generator components, which makes it suited for MT. XLE supports UTF-8 file format, and thus it is able to deal with the native script of languages that use non-Latin alphabet such as Arabic. In the XLE system, the preprocessing stages of normalization, tokenization and morphological analysis are performed by finite-state transducers which are arranged in a compositional cascade. These transducers are non-deterministic and can produce multiple outputs. After a sentence is successfully parsed, XLE show results in four windows: the first displays the phrase-structure tree (or c-structure), the second displays the f-structure, and the other two display packed representations to show, when ambiguity occurs, where the ambiguity is and what exactly is prompting it.

Arabic exhibits many complexities (Chalabi, 2000, Daimi, 2001, Fehri, 1993) which pose considerable challenges to theoretical as well as computational linguistics. It is true that some linguistic phenomena in Arabic are shared with other languages. This research shows Arabic benefited from the experiences of other ParGram languages, and how its particular characteristics were catered for within the framework. Here is a short list of the major issues involved in Arabic linguistic analysis:

1. Arabic is typographically different from the Latin character set. Arabic has 60 unique characters for letters, diacritics, punctuation marks and numbers. Furthermore, Arabic letters need to be connected together in a cursive way depending on the context in which they occur. These issues used to pose a problem when computers were limited to use of the ASCII system, but with the introduction of the Unicode system there is better handling of the character set. However, in many instances, computers still need to be Arabic enabled in order to view Arabic fonts correctly.
2. The Arabic writing direction is from right to left. Although the display of Arabic has been solved in most platforms today, there are still some applications that do not give correct representation of the writing system, such as the Mac shell which is used for XLE where the display of Arabic goes correctly from right to left but the letters are not connected, as mentioned in point 1, rendering the Arabic text unreadable. This is shown in Figure 1 for the sentence in (1). This is why we prefer to use the XLE-Web interface instead throughout this thesis.

(1) الولد أكل الموزة
al-waladu ʾakala al-mūzata
the-boy.nom ate the-banana.acc
‘The boy ate the banana.’

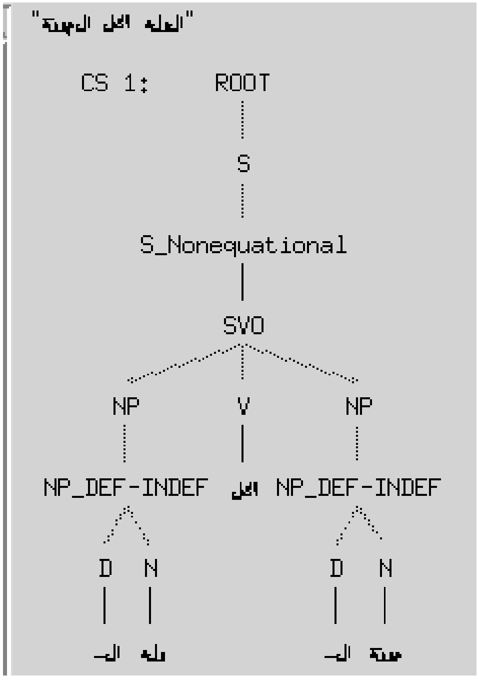


Figure 1. XLE shell interface with unconnected Arabic letters

3. Arabic has a relatively free word order. Moreover, beside the regular sentence structure of verb, subject and object, Arabic has a predicational sentence structure of a subject phrase and a predicate phrase, with no verb or copula.
4. Arabic is a highly inflectional language, which makes the morphological analysis complicated. Arabic words are built from roots rather than stems. Diacritics which help in marking the pronunciation of words with the same forms are usually omitted in modern writing.
5. Arabic is a clitic language. Clitics are morphemes that have the syntactic characteristics of a word but are morphologically bound to other words (Crystal, 1980). 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.
6. Arabic text is also characterised by the inconsistent and irregular use of punctuation marks. Punctuation marks have been introduced rather recently into the Arabic writing system, yet they are not as essential to meaning nor their use as closely regulated as is the case with English. Arabic writers shift between ideas using resumptive particles and subordinating conjunctions instead of punctuation marks.

7. Arabic is a pro-drop language. The subject can be omitted leaving any syntactic parser with the challenge to decide whether or not there is an omitted pronoun in the subject position.

1.2 System Design

In the literature there are a number of computational implementations for parsing Arabic. Daimi (2001) developed a syntactic parser for Arabic using the Definite Clause Grammar formalism. Žabokrtský and Smrž (2003) developed a dependency grammar for Arabic, with a focus on the automatic transformation of phrase-structure syntactic trees of Arabic into dependency-driven analytical ones. A probabilistic parser for Arabic is being developed at the Dublin City University based on the Arabic Penn Treebank Corpus (Al-Raheb et al., 2006). The Stanford Natural Language Processing Group[2] has developed an Arabic parser based on PCFG (Probabilistic Context-Free Grammar) using the Penn Arabic Treebank. Othman et al. (2003) developed a chart parser for analyzing Arabic sentences using Unification-based Grammar formalisms. Ramsay and Mansour (2007) wrote a grammar for Arabic within a general HPSG-like framework (Head-driven Phrase Structure Grammar) for the purpose of constructing a text-to-speech system. Within the framework of corpus linguistics, Ditters (2001) wrote a grammar for Arabic using the AGFL-formalism (Affix Grammars over a Finite Lattice).

Our system is the first Arabic parser to be built within the framework of LFG using the tools, formalisms and common inventory of the ParGram Group. Within the ParGram community grammar development is seen as a large software project (Butt et al., 1999b, Dipper, 2003) that should adhere to the techniques and design principles that are known from software engineering. One of the basic design principles is modularity. In this application each module is given a clearly defined task that it strictly adheres to it. Figure 2 shows the flow chart for an Arabic MT system based on our parser.

[2] http://nlp.stanford.edu/software/lex-parser.shtml

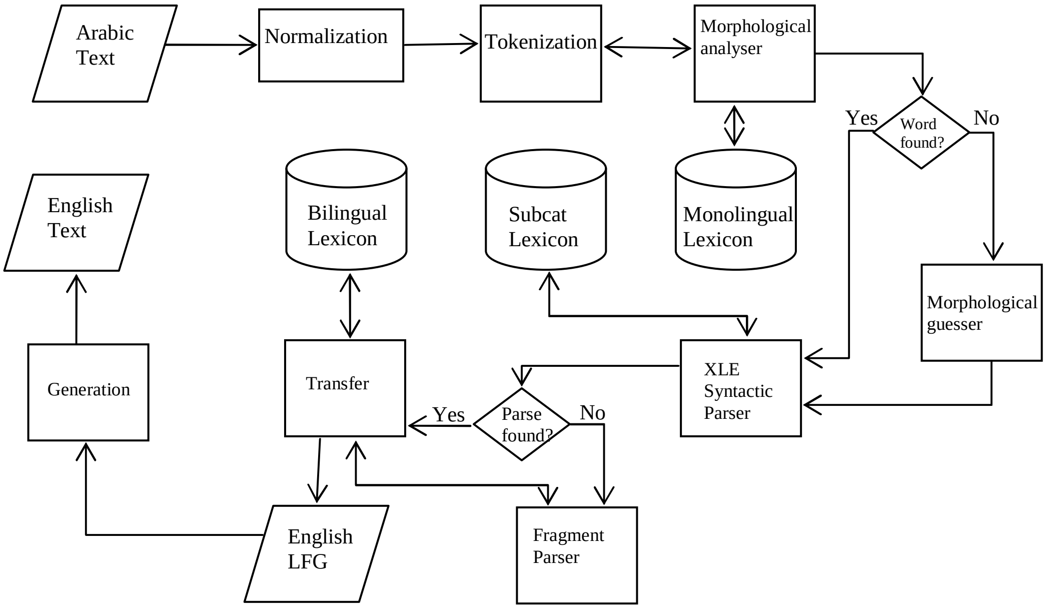


Figure 2. The architecture of the Arabic LFG Parser/MT System

The first module is the system is the normalizer whose function is to go through real-life texts and correct redundant and misplaced white spaces, diacritics and *kashida*s. It enables the system to proceed on a clean and predictable text.

The tokenizer splits the running text into tokens, so that they can be fed into a morphological transducer for processing. The tokenizer is normally responsible for demarcating words, clitics, abbreviated forms, acronyms, punctuation marks, numbers and MWEs.

The task of the morphological transducer is to provide essential morphological information for words, clitics and MWEs. It provides the grammatical category of words (part-of-speech), as well as the morpho-syntactic features related to tense, aspect, voice, mood, number, gender and person. In finite state morphology, it is not the task of the morphological transducer to order solutions, put them in packed representations or choose the most probable one. These decisions are taken later based on grammatical and semantic facts. If a word is not found in the core morphological analyser, a morphological guesser is used as a robustness technique to provide estimates.

The XLE parser uses a set of rules, notations and constraints to analyse the Arabic sentences. XLE provides a large range of tools for debugging and testing the system. When there is an ambiguity, packed morphological and grammatical features can be visually traced. If a complete analysis is not found, a partial parse is provided by the FRAGMENT grammar as a fail-safe technique.

The lexicon in this system (Subcat Lexicon) is responsible for interpreting morphological tags and stating subcategorization frames for verbs, nouns and adjectives when necessary. Lexical rules related to the passivization process are placed with each relevant verb. Idiosyncratic constraints that each verb requires are also stated.

After parsing is complete, transfer rules are applied to transform f-structures from Arabic to English. The parser in the target language is then used for generation. Researchers in the XLE translation project (Frank et al., 2001, Kay, 1999, Wedekind and Kaplan, 1996) emphasize that the transfer system does not attempt to resolve ambiguities, but transforms the packed representation (or packed ambiguities) from the source language to the target language. They consider this as an advantage as it avoids taking decisions about ambiguity handling at the wrong time, and thus discarding correct solutions too early.

1.3 Our Approach to Ambiguity

The aim of this research is to build a system that is ambiguity-conscious and ambiguity-sensitive at each level of the morphological and syntactic representation. We do not deal with ambiguity as one big problem to be treated only at the final stage after all the valid and invalid solutions have been generated, but we deal with ambiguity as a number of divisible problems spread over all the levels of processing.

We emphasize the bottom-up priority concept to ambiguity (MacDonald et al., 1994; Seidenberg et al., 1982) which states that linguistic information tend to be more effective at selecting between alternative solutions at the lower levels of the analysis and less effective at doing so at the higher levels. We believe that the impact of ambiguity in the lower levels is tremendous as it will propagates exponentially into the higher levels. For example if the morphology contains one invalid solution, this can be easily remedied by removing the invalid solution from the morphological analyser. However, if we fail to do so and try to tackle this specific ambiguity later on, we will have to overload the system by writing grammatical constraints and stating preferences with OT marks or use a probability-based disambiguation technique. The difference is that in the morphology we will be dealing with only one analysis that needs to be removed, but in the later stages the invalid analysis may have interacted with other options to create dozens of analyses, which will make the problem harder and more elusive.

Regarding ambiguity pruning at the early stages there are usually two extremes. The first extreme is to prune ambiguity early-on in the analysis and allow only one solution to surface at each processing stage. This approach usually risks discarding possible solutions on the ground of poor or insufficient evidence, thus throwing out the baby with the bath water. For instance if the morphology is required to choose one solution while it has no access to the syntactic context or semantic clues, the morphology is forced to take a decision that it is not prepared to take. An example of this approach is Ramsay and Mansour (2007) where morphological choices are made early, without all the relevant information, and hence the system has to perform backtracking in case the chosen analysis is not the correct one.

The other extreme is to allow all valid and invalid solutions to surface, as in the case of Xerox Arabic Morphological Analysis and Generation (Beesley, 1998a, Beesley, 2001) which produces a large number of rule-generated forms that have no actual place in the language. This approach complicates the ambiguity problem even further and makes ambiguity resolution even harder. The disambiguation process will not only have to choose the most likely solution but it will have to contend with scores of invalid solutions that should not be there in the first place.

The approach we have taken is a middle-course one. We discard (or avoid) invalid solutions as early as possible. For example in the morphology, if we are dealing with Modern Standard Arabic, there is no point in including classical entries and classical word senses. Another example is the inflection of verbs into passive and imperative forms. Xerox morphology allows all verbs to inflect into the passive and imperative forms and this decision makes the system overgenerate massively. For example, there is not point in allowing verbs of perception and entity-specific change of state to have an imperative form, and most of the intransitive verbs do not normally have passive forms. Buckwalter, on the other hand, allows a very small proportion of the verbs to inflect for the imperative forms. Out of 9198, only 22 verbs are allowed to have the imperative form, and this makes the system lack coverage of many valid possibilities.

In our system we took a middle course. We did not want our morphology either to overgenerate or to discard valid solutions. Therefore we reviewed all the verbs by hand bearing in mind the general criteria that intransitive verbs (with a small exceptions) do not inflect for the passive, and verbs that denote perception or entity-specific change of state do not inflect for the imperative. In this way we allowed 36% of the verbs in our morphology to inflect for the passive and 32% to inflect for the imperative. We believe that a morphological analyser should output all the valid, and only valid, solutions. This is why our evaluation experiment covers ambiguity handling as well as precision. An analysis is not considered precise if it does not include all the possible, and valid, solutions.

In the case of MWEs we tried at first to allow compositional readings along with the MWE readings and to give a positive preference mark to MWEs. However, we found in some instances that the interaction of preference marks can lead the compositional readings to surface as optimal solutions and MWE readings to be suppressed as suboptimal, as discussed in section 6.2.3. The compositional readings also cause an efficiency problem by increasing the number of solutions and parse time. Therefore we opted for pruning the compositional readings in the early stage of tokenization. However, this remains as an empirical issue and in the light of new evidence, the approach of handling MWEs could be changed.

In the grammar section we tried to make our rules produce all the valid solutions and not to prune possible solutions early on the ground of poor evidence. However, when we are met with a choice either to allow the grammar to overgenerate or write a tight constraint that selects the most probable solutions, we opt for the tight constraint and aim at adjusting the constraint in such a way that well-formed constructions are not excluded. For example Arabic allows VOS, beside VSO. The VOS word order, however, is not a frequent construction in MSA, and its occurrence is constrained by mainly two conditions. The first is when the object is a pronoun. The second is when the object is definite and the subject is indefinite. In our grammar we accommodate only one possibility of the VOS structure; that is when the object is a pronominal suffix, and in the future work we will work on the second condition.

In general to try to gain efficiency and speed, but not at the cost of accuracy and precision. We mainly concentrate on pruning invalid solutions as early as possible. However, constraints on the grammar might be too tight sometimes, but this is left as an empirical issue as constraints are usually subject to modification when new data appear.

2 Morphological Analysis and Disambiguation

This chapter explains the process of Arabic morphological analysis and the main strategies used in developing a morphological analyzer. We discuss the underspecification of POS classification in Arabic and show how this affected the accuracy of current morphological analyzers. We also illustrate the sources of ambiguity in Arabic morphology and the techniques that can be followed to manage this ambiguity.

We review Xerox Arabic Finite State Morphology and Buckwalter’s Arabic Morphological Analyzer which are two of the best known, well documented, morphological analyzers for Modern Standard Arabic (MSA). We found that there are significant problems with both systems in design as well as coverage that increase the ambiguity rate. Xerox morphology is root-based and theoretically well-motivated, yet it has uncurbed generative power that makes it produce forms that are unknown in the language. Buckwalter’s morphology is a stem-based database that lacks the generality and power of a rule-based system. Both systems include a large number of classical entries that are not part of MSA and do not occur in contemporary Arabic texts, the matter that leads to an increased number of ambiguities.

We also found out that ambiguity is increased in Buckwalter’s system by the inappropriate application of spelling relaxation rules and by overlooking rules that combine words with clitics and affixes (grammar-lexis specifications). Another source of confusion is whether to allow Arabic verbs to inflect for the imperative mood and the passive voice or not. Xerox adopted the overgeneralization that all verbs inflect for the imperative and the passive, leading it to overgenerate massively. Buckwalter’s morphology, on the other hand allowed only some verbs to have these inflections. Yet, because it did not follow a unique criteria or a systematic approach, the analysis is either underspecified or superfluous.

We show how an ambiguity-controlled morphological analyzer for Arabic is built in a rule-based system that takes the stem as the base form using finite state technology. We point out sources of genuine and spurious ambiguities in MSA, and how ambiguity in our system is reduced without compromising precision. The system is based on a contemporary corpus of news articles to ensure that the scope of the lexicon is restricted to MSA. Our morphology emphasizes the idea that inflecting all verbs in the passive and the imperative is semantically and pragmatically incorrect. Therefore, a set of broadly-defined criteria is devised to select which verbs can have a passive voice and which verbs can occur in the imperative.

In the last section, we compare our morphological analyser to Xerox and Buckwalter morphological analyzers and conduct an evaluation experiment to explore the extent to which ambiguity is controlled by the three analysers.

This chapter is an enhanced and updated version of my paper titled "An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks" (Attia, 2006a).

2.1 Development Strategies of Arabic Morphology

Arabic is known for its morphological richness and complexity (Azmi, 1988, Beesley, 1998b, Ibrahim, 2002, McCarthy, 1985, Ratcliffe, 1998). Arabic morphology has always been a challenge for computational processing and a hard testing ground for morphological analysis technologies. There are mainly two strategies for the development of Arabic morphologies depending on the level of analysis:

1. Stem-based morphologies: analyzing Arabic at the stem level and using regular concatenation. A stem is the least marked form of a word, that is the uninflected word without suffixes, prefixes, proclitics or enclitics. In Arabic, this is usually the perfective, 3rd person, singular verb, and in the case of nouns and adjectives they are in the singular indefinite form.

2. Root-based morphologies: analyzing Arabic words as composed of roots and patterns in addition to concatenations. A root is a sequence of three (rarely two or four) consonants which are called *radicals,* and the pattern is a template of vowels, or a combination of consonants and vowels, with slots into which the radicals of the root are inserted as shown in Figure 3. This process of insertion is usually called *interdigitation* (Beesley, 2001).

Root	درس drs			
Pattern	$R_1aR_2aR_3a$	$R_1aR_2R_2aR_3a$	$R_1\bar{a}R_2iR_3$	$muR_1aR_2R_2iR_3$
Stem	darasa ‘study’	darrasa ‘teach’	dāris ‘student’	mudarris ‘teacher’

Figure 3. Root and Pattern Interdigitation

There has been an intense contest between proponents and opponents of using the root as the base form. Beesley (2001) defended the “linguistic reality of Semitic roots” and cited, as a practical motivation, that traditional dictionaries are indexed by roots. It has even been maintained that “the use of Arabic roots as indexing terms substantially improves the [information] retrieval effectiveness over the use of stems” (Darwish, 2002).

However, several researchers criticized this approach. Kamir et al. (2002) assumed that the stem is the lemma, or the basic grammatical unit, in Arabic, and argued that the root is an abstract “super-lemma” that groups all the words that share a semantic field. They also maintained that the role of a root appears in word formation, or derivational morphology, while the stem is the actual manifestation of the root, and it is the stem that takes part in inflectional morphology. Dichy and Fargaly (2003) dedicated a lengthy paper to the subject and maintained that a root-and-pattern system included “huge numbers of rule-generated word-forms, which do not actually appear in the language” and that morpho-syntactic and semantic information need to be added to lexical entries at the stem level.

In our implementation we adopted the idea that a root is an abstract form that does not belong to a specific POS, but it plays a crucial part in stem formation. So using the stem as base form is far less complex in developing and maintaining, less ambiguous, and more suitable for syntactic parsers that aim at translation. The effectiveness of a root-and-pattern system in information retrieval is even doubted as some verbs like أمِنَ ’amina ‘to be safe’, أمُنَ ’amuna ‘to be honest’ and آمَنَ ’āmana ‘to believe’ have the same root but each has a different pattern and different semantic field (examples adapted from Dichy and Fargaly, 2003). So أمان ’amān ‘safety’, أمانة ’amānah ‘honesty’ and إيمان ’īmān ‘belief’ should not be made related in an information retrieval system.

2.2 The Parts of Speech Dilemma

One of the main functions of a morphological analyzer is to specify the part of speech for each word. However, reaching a clear-cut understanding of Arabic word categories has been hindered by a millennium-long underspecification of the parts of speech in Arabic. Parts of speech have been classified too broadly that they lacked the necessary details. Sibawaih (late 8th century) (1966) opens his famous book *Al-Kitab* with a classification of the parts of speech in Arabic into nouns, verbs and particles. This classification remains until the present time as a leading principle of Arabic grammar (Suleiman, 1990).

Arabic dictionaries do not list the part of speech classification, and Arabic grammar books are significantly influenced by the division of parts of speech in Arabic into nouns, verbs, and particles. For example Wright (1896/2005) uses the term noun as an umbrella etymology that encompasses six types: a substantive noun (*nomen substantivum*), adjective (*nomen adjectivum*), numeral adjective (*nomen numerale*), demonstrative pronoun (*nomen demonstrativum*), relative pronoun (*nomen conjuctivum*) and personal pronoun (*pronomen*).

Under the archetype of particles, Wright made four main divisions: prepositions, adverbs, conjunctions and interjections. Prepositions are subdivided into two categories: true prepositions such as إلى ’ilā ‘to’, and في fī ‘in’, and prepositions derived from nouns taking the accusative case (considered by traditional Arabic grammarians as adverbs) such as بين baina 'between', and تحت taḥta 'under'.

The category of adverbs is used by Wright to denote true adverbs such as فقط faqaṭ 'only', and هنا hunā 'here', and nouns taking the accusative case and functioning as adverbs, such as كثيرا kaṯīran 'frequently', and مجانا mğğānan 'freely'. Besides, Wright included sundry types of categories under 'adverbs', such as the interrogative هل hal 'is it true that'; the negative لا lā 'no/not'; the tense marker سوف sawfa 'will'; the subordinating conjunction لكن lākinna 'but'; and the conditional لولا lawlā 'if'.

In modern linguistic literature, Suleiman (1990) criticised the medieval Arab grammarians' well-known three-fold classification of Arabic parts of speech into nouns, verbs and particles, which is still a well-established hardly-contested concept in present day Arabic grammar. Suleiman refuted this tri-partite division by scrutinizing the earliest theoretician of Arabic grammar, Sibawaih, in his *Kitab*. The main thrust of Suleiman's argument is that Sibawaih did not provide any empirical or rational evidence to support the view that parts of speech are exclusively three.

In our view we consider that the tripartite division of parts of speech in Arabic serves only an archetypal classification rather than detailed listing. In a comprehensive morphological and syntactic description it is the detailed listing that is needed. It would be an enormous oversimplification to build a morphological analyzer on the assumption that parts of speech in Arabic are solely nouns, verbs and particles. Unfortunately, no detailed research has been conducted on the resolution of categorical intersection between nouns and adjectives, or on providing a comprehensive classification of function words, or the position of verbal nouns, how participles function as verbs, nouns, adjectives and adverbs, and how adverbs are formed from verbal nouns and prepositional phrases.

It is quite surprising to see many morphological analyzers today influenced by the misconception that Arabic parts of speech are exclusively nouns, verbs and particles. Xerox Arabic morphological analyzer is a good example of this limitation. In Xerox morphology, words are classified strictly into verbs, nouns and particles; no other categorical description is used. Buckwalter made a more detailed classification, but traces of generalizations are still evident in the large amount of adjectives still classified as nouns and particles still classified as function words.

In Buckwalter, we tested 996 adjectives, and 765 of them were correctly identified as adjectives, 15 were not found, while the rest of the adjectives (22%) were analysed incorrectly as nouns. The misanalysed adjectives included active participles, passive participles, and even adjectives of colour; all were analysed as nouns, as shown in Table 1.

Active Participles	Passive Participles	Adjectives of Colour
آثم ʾāṯim ‘sinful’	مأهول maʾhūl ‘populated’	أحمر ʾaḥmar ‘red’
باطل bāṭil false	مؤبد muʾabbad perpetual	أخضر ʾaḫḍar green
باهظ bāhiẓ exorbitant	مؤجل muʾağğal postponed	بني bunnī brown

Table 1. Examples of adjective analyzed as nouns in Buckwalter’s morphology

Although we admit that more research is needed into the part of speech classification in Arabic, we tried to make as much detailed description as possible in our morphology. We identified nine parts of speech categories for Arabic which proved reasonably sufficient enough to support the grammatical description of our syntactic parser. These categories are verbs, nouns, adjectives, adverbs, prepositions, determiners, conjunctions, pronouns and particles.

2.3 Morphological Ambiguity

Morphological ambiguity in Arabic is a notorious problem that has not been sufficiently addressed (Kiraz, 1998). This ambiguity represents hurdles in the way of POS taggers (Freeman, 2001) syntactic parsers, and machine translation.

For example, the greater the number of morphological analyses given for a lexical entry, the longer a parser takes in analyzing a sentence, and the greater the number of parses it produces. Overcoming ambiguity is the major challenge for NLP in Arabic (Kamir et al., 2002).

In this section we discuss sources of genuine ambiguity in Arabic, and propose the ambiguity pyramid hypothesis in which we claim that ambiguity decreases with the build-up of words by adding affixes and clitics.

2.3.1 Sources of Genuine Morphological Ambiguities in Arabic

Many words in Arabic are homographic: they have the same orthographic form, though the pronunciation is different. There are many recurrent factors that contributed to this problem. Among these factors are:

1. Orthographic alternation operations (such as deletion and assimilation) frequently produce inflected forms that can belong to two or more different lemmas. Example (2) is an extreme case of a surface form that can be interpreted as belonging to five different stems.

(2) يعد yʿd

يعِد (أعاد)	يعُد (عاد)	يعِد (وعد)	يَعُدّ (عد)	يُعِدّ (أعد)
yuʿid (ʼaʿāda)	yaʿud (ʿāda)	yaʿid (waʿada)	yaʿuddu (ʿadda)	yuʿiddu (aʿadda)
‘bring back’	‘return’	‘promise’	‘count’	‘prepare’

2. Some lemmas are different only in that one of them has a doubled sound which is not explicit in writing. Arabic Form I and Form II are different only in that Form II has the middle sound doubled.

(3) علم ʿlm

علِم	علّم
ʿalima ‘know’	ʿallama ‘teach’

3. Many inflectional operations underlie a slight change in pronunciation without any explicit orthographical effect due to lack of short vowels (diacritics). An example is the recurring ambiguity of active vs. passive vs. imperative forms.

(4)	أرسل ’rsl		
	أرسَل	أُرسِل	أرسِل
	’arsala	’ursila	’arsil
	‘sent’	‘was sent’	‘send [imperative]’

4. Some prefixes and suffixes can be homographic with each other. The prefix *ta*- can indicate 3rd person feminine or 2nd person masculine.

(5)	تكتب	تكتب
	ta-ktub	ta-ktub
	‘you.m write’	‘she writes’

Another recurring ambiguity is the person suffix –*t* which is shared by four features.

(6)	كتبت ktbt			
	كتبتُ	كتبتَ	كتبتِ	كتبتْ
	katabtu	katabta	katabti	katabat
	‘I wrote’	‘you.m wrote’	‘you.f wrote’	‘she wrote’

Similarly, the dual is always confused with the plural in the accusative case.

(7)	أمريكيين	
	أمريكيَين	أمريكيين
	’amrīkiyyain	’amrīkiyyīn
	‘American.dl’	‘American.pl’

5. Prefixes and suffixes can accidentally produce a form that is homographic with another full form word. This is termed “coincidental identity” (Kamir et al., 2002).

(8)	أسد ’asd	
	أسُدّ (أ+سد)	أسد
	’asuddu	’asadun
	‘I block’	‘lion’

Similarly, clitics can accidentally produce a form that is homographic with another full word.

(9)	علمي	
	علمي	علمي (علم + ي)
	ʿilmiyy	ʿilm-ī
	‘scientific’	‘my knowledge’

6. There are also the usual homographs of uninflected words with/without the same pronunciation, which have different meanings and usually different POS's.

(10) ذهب

ذهب	ذهب
d̲ahabun	d̲ahaba
'gold'	'go'

2.3.2 The Ambiguity Pyramid Hypothesis

The ambiguity pyramid hypothesis assumes that the rich and complex system of Arabic inflection and concatenation helps to reduce ambiguity rather than increase it. Unmarked stems are usually ambiguous but when they are inflected and/or when clitics are added, ambiguity is reduced, as shown in (11).

(11)	stem:	كتب	ktb	books / wrote / was-written
	inflected:	يكتب	ya-ktb	writes / is-written
	cliticized:	يكتبه	ya-ktb-hu	[he]-writes-it

Words from a few randomly selected sentences were morphologically analyzed at different levels. First they were analyzed as whole words, then they were analyzed after separating words from clitics, and at last they were analyzed after separating clitics and stripping off all inflectional prefixes and suffixes, that is using the base stem. The highest rate of ambiguity appeared in the stem level. The rate decreased with inflection, and decreased even further with the addition of clitics. Figure 4 illustrates that ambiguity rates decrease, on average, with the increase in word build-up.

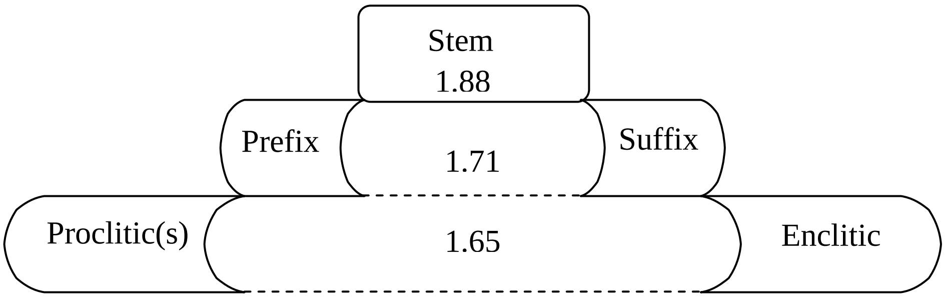


Figure 4. The ambiguity pyramid hypothesis

However, this is a hypothesis that still needs to be verified. Further testing with some other sentences contradicted these assumptions, and large scale testing on a large number of words is not possible. For example, a list of 30,000 full form words was reduced to 15,000 unique words after stripping off clitics. Comparing the ambiguity rates for two unequal sets is not indicative, as the same transducer will usually give different ambiguity rates when it is fed different ranges. So in order to verify this hypothesis, testing needs to be done on several hundred sentences, rather than words. This may not even be very meaningful, as a sentence containing 30 full form words will break down into about 50 tokens and break down further into 70 base forms. So comparing the rates at these different numbers cannot constitute strong evidence. It is also found that words with the highest scores are inflected forms.

2.4 Existing Arabic Morphological Systems

There are many morphological analyzers for Arabic; some of them are available for research and evaluation while the rest are proprietary commercial applications. Among those known in the literature are Xerox Arabic Morphological Analysis and Generation (Beesley, 1998a, Beesley, 2001), Buckwalter Arabic Morphological Analyzer (Tim Buckwalter. 2002), Diinar (Dichy and Hassoun, 1998), Sakhr (Chalabi, 2004a), and Morfix (Kamir et al., 2002). The first two are the best known and most quoted in literature, and they are well documented and available for evaluation.

2.4.1 Buckwalter Arabic Morphological Analyzer

Buckwalter Morphology is well-known in the literature and has even been considered as the “most respected lexical resource of its kind” (Hajic et al., 2005). It contains 38,600 lemmas, and is used in the LDC Arabic POS-tagger, Penn Arabic Treebank, and the Prague Arabic Dependency Treebank. It is designed as a main database of word forms interacting with other concatenation databases. Every word form is entered separately. It takes the stem as the base form, and information on the root is also provided. Buckwalter’s morphology reconstructs vowel marks and provides English glossary, and it is less ambiguous than Xerox’s. The disadvantages, however, are:

1. It is not rule-based. All word forms are entered manually. After each entry, all forms that belong to that specific entry at different inflectional levels are listed. So it does not capture generalities, and it increases the cost of maintenance.

2. The system is not suited for generation. This means that you cannot give the system a set of strings and tags in order to produce the surface forms.

3. Lack of coverage of the clitic question morpheme which can be prefixed to verbs and nouns. This was perhaps intended to reduce ambiguity, but, still, it limits coverage. For instance the examples in (12) are not found by the system.

(12)	أأقول	’a’aqūlu ‘do I say’	– not found
	أمحمد	’muḥammadun ‘Is it true that Mohammed’	– not found

4. Insufficient coverage of imperative forms: Out of 9198 verbs, only 22 verbs (0.002%) have imperative forms. This is far less than the 32% allowed in our morphology. This restricts Buckwalter’s morphology from dealing with instruction manuals, for example. Buckwalter’s system does not give the imperative senses associated with common verbs, as in (13).

(13)	حاول	ḥāwil	‘try’
	انتظر	intaẓir	‘wait’
	اضرب	iḍrib	‘hit’

5. Insufficient coverage of the passive morphology. Out of 9198 verbs, only 1404 verbs (15%) are allowed to have a passive form. In our system, 36% of verbs can have a passive form. Buckwalter’s passive forms are also restricted by tense. Only 110 of them have a passive form in the past (perfective) tense. There are even passive forms for verbs with low probability, as in (14).

(14)	يمات	yumāt	‘be made to die’
	يعاش	yuʿāš	‘be lived’

Other verbs with high probability are not allowed in the passive, such as those in (15).

(15) قابل qābala ‘meet’
استعمل ʼistaʽmala ‘use’

6. It accounts for the classical affirmative clitic ل la ‘indeed’ which is prefixed to nouns. This makes it ambiguous with the preposition which has the same form, and increases the ambiguity level.

(16) لأحزاب la-ʼaḥzāb indeed + parties

7. Some proper names are associated with senses that are no longer used in the language.

(17) حسام Husam / sword
حنيفة Hanifah / orthodox

8. Buckwalter’s system does not handle multiword expressions (MWEs). MWEs have high frequency in texts and when they are identified and analyzed correctly they add a sense of certitude to the analysis and reduce ambiguity. However, when MWEs are analyzed compositionally, they lose their meaning and add to the ambiguity problem, as component parts may be individually ambiguous. The MWE in (18) has four different analyses by Buckwalter’s system.

(18) أبي أسعد
abī ʼasʽad
my father / proud happier / make happy
‘Abu As’ad [proper name]’

9. Inclusion of classical entries. Every entry added to the lexicon of a morphological analyzer is very costly in terms of ambiguity, so terms should be extracted from contemporary data, rather than from traditional dictionaries, if they are meant to handle modern texts. There are many hints that Buckwalter and Xerox took Hans Wehr’s Arabic English Dictionary of Modern Written Arabic (Wehr, 1979) as the backbone reference. However, in the very introduction, Hans Wehr stated that the dictionary “lists classical words and phrases of elegant rhetorical style side by side with new coinages”. Buckwalter includes some roots that are totally obsolete, such as the examples in (19).

(19)	قف	qaffa	‘to be dry’
	أبد	abada	‘be untamed’
	أب	abba	‘desire’

Some forms are fossilized in contemporary usage, as their usage is limited to expressions in a certain syntactic and morphological context. However, they are included in Buckwalter’s system as full entries.

(20) لا يأبه lā yaʾbah
not care
‘He does not care.’
Root: أبه
abaha
‘be interested’

All the forms in (19) and (20) are homographic in some way with other forms that are in contemporary usage. Still, we can prove statistically that Buckwalter included classical terms by showing the Google score for some selected classical entries found Buckwalter’s morphology in Table 2. These forms are even found mostly in websites talking about grammar or morphology. The table also shows the alternative MSA forms and their comparatively high frequency occurrences in Google.

#	Meaning	Classical Word	Google	MSA Word	Google
1	sully	قلعط qalʿat	8	لطخ laṭṭaḫa	29,600
2	caulk	قلفط qalfaṭ	9	أفسد ʾafsada	205,000
3	wear	استكد ʾistakadda	4	أنهك ʾanhaka	37,100
4	fickle	غملج ġamlağ	7	متقلب mutaqallib	189,000
5	erosion	ائتكال ʾiʾtikāl	7	تآكل taʾākul	1,700,000

Table 2. Google score for Classical vs. MSA entries

10. Improper spelling relaxation rules. Buckwalter justified the inclusion of these relaxation rules by the fact that they are common in the data analyzed (Buckwalter, 2004). We reckon however, that this is not a solid justification because, firstly, we should take into account that Arabic electronic texts are relatively recent, and that not so many authors are well trained in using proofing tools. Secondly, misspelled words should be handled as special cases, or apply rules when the forms fail to receive an analysis. Applying the rules globally leads to a massive increase in the ambiguity level for correctly

spelled words. Thirdly, misspelling is even common in English. The Google score for the misspelled word “arround”, for example, is 2,530,000 and for “vedio” is 2,150,000, and this will not be deemed as a plausible ground for including these misspelled words in an English morphological analyzer. The examples in (21) show how Buckwalter analysed words with *alif* (ا) in the middle, and then applied the spelling relaxation rules to allow this *alif* to be also interpreted as *hamzah* (أ), further increasing the number of ambiguities.

(21) فاشل fāšil ‘failed’
-> فأشل fa-ʾa-šullu ‘then I paralyze’
واقف wāqif ‘standing’
-> وأقف wa-ʾa-qifu ‘and I stand’

11. Noncomprehensive treatment of the rules that govern the combination of words with clitics, or grammar-lexis specification (Abbès et al., 2004, Dichy, 2001, Dichy and Fargaly, 2003). As clitics are syntactic units, syntactic rules should apply when they combine with words. For example, when a preposition precedes a noun, the noun must be in the genitive case. Similarly, while it is acceptable for the noun to be followed by possessive pronouns, this is not acceptable for adjectives, which is not observed by Buckwalter, as shown in (22).

(22) معادي muʿādī (hostile/anti- + my)
معدي muʿdiyy (contagious/infectious + my)

Another wrong analysis is shown in (23) where a verbal noun derived from an intransitive verb is attached to an accusative pronoun clitic, which is grammatically and morphologically not acceptable.

(23) مصري muṣirr-ī (determined/insistent + my)

Similarly, names of places are usually followed by relative suffixes, not possessive pronouns, the rule which is ignored as shown in (24).

(24) عراقي ʾrāqī (Iraq + my), should be “Iraqi”
إيراني ʾirānī (Iran + my), should be “Iranian”

2.4.2 Xerox Arabic Morphological Analysis and Generation

Xerox Morphology is regarded as a system that is “based on solid and innovative finite-state technology” (Dichy and Fargaly, 2003). It adopts the root-and-pattern approach. It includes 4,930 roots and 400 patterns, effectively generating 90,000 stems. The advantages are that it is rule based with large coverage. It also reconstructs vowel marks and provides an English glossary for each word. The system inherited many disadvantages from Buckwalter’s morphology such as the lack of specifications for MWEs, and improper spelling relaxation rules. It even includes more classical entries, and lacks more grammar-lexis specifications. Example (25) shows an extreme case which violates the syntactic rule that a pronoun must be free within its binding domain, or “co-reference of the subject and of the object” (Dichy, 2001).

(25) نضربنا naḍribunā ‘we hit us’

Additional disadvantages of Xerox morphology are:

1. Overgeneration in word derivation. The distribution of patterns for roots is not even, and although each root was hand-coded in the system to select from among the 400 patterns, the task is understandably tedious and prone to mistakes.

word	transliteration	root	meaning
قال	qāl	qwl	say (verb)
		qlw	fry (active participle)
		qll	decrease (active participle)

Table 3. Overgeneration of spurious stems

The first analysis is valid, while the other two are spurious derivations that have no place in the language, and not even found in classical dictionaries.

2. Underspecification in POS classification, which makes it unsuited for serving a syntactic parser. Words are only classified into:
   - Verbs
   - Nouns, which include adjectives and adverbs.
   - Participles

- Function words, which include prepositions, conjunctions, subordinating conjunctions, articles, negative particles, and all other particles.

3. Increased rate of ambiguity. Due to the above-mentioned factors, the system suffers from a very high level of ambiguity, as it provides so many analyses (many of them spurious) for most words, as shown in (26).

(26) مصري miṣriyy 'Egyptian'
Xerox (22 solutions)
Buckwalter (10 solutions)
Attia (2 solutions)

2.5 Our System Design and Description

Our system is built using finite state technology (Attia, 2005, Attia, 2006a), and it is suitable for both analysis and generation. It is based on contemporary data (a corpus of news articles of 4.5 million words), and takes the stem as the base form. It contains 10799 lemmas (1532 verbs, 8923 nouns and adjectives, and 344 function words) and 2818 multiword expressions. The core system provides efficient coverage of MSA for its specific domain (news articles). The system is available for research and evaluation at www.attiaspace.com, along with a set of relevant finite state tools: a tokenizer, a white space normalizer, MWE transducer and a morphological guesser. The system is rule based; there is only one entry for each stem, and all inflection operations and orthographical changes are handled through xfst alternation rules. This helps in separating the task of the developer and the lexicographer. As adding new terms to the lexicon in a morphological transducer is a never ending process, the lexicographer's job is made clearer and easier.

A point of strength in the system that gives it an advantage over other morphological analyzers is the coverage of multiword expressions (Attia, 2006b). The system can efficiently handle compound names of people, places, and organizations, as shown in (27), (28) and (29), in addition to more complex expressions which can undergo inflections and lexical variations.

(27) أبو عمار
abū ʿammār (lit. father of ‘Ammar)
‘Abu ‘Ammar’

(28) بيت لحم
bait laḥim (lit. house of meat)
‘Bethlehem’

(29) مجلس الأمن
maǧlis al-ʾamn
‘Security Council’

A disadvantage of the system, however, is its limited coverage. Between Buckwalter’s 38,600 and Attia’s 13,600 entries, a good coverage, general-domain morphology is expected to be around 25,000 entries including MWEs. Our system does not handle diacritized texts. The decision to ignore diacritics was taken after examining a set of 35,000 unique words from the corpus, where only 156 words were found to carry diacritic marks, which is statistically insignificant. Other disadvantages are that it does not reconstruct diacritics, or provide English glossaries. These limitations do not affect the functionality of the morphology especially when the target is to feed a syntactic parser, yet it has been customary in Arabic morphology to provide diacritics and glossaries for illustration and pedagogical purposes.

2.5.1 Finite State Technology

Finite state technology has successfully been used in developing morphologies for many languages, including Semitic languages (Beesley, 1998b). There are a number of advantages of this technology that makes it especially attractive in dealing with human language morphologies, among these advantages are:

– The technology is fast and efficient. It can handle very huge automata of lexicons with their inflections. Compiling large networks that include several millions of paths is only a matter of seconds in a finite state calculus. Moreover, these large networks can be easily combined together to give even larger networks.
– Handling concatenative and non-concatenative morphotactics (Beesley, 1998b).

- Unicode support, which enables developers to accommodate native scripts that use non-Latin alphabets.
- Multi-platform support. Xerox finite state tools work under Windows, Linux, UNIX and Mac OS, which means that a morphological transducer developed using Xerox finite state compilers can serve applications under any of these platforms.
- A finite state system is fully reversible. So it can be used for analysis as well as generation.
- The regular expressions used in finite state closely resemble standard linguistic notations (Yona and Wintner, 2005) so the rules are reasonably readable and intelligible.

In a standard finite state system, lexical entries along with all possible affixes and clitics are encoded in the lexc language which is a right recursive phrase structure grammar (Beesley, 2001, Beesley and Karttunen, 2003). A lexc file contains a number of lexicons connected through what is known as "continuation classes" which determine the path of concatenation. In example (30) the lexicon *Proclitic* has a form *wa* which has a continuation class *Prefix*. This means that the forms in *Prefix* will be appended to the right of *wa*. The lexicon *Proclitic* has also an empty string, which means that *Proclitic* is optional and that the path can proceed without it. The bulk of lexical entries are listed under *Root* in the example.

```
(30)   LEXICON Proclitic
       wa            Prefix;
                     Prefix;
       LEXICON Prefix
       ya            Root;
       LEXICON Root
       shakara       Suffix;
       kataba        Suffix;
       LEXICON Suffix
       una           Enclitic;
       LEXICON Enclitic
       ha            #;
```

In a natural language, it usually happens that an affix or a clitic requires or forbids the existence of another affix or clitic. This is what is termed as “separated dependencies” or “long-distance dependencies” which constrain the co-occurrence of morphemes within words (Beesley and Karttunen, 2003). So Flag Diacritics were introduced as an extension to Xerox finite state implementation to serve as filters on possible concatenations to a stem. The most common form of Flag Diacritics is the unification type. Suppose we want to prevent the *Proclitic* and *Enclitic* lexicons from co-occurring. We can add a Flag Diacritic to each of them with the same feature name, but with different value, as shown in (31).

```
(31)    LEXICON Proclitic
        wa@U.Clitic.On@         Prefix;
        …
        LEXICON Enclitic
        ha@U.Clitic.Off@        #;
```

With inflections and concatenations, words usually become subject to changes or alternations in their forms. Alternations are the discrepancies between underlying strings and their surface realization (Beesley, 1998b), and alternation rules are the rules that relate the surface forms to the underlying forms. In Arabic, long vowels, glides and the glottal stop are the subject of a great deal of phonological (and consequently orthographical) alternations like assimilation and deletion. Most of the trouble a morphological analyzer faces is related to handling these issues. In our system there are about 130 replace rules to handle alternations that affect verbs, nouns, adjectives and function words when they undergo inflections or are attached to affixes and clitics. Alternation rules are expressed in finite state systems using XFST replace rules of the general form shown in (32).

```
(32)    a -> b || L _ R
```

This means that the string *a* is replaced with the string *b* when *a* occurs between the left context *L* and the right context *R*. When no context is specified the replacement operates globally. The special symbol ‘.#.’ can be used instead of *L* to express the condition when the string *a* occurs at the beginning of a word. The symbol '.#.' can also be used instead of *R* to indicate when the string *a* occurs at the end of a word. These replace rules can be composed one over the other, so that the output of one rule can be the input for another rule. This can effectively account for multi phonological and orthographical processes.

At the end we obtain a transducer with a binary relation between two sets of strings. The first set of strings is conventionally known as the lower language and contains the surface forms, and the second set of strings is the upper language and contains the lexical forms, or the analysis, as shown in (33) for the verb يشكرون yaškurūna 'they thank'.

(33) Upper Language: +verb+pres+activeشكر[šakara 'thank']+masc+pl+3
Lower Language: يشكرون[yaškurūna 'they thank']

2.5.2 Handling Arabic Morphotactics

Morphotactics is the study of how morphemes combine together to form words (Beesley, 1998b). These can be concatenative with morphemes either prefixed or suffixed to stems or non-concatenative, with stems undergoing alternations to convey morphosyntactic information. Arabic is considered as a typical example of a language that employs non-concatenative morphotactics.

Arabic words are traditionally classified into three types (Ibrahim, 2002): verbs, nouns and particles. Adjectives take almost all the morphological forms of nouns. Adjectives, for example, can be definite, and are inflected for case, number and gender.

Arabic verbs are inflected into imperfective (present), perfective (past) and imperative. Moreover, both the perfective and imperfective have two forms: the active form and the passive form. In sum, Arabic verbs are inflected to provide five forms: active perfective, passive perfective, active imperfective, passive imperfective and imperative. The base form of the verb is the perfective tense, 3rd person, singular. There are a number of indicators that tell how the base form would be inflected to give the other forms. Among these indicators are the number of letters of the base form and its template. A template (Beesley and

Karttunen, 2003) is a kind of vocalization mould in which a verb fits. Vocalism is a major factor in template shaping. Although diacritics (the manifestation of vocalism) are not present in modern writing, we still need to worry about them as they trigger other phonological and orthographical processes, such as assimilation and deletion and the re-separation (or spreading) of doubled letters.

2.5.2.1 Verbs

Possible concatenations and inflections in Arabic verbs are shown in Table 4. All elements are optional except the stem, and they can be connected together in a series of concatenations.

Proclitics		Prefix	Stem	Suffix	Enclitic
Conjunction/ question article	Complementizer	Tense/mood – number/gender	Verb	Tense/mood – number/gender	Object pronoun
Conjunctions و wa ‘and’ or ف fa ‘then’	ل li ‘to’	Imperfective tense (5)	Stem	Imperfective tense (10)	First person (2)
Question word أ ’a ‘is it true that’	س sa ‘will’	Perfective tense (1)		Perfective tense (12)	Second person (5)
	ل la ‘then’	Imperative (2)		Imperative (5)	Third person (5)

Table 4. Possible concatenations in Arabic verbs

Flag Diacritics are used to handle long-distance restrictions or what is termed “separated dependencies” for Arabic verbs. These restrictions can be considered as grammatical constraints, or grammar-lexis specifications, that govern the morphological process. These can be summarized as follows:

- The yes-no-question article أ ’a ‘it is true that’ cannot co-occur with imperatives or with the accusative case.
- The complementizer ل li ‘to’ cannot co-occur with the nominative case.
- Cliticized object pronouns do not occur either with passive or with intransitive verbs.
- Affixes indicating person and number in the present tense come in two parts one preceding and one following the verb and each prefix can co-occur only with certain suffixes.

– The imperfective, perfective and imperative have each a range of prefixes or suffixes or both which must be precisely constrained.

– A first person object pronoun cannot co-occur with a first person prefix (to account for the rule that a pronoun must be free within its binding domain), and similarly a second person object pronoun cannot co-occur with a second person prefix. This rule makes sure that the same person cannot act as subject and object at the same time.

The maximum number concatenations in Arabic verbs as shown by Table 4 above is six: one stem in addition to five other bound morphemes representing affixes and clitics. Statistically, concatenations in Table 4 give as much as 33,696 forms. In real constrained examples, some verbs, such as شكر šakara ‘to thank’, generate 2,552 valid forms. This considerable amount of form variations is a good indication of the richness and complexity of Arabic morphology.

2.5.2.2 Nouns

Possible concatenations and inflections in Arabic nouns are shown in Table 5 below. The maximum number of concatenations in Arabic nouns is five: one stem in addition to four other bound morphemes representing suffixes and clitics, bearing in mind that the genitive pronoun and the definite article are mutually exclusive.

Proclitics			Stem	Suffix	Enclitic
Conjunction/ question article	Preposition	Definite article	Noun	Gender/Number	Genitive pronoun
Conjunctions و wa ‘and’ or ف fa ‘then’	ب bi ‘with’, ك ka ‘as’ or ل li ‘to’	ال al ‘the’	Stem	Masculine Dual (4)	First person (2)
				Feminine Dual (4)	
Question word أ ʼa ‘is it true that’				Masculine regular plural (4)	Second person (5)
				Feminine regular plural (1)	Third person (5)
				Feminine Mark (1)	

Table 5. Possible concatenations in Arabic nouns

Flag Diacritics are also used to handle separated dependencies for nouns. These can be summarized as follows:

- The definite article ال al ‘the’ cannot co-occur with a genitive pronoun.
- The definite article cannot co-occur with an indefinite noun marking (*nuun* with the dual and plural or *tanween* with the singular).
- The cliticized genitive pronoun cannot co-occur with an indefinite noun marking.
- Prepositions cannot co-occur with nominative or accusative case markings.

Statistically, uncontrolled concatenations in Table 5 above give 6,240 forms. In real examples some nouns, such as معلم muʿallim ‘teacher’, generate 519 valid forms.

In our system adding nouns is made easy by selecting from a template of continuation classes which determine what path of inflection each noun is going to select, as shown in Figure 5 (transliteration and gloss are included in square brackets for illustration only).

```
LEXICON Nouns
+masc^ss^معلم[muʿallim ‘teacher’]^se^              DualFemFemplMascpl;

+masc^ss^طالب[ṭālib ‘student’]^se^                 DualFemFempl;
طالب+masc+pl:^ss^طلاب[ṭullāb ‘students’]^se^       CaseEnds;

+masc^ss^كتاب[kitāb ‘book’]^se^                    Dual;
كتاب+masc+pl:^ss^كتب[kutub ‘books’]^se^            CaseEnds;

+fem^ss^كراسة[kurrāsah ‘notebook’]^se^             DualFempl;

+fem^ss^شمس[šams ‘sun’]^se^                        Dual;
شمس+fem+pl:^ss^شموس[šumūs ‘suns’]^se^              CaseEnds;
```

Figure 5. Noun Stem Entry

These continuation class templates are based on the facts in Table 6. Table 6 shows what inflection choices are available for Arabic nouns according to gender (masculine or feminine), number (singular, dual or plural) and how the plural is formed (regular or broken plural). The table shows the variability in the choices permitted with each noun, with some nouns allowing all choices in their inventory (as shown by example 1 in the table), others selecting only one choice (as shown by examples 11–15), while the rest show a varied spectrum of choices.

	Masculine Singular	Feminine Singular	Masculine Dual	Feminine Dual	Regular Masculine Plural	Regular Feminine Plural	Broken Plural
1	جاهل ǧāḥil ‘ignorant’	جاهلة ǧāḥilah	جاهلان ǧāḥilān	جاهلتان ǧāḥilatān	جاهلون ǧāḥilūn	جاهلات ǧāḥilāt	جهلاء ǧuḥalāʾ
2	معلم muʿallim ‘teacher’	معلمة muʿallimah	معلمان muʿallimān	معلمتان muʿallimatan	معلمون muʿallimūn	معلمات muʿallimāt	✗
3	طالب ṭālib ‘student’	طالبة ṭālibah	طالبان ṭālibān	طالبتان ṭālibatān	✗	طالبات ṭālibāt	طلاب ṭullāb
4	تعليمي taʿlīmiyy ‘educational’	تعليمية taʿlīmiyyah	تعليميان taʿlīmiyyān	تعليميتان taʿlīmiyyatān	✗	✗	✗
5	امتحان ʾimtiḥān ‘exam’	✗	امتحانان ʾimtiḥānān	✗	✗	امتحانات ʾimtiḥānāt	✗
6	كتاب kitāb ‘book’	✗	كتابان kitābān	✗	✗	✗	كتب kutub
7	✗	بقرة baqarah ‘cow’	✗	بقرتان baqaratān	✗	بقرات baqarāt	بقر baqar
8	✗	همسة hamsah ‘whisper’	✗	همستان hamsatān	✗	همسات hamasāt	✗
9	✗	شمس šams ‘sun’	✗	شمسان šamsān	✗	✗	شموس šumūs
10	تنازل tanāzul ‘concession’	✗	✗	✗	✗	تنازلات tanazulāt	✗
11	خروج ḫurūǧ ‘exiting’	✗	✗	✗	✗	✗	✗
12	محمد ‘Mohammed’	✗	✗	✗	✗	✗	✗
13	✗	زينب ‘Zainab’	✗	✗	✗	✗	✗
14	✗	✗	✗	✗	✗	✗	مباحث mabāḥiṯ ‘intelligence agencies’
15	✗	✗	✗	✗	✗	استخبارات ʾistiḫbārāt ‘investigations’	✗

Table 6. Distribution of possible feminine and plural morphemes

Another problem with nouns is the issue of broken plurals (Ibrahim, 2002, Ratcliffe, 1998), which is the traditional grammarians’ term for describing the process of non-concatenative plural formation. The term was chosen to indicate that the base form of the nouns is broken either by removing one or more letters, adding one or more letters, changing vocalization or a combination of these. Arabic nouns have 30 templates which are served by 39 broken plural templates. Some templates of singular nouns can select from up to seven broken plural templates. The different plural templates were historically meant to indicate some meaning variation, such as whether the number of the plural is below or above ten, whether the noun describes a profession or an attribute, and whether the attribute is static or transient. These subtle meaning differences are no longer recognized even by well-educated native speakers.

These broken plural forms are, to a great extent, fossilized, i.e., they are not productive any more. So, the system relies only on the lexicographer’s knowledge to tell whether a particular noun is to have a regular or broken plural form. Trying to rely on the system to guess the broken plural form will make the transducer overgenerate excessively and needlessly.

2.5.2.3 Alternation Rules

Verbs are the category most affected by alternation operations. Therefore we focus here on the main conditions that trigger orthographical changes in verbs. Arabic verbs are generally classified (regarding the number of letters of the base form) into three-, four-, five- and six-letter verbs. Furthermore, trilateral verbs are traditionally classified into:

A. Strong verbs. These are the verbs that contain no weak letters. They are further classified into three categories:

1. Regular verbs. These are the verbs whose formative letters do not contain either a hamzated, doubled or weak letter.
2. Hamzated verbs. These are the verbs that contain a *hamza* (glottal stop) among its formative letters.
3. Doubled verbs. These are the verbs that are composed of two letters and the second is doubled.

B. Weak verbs. These are the verbs that contain a weak letter. A weak letter is one of three letters representing either long vowels or glides. They are ا ‘alif’ for the long vowel *ā* (which can also be represented orthographically by the letter ى ‘alif maqṣūra’). The second weak letter is و ‘wāw’ for the glide *w* or the long vowel ū. The third weak letter is ي ‘yāʾ’ for the glide *y* or the long vowel ī. Weak verbs are also classified into three categories:

1. Assimilated or miṯāl. A verb that contains an initial weak letter.
2. Hollow or aǧwaf. A verb that contains a middle weak letter.
3. Defective or nāqiṣ. A verb that contains a final weak letter.

We can extend this notion of weak and strong verbs into the four-, five- and six-letter verbs. This classification is of crucial importance in writing alternation rules. Strong regular verbs are generally not so much affected, orthographically, by inflection. The verb in (34) undergoes one alternation operation that is the deletion of the first letter, when inflected into the imperfective.

(34) استخرج ʼistaḫraǧa ‘extracted’ -> يستخرج yastaḫriǧ ‘extract’

However, more attention should be given to verbs that contain a weak, hamzated, or doubled letter at any position, as this usually requires more orthographical alterations during inflection. The verb in (35) undergoes two operations: deletion of the first letter and assimilation of the pre-final letter from ا ‘ā’ into ى ‘ī’.

(35) استقال ʼistaqāla ‘resigned’ -> يستقيل yastaqīl ‘resign’

In our lexc file, the start and end of stems are marked to provide information needed in conducting alternation operations, as shown by Figure 6 (transliteration and gloss are included in square brackets for illustration only).

```
LEXICON Verbs
^ss^شكر[šakara ‘thank’]^se^                                Transitive;
^ss^فرح[fariḥa ‘be-happy’]^se^@D.V.P@                     Intransitive;
^ss^رد[radda ‘respond’]^se^^dbl2^dbl@D.V.P@               Transitive;
^ss^أمر[ʼamara ‘order’]^se^^dbl2^dbl@D.M.I@               Transitive;
^ss^أضر[ʼaḍḍarra ‘harm’]^se^^dbl3^dbl@D.V.P@@D.M.I@       Intransitive;
^ss^امتد[ʼimtadda ‘extend’]^se^^dbl4^dbl@D.V.P@@D.M.I@    Intransitive;
^ss^تمخض[tamaḫḫaḍa ‘result-in’]^se^^dbl3^dbl@D.V.P@@D.M.I@ Intransitive;
^ss^استقر[ʼistaqarra ‘settle’]^se^^dbl5^dbl@D.V.P@@D.M.I@ Intransitive;
^ss^باع[bāʿa ‘sell’]^se^^origي^orig                       Transitive;
^ss^قال[qāla ‘say’]^se^^origو^orig                        Intransitive;
^ss^غزا[ġazā ‘fight’]^se^^origو^orig@D.V.P@@D.M.I@        Transitive;
^ss^رمى[ramā ‘throw’]^se^^origي^orig                      Transitive;
```

Figure 6. Verb stem entries

The tags are meant to provide the following information:

1. Start and end of verb stem. The multi-character symbol “^ss^” stands for stem start and “^se^” for stem end.
2. Which letter is doubled in the linear order, as the entries 4–8 in Figure 3 show. The mark “^dbl2^dbl”, for example means that the second letter is doubled.
3. If there is a long vowel that undergoes assimilation, the assimilated form needs to be explicitly stated. This is represented by the entries 10–13 in Figure 3. In traditional terms the origin of ا ‘ā’ in قال qāla ‘said’ is و ‘ū’. which means that ‘ā’ changes to ‘ū’ when the verb is inflected into the imperfective.
4. The flag diacritic “@D.V.P@” means “disallow the passive voice”, and “@D.M.I@” means “disallow the imperative mood”.

These markings are considered an intermediate language which is removed in the final stage, so that only surface strings are left on the bottom and analysis strings (or lexical strings) are left on the top of the network (Beesley, 1996).

2.5.2.4 List of Parts of Speech and Morphological Features

In our classification, there are nine parts of speech categories for Arabic: verbs, nouns, adjectives, adverbs, prepositions, determiners, conjunctions, pronouns and particles. Each of these categories has a set of morpho-syntactic features, as shown below.

Verbs: A verb has the following features:

- **Person:** first, second and third person
- **Number:** singular, dual, plural
- **Gender:** masculine, feminine
- **Voice:** active**,** passive
- **Mood:** imperative**,** declarative
- **Tense:** past, present

Nouns: A noun has the following features:

- **Gender:** masculine, female
- **Number:** singular, dual, plural
- **Case:** nominative, accusative, genitive
- **Humanness:** human, non-human
- **Additional Information:**
  - **Proper name:** person, place, organization
  - **Number:** ordinal, cardinal
  - **Date:** month, week

Adjectives: An adjective has the following features:

- **Gender:** masculine, female
- **Number:** singular, dual, plural
- **Case:** nominative, accusative, genitive

Pronouns: A pronoun has the following features:

- **Number:** singular, dual, plural
- **Gender:** masculine, female
- **Person:** first, second and third person
- **Case:** nominative, accusative, genitive
- **Relative:** for relative pronouns. They include the following features:
  - **Number:** singular, dual, plural
  - **Gender:** masculine, female
  - **Case:** nominative, accusative, genitive

Particles: A particle has the following features:

- negative
- interrogative
- **Tense:** future, past
- Complementizer
- affirmative

Determiners: A determiner has the following features:

- definite
- quantitive
- Demonstrative: for demonstrative adjectives. They have the following features:
  - **Proximity:** distal, proximal
  - **Number:** singular, dual, plural
  - **Gender:** masculine, feminine
  - **Case:** nominative, accusative, genitive

The other parts of speech, namely conjunctions, adverbs, and prepositions have no internal features.

2.5.3 Techniques followed in limiting ambiguities

We tried to make our system produce all and only the valid solutions and avoid spurious solutions. We observe well-formedness conditions of Arabic words and avoid any pruning of valid analyses. The following considerations and techniques were followed to achieve this goal.

1. Using the stem as the base form, as this approach is less likely to overgenerate. Automatic derivation from the root can be risky as it may create stems not used in the language.

2. Non-inclusion of classical words or word senses, as they add only to the size of the lexicon and the level of ambiguity. In our system words are included only if they are found in the corpus. We did not rely on classical dictionaries or word lists.

3. Observation of the rules governing the combination of words with affixes and clitics, or grammar-lexis specifications, which work as filters for spurious ambiguities (Abbès et al., 2004, Dichy and Hassoun, 1998, Dichy, 2001, Dichy and Fargaly, 2003). For example, adjectives, names of places and verbal nouns do not combine with possessive pronouns. Also verbal nouns derived from intransitive verbs do not combine with accusative

pronouns. Yet more can be done regarding the filtering of human objects from verbs that allow only non-human objects (Dichy and Fargaly, 2003) such as (36), which is still accepted by our system.

(36) قرأتهم
qaraʾtu-hum
'I read them'

There are also nouns that semantically do not allow the affixation of genitive pronouns, such as (37) which is still not properly handled by our system.

(37) كيميائي
kīmyaʾiyy
'my chemistry'

4. Specifying which verbs can have the passive forms. From 1532 verbs, only 36% are allowed to have passive forms (504 transitive verbs, and 43 intransitive verbs). Initially all transitive verbs were allowed to have a passive form and all intransitive verbs were not. Then all verbs were reviewed manually for acceptability according to the author's judgment. A sum of 198 transitive verbs was not allowed to have a passive form, while some intransitive verbs are allowed. Levin (1993) stated that intransitive, prepositional verbs can have passive constructions under constraints on the semantic roles of the arguments. In our system, verbs in the 1st and 2nd person are not allowed to have a passive form. The 1st and 2nd persons are deemed as highly unlikely forms, first, because MSA is a formal written language, and these persons are mostly used in conversations or autobiographies. Second, these persons have orthographical shapes that are identical with other forms, and writers will tend to use other syntactically equivalent structures for expressing the passive in this case. Another good idea for limiting the use of the passive would be to constrain it according to tense, as done in the Buckwalter's system.

5. Specifying which verbs can have imperative forms. Out of 1532 verbs, only 484 verbs (32%) are allowed to have an imperative form (324 Transitive verbs, 160 Intransitive verbs). According to Levin (1993), the imperative construction does not appear with verbs of perception and *admire*-type

psych-verbs. It does not also appear with verbs of entity-specific change of state. These are the “verbs that describe changes of state that are specific to particular entities”, such as *bloom*, *erode*, *corrode*, *decay*, *dry*, *stagnate*, *blossom*, *wither*, *tarnish* and *swell*. This semantic description could be extremely helpful in deciding which verbs can have an imperative form and which cannot. Building such semantics-based networks for Arabic verbs was time-consuming; therefore we had to rely on personal judgment of plausibility.

2.6 Evaluation

Our aim is to evaluate Xerox Arabic Morphological Analysis and Generation, Buckwalter Arabic Morphological Analyzer and Attia’s Arabic Morphological Transducer with respect to ambiguity. Due to the fact that a gold standard annotated corpus for Arabic is not yet available (to our knowledge), a large scale, automatic evaluation is not possible. Therefore we conducted a small-scale manual evaluation experiment to test the ambiguity rate of the three morphologies on one hand and to test precision of the two morphologies with the least ambiguity rate on the other hand.

We selected five documents from Al-Jazeera web site on 28-29/6/2006 containing a total of 950 unique words and 67 MWEs. We tested these words on each of the target morphologies, and then we conducted a detailed manual analysis for the two morphologies with the least ambiguity rate to see how accurate they were in obtaining the correct set of analyses and avoiding spurious ambiguities. We first show the precision evaluation in Table 7.

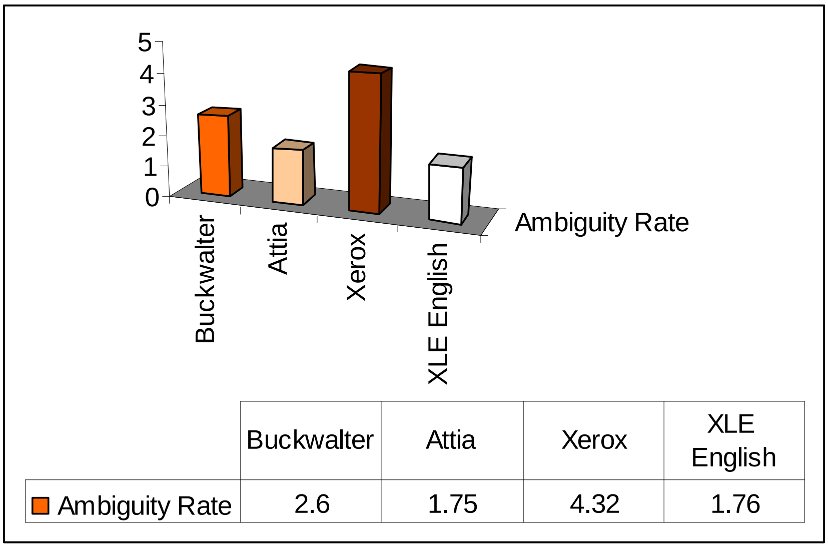
Criteria	Buckwalter	Xerox	Attia
Complete	617	-	756
Over-specified	247	-	67
Underspecified	40	-	75
Wrong Analysis	1	-	10
Over-&under-specified	20	-	5
Irrelevant	5	5	5
Not found	20	39	32
Total Solutions for 895 words (after excluding 55 not found)	2332	3871	1574

Table 7. Breakdown of evaluation results

In the table a "complete" analysis is a precise one that neither contains a spurious ambiguity nor lacks a plausible solution. An "over-specified" analysis is one that contains all plausible solutions beside one or more spurious ambiguities. It must be noted here that a spurious ambiguity is an ambiguity that falls outside the domain of the language, not a context or subject related ambiguity. An "underspecified" analysis is one that fails to account for one or more plausible solutions among the list of solutions. "Over-&under-specified" analysis denotes those solutions which contain spurious ambiguities and at the same time do not include one or more plausible solutions. The term "irrelevant" is used for misspelled words or those that do not occur alone, but usually occur as part of a MWE. Buckwalter's precision score is 64%, while Attia's Morphology achieved 79%. Although Attia's morphology is almost a quarter of the size of Buckwalter, it does not contain too many underspecified analyses. As Attia and Buckwalter achieved a relatively high score of precision at a low rate of solutions per word, it can be easily deduced that Xerox, with its high number of solutions, is over-specified for most words, and so no breakdown was perceived to be needed.

Out of curiosity, we tried to see what the ambiguity level in an English morphology is. We, however, do not intend to say that Arabic ambiguity level should be similar to English or that English can be used as a baseline for Arabic, as ambiguity is a language-specific issue and comparing ambiguity between two languages is not possible. English ambiguity rate is tested using XLE morphological transducer (Butt et al., 2002) on 979 words and received 1732 solutions, giving an ambiguity rate of 1.76.

In order to measure the ambiguity rate in the three morphologies in our experiment, all words that were not known to any of the morphologies (that was a total of 55 words) were removed from the test list, which was reduced to 895. The ambiguity rates for the three morphologies are shown in Figure 7. A total of 67 MWEs were excluded from overall evaluation, as they are not supported on Buckwalter or Xerox. Attia, however, recognized 25 MWEs, that is 37% coverage.



	Buckwalter	Attia	Xerox	XLE English
■ Ambiguity Rate	2.6	1.75	4.32	1.76

Figure 7. Comparison of the ambiguity rates in three morphologies

As Figure 7 shows, Attia's Morphology outperforms both Buckwalter's and Xerox Morphologies in curbing ambiguities. Error review shows that the sources of spurious ambiguities in Buckwalter and Xerox morphologies are summarized mainly in the following three points:

1. Inclusion of classical terms.
2. Incompliance with the rules of grammar-lexis relations.
3. Improper application of spelling relaxation rules.

We conclude that the rich and complex morphology of Arabic does not automatically mean that it is highly ambiguous. The analysis and evaluation conducted in this research shows that most of the ambiguities produced by Xerox Arabic Finite State Morphology and Buckwalter Arabic Morphological Analyzer are spurious ambiguities caused by the inclusion of classical entries, rule-created overgenerated stems with no actual place in the language, overlooking word-clitic combination rules (or grammar-lexis specifications), and overdoing spelling relaxation rules. By avoiding these pitfalls a more focused, less ambiguous morphological analyzer can be developed.

3 Tokenization

Tokenization is a necessary and non-trivial step in natural language processing. The function of a tokenizer is to split a running text into tokens, so that they can be fed into a morphological transducer or POS tagger for further processing. The tokenizer is responsible for defining word boundaries, demarcating clitics, multiword expressions, abbreviations and numbers.

In this chapter we describe a rule-based system that handles tokenization as a well-rounded process with a preprocessing stage (white space normalizer), and a post-processing stage (token filter). We also show how it interacts with morphological transducers, and how ambiguity is resolved.

Tokenization is a significant issue in natural language processing as it is “closely related to the morphological analysis” (Chanod and Tapanainen, 1996). This is even more the case with languages with rich and complex morphology such as Arabic. In the case of Arabic, where a single word can comprise a stem and up to three clitics, morphological knowledge needs to be incorporated into the tokenizer.

Clitics are syntactic units that do not have free forms but are instead attached to other words. Deciding whether a morpheme is an affix or a clitic can be confusing. However, we can generally say that affixes carry morpho-syntactic features (such as tense, person, gender or number), while clitics serve syntactic functions (such as negation, definition, conjunction or preposition) that would otherwise be served by an independent lexical item. Therefore tokenization is a crucial step for a syntactic parser that needs to build a tree from syntactic units. An example of a clitic in English is the contracted form *n’t* in *He didn’t go*.

Arabic clitics, however, are not as easily recognizable. Clitics use the same alphabet as that of words, with no demarcating mark, and they can be concatenated one after the other. Without sufficient morphological knowledge, it is impossible to detect and mark clitics. Here we show different levels of implementation of the Arabic tokenizer, according to the levels of linguistic depth involved.

Arabic Tokenization has been described in various researches and implemented in many solutions as it is a required preliminary stage for further processing. These solutions include morphological analysis (Beesley, 2001, Buckwalter, 2002), diacritization (Nelken and Shieber, 2005), Information Retrieval (Larkey and Connell, 2002), and POS Tagging (Diab et al., 2004, Habash and Rambow, 2005). None of these projects, however, describe tokenization as a standalone solution or show how ambiguity is filtered and MWEs are treated.

In our research, tokenization is handled in a rule-based system as an independent process. We show how the tokenizer interacts with other transducers. We also show how incorrect tokenizations are filtered out, and how undesired tokenizations are marked. In Chapter 4, we show how MWEs are identified and delimited. All tools in this research are developed in Finite State Technology (Beesley and Karttunen, 2003). These tools have been developed to serve an Arabic Lexical Functional Grammar parser using XLE (Xerox Linguistics Environment) platform as part of the ParGram Project (Butt et al., 2002).

This chapter is an updated version of my paper titled “Arabic Tokenization System” (Attia, 2007).

3.1 Arabic Tokens

A *token* is the minimal syntactic unit; it can be a word, a part of a word (or a clitic), a multiword expression, or a punctuation mark. A tokenizer needs to know a list of all word boundaries, such as white spaces and punctuation marks, and also information about the token boundaries inside words when a word is composed of a stem and clitics. Throughout this research full form words, i.e. stems with or without clitics, as well as numbers will be termed *main tokens*. All main tokens are delimited either by a white space or a punctuation mark. Full form words can then be divided into *sub-tokens*, where clitics and stems are separated.

3.1.1 Main Tokens

A tokenizer relies mainly on white spaces and punctuation marks as delimiters of word boundaries (or main tokens). Additional punctuation marks are used in Arabic such as the comma '،', question mark '؟' and semicolon '؛'. Numbers are also considered as main tokens. A few Arab countries use the Arabic numerals as in English, while most Arab countries use the Hindi numerals such as ٢ '2' and ٣ '3'. Therefore a list of all punctuation marks and number characters must be fed to the system to allow it to demarcate main tokens in the text.

3.1.2 Sub-Tokens

Arabic morphotactics allows words to be prefixed or suffixed with clitics (Attia, 2006a). Clitics themselves can be concatenated one after the other. Furthermore, clitics undergo assimilation with word stems and with each other, which makes them even harder to handle in any superficial way. A verb can comprise up to four sub-tokens (a conjunction, a complementizer, a verb stem and an object pronoun) as illustrated by Figure 8.

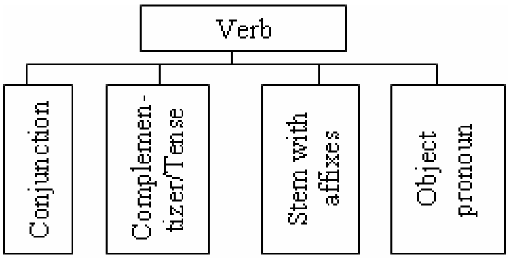


Figure 8. Possible sub-tokens in Arabic verbs

Similarly a noun can comprise up to four sub-tokens. Although Figure 9 shows five sub-tokens, the definite article and the genitive pronoun are mutually exclusive.

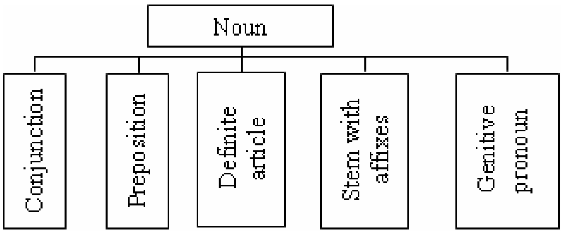


Figure 9. Possible sub-tokens in Arabic nouns

Moreover there are various rules that govern the combination of words with affixes and clitics. These rules are called grammar-lexis specifications (Abbès et al., 2004, Dichy, 2001, Dichy and Fargaly, 2003). An example of these specifications is a rule that states that adjectives and proper nouns do not combine with possessive pronouns.

3.2 Tokenization Solutions

There are different levels at which an Arabic tokenizer can be developed, depending on the depth of the linguistic analysis involved. During our work with the Arabic grammar we developed three different solutions, or three models, for Arabic tokenization. These models vary greatly in their robustness, compliance with the concept of modularity, and the ability to avoid unnecessary ambiguities.

The tokenizer relies on white spaces and punctuation marks to demarcate main tokens. In demarcating sub-tokens, however, the tokenizer needs more morphological information. This information is provided either deterministically by a morphological transducer, or non-deterministically by a token guesser. Eventually both main tokens and sub-tokens are marked by the same token boundary, which is the sign '@' throughout this research. The classification into main and sub-tokens is a conceptual idea that helps in assigning the task of identification to different components.

Identifying main tokens is considered a straightforward process that looks for white spaces and punctuation marks and divides the text accordingly. No further details of main tokens are given beyond this point. The three models described below are different ways to identify and divide sub-tokens, or clitics and stems within a full form word.

3.2.1 Model 1: Tokenization Combined with Morphological Analysis

In this implementation the tokenizer and the morphological analyzer are one and the same. A single transducer provides both morphological analysis and tokenization. Examples of the tokenizer/morphological analyser output are shown in (38). The '+' sign precedes morphological features, while the '@' sign indicates token boundaries.

(38) وليشكر (waliyaškur: and to thank)
و+conj@ل+comp@+verb+pres+sgشكر@

This sort of implementation is the most linguistically motivated. This is also the most common form of implementation for Arabic tokenization (Habash and Rambow, 2005). However, it violates the design concept of modularity which requires systems to have separate modules for undertaking separate tasks. For a syntactic parser that requires the existence of a tokenizer besides a morphological analyzer (such as XLE), this implementation is not workable, and either Model 2 or Model 3 is used instead.

3.2.2 Model 2: Tokenization Guesser

In this model tokenization is separated from morphological analysis. The tokenizer only detects and demarcates clitic boundaries. Yet information on what may constitute a clitic is still needed. This is why two additional components are required: a clitics guesser to be integrated with the tokenizer, and a clitics transducer to be integrated with the morphological transducer.

Clitics Guesser. We developed a guesser for Arabic words with all possible clitics and all possible assimilations. See Beesley and Karttunen (2003) on how to create a basic guesser. The core idea of a guesser is to assume that a stem is composed of any arbitrary sequence of Arabic alphabets, and this stem can be prefixed or/and suffixed with a limited set of tokens. This guesser is then used by the tokenizer to mark clitic boundaries. Due to the nondeterministic nature of a guesser, there will be increased tokenization ambiguities, as in (39) (only correct analysis is provided with transliteration and gloss).

(39) وللرجل (walirrağul: and to the man)
و@ل@ال@رجل@ wa@li@al@rağul@ and@to@the@man@
و@ل@الرجل@
و@للرجل@
وللرجل@

Clitics Transducer. We must note that Arabic clitics do not occur individually in natural texts. They are always attached to words. Therefore a specialized small-scale morphological transducer is needed to handle these newly separated forms. We developed a lexc transducer for clitics only, treating them as separate words. The purpose of this transducer is to provide analysis for morphemes that do not occur independently.

(40) و[wa- ‘and’]+conj
ل[li- ‘to’]+prep
ال[al- ‘the’]+art+def

This small-scale specialized transducer is then unioned (or integrated) with the main morphological transducer. Before making the union it is necessary to remove all paths that contain any clitics in the main morphological transducer to eliminate redundancies.

In our opinion this is the best model, the advantages are robustness as it is able to deal with any words whether they are known to the morphological transducer or not, and abiding by the concept of modularity as it separates the process of tokenization from morphological analysis.

There are disadvantages, however, for this model, and among them is that the morphological analyzer and the syntactic parser have to deal with increased tokenization ambiguities. The tokenizer is highly non-deterministic as it depends on a guesser which, by definition, is non-deterministic. For a simple sentence of three words, we are faced with eight different tokenization solutions. Nonetheless, this influx of ambiguities can be handled as will be explained later.

3.2.3 Model 3: Tokenization Dependent on the Morphological Analyser

In the above solution, the tokenizer defines the possible Arabic stem as any arbitrary sequence of Arabic letters. In this solution, however, word stems are not guessed, but taken as a list of actual words. A possible word in the tokenizer in this model is any word found in the morphological transducer. The morphological transducer here is the same as the one described in Model 1 but with one difference, that is the output does not include any morphological features, but only token boundaries between clitics and stems.

This is a much more deterministic tokenizer that handles clitics properly. The main downfall is that only words found in the morphological transducer are tokenized. It is not robust, yet it may be more convenient during grammar debugging, as it provides much fewer analyses than model 2. Here spurious ambiguities are successfully avoided.

(41) وللرجل (walirrağul: and to the man)
و@ل@ال@رجل@ wa@li@al@rağul@ and@to@the@man@

One advantage of this implementation is that the tool becomes more deterministic and more manageable in debugging. Its lack of robustness, however, makes it mostly inapplicable as no single morphological transducer can claim to comprise all the words in a language. In our XLE grammar, this model is only 0.05% faster than Model 2. This is not a statistically significant advantage compared to its limitations.

3.2.4 Normalization

Normalization is a preliminary stage to tokenization where preliminary processing is carried out to ensure that the text is consistent and predictable. In this stage, for example, the decorative elongation character, *kashida*, and all diacritics are removed. Redundant and misplaced white spaces are also corrected, to enable the tokenizer to work on a clean and predictable text.

In real-life data spaces may not be as regularly and consistently used as expected. There may be two or more spaces, or even tabs, instead of a single space. Spaces might even be added before or after punctuation marks in the wrong manner. Therefore, there is a need for a tool that eliminates inconsistency in using white spaces, so that when the text is fed into a tokenizer or morphological analyzer, words and expressions can be correctly identified and analyzed. Table 8 shows where spaces are not expected before or after some punctuation marks.

No Space Before	No Space After
)	(
}	{
]	[
”	“

Table 8. Space distribution with some punctuation marks

We have developed a white space normalizer whose function is to go through real-life texts and correct mistakes related to the placement of white spaces. When it is fed an input such as the one in (42a) in which additional spaces are inserted and some spaces are misplaced, it corrects the errors and gives the output in (42b):

(42) a. نشر الديمقراطية (في الشرق الأوسط)سيقود إلى   السلام .
našru ad-dīmuqrāṭiyyh (fī aš-šarq al-ʾawsaṭ)sayaqūdu ʾilā   as-slām .
Spreading democracy (in the Middle East)will lead to   peace .
b. نشر الديمقراطية (في الشرق الأوسط) سيقود إلى السلام.
našru ad-dīmuqrāṭiyyh (fī aš-šarq al-ʾawsaṭ) sayaqūdu ʾilā as-slām.
Spreading democracy (in the Middle East) will lead to peace.

3.3 Resolving Ambiguity

There are different types of ambiguity. There are spurious ambiguities created by the guesser. There are also ambiguities which do not exist in the text before tokenization but are only created during the tokenization process. Finally there are real ambiguities, where a form can be read as a single word or two sub-tokens, or where an MWE has a compositional reading. These three types are treated by the following three subsections respectively.

3.3.1 Discarding Spurious Ambiguities

Tokenization Model 2 is chosen as the optimal implementation due to its modularity, efficiency and robustness, yet it is highly nondeterministic and produces a large number of spurious ambiguities. Therefore, a morphological transducer is needed to filter out the tokenization paths that contain incorrect sub-tokens. Recall example (39) which contained the output of the nondeterministic tokenizer. In (43) below, after the output is fed into a morphological transducer, only one solution is accepted and the rest are discarded, as underlined words do not constitute valid stems.

(43) وللرجل (walirrağul: and to the man)
و@ل@ال@رجل@ wa@li@al@rağul@ and@to@the@man@ - Passed.
و@ل@الرجل@ - Discarded.
و@للرجل@ - Discarded.
وللرجل@ - Discarded.

3.3.2 Handling Tokenization Ambiguities

Among the functions of a tokenizer is to separate clitics from stems. Some clitics, however, when separated, become ambiguous with other clitics and also with other free forms. For example the word كتابهم kitābuhum has only one morphological reading (meaning *their book*), but after tokenization كتاب@هم there are three different readings, as the second token هم can either be a clitic genitive pronoun (the intended reading) or a free pronoun meaning *they* (the overall meaning is *a book*, *they*) or a noun meaning *worry* (forming the compound *book of worry*).

This problem is solved by inserting a mark that precedes enclitics and follows proclitics to distinguish them from each other as well as from free forms (Ron M. Kaplan and Martin Forst, personal communications, Oxford, UK, 20 September 2006). The mark we choose is the Arabic elongation short line called *kashida* which is originally used for graphical decorative purposes and looks natural with most clitics. To illustrate the usage, a two-word string (44a) will be rendered without *kashida*s as in (44b), and a single-word string that contains clitics (45a) will be rendered with a distinctive *kashida* before the enclitic pronoun as in (45b). This indicates that the pronoun is attached to the preceding word and not standing alone.

(44) a. كتاب هم
kitābu hum/hammin
'book of worry/a book, they'
b. كتاب@هم

(45) a. كتابهم
kitābu-hum
'their book'
b. كتاب@ـهم

This implementation will also resolve a similar ambiguity, that is ambiguity arising between proclitics and enclitics. The proclitic preposition ك ka 'as' always occurs initially. There is a homographic enclitic accusative pronoun ك ka ‘you’ that always occurs in the final position. This can create ambiguity in instances such as the made-up sentence in (46a). The sentence has the initial tokenization of (46b) without a *kashida,* and therefore the central token becomes ambiguous as it can now be attached either to the preceding or following word leading either to the readings in (46a) or (46c). The *kashida* placement, however, resolves this ambiguity as in (46d). The *kashida* is added after the token, indicating that it is attached to the following word and now only the reading in (46a) is possible.

(46) a. أعطيت كالأمير
ʼaʿṭaitu ka-l-ʼamīr
‘I gave like the prince’
b. أعطيت@ك@الأمير
c. أعطيتك الأمير
ʼaʿṭaitu-ka al-ʼamīr
‘I gave you the prince’
d. أعطيت@كـ@الأمير

3.3.3 Handling Real Ambiguities

Some tokenization readings are genuine, yet highly infrequent and undesired in real-life data. These undesired readings create spurious ambiguities, as they are confused with more common and more acceptable forms. For example the Arabic preposition بعد baʿd ‘after’ has a possible remote reading if split into two tokens بـ@عد, which is made of two elements: بـ bi ‘with’ and عد ʿadd ‘counting’, meaning ‘by counting’. Similarly بين baina ‘between’ has the possible remote reading بـ@ين, which is made of two tokens as well: بـ bi ‘with’ and ين yin ‘Yen’, meaning ‘by a Yen’.

The same problem occurs with MWEs. The optimal handling of MWEs is to treat them as single tokens and leave internal spaces intact. Yet a nondeterministic tokenizer allows MWEs to be analysed compositionally as individual words. So the MWE حظر التجول ḥaẓr at-tağawwul ‘curfew’ has two analyses, as in (47), although the compositional reading in (47b) is undesired.

(47) a. حظر التجول@ ḥaẓr at-tağawwul ‘curfew’
b. حظر@التجول
ḥaẓr ‘forbidding’ @ at-tağawwul ‘walking’

The solution to this problem is to mark the undesired readings. This is implemented by developing a filter, or a finite state transducer that contains all possible undesired tokenization possibilities and attaches the “+undesired” tag to each one of them.

Undesired tokens, such as بـ@ين and بـ@عد, explained above, can be included in a custom list in the token filter. As for MWEs, the token filter imports a list from the MWE transducer and replaces the spaces with the token delimiter ‘@’ to denote the undesired tokenization solutions. The token filter then matches the lists against the output of the tokenizer. If the output contains a matching string a mark is added. Notice how (48b) is marked with the “+undesired” tag.

(48) a. حظر التجول@ [ḥaẓr at-tağawwul ‘curfew’]
b. حظر@التجول+undesired

This transducer or filter is composed on top of the core tokenizer. The overall design of the tokenizer and its interaction with other finite state components is shown in Figure 10. We need to note that the tokenizer, in its interaction with the morphological transducer and the MWE transducer, does not seek morpho-syntactic information, but it queries for lists and possible combinations.

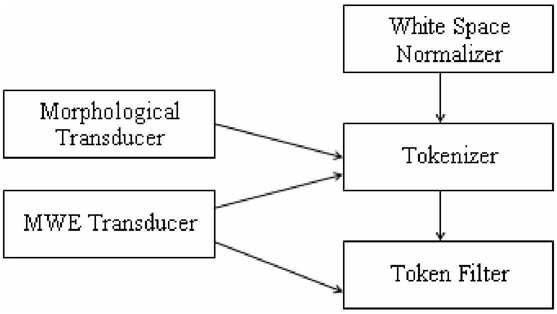


Figure 10. Design of the Arabic Tokenizer

We conclude that tokenization is a process that is closely connected to and dependent on morphological analysis. In our research we show how different models of tokenization are implemented at different levels of linguistic depth. We also explain how the tokenizer interacts with other components, and how it resolves complexity and filters ambiguity. By applying token filters we gain control over the tokenization output.

4 Handling Multiword Expressions

Multiword expressions (MWEs) are so pervasive in all languages that they cannot be ignored in any plausible linguistic analysis. When neglected, MWEs put great hurdles in the way of syntactic parsing, machine translation and information retrieval systems. On the other hand, when they are properly accommodated, syntactic ambiguity and parse time are reduced, and more importantly, a degree of certitude is given to the syntactic analysis, as well as to machine translation output. MWEs vary in syntactic category, structure, the degree of semantic opaqueness, the ability of one or more constituents to undergo inflection and processes such as passivization, and the possibility of having intervening elements. Therefore, there is no straight-forward way of dealing with them. This research shows how some MWEs can be dealt with in as early stage as the tokenization, while others are recognized only by the syntactic parser.

There was previously a tendency to ignore MWEs in linguistic analysis due to their idiosyncrasy. However, it is now recognized that MWEs have a high frequency in day-to-day interactions (Venkatapathy, 2004), that they are in the same order of magnitude as the speaker's lexicon of single words, that they account for 41% of the entries in WordNet 1.7 (Fellbaum, 1998, Sag et al., 2002), that phrasal verbs account for "about one third of the English verb vocabulary" (Li et al., 2003), and that technical domains rely heavily on them. This makes it imperative to handle MWEs if we want to make large-scale, linguistically-motivated, and precise processing of the language.

MWEs constitute serious pitfalls for machine translation systems and human translators as well (Volk, 1998). When they are translated compositionally, they give textbook examples of highly intolerable, blind and literal translation. It also underestimates the problem to assume that it only concerns translation, and that it should be handled during higher phases of processing such as transfer. In fact MWEs require deep analysis that starts as early as the normalization and tokenization, and goes through morphological analysis and into the syntactic rules. The focus of this section is to explain how MWEs can be accommodated in each step in the preprocessing and the processing stages. The advantages of handling MWEs in the pre-processing stage are manageability of translation, avoidance of needless analysis of idiosyncratic structures, reduction of parsing ambiguity, and reduction of parse time (Brun, 1998). This is why there are growing calls to construct MWE dictionaries (Guenthner and Blanco, 2004), lexicons (Calzolari et al., 2002), and phrasets (sets of phrases) (Bentivogli and Pianta, 2003).

We show how several devices can be applied to handle MWEs properly at several stages of processing. All the solutions are applied to Arabic; yet, most of the solutions are general and are applicable to other languages as well.

This chapter is an enhanced and updated version of my paper titled "Accommodating Multiword Expressions in an Arabic LFG Grammar" (Attia, 2006b).

4.1 Definition

MWEs encompass a wide range of linguistically related phenomena that share the criterion of being composed of two words or more, whether adjacent or separate. MWEs have been defined as "idiosyncratic interpretations that cross word boundaries (or spaces)" (Sag et al., 2002). In an MWE, the structure and the semantics of the expression are dependant on the phrase as a whole, and not on its individual components (Venkatapathy, 2004). MWEs cover expressions that are traditionally classified as idioms (e.g. *down the drain*), prepositional verbs (e.g. *rely on*), verbs with particles (e.g. *give up*), compound nouns (e.g. *book cover*) and collocations (e.g. *do a favour*).

The term *multiword* itself has been challenged as "vague" (Alegria et al., 2004) if you follow the conventional definition of a word as a string of letters between two delimiters (spaces, tabs, punctuation marks, etc.). There are languages that do not use spaces between words, such as Japanese. Compound nouns in German are written without spaces. Arabic has a group of clitics (pronouns, prepositions, definite article, etc.) that typically attach themselves to other words. Therefore, we need either to change the term *multiword* to *multitoken*, or more conveniently redefine *word* in this context to mean tokens that convey grammatical functions that can either be separated by spaces or attached to other words.

Although there is not a clear-cut definition with which we can decide what expressions can be considered MWEs, there is a set of criteria (adapted from (Baldwin, 2004, Calzolari et al., 2002, Guenthner and Blanco, 2004)) when one or more of which applies, the expression can safely be considered as an MWE.

1. Lexogrammatical fixedness. The expression has come to a rigid or frozen state. This fixedness can be identified through a number of tests. Components of the expression must be immune to the following operations:
   a. Substitutability. The word *many* in (49) cannot be substituted with its synonym *several*.

   (49) *many thanks* -> * *several thanks*

   b. Deletion. The adjective in (50) cannot be deleted.

   (50) *black hole* -> * *the hole*

   c. Category transformation. The adjective in (51) cannot be changed to a noun.

   (51) *black hole* -> * *the blackness of the hole*

   d. Permutation. A noun-noun compound can usually be expressed by a noun-preposition-noun as in (52), but not in the case of MWEs as in (53) and (54).

   (52) *the hospital staff* -> *the staff of the hospital*
   (53) *life guard* -> * *the guard of life*
   (54) *kiss of life* -> * *life kiss*

2. Semantic non-compositionality. The meaning of the expression is not derived from the meaning of the component parts.

   (55) *kick the bucket* = die

3. Syntactic irregularity. The expression exhibits a structure that is inexplicable by regular grammatical rules.

(56) *long time, no see*

(57) *by and large*

4. Single-word paraphrasability. The expression can be paraphrased by a single word.

(58) *give up = abandon*

5. Translatability into a single word or when their translation differs from a word to word translation (Brun, 1998). In various projects a corpus of translated texts is used to judge or detect MWEs (Butt et al., 1999b, Nerima et al., 2003, Smadja et al., 1996). Sometimes a unilingual analysis may be confused about whether an expression is a regular combination of words or an MWE. Translation usually helps to show expressions in perspective.

(59) *looking glass = مرآة mir'āh* (Arabic)

4.2 Classification of Multiword Expressions

In order for an expression to be classified as an MWE, it must show a degree of semantic non-compositionality and/or a degree of morpho-syntactic inflexibility. MWEs are classified with regard to their semantic compositionality into lexicalized and institutionalized expressions. Moreover, they are classified with regard to their flexibility into fixed, semi-fixed and syntactically flexible expressions (adapted from (Sag et al., 2002)).

4.2.1 Compositional vs. Non-Compositional MWEs

Semantic compositionality, sometimes termed decomposability, is "a means of describing how the overall sense of a given idiom is related to its parts" (Sag et al., 2002). An example of non-compositionality is the expression *kick the bucket*, where the meaning 'die' has no relation to any word in the expression. An example of compositional expressions is the compound noun *book cover*, where the meaning is directly related to the component parts. Unfortunately, it can be very elusive to decide whether an expression is compositional or not. Most of the time "one cannot really make a binary distinction between compositional and non-compositional MWEs” (Venkatapathy, 2004). In fact, MWEs occupy a continuum in a large scale. At one end of the scale there are those expressions that are highly opaque and non-compositional. Here the meaning is not traceable to any of the component parts, such as *kick the bucket*. In the middle of the scale there are those where one or more words are used in an idiosyncratic sense, or use “semantics unavailable outside the MWE” (Baldwin et al., 2003), such as *spill the beans*. At the other end of the scale there are those who are highly compositional, such as *book cover, traffic light*s, *health crisis* and *party meeting*.

Non-compositional expressions, or, more accurately, expressions that show any degree of non-compositionality, are termed *lexicalized* and are automatically eligible to be considered as MWEs. However, in order for compositional expressions to be included in an MWE lexicon, they need to be conventionalized or *institutionalized*. This means that these expressions have come to such a frequent use that they block the use of other synonyms and near synonyms (Nerima et al., 2003). When words co-occur in a statistically meaningful way like this they are called *collocations*. Relying on this criterion, expressions such as *book cover* and *traffic lights* can be safely added to an MWE lexicon, while *health crisis* and *party meeting* cannot.

Collocations also include frozen modifiers (Guenthner and Blanco, 2004). There are two forms of frozen modifiers. Sometimes a noun is systematically modified by an adjective or adverb that indicates “intensity, anti-intensity, praise and anti-praise” (Guenthner and Blanco, 2004). Examples are *bad weather, heavy rain, bitter cold,* and *scorching heat.* The second form of frozen modifiers is the set of modifiers whose use in the language has died out except with specified nouns. These modifiers are morphologically rigid and their use is extremely restricted. Examples from Arabic are:

(60) حرب شعواء
ḥarbun šaʿwāʾun
war large-scale
‘large-scale war’

(61) حرب ضروس
ḥarbun ḍarūsun
war exhausting
‘exhausting war’

(62) ظلام دامس
ẓalāmun dāmisun
darkness gloomy
‘gloomy darkness’

Nowhere in Modern Standard Arabic can we find the adjectives šaʿwāʾun, ḍarūsun and dāmisun except as modifiers for the nouns specified. Even the gender and number varieties of the adjectives are never used or heard.

4.2.2 Flexible vs. Inflexible MWEs

With regard to syntactic and morphological flexibility, MWEs are classified into three types: fixed, semi-fixed and syntactically flexible (Baldwin, 2004, Oflazer et al., 2004, Sag et al., 2002).

4.2.2.1 Fixed Expressions

These expressions are lexically, syntactically and morphologically rigid. An expression of this type acts just like a single word that happens to contain spaces, such as الشرق الأوسط aš-šarq al-ʾawsat ‘the Middle East’ and بيت لحم bait lahim ‘Bethlehem’. Some expressions are frozen at the level of the sentence, sometimes termed “frozen texts” (Guenthner and Blanco, 2004). These include proverbs such as *Buy cheap, buy twice* and *A bird in hand is worth two in the bush*, and pragmatically fixed expressions such as *Good morning* and *We haven’t got all day*.

4.2.2.2 Semi-Fixed Expressions

These expressions undergo morphological and lexical variations, but still the expression components are adjacent. They can be neither reordered nor separated by external elements.

The variations that can affect semi-fixed expressions include:

1. Morphological variations that express person, number, tense, gender, etc. Examples are:

(63) a. فترة انتقالية
fatratah ʾintiqāliyyah
period.sg.fem translational.sg.fem
'transitional period'

b. فترتان انتقاليتان
fatratān ʾintiqāliyyatān
period.dual.fem translational.dual.fem
'two transitional periods'

2. Lexical variations. This includes the case when a position is filled by a choice from the set of reflexive pronouns (e.g. *prostrate himself/herself*), or when one word can be replaced by another (e.g. *to sweep something under the carpet/rug*).

(64) على وجه/ظهر الأرض/البسيطة
ʿalā waǧhi/ẓahri al-ʾarḍi/al-basīṭati
on face/back the-earth/the-land (on the face of the earth)

4.2.2.3 Syntactically Flexible Expressions

These are the expression that can either undergo reordering, such as passivization (e.g. *the cat was let out of the bag*), or allow external elements to intervene between the components such as (65b), where the adjacency of the MWE is disrupted.

(65) a. دراجة نارية
darraǧah nāriyyah
bike fiery
'motorbike'

b. دراجة الولد النارية
darraǧat al-walad an-nāriyyah
bike the-boy the-fiery
'the boy's motorbike'

4.3 Collecting Multiword Expressions

Although many monolingual and bilingual electronic dictionaries of single entries have been made available for different languages, few such lexicons have been constructed for MWEs. Hence comes the need to identify and collect MWEs before starting to process the text. Many projects have dealt with the automatic extraction of MWEs (Agarwal et al., 2004, Butt et al., 1999b, Deane, 2005, Nerima et al., 2003, Smadja et al., 1996) from texts. In order be able to conduct automatic extraction you need to work with a number of tools such as a tagger, parser, and a corpus of translated texts.

In our project some MWEs are collected manually. The rest are extracted semi-automatically from the Arabic corpus using a concordance tool. A list of terms that frequently occur as part of an MWE is built. These terms are then tracked in a concordance and the output is sorted and filtered. For example, some words are frequently found in a compound name, such as جمهورية ğumhūriyyah ‘republic’, which helped gather 54 expressions, حزب ḥizb ‘party’ helped gather 258 expressions, منظمة munaẓẓamah ‘organization’ helped gather 163 expressions, and مجلس mağlis ‘council’ helped gather 124 expressions.

(66) جمهورية مصر العربية
ğumhūriyyatu miṣra al-ʿarabiyyah
republic Egypt the-Arab
‘The Arab Republic of Egypt’

(67) منظمة الصليب الأحمر
munaẓẓamatu aṣ-ṣalībi al-ʾaḥmar
organization the-cross the-red
‘Red Cross Organization’

(68) مجلس التعاون الخليجي
mağlisu at-taʿāwuni al-ḫalīğiyy
council the-cooperation the-gulf
‘Gulf Cooperation Council’

Some proper names in Arabic are composed of two parts, the first is the word عبد (ʿabd [lit. servant]) preceding one word from a fixed set of “divine attributes” This helped in collecting 90 compound names.

(69) عبد الرحمن
ʿabdu ar-raḥman
Abd al-Rahman (lit. Servant of the Merciful)

Adverbs of manner in Arabic are generally formed by adding an adjective after the expressions such as بطريقة bi-ṭarīqatin ‘in a way’, which helped in collecting 95 adverbs, and بشكل bi-šaklin ‘in a form’ helped in collecting 259 expressions.

(70) بطريقة قانونية
bi-ṭarīqatin qānūniyyah
in-way legal
'legally'

(71) بشكل نهائي
bi-šaklin nihāʾiyy
in-form final
'finally'

Kin terms, such as بن ʾibn 'son-of' and أبو abū 'father-of', also form compound proper nouns. The kin term أبو abū 'father-of' helped in collecting 291 compound names.

(72) أبو مازن
ʾabū māzin
'Abu Mazen'

(73) بن لادن
bin lādin
'Bin Laden'

In some instances it might seem that a grammatical rule can be written to build a compound noun or proper noun so that generalities can be captured. The consequence, however, is that ambiguities will not be avoided. Another advantage of making this list is that the correct equivalent in a target language can easily be provided for translation.

4.4 Handling MWEs

In this section we show how an MWE transducer is built to complement the morphological transducer, and how the MWE transducer interacts with other preprocessing components. We also show how the grammar is responsible for detecting and interpreting syntactically flexible expressions.

4.4.1 Building the MWE Transducer

A specialized two-sided transducer is build for MWEs using a finite state regular expression (Beesley and Karttunen, 2003) to provide correct analysis on the lexical side (upper side) and correct generation on the surface side (lower side). This transducer covers two types of MWEs: fixed and semi-fixed expressions, leaving syntactically-flexible expressions to be handled by the grammar. This entails that the MWE transducer will not handle verbs at all (in the case of Arabic), and will not handle nouns that allow external elements to intervene. In order for the transducer to account for the morphological flexibility of some components, it consults the core morphological transducer (Attia, 2005, Attia, 2006a) to get all available forms of words. We now show how the MWE is enabled to search through the core morphological transducer. First the morphological transducer is loaded and put in a defined variable:

(74) load ArabicTransducer.fst
define AllWords

For the word وزير wazīr 'minister', the transducer has the following upper and lower structures (finite state terminology for input and output).

(75) +noun [وزير (wazīr 'minister')]+masc+sg
وزير (wazīr 'minister')

In order to get all different forms of the word (number and gender variations) we compose the following rule above the finite state network (or transducer):

(76) $[?* "[" {وزير} "]" ?*] .o. AllWords

The sign "$", in finite state notation, means only paths that contain the specified string, and "?*" is a regular expression that means any string of any length. This gives us all surface forms that contain the wanted stem.

4.4.1.1 Arabic Multiword Nouns

Fixed compound nouns are entered in the lexicon as a list of words with spaces. Example (77) shows how the compound noun حفظ الأمن ḥifz al-ʾamn 'peace keeping' is coded in a finite state regular expression.

(77) ["+noun" "+masc" "+def"]:{حفظ} sp {الأمن}

The string "sp" here indicates a separator or space between the two words, so that each word can be identified in case there is need to access it. Compound proper names, including names of persons, places and organizations, are also treated in the same way.

Semi-fixed compound nouns that undergo limited morphological/lexical variations are also entered in the lexicon with the variations explicitly stated. Example (78) shows the expression نزع سلاح nazʿ silāh ‘lit. removing a weapon: disarming’ which can have a variation by prefixing a definite article to the second compound.

(78) ["+noun" "+masc"]:{نزع} sp ("+def":{ال}) {سلاح}

The regular expression in (79) shows an instance of lexical variation. The expression مدعى عليه muddaʿa ʿalaih ‘defendant’, lit. ‘the charged against him’, can have a selection from a fixed set of third person pronouns to indicate the number and gender of the noun.

(79) ["+noun"]:0 ("+def":{ال}) {مدعى} sp {علي}
[“+sg” “+masc”:ه | “+sg” “+fem”:{ها} | “+dual”:{هما}
| “+pl” “+masc”:{هم} | “+pl” “+fem”:{هن}]

As for semi-fixed compound nouns that undergo full morphological variations, a morphological transducer is consulted to get all possible variations.

First we need to explain how Arabic compound nouns are formed and what morphological variations they may have. They are generally formed according to the re-write rule in (80):

(80) NP[_Compound] -> [N N* A*] & ~N

This means that a compound noun can be formed by a noun optionally followed by one or more nouns optionally followed by one or more adjectives. The condition “&~N” is to disallow the possibility of a compound noun being composed of a single noun. In an N_N construction, the first noun is inflected for number and gender, while the second is inflected for definiteness. When the compound noun is indefinite there is no article attached anywhere, but when it is definite, the definite article ال al ‘the’ is attached only to the last noun in the structure. The regular expression in (81) shows how the compound وزير الخارجية wazīr al-ḫariğiyyah ‘foreign minister’ is formatted:

(81) $[?* "[" {وزير} "]" ?*] .o. AllWords sp ("+def":{ال}) {خارجية}

In an N_A structure the noun and adjective are both inflected for number and gender and can take the definite article. The regular expression in (82) shows the format of the expression سيارة مفخخة sayyārah mufaḫḫaḫah ‘lit. trapping car: car bomb’.

(82) $[?* "[" {سيارة} "]" ?*] .o. AllWords sp $[?* "[" {مفخخ} "]" ?*] .o. AllWords

This regular expression, however, is prone to overgenerate allowing for a masculine adjective to modify a feminine noun in contradiction to agreement rules. This is why paths need to be filtered by a set of combinatorial rules (or local grammars). The rules in (83) discard from the network paths that contain conflicting features:

(83) ~$["+dual" <> ["+sg" | "+pl"] /?*]
.o. ~$["+fem" <> "+masc" /?*]

The expression “~$” means ‘does not contain’, “<>” means ‘order is not important’ and “/?*” means ‘ignore noise from any intervening strings’.

After the words are combined correctly, they need to be analyzed correctly as well. First we do not want features to be repeated in the upper language. In the example (84a), the noun sayyārah ‘car’ is analyzed as ‘+fem+sg’, the adjective mufaḫḫaḫah ‘trapping’ repeats the same features ‘+fem+sg’. Second we do not want features to be contradictory. The first word is analyzed as ‘+noun’ while the second is analyzed as ‘+adj’. This is shown by the representation in (84b).

(84) a. سيارة مفخخة
sayyārah mufaḫḫaḫah
car.noun.fem.sg trapping.adj.fem.sg (bomb car)

b. +noun+fem+sgسيارة +adj+fem+sg مفخخة
سيارة مفخخة

Therefore, we need to remove all redundant features from non-head components, and the rules in (85) serve this purpose.

```
(85)     "+sg" -> [] || sp ?* _
         .o. "+fem" -> [] || sp ?* _
         .o. "+adj" -> [] || sp ?* _
         .o. "+noun" -> [] || sp ?* _
```

When these rules are applied to the upper language in the transducer, they remove all specified features from non-initial words, leaving features unique and consistent.

(86)+noun+fem+sgسيارة [sayyārah ‘car’] مفخخة [mufaḫḫaḫah ‘trapping’]
سيارة [sayyārah ‘car’]   مفخخة [mufaḫḫaḫah ‘trapping’]

Special attention, however, is given to cases where some features are drawn from the non-initial nouns like definiteness in (81) above and the features of number and gender in (79).

4.4.1.2 Adjectives, Adverbs and Others

Adjectives are treated to a great extent like semi-fixed expressions, as they can undergo morphological variations, as in (87).

(87) a. قصير النظر
qaṣīr   an-naẓar
short.masc.sg the-sight
‘short-sighted’

b. قصيرات النظر
qaṣīrat   an-naẓar
short.fem.pl   the-sight
‘short-sighted’

Some adverbs have regular forms and can be easily classified and detected. They are usually composed of a preposition, noun and a modifying adjective. The preposition and the noun are relatively fixed while the adjective changes to convey the meaning of the adverb. Examples are given in (88).

(88) a. بشكل جذري
bi-šaklin ǧaḏri
in-form fundamental
‘fundamentally’

b. بطريقة عشوائية
bi-ṭarīqatin ʿašwāʾiyyah
in-way   random
‘randomly’

Some MWEs, however, are less easily classified. They include expressions that function as linking words:

(89) وعلى هذا
wa-ʿalā hadā
and-on this
'whereupon'

They also include a list of highly repetitive complete phrases:

(90) ومما يذكر أن
wa-mi-mma yudkaru ʾanna
and-from-what mention.pass that
'It is mentioned that'

4.4.1.3 One String MWE

Some multiword expressions in Arabic are composed of words with clitics with no visible spaces in between. They look like single words but if they are to be treated by the morphological analyzer alone they will be analyzed compositionally and lose their meaning. Examples are:

(91) a. بالتالي
bi-t-tālī
Meaning: consequently
Compositional meaning: with the following

b. كذلك
kadālika
Meaning: also
Compositional meaning: as that

4.4.2 Interaction with the Tokenizer

The function of a tokenizer is to split a running text into tokens, so that they can be fed into a morphological transducer for processing. The tokenizer is normally responsible for demarcating words, clitics, abbreviated forms, acronyms, and punctuation marks. The output of the tokenizer is a text with a mark after each token; '@' sign in XLE case. Besides, the tokenizer is responsible for treating MWEs in a special way. They should be treated as single tokens with the inner spaces preserved.

One way to allow the tokenizer to handle MWEs is to embed the MWEs in the Tokenizer (Beesley and Karttunen, 2003). Yet a better approach, described by Karttunen et al. (1996), is to develop one or several multiword transducers or “staplers” that are composed on the tokenizer. We will explain here how this is implemented in our solution, where the list of MWEs is extracted from the MWE transducer and composed on the tokenizer. Let’s look at the composition regular expression:

```
(92) 1   singleTokens.i
     2   .o. ?* 0:"[[[" (MweTokens.l) 0:"]]]" ?*
     3   .o. "@" -> " " || "[[[" [Alphabet* | "@"*]  _
     4   .o. "[[[" -> [] .o. "]]]" -> []].i;
```

Single words separated by the ‘@’ sign are defined in the variable *singleTokens* and the MWE transducer is defined in *MweTokens*. In the MWE transducer all spaces in the lower language are replaced by “@” so that the lower language can be matched against *singleTokens*. In line 1 the *singleTokens* is inverted (the upper language is shifted down) by the operator “.i” so that composition goes on the side that contains the relevant strings. From the MWE transducer we take only the lower language (or the surface form) by the operator “.l” in line 2. Single words are searched and if they contain any MWEs, the expressions will (optionally) be enclosed by three brackets on either side. Line 3 replaces all “@” signs with spaces inside MWEs only. The two compositions in line 4 remove the intermediary brackets.

Let’s now show this with a working example. For the phrase in (93), the tokenizer first gives the output in (94). Then after the MWEs are composed with the tokenizer, we obtain the result in (95) with the MWE identified as a single token.

(93) ولوزير خارجيتها
wa-li-wazīr ḫāriǧiyyati-hā
and-to-minister foreign-its
‘and to its foreign minister’

(94) و@ل@وزير@خارجية@ها@
(approx. and@to@foreign@minister@its@)

(95) و@ل@وزير خارجية@ها
(approx. and@to@foreign minister@its@)

4.4.3 Interaction with the White Space Normalizer

Spaces are a crucial element in identifying MWEs. Yet in real-life data, which is prone to errors, spaces may not be regularly and consistently used as expected. There may be two or more spaces, or even tabs, instead of a single space. Moreover, spaces might be added before or after punctuation marks in the wrong manner. The function of the white space normalizer is to go through real-life text and, if they contain mistakes related to the distribution of white spaces, correct (or normalize) these errors, as shown in section 3.2.4.

4.4.4 Interaction with the Grammar

As for fixed and semi-fixed MWEs that are identified both by the tokenizer and the morphological analyzer, they are represented in Lexical Functional Grammar (LFG) as a single word.

(96) a. جنود حفظ الأمن
ǧunūd hifẓi al-ʾamn
soldiers keeping the-peace
'peace keeping soldiers'

b. C-Structure

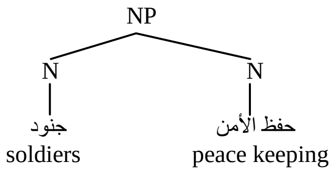


Figure 11. C-structure of an MWE NP

c. F-Structure

PRED 'جنود[soldiers]'
MOD [PRED 'حفظ الأمن[peace keeping]'
DEF +, GEND masc, NUM sg, PERS 3]
DEF +, GEND masc, NUM pl, PERS 3

Figure 12. F-structure of an MWE NP

When MWEs are syntactically flexible, by either allowing reordering such as passivization or allowing intervening elements such as phrasal verbs, they are handled by the syntactic parser. As passivization in Arabic is not made by configurational restructuring of the sentence, but rather by morphological inflection of verbs, we can say that Arabic shows only one instance of syntactic flexibility in MWEs, that is allowing intervening elements.

Syntactically flexible MWEs are handled through lexical rules where one word selects another word or preposition, and that word's semantic value is determined by the selected element. We will show how this is accommodated in LFG in two examples: adjective noun constructions and prepositional verbs.

When a noun is modified by an adjective it usually allows for genitive nouns or pronouns to come in between, even if the expression is highly non-compositional, as shown by the examples in (97):

(97) a. دراجة نارية
darrāǧah nāriyyah
bike fiery
'motorbike'

b. هذه دراجة الولد الصغير النارية
hāḏihi darrāǧatu al-waladi aṣ-ṣaġīri an-nāriyyati
this bike the-boy the-young the-fiery
'This is the young boy's motorbike.'

c. C-Structure of the NP in sentence (97b)

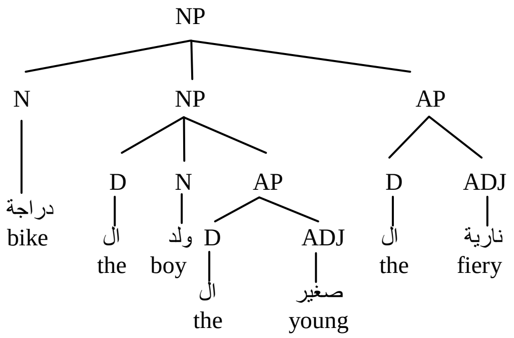


Figure 13. C-structure of an MWE NP

d. F-Structure of the NP in sentence (97b)

```
PRED 'دراجة[bike]'
MOD       [PRED ' ولد[boy]'
           DEF +, GEND masc, NUM sg, PERS 3
           ADJUNCT    [PRED 'صغير [young]'
                        DEF +, GEND masc,
                        NUM sg, PERS 3]
           DEF +, GEND masc, NUM sg, PERS 3]
ADJUNCT {[PRED ' ناري[fiery]'
          DEF +, GEND fem, NUM sg, PERS 3]}
DEF +, GEND fem, NUM sg, PERS 3,
TRANS motorbike
```

Figure 14. F-structure of an MWE NP

This is done by allowing the lexical entry of the noun to select its modifier, as shown by the following lexical rule:

(98) دراجة [bike]   N {(^ PRED='دراجة [bike]'
(^ ADJUNCT PRED)=c 'ناري [fiery]' (^ TRANS)=motorbike
| (^ PRED='bike' (^ ADJUNCT PRED)~= 'ناري [fiery]' (^ TRANS)=bike}.

This means that the translation, or the semantic value, of the noun changes according to the value of the adjunct, or the adjectival modifier.

Prepositional verbs in Arabic allow for subject to intervene between verb and object. This is why they need to be handled in the syntax.

(99) a. اعتمد الولد على البنت
ʾiʿtamada al-waladu ʿalā al-binti
relied the-boy on the-girl
'The boy relied on the girl.'

b. C-Structure

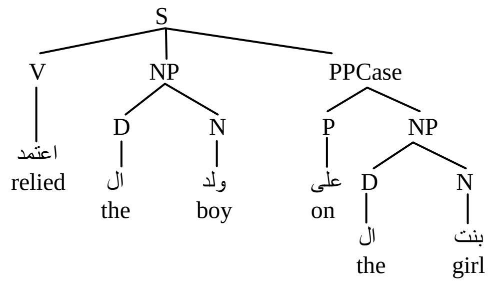


Figure 15. C-structure of an MWE NP

c. F-Structure

```
PRED  'اعتمد[rely]<(^ SUBJ)(^ OBJ)'
SUBJ        [PRED 'ولد[boy]'
             SPEC [DET [DET-TYPE def]]
             DEF +, GEND masc, NUM sg, PERS 3
 OBJ         PRED 'بنت[girl]'
             SPEC [DET [DET-TYPE def]]
             DEF +, GEND fem, NUM sg, PERS 3,
             PFORM على[on]
```

Figure 16. F-structure of an MWE NP

In the c-structure the prepositional verb looks like a verb followed by a PP. In the f-structure, however, the PP functions as the object of the verb. The semantic value, or PRED, of the preposition is removed. The preposition functions only as a case assigner and a feature marker to the main object, but it does not subcategorize for an object itself, as shown in (100).

(100) على [on]   P (^ PFORM)=على [on] (^ PCASE)=gen.

The lexical entry of the verb (101) states that the verb subcategorizes for an object with a certain value for the PFORM feature. This means that the object must be preceded by a specified preposition.

(101)   اعتمد [rely]   V (^ PRED)='اعتمد [rely]<(^ SUBJ)
                                   (^ OBJ)>' (^ OBJ PFORM)=c على [on].

We conclude that in order to accommodate MWEs there is no alternative to integrating them in the processing and preprocessing stages. MWEs are too significant to ignore in any viable linguistic analysis. When MWEs are properly dealt with, they reduce parse ambiguities and give a noticeable degree of certitude to the analysis and machine translation output.

5 Arabic Sentence Structure in LFG

In this chapter we try to formulate a description of main syntactic structures in Arabic within the LFG framework. The challenge is that a complete description of Arabic is not yet available, let alone in the domain of LFG. Therefore, in some instances we provide solutions, while in other instances we pose open questions that need further research and investigation.

We start this chapter with an account of the main characteristics of the Arabic language. Then we move on to describe the main clausal architecture and sentence types in Arabic, and how they can be accounted for in LFG. In the subsequent section we investigate agreement in Arabic, and show how Arabic is a language with alternate agreement and how agreement is best accounted for within the phrase structure rules. Then we explore functional control and long-distance dependencies in Arabic, and show how agreement and resumptive pronouns are used to mark the relation between the position of the filler and the position of the gap. Finally we provide a detailed investigation of the approaches to analysing copula constructions in LFG and argue for the need of a unified representation of the universal predicational construction.

5.1 About Arabic

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

1. Arabic is syntactically flexible. It has a relatively free word order: the orders SVO, VSO, VOS are all acceptable sentence structures. Daimi (2001) also emphasised that Arabic shows a high syntactical flexibility, such as the omission of some arguments associated with verbs, the sharpness of pronominalization phenomena where the pronouns usually indicate the original positions of words before their fronting or omission (as in the case of pro-drop and resumptive pronouns), and in many cases an agent noun can function in place of a verb.

2. Beside the regular sentence structure of verb, subject and object, Arabic has an equational sentence structure of a subject phrase and a predicate phrase, without a verb or copula.
3. Arabic is a highly inflectional language, the matter that makes Arabic morphological analysis complicated. Arabic words are built from roots rather than stems. The morphological complexity and ambiguity directly influences the level of syntactic ambiguity.
4. Arabic writing involves diacritization, which is largely ignored in modern texts, the matter that makes morphological analysis yet more difficult. Chalabi (2000) even claims that the absence of diacritization in Arabic poses a computational complexity “one order of magnitude bigger than handling Latin-based language counterparts”. In our implementation, the loss of diacritics in Arabic is treated as spelling ambiguity, i.e. words with the same spelling but different pronunciations and different meanings. This phenomenon is present in other language with a greater or lesser magnitude. English has a limited number of homographs (e.g., *wound*, *bass* or *wind*), but still English has a lexical ambiguity problem with a comparable impact. MacDonald et al. (1994) claimed that almost all words in the English lexicon exhibit a nonzero degree of ambiguity of one sort or the other.
5. Arabic is a clitic language. Clitics are morphemes that have the syntactic characteristics of a word but are morphologically bound to other words (Crystal, 1980). 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 the one word sentence in (102a) contains a complete syntactic structure as shown in (102b).

(102) a. أعطيتمونيها
ʾaʿṭaitumūnīhā

b. ʾaʿṭaitum ū nī hā
gave.pl you.pl me it
‘You gave it to me.’

6. Written Arabic is also characterised by the inconsistent and irregular use of punctuation marks. Punctuation marks have been introduced fairly recently into the Arabic writing system, yet they are not as essential to meaning or as strictly observed as is the case with English. Arabic writers shift between ideas using resumptive particles and subordinating conjunctions instead of punctuation marks. In MSA, however, due to the influence of translation which, to some extent, transfers punctuation marks from the source languages, and due to the tendency of modern writers to use punctuation marks more consistently, Arabic has come to see more punctuation. In our corpus we found that the period is a convenient criterion for demarcating the sentence boundary, as it is used as expected most of the time. Yet, even in modern writing it is still hard to rely on the period alone to demarcate the sentence boundary. In our corpus of 209,949 sentences of news articles, 14,218 sentences (7%) exceed 40 words in length. The longest sentence reached 144 words. In domains other than the news we even found that the longest sentence reached 803 words. By looking closely at these sentences we found that commas and resumptive particles are consistently used instead of periods. In accounting for this fact, Daimi (2001) remarked that Arabic is distinguished by its high context sensitivity with the desire to exhibit the different synthetic coherence relations. He also noted that Arabic sentences are usually embedded or connected by copulatives, exceptives (particles that denote exception), resumptives and adversative particles. This is why it is difficult to identify the end of an Arabic sentence.
7. Arabic is a pro-drop language. The subject in the sentence can be omitted leaving any syntactic parser with the challenge to decide whether or not there is an omitted pronoun in the subject position.
8. There is no agreed upon and complete formal description of Arabic available yet (Daimi, 2001). Many aspects of Arabic are not investigated satisfactorily, such as topicalization, agreement, and long-distance dependencies. There is even no agreement among researchers on the basic sentence structure in Arabic.

5.2 Arabic Basic Sentence Structure

Arabic has intricate, complex and multi-faceted syntactic structures which led researchers to propose differing representations. The examples (103)–(106) are instances of the basic clausal structures, yet with no agreed representation.

(103) الشمس مشرقة (Verbless copula sentence)
aš-šamsu mušriqatun
the-sun.sg.fem bright.sg.fem
'The sun is bright.'

(104) كان الرجل كريما (Copula sentence with an explicit copula verb)
kāna ar-rağulu karīman
was the-man.sg.masc generous.sg.masc
'The man was generous.'

(105) أكل الولد التفاحة (VSO sentence)
ʾakala al-waladu at-tuffāḥata
ate the-boy.nom the-apple.acc
'The boy ate the apple.'

(106) الولد أكل التفاحة (SVO sentence)
al-waladu ʾakala at-tuffāḥata
the-boy.nom ate the-apple.acc
'The boy ate the apple.'

There is a long history of attempts to describe Arabic syntactic structures. Wright (1896/2005) pointed out that a *nominal sentence* according to the Arab grammarians is one which begins with the subject, whether the predicate is another noun, a prepositional phrase or a verbal predicate. A *verbal sentence* on the other hand is one which starts with a verb.

Cantarino (1974) divided the Arabic sentence into a nominal sentence in which only nominal elements are used as constituents and a verbal sentence which includes a verb as a constituent.

Ryding (2005) and Buckley (2004) classified Arabic sentences into equational (or verbless) sentences, and verbal sentences (those containing a verb).

In the transformational-generative traditions the focus was on whether the original word order in Arabic is VSO or SVO (Anshen and Schreiber, 1968, Fehri, 1993). Verbless sentences were also considered as derived constructions

(Fehri, 1993). However, within LFG we do not have to concern ourselves with this issue, as there are no assumptions about underlying structures. All we need is to provide an adequate description for the c-structure and f-structure of all possible sentence constructions.

Ditters (2001) based his description of the sentence structure in MSA on the distinction between nominal and verbal sentences in the traditional sense that a verbal sentence is one which starts with a verb, while a nominal sentence is one which starts with a noun phrase (NP). If a sentence starts with an NP, the initial NP fulfils a topic function, while the comment function is fulfilled by another NP, an adjective phrase (ADJP), adverb phrase (ADVP), prepositional phrase (PP), or verb phrase (VP).

Badawi et al. (2004) divided the Arabic kernel sentences into three types. The first type is equational sentences, which consist of subject and predicate only, and contain no verbal copula or any other verbal elements. The second type is verbal sentences, which consist of a verb, always in the first position with the agent usually in the second position and the other complements usually in the third position. The third type is the topic–comment structure. In this sort of structure the topic is an NP in the initial position and the comment is an entire clause (either an equational or verbal sentence, or another topic–comment sentence) anaphorically linked to the topic.

Many researchers, as shown above, considered a predicational sentence “equational” if no copula verb is used, and “verbal” if a copula verb is used. This approach fails to properly account for the copula constructions which are composed of subject and predicate whether the copula is overt or non-overt. Marshad and Suleiman (1991) avoided this pitfall and considered that equational sentences are those following the structure of subject and predicate, whether a copula verb is contained or not, as the copula verb in these constructions is semantically vacuous.

It is quite peculiar that while traditional Arabic grammarians agreed on a way to classify sentences in their language, it is hard to find any sort of unanimity in the

Western academia regarding the classification of Arabic clausal structures. Ryding (2005) has successfully identified the source of this divergence in that the criteria of the classification are different in the Western enquiries from those applied in the Arabic indigenous thought. She clarifies that traditional Arabic grammarians divide the sentences into nominal and verbal depending on the nature of the first word in the sentence. If the first word is a noun, the sentence is nominal, and if it is a verb, the sentence is verbal. Ryding goes on to explain that in the West, however, researchers adopted a different criterion: the “distinction is based on whether or not the sentence contains a verb.” If the sentence contains a verb, it is verbal, and if it does not contain a verb, it is equational.

We believe that both criteria are valid and both are required for sound analysis of the Arabic sentence. On the one hand we need to know the constituent structure of the sentence, and on the other hand we need to know whether the sentence begins with a noun or a verb. Both criteria have their application in the grammar writing. In our grammar we found that both views are indispensable. The first view is useful in describing the sentence phrasal hierarchical construction, i.e. what elements are used in the composition of a sentence. The second view is useful in understanding sentential contextual constraints regarding what type of sentence is allowed after complementizers and discourse markers. For example, a nominal sentence in the traditional sense (a sentence starting with an NP) is required after the affirmative إن *ʾinna*, the complementizer أنّ *ʾanna* and the subordinating conjunction لكن *lākinna*, while a verbal sentence in the traditional sense (a sentence starting with a verb) is required after the complementizer أنْ *ʾan*.

To avoid terminological confusion we will reserve the terms *nominal* and *verbal* for the traditional senses, and we introduce the terms *equational* and *non-equational* to describe the phrase structure. In our grammar the division into equational and non-equational constructions appears in the phrase structure as nodes of the tree, while the division into verbal and nominal sentences appears in the f-structure as a feature–value matrix, as will be expounded in the following subsections.

Broadly speaking, we divide the sentences into equational and non-equational. The non-equational sentences are subdivided into VSO, SVO, and VOS, and they are verbal when the verb occurs initially and nominal when an NP occurs initially. Equational sentences are copula constructions and they can be verbal if a copula occurs in the initial position, otherwise they are nominal. Our classification of Arabic sentences into equational and non-equational construction is useful in outlining the constituent structure of the Arabic sentences, while the internal division into nominal and verbal clauses is crucial in accounting for subordination and embedding. While some complementizers require nominal sentences, others require verbal sentences.

5.2.1 Equational Sentences

An equational sentence consists of two parts: a subject phrase and a predicate phrase. The subject is an NP and the predicate can be an NP, ADJP, ADVP, PP, or Complement Phrase (CP), as shown in the examples (107)–(112). The subject is usually definite and the predicate is usually indefinite and it is the shift from definiteness to indefiniteness that marks the transition from subject to predicate. When the predicate is an adjective or a noun it has to agree with the subject in number and gender.

(107) الشمس مشرقة (ADJP predicate with fem subject)
aš-šamsu mušriqatun
the-sun.sg.fem bright.sg.fem
‘The sun is bright.’

(108) الرجل كريم (ADJP predicate with masc subject)
ar-rağulu karīmun
the-man.sg.masc generous.sg.masc
‘The man is generous.’

(109) الكتاب هنا (ADVP predicate)
al-kitābu hunā
the-book here
‘The book is here.’

(110) أخي طبيب (NP predicate)
ʾaḫ-ī ṭabībun
brother.sg.masc-my doctor.sg.masc
‘My brother is a doctor.’

(111) الرجل في الدار (PP predicate)
ar-rağulu fī ad-dāri
the-man in the-house
'The man is in the house.'

(112) الحقيقة أن الحرب تؤدي إلى الهلاك (CP predicate)
al-ḥaqīqatu ʾanna al-ḥarba tuʾddī ʾilā al-halāk
the-fact that the-war leads to destruction.
'The fact is that war leads to destruction.'

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, as in (113).

(113) في الدار رجل
fī ad-dāri rağulun
in the-house man
'In the house there is a man.'

In Arabic, a copula verb is not used when the sentence is in the present tense. However, the copula must be overtly expressed in the past and future tenses, and in the present when the sentence is negated, as in (114)–(116).

(114) كان الرجل كريما
kāna ar-rağulu karīman
was the-man.sg.masc generous.sg.masc
'The man was generous.'

(115) سيكون التقرير جاهزا
sayakūnu at-taqrīru ğāhizan
will-be the-report.sg.masc ready.sg.masc
'The report will be ready.'

(116) ليس الرجل كريما
laisa ar-rağulu karīman
is-not the-man.sg.masc generous.sg.masc
'The man is not generous.'

In the generative framework Arabic verbless copula constructions are considered as derived from constructions which contain a copula after the application of a copula deletion rule (Marshad and Suleiman, 1991). Within the LFG paradigm, where derivations and empty categories are not allowed, the idea is expressed by a special notation in the phrase structure which assumes a non-overt copula, an empty string of a category symbolized by 'ε' (Dalrymple et al., 2004).

As we explained in the introduction to this chapter, we need two pieces of information to account for the sentence structure in Arabic. First we need to know the constituent structure of the sentence. Second we need to know whether the sentence is initiated by a noun or a verb. When the copula verb is overt it usually takes the initial position in the sentence, in which case the sentence clausal type (*comp-type* in our grammar notation) is *verbal*. If it comes after the subject, the sentence clausal type is *nominal*. The phrase structure rules, with functional annotations, expressing these facts for the equational sentence in LFG notation is stated in Figure 17.

$$\text{S\_Equational} \rightarrow \left\{ \begin{array}{ll} \begin{array}{l}\text{NP}\\ (\uparrow \text{SUBJ})=\downarrow\end{array} & \left\{\begin{array}{l}\text{Copula}\\ \uparrow=\downarrow\\ (\uparrow \text{Comp-Type})=\text{nominal}\\ \varepsilon\\ \uparrow=\downarrow\\ (\uparrow \text{Comp-Type})=\text{nominal}\end{array}\right\} \\ \begin{array}{l}\text{Copula}\\ \uparrow=\downarrow\\ (\uparrow \text{Comp-Type})=\text{verbal}\end{array} & \begin{array}{l}\text{NP}\\ (\uparrow \text{SUBJ})=\downarrow\end{array}\end{array}\right\} \left\{\begin{array}{l}\text{NP}\\ (\uparrow \text{PREDLINK})=\downarrow\\ \text{ADJP}\\ (\uparrow \text{PREDLINK})=\downarrow\\ \text{ADVP}\\ (\uparrow \text{PREDLINK})=\downarrow\\ \text{PP}\\ (\uparrow \text{PREDLINK})=\downarrow\\ \text{CP}\\ (\uparrow \text{PREDLINK})=\downarrow\end{array}\right\}$$

Figure 17. Phrase structure of the equational sentence in Arabic

Japanese has a structure somehow similar to the Arabic verbless sentences. Within ParGram (Butt et al., 2002), the Japanese sentences which are composed of an NP and an adjective, the adjective is considered to be the main predicate of the sentence. If we adopt the Japanese sentence analysis to the Arabic sentence in (117), we will have an f-structure analysis as shown in Figure 18 below.

(117) الشمس مشرقة
aš-šamsu mušriqatun
the-sun.sg.fem bright.sg.fem
'The sun is bright.'

$$\left[\begin{array}{ll} \text{PRED} & \text{`bright<SUBJ >`} \\ \text{GEND fem, CASE nom} & \\ & \\ \text{TNS-ASP} & \left[\begin{array}{ll}\text{TENSE} & \text{pres}\\ \text{MOOD} & \text{indicative}\end{array}\right] \\ & \\ \text{SUBJ} & \left[\begin{array}{ll}\text{PRED} & \text{`sun`}\\ \multicolumn{2}{l}{\text{NUM sg, GEND fem,}}\\ \multicolumn{2}{l}{\text{CASE nom, DEF +}}\end{array}\right]\end{array}\right]$$

Figure 18. A putative f-structure of a nominal Arabic sentence

However, we assume that this analysis is not linguistically motivated for Arabic (although it could be linguistically motivated for Japanese adjectives which bear paradigmatic verbal characteristics). There is no evidence here to support the idea that the adjective is either the main predicate or that it subcategorizes for a subject. Moreover, external governors, as in example (118), can precede the whole structure and assign a different case (accusative case in the example) to the subject. If an external governor can assign case to the subject, this means that the adjective cannot be a main predicate or a case assigner.

(118) إن الرجل كريم
ʾinna ar-rağula karīmun
indeed the-man.acc.sg.masc generous.sg.masc
'The man was generous.'

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, copula sentences in Hebrew, a Semitic language with a structure very similar to that of Arabic, are analysed as mixed category which are categorically nominal and functionally verbal (Falk, 2004). This makes Arabic nominal sentences eligible for an f-structure analysis where a null predicator subcategorizes for SUBJ and PREDLINK. Figure 19 shows the c-structure, and Figure 20 shows the f-structure of the copula sentence in (117) as analysed by our parser. More on the analysis of copula constructions in LFG will follow in section 5.5.

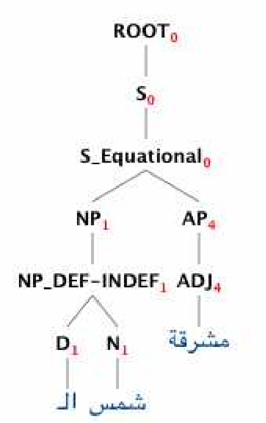
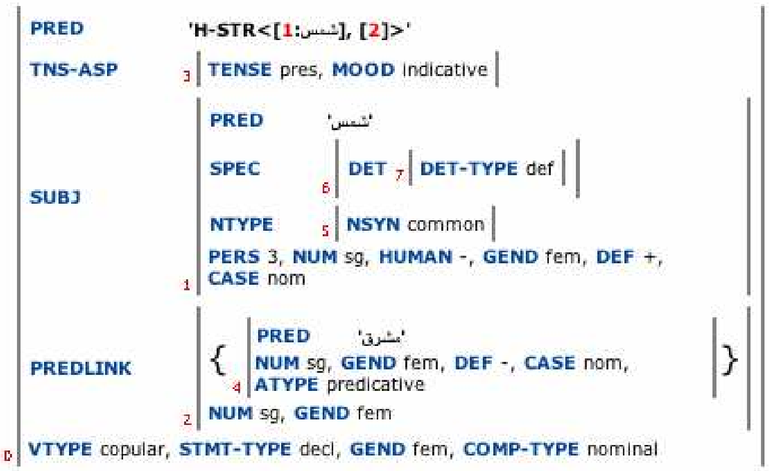


Figure 19. C-structure of a copula Arabic sentence



```
  PRED       'H-STR<[1:شمس], [2]>'
  TNS-ASP   3| TENSE pres, MOOD indicative |
              | PRED       'شمس'
              | SPEC       | DET 7| DET-TYPE def | |
  SUBJ        |           6|
              | NTYPE     5| NSYN common |
              | PERS 3, NUM sg, HUMAN -, GEND fem, DEF +,
             1| CASE nom
              |   | PRED     'مشرق'
  PREDLINK    | { | NUM sg, GEND fem, DEF -, CASE nom,    | }
              |  4| ATYPE predicative
             2| NUM sg, GEND fem
0| VTYPE copular, STMT-TYPE decl, GEND fem, COMP-TYPE nominal
```

Figure 20. F-structure of a copula Arabic sentence

5.2.2 Non-Equational Sentences

Non-equational sentences are sentences where a non-copula verb functions as the main predicator in the construction. In Arabic there are generally three accepted word orders: VSO, SVO and VOS, as shown in the examples in (119), (120) and (121) respectively.

(119) أكل الولد التفاحة (VSO sentence)
ʾakala al-waladu at-tuffāḥata
ate the-boy.nom the-apple.acc
'The boy ate the apple.'

(120) الولد أكل التفاحة (SVO sentence)
al-waladu ʾakala at-tuffāḥata
the-boy.nom ate the-apple.acc
'The boy ate the apple.'

(121) أكل التفاحة الولد (VOS sentence)
ʾakala at-tuffāḥata al-waladu
ate the-apple.acc the-boy.nom
'The boy ate the apple.'

As we explained in the introduction to this chapter, we need to account for two pieces of information: first we need to know the phrasal structure of the sentence, and second we need to know whether the sentence starts with a noun or a verb. When the verb is initial the sentence clausal type (expressed as *comp-type* in the grammar notation) is *verbal*, otherwise it is *nominal*. The classification into nominal and verbal clausal types is helpful as some complementizers and focus markers select nominal sentences while others select verbal sentences. The phrase structure rule expressing these facts for the non-equational sentence in LFG notation is stated in Figure 21.

$$
\begin{array}{llll}
\text{S\_Non-Equational} & \rightarrow & \left\{ \begin{array}{c} \text{VSO} \\ \text{SVO} \end{array} \right\} & \\
\text{VSO} & \rightarrow & \begin{array}{l} \text{V} \\ \uparrow=\downarrow \\ (\uparrow \text{Comp-Type})=\text{verbal} \end{array} \quad \begin{array}{l} \text{NP} \\ (\uparrow \text{SUBJ})=\downarrow \end{array} & \left(\begin{array}{l} \text{NP} \\ (\uparrow \text{OBJ})=\downarrow \end{array} \right) \\
\text{SVO} & \rightarrow & \begin{array}{l} \text{NP} \\ (\uparrow \text{SUBJ})=\downarrow \end{array} \quad \begin{array}{l} \text{V} \\ \uparrow=\downarrow \\ (\uparrow \text{Comp-} \\ \text{Type})=\text{nominal} \end{array} & \left(\begin{array}{l} \text{NP} \\ (\uparrow \text{OBJ})=\downarrow \end{array} \right)
\end{array}
$$

Figure 21. Phrase structure of the non-equational sentence in Arabic

For sentence (119) above, the phrase structures rule will yield the parse tree in Figure 22 and the f-structure in Figure 23.

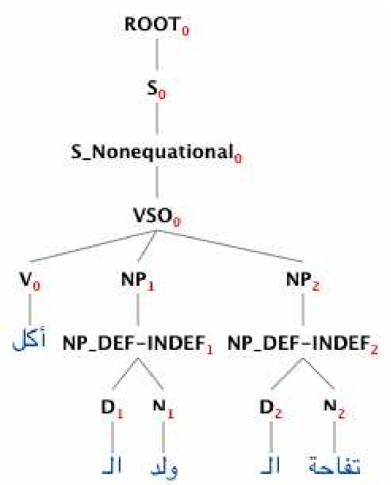
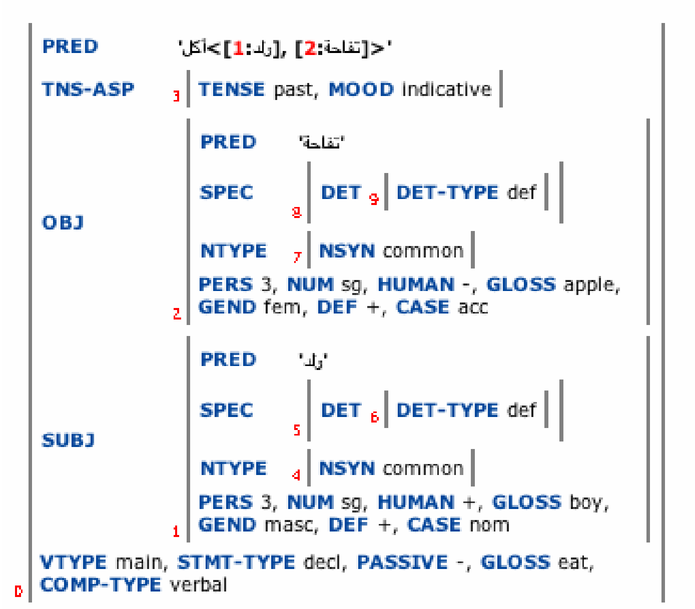


Figure 22. C-structure of a VSO Arabic sentence



```
  PRED        'أكل<[1:ولد], [2:تفاحة]>'
  TNS-ASP   3 TENSE past, MOOD indicative
                PRED     'تفاحة'
                SPEC   8 DET 9 DET-TYPE def
  OBJ           NTYPE  7 NSYN common
                PERS 3, NUM sg, HUMAN -, GLOSS apple,
            2   GEND fem, DEF +, CASE acc
                PRED     'ولد'
                SPEC   5 DET 6 DET-TYPE def
  SUBJ          NTYPE  4 NSYN common
                PERS 3, NUM sg, HUMAN +, GLOSS boy,
            1   GEND masc, DEF +, CASE nom
  VTYPE main, STMT-TYPE decl, PASSIVE -, GLOSS eat,
0 COMP-TYPE verbal
```

Figure 23. F-structure of a VSO Arabic sentence

There is evidence, however, to indicate that the use of the VOS word order is restricted in Modern Standard Arabic. The structure is possible in limited cases, but in our grammar we accommodated only one possibility of the VOS structure, that is when the object is a pronominal suffix, as in (122).

(122) شكرهم الولد
šakara-hum al-waladu
thanked-them the-boy
'The boy thanked them.'

Moreover, in SVO word order, there is another different possible analysis, that is to consider the subject as the topic phrase and the rest of the sentence as the comment phrase in which case the subject of the verb is an elliptic pronoun that refers back to the subject. This analysis accounts for the fact that when the subject comes initially the verb must agree in number, gender and person; but when the subject follows the verb the verb agrees with the subject in gender and person only. More details on the discussions on this construction will be provided in section 5.2.3.

5.2.2.1 Pro-Drop in Arabic

Arabic is a pro-drop language. The pro-drop theory (Baptista, 1995, Chomsky, 1981) stipulates that a null category (*pro*) 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.

In Arabic the subject can be explicitly stated as an NP or implicitly understood as a pro-drop. Arabic has a rich agreement morphology. Arabic verbs conjugate for number, gender and person, which enables the reconstruction of the missing subject. This is shown by example (123) which has the c-structure in Figure 24 and the f-structure in Figure 25.

(123) يأكلون التفاح
ya'kulūna at-tuffāḥa
eat.pl.masc the-apples
'They eat the apples.'

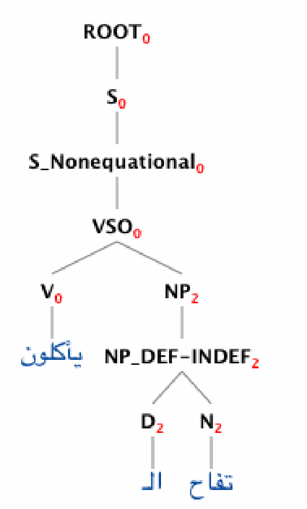
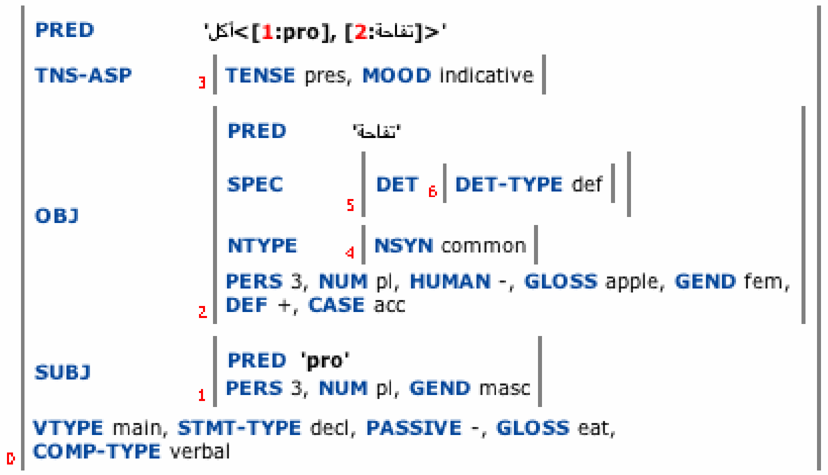


Figure 24. C-structure of a pro-drop sentence



```
PRED        'أكل<[1:pro], [2:تفاحة]>'
TNS-ASP    3 [ TENSE pres, MOOD indicative ]
              [ PRED    'تفاحة'
              | SPEC   5 [ DET 6 [ DET-TYPE def ] ]
OBJ           | NTYPE  4 [ NSYN common ]
              | PERS 3, NUM pl, HUMAN -, GLOSS apple, GEND fem,
           2  [ DEF +, CASE acc
              [ PRED 'pro'
SUBJ       1  [ PERS 3, NUM pl, GEND masc ]
  VTYPE main, STMT-TYPE decl, PASSIVE -, GLOSS eat,
0 COMP-TYPE verbal
```

Figure 25. F-structure of a pro-drop sentence

According to Hoyt (2004), Arabic is a pro-drop language, in the sense that the agreement morphemes on the verbs can be interpreted as pronouns, and that person and number agreement forms are specifically pronominal in this regard. He also pointed out that these pronominal morphemes can only appear if the subject is non-overt or if it occurs in the pre-verbal position. If the subject occurs in the post-verbal position, the verb matches only a subset of the subject's agreement features, and the agreement morphemes in this case are not pronominal.

Chalabi (2004b) maintained that 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. The second challenge, after deciding that there is a null pronoun in the subject position, is to resolve the pronoun reference.

The challenge to decide whether there is a pro-drop or not comes from the fact that many verbs in Arabic can be both transitive and intransitive. In case these verbs are followed by only one NP the ambiguity arises, as in (124).

(124) أكلت الدجاجة
ʾakalat ad-dağāğah
ate.fem the-chicken

In (124) 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 meaning *she* and understood by the feminine mark on the verb (in which case the meaning will be ‘she ate the chicken’). This ambiguity is caused by two facts: first there a possibility for a pro-drop subject following Arabic verbs, second the verb *akala* ‘eat’ can be both transitive and intransitive. This ambiguity results in two f-structures as shown in Figure 26 and Figure 27. In the pro-drop case, person, number and gender morphosyntactic features on the verb are used to reconstruct the number, gender and person features for the “pro” subject.

$$\left[\begin{array}{ll}
\text{PRED} & \text{`eat<(}\uparrow\text{ SUBJ) (}\uparrow\text{OBJ)>`} \\
\text{STMT-TYPE} & \text{declarative} \\
\text{TNS-ASP} & \left[\begin{array}{ll} \text{TENSE} & \text{past} \\ \text{MOOD} & \text{indicative} \end{array}\right] \\
\text{SUBJ} & \left[\begin{array}{l} \text{PRED} \quad \text{`pro`} \\ \text{NUM sg, GEND fem} \end{array}\right] \\
\text{OBJ} & \left[\begin{array}{l} \text{PRED} \quad \text{`chicken`} \\ \text{NUM sg, GEND fem, CASE acc} \end{array}\right]
\end{array}\right]$$

Figure 26. F-structure with a pro-drop

$$\left[\begin{array}{ll}
\text{PRED} & \text{`eat<(}\uparrow\text{ SUBJ) >`} \\
\text{STMT-TYPE} & \text{declarative} \\
\text{TNS-ASP} & \left[\begin{array}{ll} \text{TENSE} & \text{past} \\ \text{MOOD} & \text{indicative} \end{array}\right] \\
\text{SUBJ} & \left[\begin{array}{l} \text{PRED} \quad \text{`chicken`} \\ \text{NUM sg, GEND fem, CASE nom} \end{array}\right]
\end{array}\right]$$

Figure 27. F-structure with no pro-drop

The second challenge, after deciding that there is a null pronoun in the subject position, is to recover the pronoun. In the example in (125) the verb is transliterated without any vowels. Examples (126a–d) show possible vowelization of the verb indicating different possible pronouns. The syntactic ambiguity arises from the morphological ambiguity where a single suffix in the verb can have multiple pronominal interpretations.

(125) ذهبت إلى الحديقة
ḏhbt ʾilā al-ḥadīqati
went to the-garden

(126) a. ḏahabat ʾilā al-ḥadīqati
went.sg.fem to the-garden
‘She/it went to the garden.’

b. ḏahabtu ʾilā al-ḥadīqati
went.sg.1 to the-garden
‘I went to the garden.’

c. ḏahabta ʾilā al-ḥadīqati
went.sg.masc.2 to the-garden
‘You went to the garden.’

d. ḏahabti ʾilā al-ḥadīqati
went.sg.fem.2 to the-garden
‘You went to the garden.’

5.2.3 On Topic–Comment Constructions

In our grammar we analyze the initial NP which is followed by a verb as a subject. For the sentence in (127) our parser will output the parse tree in Figure 28 and the f-structure in Figure 29.

(127) الولد أكل التفاحة (SVO sentence)
al-waladu ʾakala at-tuffāḥata
the-boy.nom ate the-apple.acc
‘The boy ate the apple.’

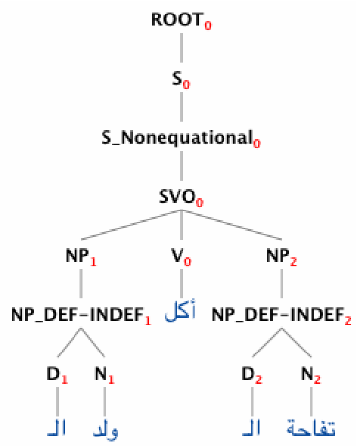
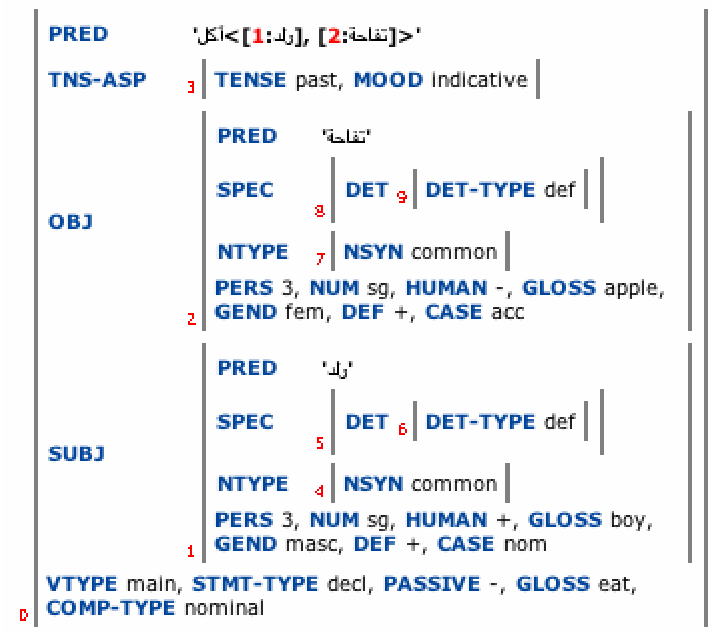


Figure 28. C-structure of an SVO Arabic sentence



```
  PRED       'أكل<[1:ولد], [2:تفاحة]>'
  TNS-ASP  3 |TENSE past, MOOD indicative|
             |PRED     'تفاحة'                           |
             |SPEC    |DET 9|DET-TYPE def||              |
             |       8|                   |              |
  OBJ        |NTYPE  7|NSYN common|                      |
             |PERS 3, NUM sg, HUMAN -, GLOSS apple,      |
           2 |GEND fem, DEF +, CASE acc                  |
             |PRED     'ولد'                             |
             |SPEC    |DET 6|DET-TYPE def||              |
             |       5|                   |              |
  SUBJ       |NTYPE  4|NSYN common|                      |
             |PERS 3, NUM sg, HUMAN +, GLOSS boy,        |
           1 |GEND masc, DEF +, CASE nom                 |
  VTYPE main, STMT-TYPE decl, PASSIVE -, GLOSS eat,
0 COMP-TYPE nominal
```

Figure 29. F-structure of an SVO Arabic sentence

However, we found out that this analysis cannot account for all the facts, variations and complexities involved when the NP comes in the initial position of a non-equational construction. This structure has been the subject of a lot of debate in the literature and we would like here to contemplate the relevant issues and arguments and give some thought as to what possible solutions can be implemented within the framework of LFG.

Badawi et al. (2004) championed the idea that beside verbal and equational sentences there is a third type which is the topic–comment structure. In this structure the topic is an NP in the initial position and the comment is an entire clause (either an equational or verbal sentence, or another topic–comment sentence) anaphorically linked to the topic.

Jouitteau and Rezac (2006) assumed that the preverbal subject in Arabic is a topic linked to an empty pronominal in the subject position.

Hoyt (2006) provided a detailed analysis of the Arabic nominal clauses where the NP occupies the initial position and the predicate is a complete verbal clause containing a pronoun which is bound by the initial NP. He divided this type into two further subtypes:

1. **Non-subject initial NP.** The initial NP cannot be interpreted as the subject, but is interpreted as an object (128a), an oblique argument (128b), or possibly as an argument of a more deeply embedded clause (128c) (all examples from Hoyt, 2006). Clauses of this type are analyzed as involving left-dislocation of the initial NP to a position outside of the clause where it fills a discourse role (topic or focus). It is linked to a binding pronoun occupying a position within the predicate sentence.

(128) a. هند سمعها محمد
hindun samiʿa-hā muḥammadun
Hind.fem.sg.nom hear.past.3.masc.sg-her Mohammed.nom
'Hind, Mohammad heard her.'

b. الشارع قابلت سالما فيه
aš-šāriʿ qābaltu sāliman fī-hi
the-street, meet.past.1.sg Salim.acc in-it
'The street, I meet Salim on it.'

c. فاطمة اشتريت كتابها أمس
fāṭimatun ʾištaraitu kitāba-hā ʾamsa
Fatima.nom buy.past.1.sg book.acc-her yesterday
'Fatima, I bought her book yesterday.'

2. **Subject initial NP.** The initial NP can be interpreted as the subject of the verb and the verb carries agreement morphology with the pre-verbal NP, as shown in example (129).

(129) الأولاد لعبوا كرة القدم (Hoyt, 2006)
al-ʾawlādu laʿibū kurata al-qadami
the-boys.masc.pl.nom play.past.3.masc.pl ball the-foot
'The boys played football.'

Hoyt (2006) explained that there are two approaches to analyzing this construction. One approach considers the initial NP as left dislocated from the subject position. This NP is now occupying the external position of a topic. This approach accounts for two facts. The first fact is that initial subjects control full agreement on the verb, while a post-verbal subject controls only gender agreement. The second fact is that the initial NP must precede a fronted element such as question-words, as in (130).

(130) الطلاب متى ذهبوا إلى العراق (Hoyt, 2006)
aṭ-ṭullābu matā ḏahabū ʾilā al-ʿirāqi?
the-students.masc.pl.nom when go.past.3.pl.masc to the-Iraq
'The students, when did they go to Iraq?'

> The second approach assumes that the initial NP is a pre-verbal subject, because the dependency between the initial NP and the pronominal agreement on the verb is more local than are the dependencies between non-subject initial NPs and their binding pronouns.

As a further evidence that the initial NP in SVO sentences is not a true subject, Hoyt (2004) made an interesting comparison between SVO agreement and anaphoric agreement. He concluded that agreement marking patterns in the SVO word orders are identical to patterns of agreement between anaphoric pronouns and their antecedents. An anaphoric pronoun agrees with its antecedent in gender, person, and number, and if the antecedent is a conjoined NP, the same person, gender and number resolution rules apply as in the agreement between verbs and pre-verbal conjoined subjects. Hoyt (2004) deduced from this that the dependency between subject and verb in SVO word order is a semantic dependency.

Suleiman (1989) emphasised that the preferred word order in Arabic is VSO. But when either the subject or the object is placed before the verb, these shifts in word order are semantically marked and motivated by a desire to express additional meanings. He assumed that the purpose of preposing is emphasizing the fronted element and giving it more weight.

The literature above shows three different tendencies in dealing with the initial NP in SVO sentences: the first is treating it as a dislocated (topicalized) element that has been fronted for semantic considerations. The second is to treat it as a preverbal subject. The third is to treat it as part of basic clausal construction involving a TOPIC and a COMMENT. The TOPIC here is not considered a discourse function, but a primitive grammatical function. Our preference is to consider the initial NP as part of a basic grammatical construction (topic–comment construction), albeit semantic and pragmatic considerations may be involved. In this instance we can say that the semantics of the language penetrates the syntactic structure. Hoyt (2006) emphasised that Arabic is a *discourse configurational* language, in the sense that the initial subject functions to encode discourse relations in addition to thematic relations.

The solution we choose is to consider the initial NP as TOPIC and the following sentence as COMMENT. The problem with this option is that TOPIC and COMMENT are not recognized as governable grammatical functions in the LFG literature, which makes the implementation of a topic–comment construction a non-standard analysis in LFG.

Rosén (1996) brought up the issue early in the LFG paradigm and explained that the topic–comment construction is an important sentence type in languages such as Japanese, Mandarin, and Vietnamese. She explained that topics in these languages differ from the English topics in that a topic does not necessarily correspond to a gap. She analysed three types of topic–comment constructions in Vietnamese and argued that a uniform analysis for all three constructions may be achieved by using both TOPIC and COMMENT functions in the f-structure.

The gap in Arabic topics is mostly filled by a pronoun, which justifies that the TOPIC and COMMENT functions will provide a plausible representation. This proposed solution will account for the difference between two constructions. The first construction, as shown in example (131), is a true topicalization where the object is fronted and the accusative case marking is still preserved and there is an unfilled gap in the sentence. The f-structure for this construction is provided in Figure 30. The second type is a topic–comment construction, as shown in example (132), where the fronted noun is now in the nominative case and the gap is filled by a pronoun. The f-structure for this construction is provided in Figure 31. In (131) the topic functionally controls the gap, while in (132) the topic is anaphorically linked to the pronoun which fills the object position. The TOPIC in (131) is a discourse function while in (132) it is a primitive grammatical function.

(131) التفاحة أكل الولد
at-tuffāḥata ʾakala al-waladu
the-apple.acc ate the-boy
‘The apple, the boy ate.’

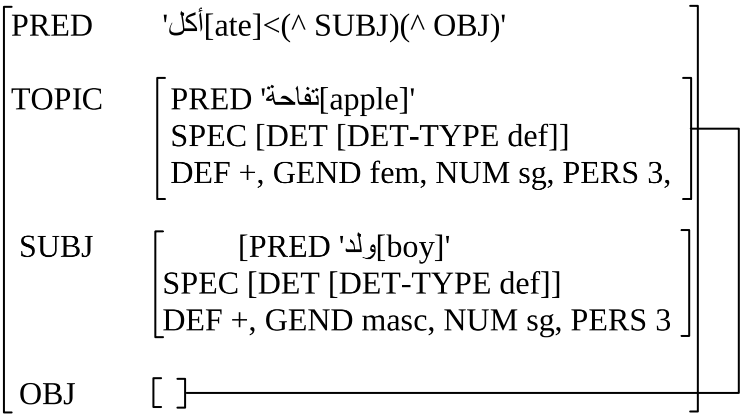

PRED 'أكل[ate]<(^ SUBJ)(^ OBJ)'
TOPIC
PRED 'تفاحة[apple]'
SPEC [DET [DET-TYPE def]]
DEF +, GEND fem, NUM sg, PERS 3,
SUBJ
[PRED 'ولد[boy]'
SPEC [DET [DET-TYPE def]]
DEF +, GEND masc, NUM sg, PERS 3
OBJ
[ ]

Figure 30. F-structure of TOPIC as a discourse function

(132) التفاحة أكلها الولد
at-tuffāḥatu ʼakala-ha al-waladu
the-apple.nom ate-it the-boy
‘The apple, the boy ate it.’

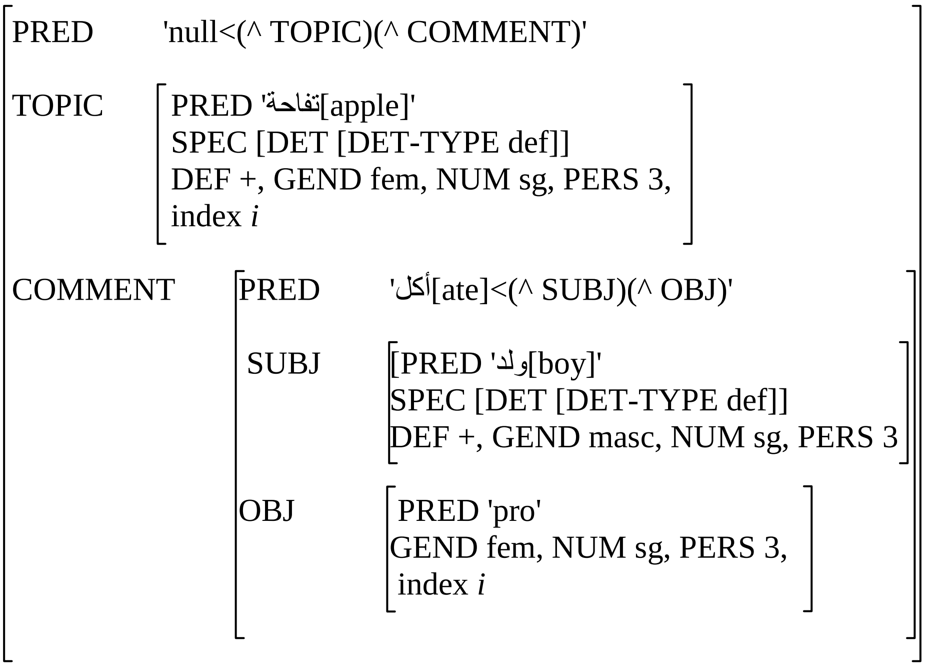

PRED 'null<(^ TOPIC)(^ COMMENT)'
TOPIC
PRED 'تفاحة[apple]'
SPEC [DET [DET-TYPE def]]
DEF +, GEND fem, NUM sg, PERS 3,
index i
COMMENT
PRED 'أكل[ate]<(^ SUBJ)(^ OBJ)'
SUBJ
[PRED 'ولد[boy]'
SPEC [DET [DET-TYPE def]]
DEF +, GEND masc, NUM sg, PERS 3
OBJ
PRED 'pro'
GEND fem, NUM sg, PERS 3,
index i

Figure 31. F-structure of TOPIC as a grammatical function

Another justification for opting for a topic–comment analysis is the frequency and variability of the construction in Arabic. The topic is a noun, and the comment can be an equational sentence, as shown in (133); a non-equational sentence, as shown in (134); or another topic–comment sentence that is anaphorically linked to the topic by a binding pronoun, as shown in (135) and (136). The topic can be linked to a subject position in the comment sentence, as shown in (134); an object position, as shown in (137); an object of oblique, as shown in (138); or another embedded phrase in the sentence, as shown in (139).

(133) الحجرة أثاثها جميل
al-ḥuǧratu ʾaṯaāṯu-hā ǧamīlun
the-room furniture-its beautiful
‘The room, its furniture is beautiful.’

(134) الطالب يقرأ الكتاب
aṭ-ṭālibu yaqraʾu al-kitāb
the-student read the-book
‘The student reads the book.’

(135) أما الطرق الأخرى فكلها يودي إلى الفشل (Badawi et al., 2004)
ʾammā aṭ-ṭuruqu al-ʾuḫrā fakulluhā yūʾaddī ʾilā al-fašali
as-for the-roads the-other all lead to the-failure
‘As for the other roads, all of them, they lead to failure.’

(136) أما وزارة الصحة فمسؤوليتها لا شك فيها (Badawi et al., 2004)
ʾammā wazāratu aṣ-ṣiḥḥati famasʾūliyyatu-hā lā šakka fī-hā
as-for the-ministry the-health responsibility-its no doubt in-it
‘As for the Ministry of Health, its responsibility, there is no doubt about it.’

(137) التفاحة أكلها الولد
at-tuffāḥatu ʾakala-ha al-waladu
the-apple ate-it the-boy
‘The apple, the boy ate it.’

(138) الرجل اعتمدت عليه
ar-raǧulu iʿtamadtu ʿalai-hi
the-man rely.past.1.sg on-him
‘The man, I relied on him.’

(139) التقرير تم إعداده
at-taqrīru tamma ʾiʿdādu-hu
the-report completed preparing-it
‘The report, its preparation has been completed.’

5.3 Agreement

Arabic has rich agreement morphology which allows it to show agreement relations between various elements in the sentence. There are five morphosyntactic features involved in agreement in Arabic: number (singular, dual and plural), gender (feminine and masculine), person (1st, 2nd, and 3rd), case (nominative, accusative and genitive) and definiteness (definite and indefinite). The strongest agreement relation is that between a noun and adjective where four of the five agreement features are involved: number, gender, case and definiteness. Examples (140)–(144) show different type of agreement relationships.

(140) هذا الرجل (noun – demonstrative pronoun: number, gender)
haḏā ar-rağulu
this.sg.masc the-man.sg.masc
‘this man’

(141) رأيت الرجلين الكريمين
(noun – adjective: number, gender, case, definiteness)
raʾaitu ar-rağulaini al-karīmaini
I-saw the-man.dual.acc.def the-generous.dual.masc.acc.def
‘I saw the two generous men.’

(142) الطالبتان اللتان نجحتا (noun – relative pronoun: number, gender, case)
aṭ-ṭālibatāni allatāni nağaḥatā
the-student.dual.fem.nom who.dual.fem.nom succeed.past.dual.fem.3
‘The two students who succeeded’

(143) الطالبات ذاكرن دروسهن (noun – pronoun: person, number, gender)
aṭ-ṭālibātu ḏākarna durūsa-hunna
the-student.pl.fem.3.nom study.past.pl.fem.3 lessons-their.pl.fem.3
‘The students studied their lessons.’

(144) الرجل كريم (subject – predicate: number, gender)
ar-rağulu karīmun
the-man.sg.masc generous.sg.masc
‘The man is generous.’

Regarding verb–subject agreement, when subjects are in the pre-verbal position, verbs have full (rich) agreement as they are required to agree with their subjects in number, gender and person, as in (145).

(145) البنات ذهبن إلى الحديقة
al-banātu ḏahabna ʾilā al-ḥadīqati
the-girl.pl.fem.3 go.past.pl.fem.3 to the-garden
‘The girls went to the garden.’

Contrastively if subjects are in the post-verbal position, verbs show partial (weak or poor) agreement, as verbs agree with their subjects in gender and person only, as in (146). Verbs take the default singular form whether subjects are singular, dual or plural.

(146) ذهبت البنات إلى الحديقة
ḏahabat al-banātu ʾilā al-ḥadīqati
go.past.sg.fem.3 the-girl.pl.fem.3 to the-garden
‘The girls went to the garden.’

The feature of humanness plays an important rule in agreement in Arabic. With non-human plural nouns, verbs are invariably in the singular and feminine, as shown in (147).

(147) القطط تشرب اللبن
al-qiṭaṭu tašrabu al-labana
the-cat.pl.fem.nom.3 drink.sg.fem.3 the-milk
‘The cats drink milk.’

Sometimes in subject–predicate constructions the morphosyntactic agreement is replaced by a semantic agreement. In the example in (148) the subject is plural and the predicate is singular, but they are semantically compatible.

(148) هؤلاء هم السبب في هزيمتنا
hāʾulāʾi humu as-sababu fī hazīmati-nā
These.masc.sg they the-reason.masc.sg in defeat-our
‘These people are the reason behind our defeat.’

Regarding the definition of agreement, Ryding (2005) states that agreement or concord is the feature compatibility between words in a phrase or clause. Agreement is formally defined by Corbett (2001) as “systematic covariance between a semantic or formal property of one element and a formal property of another.” Corbett (2001) used the terms “controller” to refer to the element which determines the agreement, “target” to refer to the element whose form is determined by agreement, and “domain” to refer to the syntactic environment in which agreement occurs.

Corbett (2001) maintained that the relationship in agreement is asymmetrical in general because the target cannot match all the features of the controller. Androutsopoulou (2001, p. 40) provided a formal definition of the principle of asymmetric agreement as:

> In an agreement relation between two elements α and β, where α is the head and β is the specifier, the set of agreeing features of β must be a subset of the set of agreeing features of α.

Platzack (2003) classified languages into “uniform agreement” languages and “alternate agreement” languages. He stated that Standard Arabic is a language with alternate agreement, where the verb shows full agreement in person, gender and number when the subject is in front of it, but partial agreement (only person and gender) when the subject follows the verb.

Corbett (2001) pointed out that a common approach to dealing with agreement is unification, in which agreement is considered as a process of cumulating partial information from both the controller and the target. He gave the French example in (149).

(149) Je suis content /contente
I be.1.sg pleased.sg.masc/ pleased.sg.fem
‘I am pleased’ (man/woman taking)

According to Corbett we have two feature structures: one for the personal pronoun and the verb (150a) and the second for the predicative adjective (150b).

(150) a. $\begin{bmatrix}\text{number: sg}\\ \text{person: 1}\end{bmatrix}$

b. $\begin{bmatrix}\text{number: sg}\\ \text{gender: fem}\end{bmatrix}$

Corbett (2001) considered that these feature structures are compatible and hence can be unified, giving the structure in (151):

(151) $\begin{bmatrix}\text{number: sg}\\ \text{person: 1}\\ \text{gender: fem}\end{bmatrix}$

However, we believe that unification will not be very efficient in accounting for agreement in Arabic. In the Arabic example in (152) the verb is singular and the subject is plural and the unification will fail in this case.

(152) ذهب الأولاد إلى المدرسة
ḏahaba al-ʾawlādu ʾilā al-madrasati
go.past.sg.masc.3 the-boy.pl.masc.3 to the-school.
‘The boys went to school.’

A possible workaround might be to make the singular feature of the verb as a default non-obligatory feature.

ذهب V {(↑ NUM) (↑ NUM) ~= sg
| (↑ NUM)=sg}

This solution will effectively work for (152), yet this will make the feature lose its constraining power, and there will be no way to account for the ungrammaticality of the sentence in (153), where the verb must agree in number with the plural pre-verbal subject. In this example the incompatibility between the subject and the verb will go undetected. This shows that in Arabic agreement cannot be specified satisfactorily through unification.

(153) * الأولاد ذهب إلى المدرسة
* al-ʾawlādu ḏahaba ʾilā al-madrasati
the-boy.pl.masc.3 go.past.sg.masc.3 to the-school.
‘The boys went to school.’

Arabic verb–subject agreement has a complex system of variability which cannot be modelled in terms of unification or constraints. Arabic is a language with alternate agreement. In VSO word order the verb agrees with the subject in gender and person, and is invariably in the singular, whether the subject is singular, dual or plural. In SVO word order the verb must agree with the subject NP in gender, number and person.

Within the LFG-XLE framework, Hoyt (2004) described a grammar for modelling the morphosyntax of verbal agreement in Modern Standard Arabic. Hoyt (2004) showed that the variability of subject–verb agreement in Arabic poses a problem for a unification-based approach. Therefore he proposed the projection of a semantic layer represented as s-structure which interacts with the f-structure to control the agreement features.

Here we propose that an additional layer is not necessary to represent the agreement features in Arabic and that they can be handled within the two basic representations: c-structures and f-structures. Agreement in Arabic is determined by word order and this is why we think that agreement must be specified by the phrase structure rules. Initially, the agreement features of the verbs can be temporarily stored in an independent structure. Later the relationship between the subject and the verb is resolved through functional equations on the phrase structure according to the position of the subject to the verb, i.e. whether it precedes or follows the subject.

To show how this solution is implemented, lets first look at the two examples in (154) and (155) where the verb is singular in one instance and plural in the other.

(154) لعب الأولاد
la'iba al-'awlādu
play.past.sg.masc the-boy.pl.masc.3
'The boys played.'

(155) الأولاد لعبوا
al-'awlādu la'ibū
the-boy.pl.masc play.past.pl.masc.3
'The boys played.'

To start with, we make the lexical entry of the verb لعب la'ib 'play' specify the features of a temporary f-structure AGR, rather than the features of SUBJ. Within AGR, the verb stores the values for number, gender and person.

لعب V (↑ PRED)='لعب'
(↑ AGR NUM)=sg
(↑ AGR GEND)=masc
(↑ AGR PERS)=3

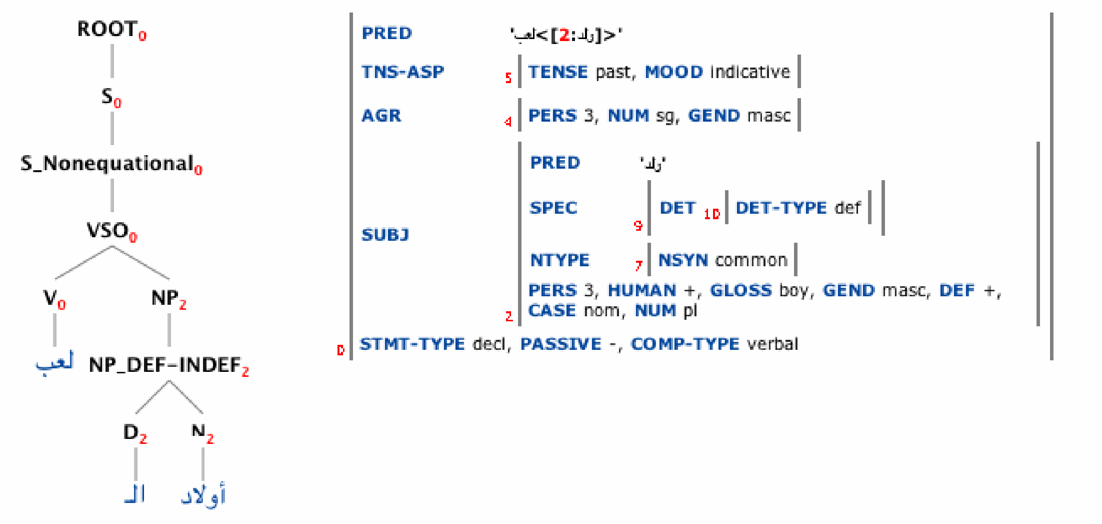
لعبوا V (↑ PRED)='لعب'
(↑ AGR NUM)=pl
(↑ AGR GEND)=masc
(↑ AGR PERS)=3

Then, functional equations are inserted in the phrase structure rules to select which features are relevant in agreement according to the position of the subject in relation to the verb.

SV →	NP	V
	(↑ SUBJ)=↓	↑=↓
		(↑AGR GEND)=(↑ SUBJ GEND)
		(↑AGR NUM)=(↑ SUBJ NUM)
		(↑AGR PERS)=(↑ SUBJ PERS)

VS →	V	NP
	↑=↓	(↑ SUBJ)=↓
	(↑AGR NUM)=sg	
	(↑AGR GEND)=(↑ SUBJ GEND)	
	(↑AGR PERS)=(↑ SUBJ PERS)	

According to the equations above, when the verb follows the subject it agrees with it in number, gender and person, while it agrees in gender and person only when it precedes it. This shows how agreement is resolved by storing the agreement features in a temporary reservoir and using phrase structure rules annotated with functional equations to distribute the agreement features. Figure 32 and Figure 33 show the c-structure and f-structure representations for the sentences in (154) and (155) above.



PRED 'لعب<[2:ولد]>'
TNS-ASP 5 TENSE past, MOOD indicative
AGR 4 PERS 3, NUM sg, GEND masc
SUBJ 2 PRED 'ولد'
SPEC 9 DET 10 DET-TYPE def
NTYPE 7 NSYN common
PERS 3, HUMAN +, GLOSS boy, GEND masc, DEF +, CASE nom, NUM pl
0 STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal

Figure 32. C-structure and f-structure of a VS sentence

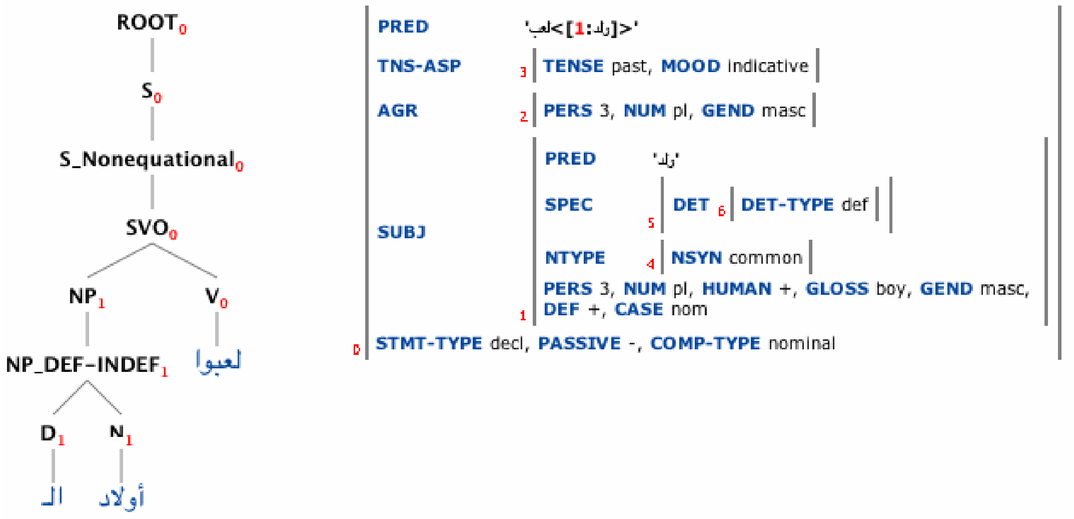

ROOT0
S0
S_Nonequational0
SVO0
NP1
V0
NP_DEF-INDEF1
لعبوا
D1
N1
الـ
أولاد
PRED 'لعب<[1:ولد]>'
TNS-ASP 3 TENSE past, MOOD indicative
AGR 2 PERS 3, NUM pl, GEND masc
SUBJ
PRED 'ولد'
SPEC 5 DET 6 DET-TYPE def
NTYPE 4 NSYN common
1 PERS 3, NUM pl, HUMAN +, GLOSS boy, GEND masc, DEF +, CASE nom
0 STMT-TYPE decl, PASSIVE -, COMP-TYPE nominal

Figure 33. C-structure and f-structure of an SV sentence

5.4 Functional Control and Long-Distance Dependencies

Raising and control in English are treated mostly in terms of *structure sharing* (Asudeh, 2005). In these constructions the subject or object of the matrix clause controls the subject of the embedded clause. However, we believe that this is not necessarily applicable for Arabic in which embedded clauses reconstruct their subjects mostly as a pro-drop. Al-Haq (1992) assumed that in Jordanian Arabic there are closed functions rather than open functions. We also assume that the functional control relationship in Arabic is more of an obligatory anaphoric control represented in closed functions rather than structure sharing represented in open functions. In Arabic, the control target can be totally absent or have some sort of realization, such as agreement features attached to the verb in the subordinate clause. It can also be realized as a resumptive pronoun, as in the case in long-distance dependencies.

Functional control can be either lexically determined, as in the case of raising and equi constructions, or structurally determined, as in the case of open adjuncts and long-distance dependencies. In raising and equi constructions the lexical entry of the verb specifies the control relationship, but in open adjuncts and long-distance dependencies, it is the phrase structure rules that specify the control relationships between the matrix and the subordinate clauses in the sentence.

The purpose of this section is to investigate the nature of control in Modern Standard Arabic within the framework of Lexical-Functional Grammar and to provide practical solutions to the different aspects of the phenomenon.

5.4.1 Raising

The main argument we are going to introduce under this subsection is that raising verbs which take non-verbal complements should be treated as quasi-copulas, not as raising verbs. This can generally be applied to English as well as Arabic.

In an English raising sentence, such as *He seems to work hard*, there are two verbs, but there is only one thematic role involved. The subject is an argument in the subordinate clause (*work hard*) but not an argument in the matrix clause (*seem*) (Asudeh, 2005, Falk, 2001, Lødrup, 2006). The complement of *seem* is a functionally controlled open function XCOMP. The identity of the subject in the subordinate clause is resolved by a functional control equation on the lexical entry of the raising verb, as shown in (156), reproduced from Asudeh (2005). This control equation makes XCOMP's subject equivalent to the matrix subject. The subject is not semantically selected by the verb *seem*, and this is why the SUBJ function is located outside the angle brackets in the verbs a-structure.

(156) seem V (↑ PRED)= 'seem<(↑ XCOMP)> (↑ SUBJ)'
(↑ XCOMP SUBJ) = (↑ SUBJ)

In other words, raising verbs, such as *seem* and *expect*, take a whole proposition as an argument; they selects a "propositional-theme", as in (157a) and (157b), reproduced from (Lødrup, 2006).

(157) a. seem ___ < propositional-theme > (Raising to subject)
b. expect < experiencer propositional-theme > ___ (Raising to object)

In English raising sentences, the verbal complement can either be a *to*-infinitive, or infinitive without *to*. The controller in the matrix sentence can either be subject (raising to subject) or object (raising to object), as shown in the examples in (158).

(158) a. I saw him work hard. (Infinitive complement)
b. He seems to study hard. (*to*-infinitive complement)
c. He seems to work hard (Raising to subject)
d. I expect him to work hard (Raising to object)

Similarly in Arabic raising sentences, the complement can either be preceded or not preceded by a complementizer ʼan ‘to’. The controller in the matrix sentence can either be subject (raising to subject) or object (raising to object), as shown in the examples (159)–(162).

(159) أوشك الولد أن ينام (Complement with a complementizer)
ʼawšaka al-waladu ʼan yanāma
was-nearly the-boy to sleep
‘The boy nearly slept.’

(160) أصبح الطالب يحب القراءة (Complement without a complementizer)
ʼaṣbaḥa aṭ-ṭālibu yuḥibbu al-qirāʼata
became the-student love the-reading
‘The student has come to love reading.’

(161) ظل الطالب يذاكر (Raising to subject)
ẓalla aṭ-ṭālibu yuḏākiru
remained the-student study
‘The student remained studying.’

(162) ظننت الطالب يذاكر (Raising to object)
ẓanantu aṭ-ṭāliba yuḏākiru
thought-I the-student study
‘I thought the student is studying.’

As a further division, Lødrup (2006) pointed out that raising verbs can have verbal complements or non-verbal complements. Examples in (163) are reproduced from Lødrup (2006).

(163) a. Peter seems to study hard. (Verbal complement)
b. John seems nice. (Adjectival complement)
c. The pills made him a monster. (NP complement)
d. She seems in a bad mood. (PP complement)

In English raising constructions both verbal and non-verbal constructions are treated as the same and both are represented as an open XCOMP function. However, we maintain that the two types of predications are totally different. While the verbal complement naturally selects for a subject and it is quite logical to treat it as a raising construction, it is hard to prove that ADJPs, ADVPs, NPs and PPs can subcategorize for a subject. As a workaround, Bresnan (2001) (cited by Lødrup, 2006) tried to equip nouns and prepositions with subject by the application of lexical rules.

> 'monster' => 'be-a-monster<(↑ SUBJ)>'
> 'in<(↑ OBJ)>' => 'be-in-state-of<(↑ SUBJ) (↑ OBJ)>'

This analysis, however, does not look very linguistically motivated. The verb's power to project onto the sentence structure cannot in any way be rivalled by any other lexical item. Verbs are the "inherent predicators" (Avgustinova and Uszkoreit, 2003), and they are the uncontested predicators in the general case (Bresnan, 1995). Verbs function in basically different relationships from other constituents. In the verb–subject clauses, the subject is generally the doer of the action which in most cases carries the roles 'volitional' and 'agentive'.

Our proposed solution is to treat *seem* with non-verbal complements as a quasi-copula that links a subject and a predicate. The difference between *he seems to go* and *he seems happy* is the same as the difference between *he goes* and *he is happy* which are completely different syntactic structures. The first is a verbal construction while the second is a predicational construction.

The sentences in (164) are syntactically equivalent, as all the verbs function as quasi copulas. Therefore we can assume that when *seem* takes a non-verbal complement it does not function as a raising verb but rather as a quasi copula verb.

(164) a. It seems nice.
b. It looks nice.
c. It tastes nice.
d. It smells nice.

The division of raising complements into verbal and non-verbal is also applicable in Arabic. There is a class of raising verbs that take both verbal and non-verbal complements. These are called كان وأخواتها kān wa-ʾaḫawātuhā ‘kāna and its sisters’, or نواسخ المبتدأ والخبر nawāsiḫu al-mubtadaʾ wa-al-ḫabar ‘governors of copula constructions’ such as كان kāna ‘was’, أصبح ʾaṣbaḥa ‘became’, أمسى ʾamsā ‘turned out to be’, صار ṣāra ‘became’, ظل ẓalla ‘remained’ and ليس laisa ‘is not’, as shown in the examples (165) and (166). There is another class of verb that take only verbal complements known in Arabic grammar as أفعال المقاربة ʾafʿālu al-muqārabati ‘verbs of nearness’, such as كاد kāda ‘became near’ and أوشك ʾawšaka ‘became near’, as in (167). And finally there are some verbs that take a non-verbal complement only such as (168).

(165) أصبح الطالب سعيدا (Verbal/non-verbal complements)
ʾaṣbaḥa aṭ-ṭālibu saʿīdan
became the-student happy
‘The student became happy.’

(166) أصبح الطالب يحب القراءة (Verbal/non-verbal complements)
ʾaṣbaḥa aṭ-ṭālibu yuḥibbu al-qirāʾata
became the-student love the-reading
‘The student became to love reading.’

(167) كاد الولد أن ينام (Verbal complements)
kāda al-waladu ʾan yanāma
was-nearly the-boy to sleep
‘The boy nearly slept.’

(168) يبدو الطالب سعيدا (Non-verbal complements)
yabdū aṭ-ṭālibu saʿīdan
seem the-student happy
‘The student seems happy.’

The XCOMP analysis can be implemented for Arabic. For the example in (169) we can have the f-structure in Figure 34.

(169) أصبح الرجل كريما
ʾaṣbaḥa ar-rağulu karīman
became the-man generous
‘The man became generous.’

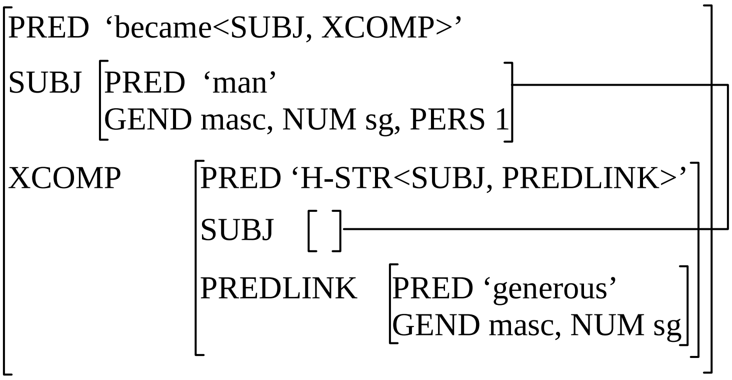

PRED ‘became<SUBJ, XCOMP>’
SUBJ
PRED ‘man’
GEND masc, NUM sg, PERS 1
XCOMP
PRED ‘H-STR<SUBJ, PREDLINK>’
SUBJ [ ]
PREDLINK
PRED ‘generous’
GEND masc, NUM sg

Figure 34. F-structure of XCOMP analysis for Arabic raising verb

However, our argument is that in Arabic some verbs function as raising verbs when they are followed by verbal complements. In this case the complement is in functional control relationship with the matrix clause.

(170) كان الطالب يحب القراءة
kāna aṭ-ṭālibu yuḥibbu al-qirāʾata
was the-student love.pres.3.masc.sg the-reading
‘The student used to love reading.’

(171) أصبح الطالب يحب القراءة
ʾaṣbaḥa aṭ-ṭālibu yuḥibbu al-qirāʾata
became the-student love.pres.3.masc.sg the-reading
‘The student became to love reading.’

The c-structure and f-structure for the example in (170) are shown in Figure 35 and Figure 36 respectively.

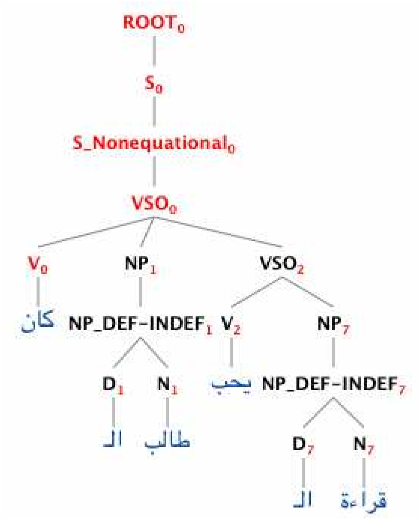
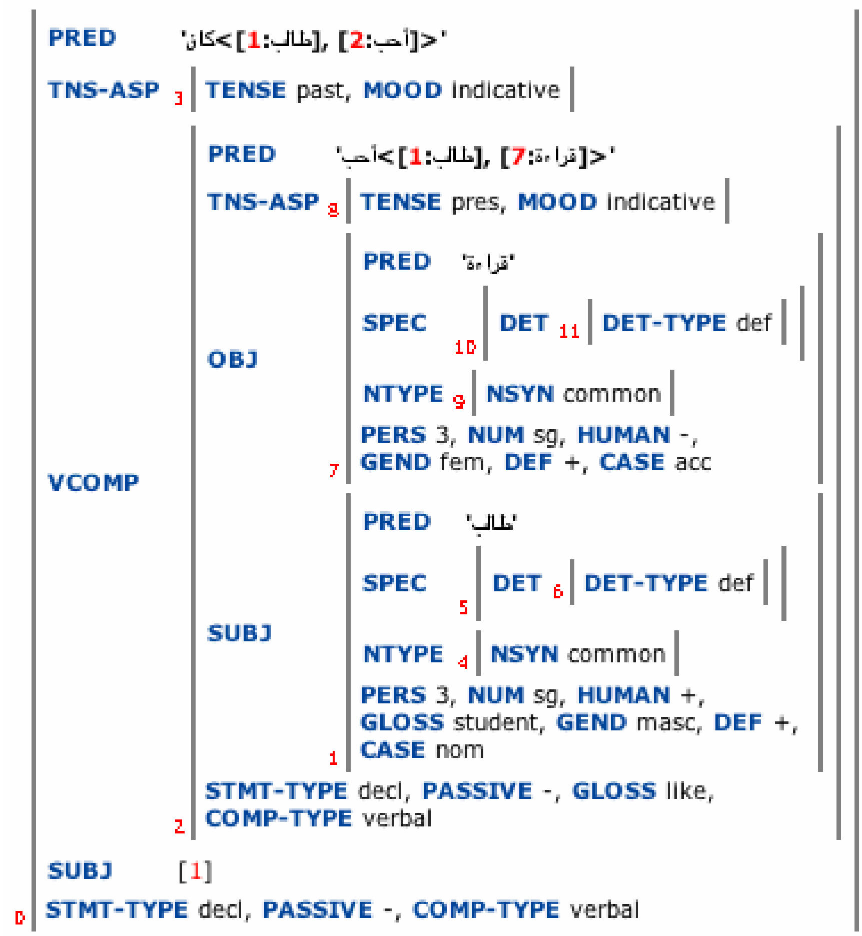


Figure 35. C-structure of an Arabic raising sentence



```
PRED      'كان<[1:طالب], [2:أحب]>'
TNS-ASP 3 | TENSE past, MOOD indicative |
            PRED      'أحب<[1:طالب], [7:قراءة]>'
            TNS-ASP 8 | TENSE pres, MOOD indicative |
                        PRED  'قراءة'
                        SPEC    | DET 11 | DET-TYPE def | |
                             10
            OBJ         NTYPE 9 | NSYN common |
                        PERS 3, NUM sg, HUMAN -,
                      7 GEND fem, DEF +, CASE acc
VCOMP                   PRED  'طالب'
                        SPEC    | DET 6 | DET-TYPE def | |
                              5
            SUBJ        NTYPE 4 | NSYN common |
                        PERS 3, NUM sg, HUMAN +,
                        GLOSS student, GEND masc, DEF +,
                      1 CASE nom
            STMT-TYPE decl, PASSIVE -, GLOSS like,
          2 COMP-TYPE verbal
SUBJ      [1]
0 STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal
```

Figure 36. F-structure of an Arabic raising sentence

Otherwise in situations when these verbs are followed by non-verbal complements, they are considered as copula and quasi-copula verbs.

(172) كان الطالب سعيدا
kāna aṭ-ṭālibu sa῾īdan
was the-student happy
'The student was happy.'

(173) أصبح الطالب سعيدا
᾿aṣbaḥa aṭ-ṭālibu sa῾īdan
became the-student happy
'The student became happy.'

The c-structure and f-structure for the example in (173) are shown in Figure 37 and Figure 38 respectively.

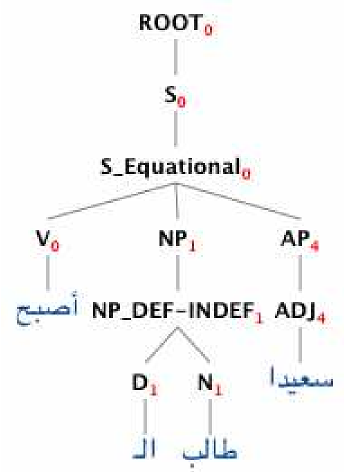
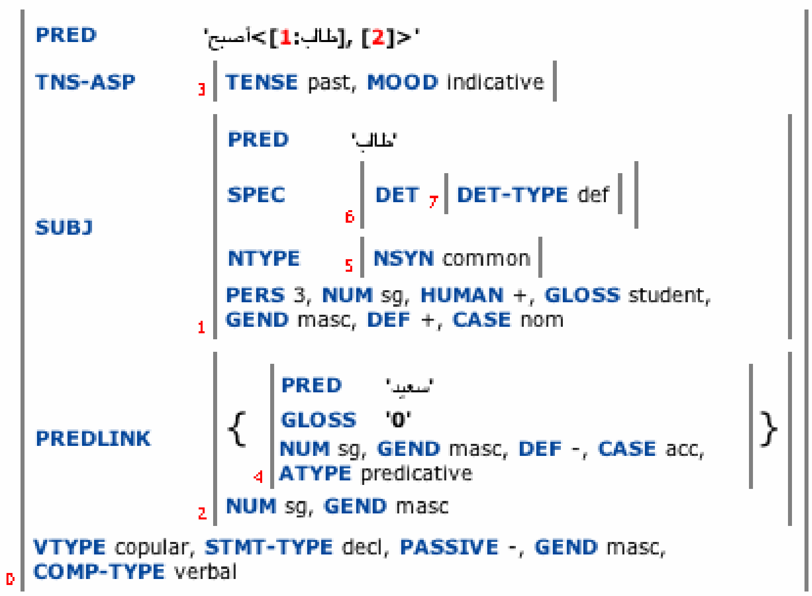


Figure 37. C-structure of an Arabic sentence with a quasi copula verb



```
PRED          'أصبح<[1:طالب], [2]>'
TNS-ASP     3 TENSE past, MOOD indicative
                PRED      'طالب'
                SPEC    6 DET 7 DET-TYPE def
SUBJ            NTYPE   5 NSYN common
                PERS 3, NUM sg, HUMAN +, GLOSS student,
              1 GEND masc, DEF +, CASE nom
                   PRED   'سعيد'
PREDLINK        {  GLOSS  '0'                                  }
                   NUM sg, GEND masc, DEF -, CASE acc,
                 4 ATYPE predicative
              2 NUM sg, GEND masc
VTYPE copular, STMT-TYPE decl, PASSIVE -, GEND masc,
0 COMP-TYPE verbal
```

Figure 38. F-structure of an Arabic sentence with a quasi copula verb

5.4.2 Equi

In equi sentences, such as *He tries to work hard*, the subject has two thematic roles. It is a thematic argument of the main verb and also a thematic argument of the complement (Falk, 2001, Lødrup, 2006).

Falk (2001) argues that the complement of *try* is a functionally controlled open function XCOMP based on the rough generalization that "obligatory control constructions involve functional control and nonobligatory control constructions involve anaphoric control."

The identity of the subject in the subordinate clause is resolved by a functional control equation on the lexical entry of the equi verb, which makes XCOMP's subject equivalent to the matrix subject (Falk, 2001).

(174) try   V   (↑ PRED)= 'try<(↑ SUBJ) (↑ XCOMP)>'
                  (↑ XCOMP SUBJ) = (↑ SUBJ)

Dalrymple (2001) however, assumed that English equi verbs exemplify anaphoric control, while English raising verbs exhibit functional control.

In analyzing Serbo-Croatian, Asudeh (2000) assumed that as Serbo-Croatian is a pro-drop language the control in equi in Serbo-Croatian sentences is different from English sentences. In Serbo-Croatian the complement has its own subject, and rather than the subject being structure-shared with another GF, the subject of the matrix sentence is co-indexed with the subject of the subordinate sentence.

The same argument is valid for Arabic which is also a pro-drop language. The morphosyntactic morphemes on the verbs allow them to reconstruct their own subjects, and this is why the complements in Arabic equi sentences can be considered as COMPs, not XCOMPs. The subject in Arabic equi sentences is not structure shared between the controller (subject of the matrix clause) and the control target (subject of the subordinate clause). Instead, the relationship between the two elements is anaphoric, expressed through co-indexation and obligatory compatibility of the agreement features (full agreement in number, gender and person).

Equi complements in English can either be *to*-infinitives or gerunds, as shown in the examples in (175).

(175) a. I promised him to go.   (*to*-infinitive)
      b. He tried switching the phone off.   (gerund)

The controller can either be subject or object in the matrix sentence, as shown in the examples in (176).

(176) a. He tried to go. (Subject Controller)
b. I persuaded him to go. (Object Controller)

Similarly complements in Arabic equi constructions can be verbs preceded or not preceded by a complementizer ʼan ‘to’, or verbal nouns, as shown in the examples (177), (178) and (179). The second example is an instance of a class of verbs known in Arabic grammar as أفعال الشروع ʼafʿālu aš-šurūʿi ‘verbs of starting the action’ such as شرع šaraʿa ‘started’, أخذ ʼaḫaḏa ‘started’ and جعل ğaʿala ‘kept’.

(177) وعدته أن أذهب (Complement with a complementizer)
waʿadtu-hu ʼan ʼaḏhaba
promised-him to go
‘I promised him to go.’

(178) أخذ المدير يدرس القرار (Complement without a complementizer)
ʼaḫaḏa al-mudīru yadrusu al-qarāra
kept the-manager study the-decision
‘The manager kept studying the decision.’

(179) حاول إصلاح الماكينة (Verbal–noun complement)
ḥāwala ʼiṣlāḥa al-mākīnti
tried fixing the-machine
‘He tried fixing the machine.’

The controller also can be subject or object in the matrix sentence, as shown by the examples (180) and (181) respectively.

(180) حاول أن ينام (Subject Controller)
ḥāwala ʼan yanāma
tried to sleep
‘He tried to sleep.’

(181) أقنعته أن يذهب (Object Controller)
ʼaqnaʿtu-hu ʼan yaḏhaba
convinced-him to go
‘I convinced him to go.’

In Arabic the relationship is established as anaphoric control rather than functional control. The subject of the subordinate clause is established as a pro-drop (unexpressed pronoun) and the subordinate verb provides gender and number information about the subject. The control relationship, presented in (182), equates the number and gender of the subject of the subordinate clause with those of the subject of the matrix clause.

(182) حاول [try] V (↑ PRED)= ‘حاول<(↑ SUBJ) (↑ COMP)>’
(↑ COMP SUBJ NUM) = (↑ SUBJ NUM)
(↑ COMP SUBJ GEND) = (↑ SUBJ GEND)
(↑ COMP SUBJ PERS) = (↑ SUBJ PERS)

The c-structure and f-structure for the example in (180) are shown in Figure 39 and Figure 40 respectively.

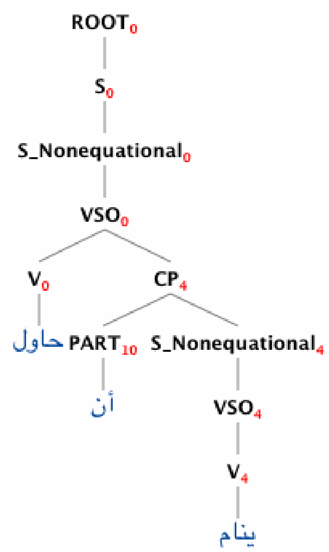
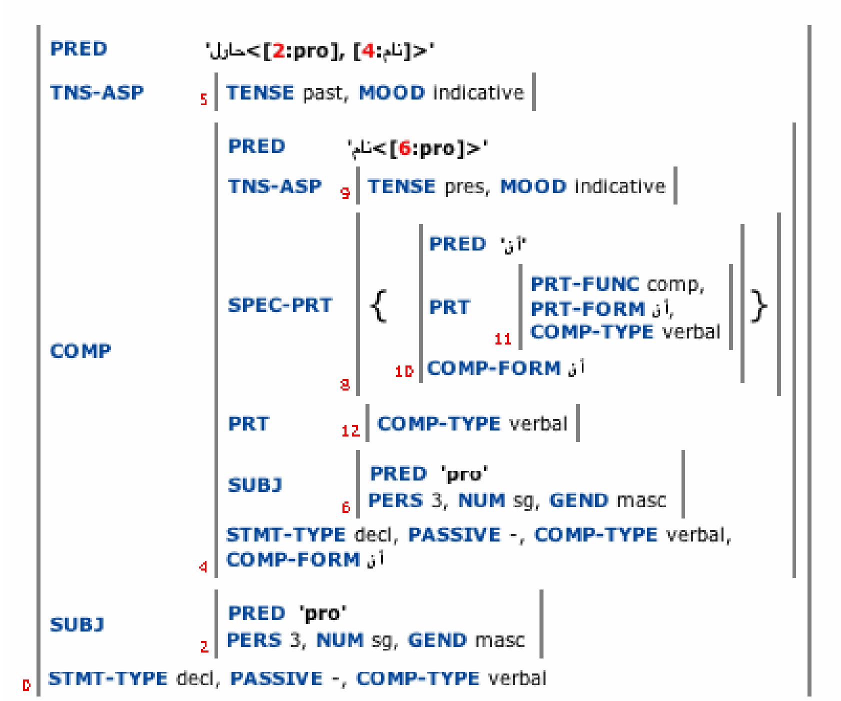


Figure 39. C-structure of an Arabic equi sentence



```
PRED       'حاول<[2:pro], [4:نام]>'
TNS-ASP   5[TENSE past, MOOD indicative]
           PRED      'نام<[6:pro]>'
           TNS-ASP  9[TENSE pres, MOOD indicative]
                         PRED  'أن'
                                  PRT-FUNC comp,
           SPEC-PRT  {   PRT      PRT-FORM أن,
                               11 COMP-TYPE verbal  }
                      10 COMP-FORM أن
COMP               8
           PRT      12[COMP-TYPE verbal]
           SUBJ       PRED 'pro'
                    6 PERS 3, NUM sg, GEND masc
           STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal,
          4 COMP-FORM أن
SUBJ       PRED 'pro'
          2 PERS 3, NUM sg, GEND masc
0 STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal
```

Figure 40. F-structure of an Arabic equi sentence

5.4.3 Control in Adjuncts

While control in raising and equi constructions is lexically determined (control is defined by functional annotations on the lexical entries of verbs), control in open adjuncts is described as “structurally determined” (Sells, 1985) or “constructionally induced” (Lødrup, 2006). This control relation is expressed by functional annotations on the phrase structure rules. Open adjuncts are adjuncts which take their subject from outside their clauses. For example the clause-initial adjectival adjunct, XADJUNCT, in (183) is controlled by the subject of the clause. This control relation is determined by the phrase structure rule in (184).

(183) Sure of winning, Mary entered the competition yesterday. (Sells, 1985)

(184) S →	(AP)	XP	VP	(Sells, 1985)
	(↑ XADJUNCT) = ↓	(↑ SUBJ) = ↓	↑ = ↓	
	(↑ SUBJ) = (↓ SUBJ)			

English XADJUNCTs can either be adjectival phrases or participial phrases (active or passive), as shown in the examples in (185). In these cases the SUBJ of the adjunct clause is functionally controlled by the SUBJ of the matrix clause.

(185) a. He went away, proud of himself. (Adjectival XADJUNCT)
b. Going to school, Peter lost his bag. (Active participle XADJUNCT)
c. Defeated in the race, he decided to quit. (Passive participle XADJUNCT)

Arabic open adjuncts can be headed by an adjective, active participle, patient participles, or verbal nouns, as shown in the examples in (186). Adjuncts here are adverbial, expressing either manner or resumption.

(186) a. عاد فخورا بانتصاراته (Adjectival XADJUNCT)
ʿāda faḫūran bi-ʾintiṣārāti-hi
returned proud of-victories-his
‘He returned, proud of his victories.’

b. معربا عن أسفه، قدم استقالته (Active participle XADJUNCT)
muʿriban ʿan ʾasafi-hi, qaddama ʾistiqalata-hu
expressing of regret-his, offered resignation-his
Expressing his regret, he offered his resignation.

c. عاد إلى البيت منهارا (Passive participle XADJUNCT)
ʿāda ʾilā al-baiti munhāran
came to the-home devastated
'He came home devastated.'

d. زار زعماء المعارضة بحثا عن الدعم (Verbal noun XADJUNCT)
zāra zuʿamāʾa al-muʿāraḍati baḥṯan ʿan ad-daʿmi
visited leaders the-opposition searching for the-support
'He visited opposition leaders, searching for support'

Dalrymple (2001) explains that the English open function XADJUNCT has an open SUBJ position functionally controlled by the SUBJ of the matrix clause, and the same f-structure fills both functions. For the sentence in (187) she proposed the f-structure in Figure 41.

(187) Walking the dog, Chris saw David.

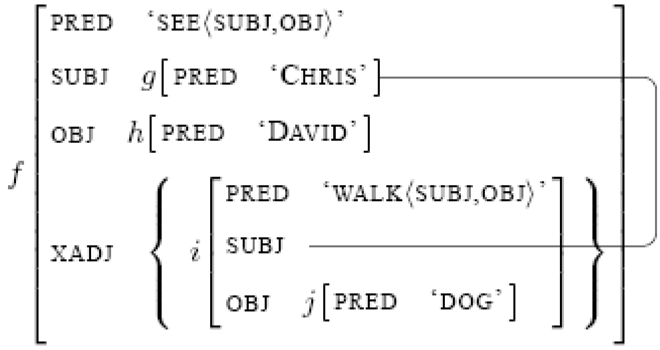


Figure 41. F-structure of an English sentence with an XADJUNCT

This does not necessarily apply for other languages. Dalrymple (*ibid.*) took the example in (188) from Warlpiri as evidence that some adjuncts participate in *obligatory anaphoric control*, where an unexpressed pronominal argument of a clausal adjunct is anaphorically controlled by an argument of the matrix clause.

(188) *karnta* *ka-rla* *wangka-mi* *ngarrka-ku* *[ngurra-ngka-rlu*
woman.ABS PRES-DAT speak-NONPAST man-DAT camp-LOC-ERG
jarnti-rninja-kurra-(ku)]
trim-INF-COMP-(DAT)
'The woman is speaking to the man (while he is) trimming it in camp.'

Dalrymple stated that in the above example the OBJ of the matrix clause anaphorically controls the SUBJ of the adjunct clause. She considers the example as involving anaphoric rather than functional control.

Similar to Warlpiri, Arabic adjuncts (apart from verbal nouns) can be considered as *closed* adjuncts since they are semantically complete, containing within them all the elements required for logical interpretation of the subject. Adjectives and participles are inflected for number and gender, which allows the establishment of the relation between the subject of the subordinate clause and the subject of the matrix clause anaphorically through co-indexing.

The c-structure and f-structure for the example in (189) are shown in Figure 42 and Figure 43 respectively.

(189) معربا عن أسفه، قدم استقالته
mu'riban 'an 'asafi-hi, qaddama 'istiqalata-hu
expressing of regret-his, offered resignation-his
Expressing his regret, he offered his resignation.

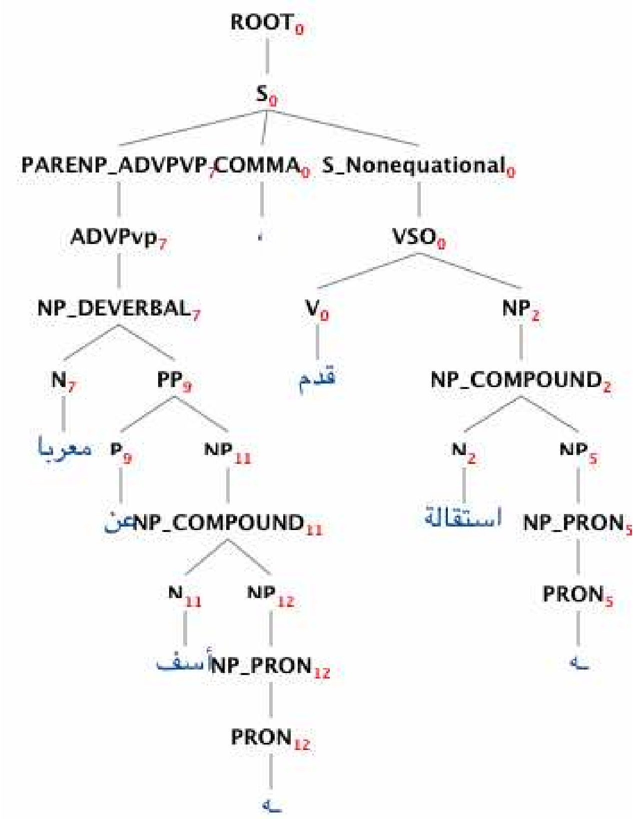
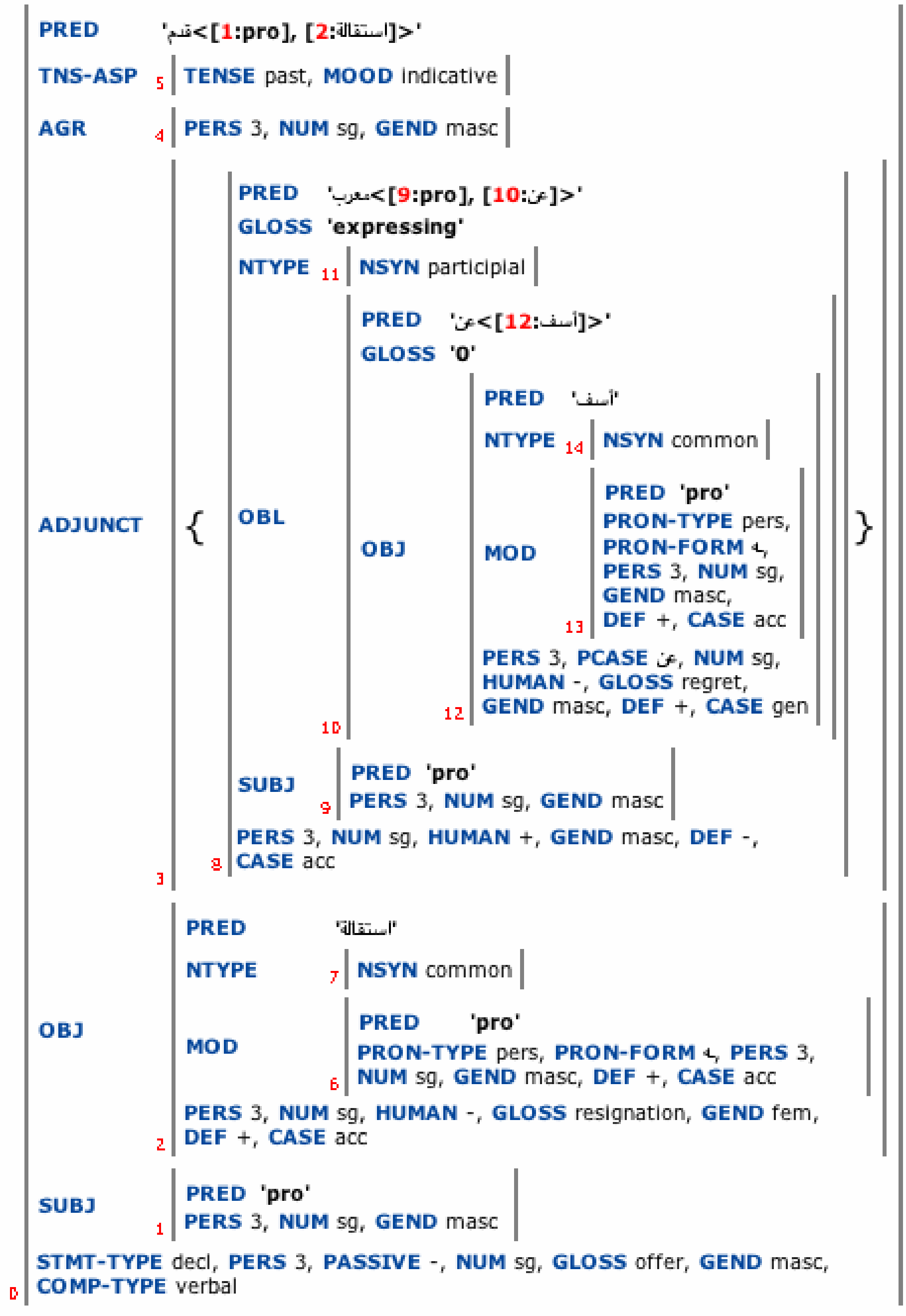


Figure 42. C-structure of Arabic control in adjunct phrase



```
PRED        'قدم<[1:pro], [2:استقالة]>'
TNS-ASP   5 [TENSE past, MOOD indicative]
AGR       4 [PERS 3, NUM sg, GEND masc]
ADJUNCT   { PRED   'معرب<[9:pro], [10:عن]>'
            GLOSS  'expressing'
            NTYPE 11 [NSYN participial]
            OBL    PRED   'عن<[12:أسف]>'
                   GLOSS  '0'
                   OBJ    PRED   'أسف'
                          NTYPE 14 [NSYN common]
                          MOD    PRED 'pro'
                                 PRON-TYPE pers,
                                 PRON-FORM ه,
                                 PERS 3, NUM sg,
                                 GEND masc,
                              13 DEF +, CASE acc
                          PERS 3, PCASE عن, NUM sg,
                          HUMAN -, GLOSS regret,
                       12 GEND masc, DEF +, CASE gen
                10
            SUBJ   PRED 'pro'
                 9 PERS 3, NUM sg, GEND masc
            PERS 3, NUM sg, HUMAN +, GEND masc, DEF -,
          8 CASE acc                                    }
          3
OBJ         PRED       'استقالة'
            NTYPE    7 [NSYN common]
            MOD        PRED      'pro'
                       PRON-TYPE pers, PRON-FORM ه, PERS 3,
                     6 NUM sg, GEND masc, DEF +, CASE acc
            PERS 3, NUM sg, HUMAN -, GLOSS resignation, GEND fem,
          2 DEF +, CASE acc
SUBJ        PRED 'pro'
          1 PERS 3, NUM sg, GEND masc
STMT-TYPE decl, PERS 3, PASSIVE -, NUM sg, GLOSS offer, GEND masc,
0 COMP-TYPE verbal
```

Figure 43. F-structure of Arabic control in adjunct phrase

5.4.4 Long-Distance Dependencies

There are three instances of long-distance dependencies: topicalization constructions, relative clauses and wh-questions.

(190) التفاحة أكل الولد (Topicalized)

at-tuffāḥata	ʾakala	al-waladu
the-apple	ate	the-boy

‘The apple, the boy ate.’

(191) حول هذه المسألة دار كثير من الجدل (Topicalized)
ḥawla haḏihi al-masʾalata dāra kaṯīrun mina al-ğadali
around this the-question revolved much of the-controversy
‘Around this question revolved much of the controversy.’

(192) من ضرب زيدا؟ (Question)
man ḍaraba zaidan?
who hit.sg.masc Zaid
‘Who hit Zaid?’

(193) الرجل الذي فاز بالجائزة صديقي (Relative)
ar-rağulu allaḏī fāza bi-l-ğāʾizati ṣadīq-ī
the-man who won of-the-prize friend-my
‘The man who won the prize is my friend.’

Dalrymple (2001) defined long-distance dependencies as “constructions in which a displaced constituent bears a syntactic function usually associated with some other position in the sentence.” In these constructions, the displaced (or extracted) constituent controls two positions and plays two roles simultaneously: it is the TOPIC or FOCUS of the sentence (the filler position) and it has also another grammatical function within the sentence such as OBJ, SUBJ or OBL (the gap position), and this is considered as the position from which it has been extracted. The relation between the two positions must be controlled according to the Extended Coherence Condition:

> Extended Coherence Condition (Dalrymple, 2001):
> FOCUS and TOPIC must be linked to the semantic predicate argument structure of the sentence in which they occur, either by functionally or by anaphorically binding an argument.

These constructions are called “long-distance dependencies” and “unbounded dependencies” because the distance between the initial position, the filler, and the grammatical function from which it has been extracted, the gap, can be potentially unlimited (Austin, 2001).

As explained by Austin (2001), long-distance dependencies are accounted for in the LFG literature in terms of “functional uncertainty”, where a functional equation, shown in (194), identifies the initial element bearing a discourse function (DF) such as TOPIC or FOCUS with a grammatical function (GF) such as SUBJECT or OBJECT later in the sentence. The path of this identification can be long and passes through (in English) any number of COMPlement clauses.

(194) (↑ DF) = (↑ COMP* GF) (Austin, 2001)

5.4.4.1 Island Constraints

Island constraints are defined by Falk (2001) as the “restrictions on the relation between filler and gap in long-distance dependency constructions”. In English we can give three examples of island constraints: complex NP constraint, as in the examples in (195); SUBJ constraint, as in (196); and ADJUNCT constraint, as shown in (197).

(195) a. *What did you deny [the claim that you put __ on the shelf]?
b. *This is the book which I saw [the woman who wrote __].

(196) a. *What do you think that [to put __ on the shelf] would be a good idea?
b. *Which person does [a picture of __] looks nice?

(197) *Which picture did they blush [when they saw ___]?

5.4.4.2 Resumptive Pronouns

Resumptive pronouns are defined as pronouns that are used in some languages to mark the lower end of a long-distance dependency (Falk, 2002). Resumptive pronouns fill the gaps in the domain of extraction, and like gaps, resumptive pronouns are linked to a discourse function. The Extended Coherence Condition allows an anaphoric link. Dalrymple (2001) pointed out that some languages signal the domain of extraction in long-distance dependency constructions by means of special morphological or phonological forms.

Resumptive pronouns are reported in many languages such as Turkish (Meral, 2004), Irish (Vaillette, 2002), Palauan (Georgopoulos, 1991), Welsh (Willis, 2000), Hebrew (Falk, 2002) and Arabic. The distribution of the resumptive pronouns in Arabic shows that in some syntactic positions they are required while in others they are optional or even prohibited.

The distribution of resumptive pronouns in Arabic can be summarized as follows. With questions, resumptive pronouns are not allowed, as shown by (198). With topicalized constructions, resumptive pronouns are required (in Classical Arabic objects, bearing the accusative case, are fronted without the need for a resumptive pronoun) , as in (199). With relative constructions, resumptive pronouns are not allowed when extracting from the immediate subject position, as in (200), but they are optional when extracting from the object position, as in (201) and (202). However, they are required with object of oblique, as in (203) and in long paths, as in (204).

(198) ماذا أكلها الرجل؟ *
māḏā ʾakala-ha ar-rağulu?
what ate-it the-man?
* 'What did the man eat it?'

(199) هذا المعلم يقدره الطلاب
haḏā al-muʿallim yuqddiru-hu aṭ-ṭullābu
this the-teacher appreciate-him the-students
'This teacher, the students appreciate.'

(200) الرجل الذي أكل التفاحة
ar-rağulu allaḏī ʾakala at-tuffāḥata
the-man who ate the-apple
'the man who ate the apple'

(201) التفاحة التي أكل الرجل
at-tuffāḥata allatī ʾakala ar-rağulu
the-apple which ate the-man
'the apple which the man ate'

(202) التفاحة التي أكلها الرجل
at-tuffāḥatu allatī ʾakala-hā ar-rağulu
the-apple which ate-it the-man
'the apple which the man ate'

(203) الولد الذي يعتمد عليه الرجل
al-waladu allaḏī yaʿtamidu ʿalai-hi ar-rağulu
the-boy who relies on-him the-man
'the boy on whom the man relies'

(204) الرجل الذي ز عمت البنت أنه أكل التفاحة
ar-rağulu allaḏī zaʿamat al-bintu ʾanna-hu ʾakala at-tuffāḥata
the-man who claimed the-girl that-he ate the-apple
'the man who the girl claimed that he ate the apple'

In other languages resumptive pronouns might have different distribution. In Hebrew resumptive pronouns are only used in relative clauses, and disallowed in questions (Falk, 2002).

Resumptive pronouns in Arabic are correlated with the applicability of island constraints. Where resumptive pronouns are not used long-distance dependencies are subject to island constraints, as shown in (205), but when resumptive pronouns are used, the constructions are not subject to island constraints. Example (206) shows how resumptive pronouns can cross the complex NP constraint, while example (207) shows how they cross the SUBJ constraint, and finally example (208) shows how they cross the ADJUNCT constraint.

(205) * ماذا هناك ادعاء أن الرجل سرق؟
*māḏā hunāka ʾiddiʿāʾun ʾanna ar-rağulu saraqa?
what there claim that the-man stole?
‘*What there is a claim that the man stole __ ?’

(206) الرجل الذي هناك ادعاء أنه سرق المال
ar-rağulu allaḏī hunāka ʾiddiʿāʾun ʾanna-hu saraqa al-mala
the-man who there claim that-he stole the-money
‘This man who there is a claim that he stole the money.’

(207) الرجل الذي فازت صورته بالجائزة
ar-rağulu allaḏī fāzat ṣūratu-hu bi-l-ğāʾizati
the-man who won picture-his of-the-prize
‘the man whose picture won the prize’

(208) الأعداء الذين مات أبوك وهو يحاربهم
al-ʾaʿdāʾu allaḏīna māta ʾabū-ka wa-hwa yuḥāribu-hum
the-enemies who died father-your and-he fight-them
‘the enemies who your father died while fighting them’

Regarding the syntactic analysis of constructions with resumptive pronouns, Falk (2002) concluded that resumptive pronouns participate in long-distance dependency constructions, and that they are not licensed in the normal way by functional uncertainty equations, but rather by establishing a referential (anaphoric) identity between the two positions. He considered that this analysis is able to account for the similarities and differences between gaps and resumptive pronouns.

In our Arabic parser we adopted Falk's (2002) analysis of the resumptive pronouns. When resumptive pronouns are used to mark the lower end in long-distance dependencies, the relationship is established anaphorically through matching the agreement features between the filler and the resumptive pronouns. In other instances of long-distance dependencies where resumptive pronouns are not used the relationship is marked through functional uncertainty equations which allows the filler to control the position of the gap.

We will explain this with some details in one instance of long-distance dependencies in Arabic; that is relative clauses. In our grammar, when the extraction is from the subject position and no other syntactic function cuts the path, resumptive pronouns are not allowed and the relation is expressed by functional identity between the two positions, as shown by the functional equations in (209).

(209) (^ TOPIC-REL)=(^ SUBJ)

The above equation will be able to handle clauses such as the one in (210) and give the analysis in Figure 44 and Figure 45.

(210) الطريق الذي يقود الفلسطينيين إلى السلام
aṭ-ṭarīqu allaḏī yaqūdu al-filisṭīniyyīn ʾilā as-salām
the-road which lead the-Palestinians to the-peace
'the road which leads the Palestinians to peace'

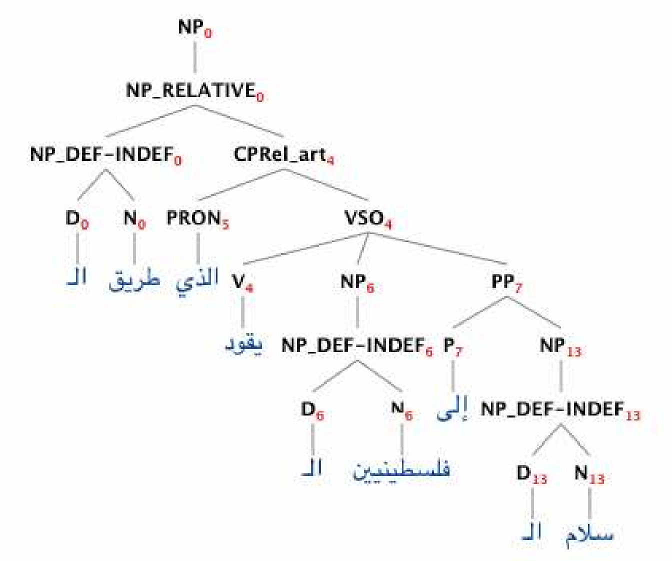


Figure 44. C-structure of an Arabic relative clause

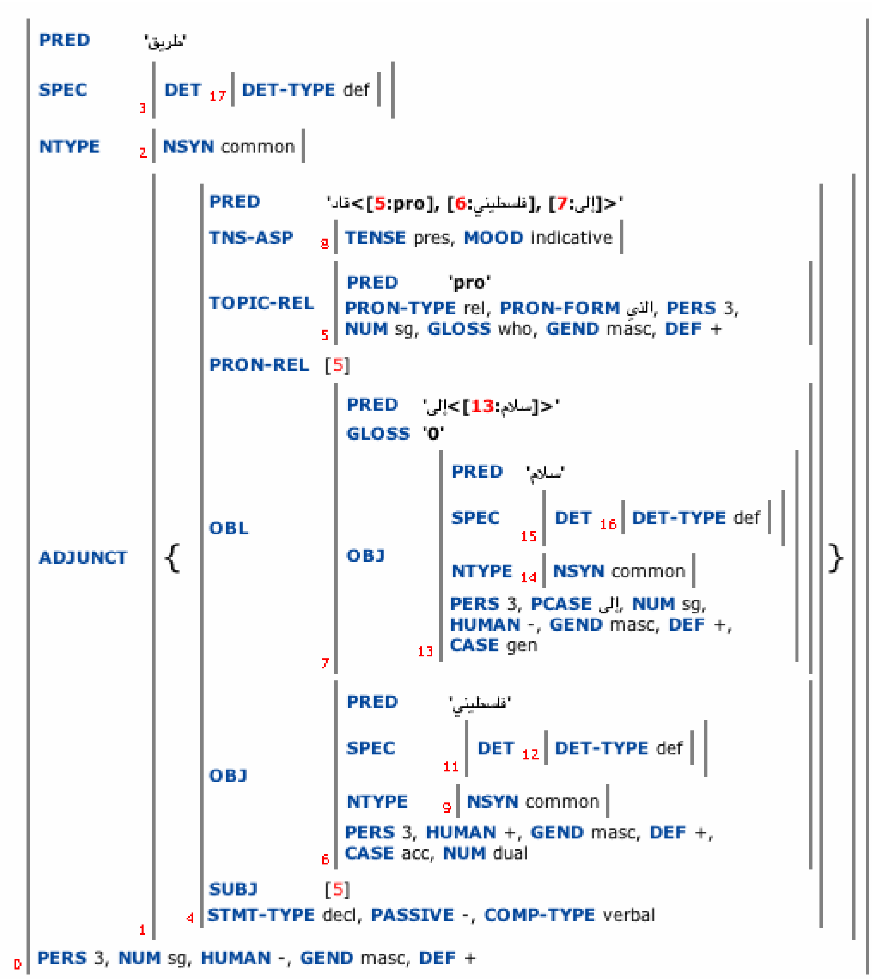

PRED 'طريق'
SPEC 3 DET 17 DET-TYPE def
NTYPE 2 NSYN common
ADJUNCT
PRED 'قاد<[5:pro], [6:فلسطيني], [7:إلى]>'
TNS-ASP 8 TENSE pres, MOOD indicative
TOPIC-REL
PRED 'pro'
PRON-TYPE rel, PRON-FORM الذي, PERS 3,
5 NUM sg, GLOSS who, GEND masc, DEF +
PRON-REL [5]
OBL
PRED 'إلى<[13:سلام]>'
GLOSS '0'
OBJ
PRED 'سلام'
SPEC 15 DET 16 DET-TYPE def
NTYPE 14 NSYN common
PERS 3, PCASE إلى, NUM sg,
HUMAN -, GEND masc, DEF +,
13 CASE gen
7
OBJ
PRED 'فلسطيني'
SPEC 11 DET 12 DET-TYPE def
NTYPE 9 NSYN common
PERS 3, HUMAN +, GEND masc, DEF +,
6 CASE acc, NUM dual
SUBJ [5]
4 STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal
1
0 PERS 3, NUM sg, HUMAN -, GEND masc, DEF +

Figure 45. F-structure of an Arabic relative clause

In all other instances the relation in relative clauses is expressed by equality of the morpho-syntactic features of the two elements: gap and resumptive pronoun, as shown by the functional equation in (211).

(211) (^ GF* GF PRED)=c 'pro'
(^ TOPIC-REL NUM)=(^ GF* GF NUM)
(^ TOPIC-REL GEND)=(^ GF* GF GEND)

The domain of extraction in relative clauses in which resumptive pronouns are required can be virtually anywhere in the sentence, as shown by the examples (212)–(223). For the sentence in (212) we show the c-structure and f-structure representations in Figure 46 and Figure 47 respectively.

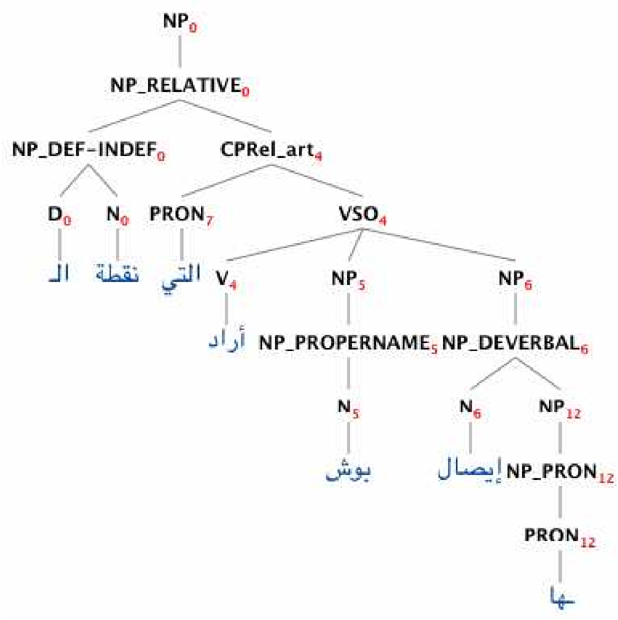
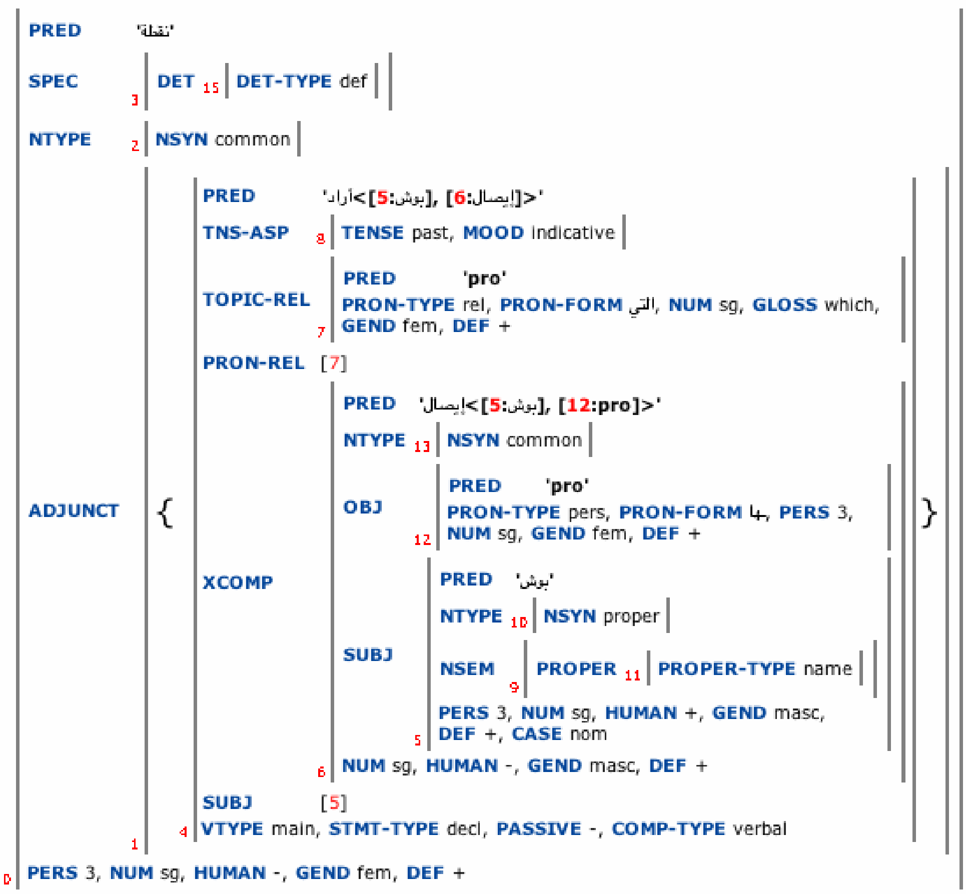


Figure 46. C-structure of an Arabic relative clause



```
PRED        'نقطة'
SPEC      3 [DET 15 [DET-TYPE def]]
NTYPE     2 [NSYN common]
ADJUNCT     { [PRED       'أراد<[5:بوش], [6:إيصال]>'
              TNS-ASP   8 [TENSE past, MOOD indicative]
              TOPIC-REL   [PRED        'pro'
                           PRON-TYPE rel, PRON-FORM التي, NUM sg, GLOSS which,
                        7  GEND fem, DEF +]
              PRON-REL  [7]
              XCOMP       [PRED   'إيصال<[5:بوش], [12:pro]>'
                           NTYPE 13 [NSYN common]
                           OBJ      [PRED      'pro'
                                     PRON-TYPE pers, PRON-FORM ـها, PERS 3,
                                  12 NUM sg, GEND fem, DEF +]
                           SUBJ     [PRED  'بوش'
                                     NTYPE 10 [NSYN proper]
                                     NSEM   9 [PROPER 11 [PROPER-TYPE name]]
                                     PERS 3, NUM sg, HUMAN +, GEND masc,
                                   5 DEF +, CASE nom]
                        6  NUM sg, HUMAN -, GEND masc, DEF +]
              SUBJ      [5]
          1 4 VTYPE main, STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal] }
0 PERS 3, NUM sg, HUMAN -, GEND fem, DEF +
```

Figure 47. F-structure of an Arabic relative clause

(212) النقطة التي أراد بوش إيصالها (XCOMP OBJ)
an-nuqṭatu allatī ʾarāda būš ʾīṣala-hā
the-point which wanted Bush conveying-it
‘the point which Bush wanted to convey’

(213) الدول التي اتهمها بالديكتاتورية (OBJ)
ad-duwalu allatī ʾitahama-hā bil-dīktātūriyyah
the-countries which accused-it with-dictatorship
‘the countries which he accused of dictatorship’

(214) الرجل الذي يتواجد ابنه في مدينة الفلوجة (SUBJ MOD)
ar-rağulu allaḏī yatawāğadu ʾibnu-hu fī madīnati al-fallūğah
the-man who exist son-his in city the-Fallujah
‘the man whose son is in Fallujah’

(215) القرارات التي تدعو الأمم المتحدة فيها إسرائيل إلى احترام حقوق الفلسطينيين (ADJUNCT OBJ)
al-qarārātu allatī tadʿū al-ʾumamu al-muttaḥidatu fī-hā ʾisrāʾīl
the-resolutions which call the-nations the-united in-it Israel
ʾilā ʾiḥtirāmi ḥuqūqi al-filisṭīniyyīn
to respecting rights the-Palestinians
‘the resolutions in which the United Nations calls on Israel to respect the rights of the Palestinians’

(216) المفاهيم التي يدعو لإرسائها (OBL OBJ)
al-mafāhīmu allatī yadʿū li-ʾirsāʾi-hā
the-concepts which call to-establishing-it
‘the concepts which he calls for establishing it’

(217) الهدنة التي وافقت عليها جميع الفصائل (OBL OBJ)
al-hudnatu alltī wāfaqat ʿalai-hā ğamīʿu al-faṣāʾili
the-truce which agreed on-it all the-factions
‘the truce on which all the factions agreed’

(218) الولايات المتحدة التي قال إنها تدعي الحرص على السلام (COMP SUBJ)
al-wilāuātu al-muttaḥidatu allatī qala ʾinna-hā taddaʿī
the-states the-united which said that-it pretend
al-ḥirṣa ʿalā as-salām
the-keenness on the-peace
‘the United States which he said that it pretends keenness on peace’

(219) الدولة التي قال إن سياساتها أذكت روح العداء للغرب (COMP SUBJ MOD)
ad-dawlatu allatī qala ʾinna siyāsāti-hā ʾaḏkat rūḥa
the-country which said that policies-its fostered spirit
al-ʿadāʾi li-l-ġarbi
the-enmity to-the-west
‘the countries which he said that its policies fostered enmity to the West’

(220) تونس التي قال أنه واكب تجربتها الإصلاحية (COMP SUBJ MOD)
tūnis allatī qala ʾanna-hu wākaba taǧribata-hā al-ʾiṣlāḥiyyata
Tunisia which said that-he witnessed experience-its the-reformative
‘Tunisia whose reform experience he said that he witnessed’

(221) الدولة التي قال إن الشيعة يكرهونها (COMP OBJ)
ad-dawlatu allatī qala ʾinna aš-šīʿata yakrahūn-hā
the-country which said that the-Shiites hate-it
‘the country which he said that the Shiites hate’

(222) السلام الذي قال إنه يحرص عليه (COMP OBL)
as-salāmu allaḏī qala ʾinna-hu yaḥriṣu ʿalai-hi
the-peace which said that-he care for-it
‘the peace which he said that he cares for’

(223) التصريحات التي قال فيها إن إسرائيل لن تتخلى عن مبادئها (COMP ADJUNCT)
at-taṣrīḥātu allatī qala fī-hā ʾinna ʾisrāʾīla lan
the-statements which said in-it that Israel will-not
tataḫllā ʿan mabādiʾi-hā
abandon from principles-its
‘the statements in which he said that Israel will not abandon its principles’

5.5 Unified Analysis of Copula Constructions in LFG

In this section we maintain that a unified analysis of the copula constructions in LFG is necessary to capture syntactic generalizations. We discuss the various options available in the LFG literature and investigate their feasibility, in order to arrive at the most appropriate representation. In doing so, we make use of the concepts and mechanisms already available in the framework of LFG without violating any fixed conditions or breaking with any established conventions. In this introduction we quickly review the three strategies used in LFG to represent copula constructions. In the next section we explain why a unified analysis is motivated. Then we explain the implications of adjectives in the copula constructions. Next we proceed into a detailed analysis of each strategy and provide our reasons for choosing one analysis and discarding the others.

The treatment of copula constructions in LFG has been outlined by Nordlinger and Sadler (2006), Dalrymple et al. (2004) and Rosén (1996). Although there is no controversy regarding the c-structure analysis of copula constructions in LFG, different strategies have been proposed for the f-structure representation of these constructions.

One possibility as outlined by Nordlinger and Sadler (2006) is what they termed as the “single-tier analysis” where the predicate functions as the sentential head and selects for a subject. The example they mentioned is from Russian:

(224) Ona vrač.
3sg.fem.nom doctor.sg.nom
`She is a doctor.'

F-structure

$$\left[\begin{array}{ll} \text{PRED} & \text{‘doctor<(}\uparrow\text{SUBJ)>’} \\ \text{CASE} & \text{nom} \\ \text{NUM} & \text{sg} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{‘pro’} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{fem} \\ \text{PERS} & 3 \\ \text{CASE} & \text{num} \end{array}\right] \end{array}\right]$$

Figure 48. F-structure of a Russian copula sentence

Nordlinger and Sadler point out that this analysis is also possible, at least in theory, for languages which have overt copulas such as English. For example in the sentence *He is famous* the adjective *famous* can select for the subject and the copula *is* functions only as a tense marker. They also make it clear that in the LFG literature the tendency in analyzing languages with explicit copulas is to adopt one version or the other of the double-tier analysis.

The double-tier analysis is another possibility for representing the copula construction. In this approach both the subject and the predicate function as arguments within the structure. Dalrymple et al. (2004) made a more detailed discussion of this type by dividing it into two significantly different variants. The first is to consider the predicate as a closed complement PREDLINK and the second is to consider it as an open complement XCOMP. Figure 49 shows all possible analyses of copula constructions in LFG.

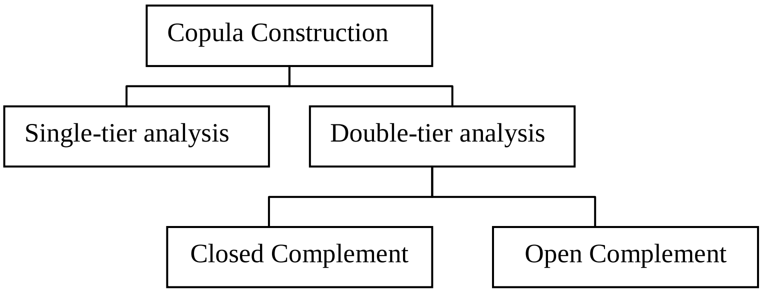


Figure 49. Possible Analyses of Copula Constructions in LFG

In the closed complement analysis, the main predicate of the sentence is provided by the copula. Figure 50 shows the double tier, closed function analysis of the English sentence in (225).

(225) She is a doctor.

$$\left[\begin{array}{ll} \text{PRED} & \text{`is}\langle(\uparrow\text{SUBJ})\ (\uparrow\text{PREDLINK})\rangle\text{'} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{`pro'} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{fem} \\ \text{PERS} & 3 \end{array}\right] \\ \text{PREDLINK} & \left[\begin{array}{ll} \text{PRED} & \text{`doctor'} \\ \text{NUM} & \text{sg} \end{array}\right] \end{array}\right]$$

Figure 50. A double-tier, closed-complement f-structure representation

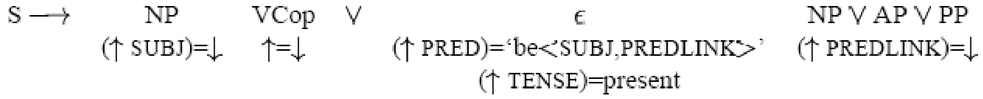
For languages with no overt copula the main predicate is provided by special annotations on phrase structure rules. Nordlinger and Sadler (2006) provide a reasonable account of where this unseen predicator is coming from. They argue that the main predicator is not an elided copula but a higher structure that governs the whole sentence:

> … these verbless clauses have a more hierarchical f-structure in which the f-structure of the non-verbal predicate functions as an argument within a higher f-structure which itself has a PRED, but where there is no overt syntactic element corresponding to this predicate in the c-structure. (Nordlinger and Sadler, 2006)

For the Russian example in (226a), we have the phrase structure rules in (226b) which produce the f-structure in Figure 51, all adapted from Dalrymple et al. (2004).

(226) a. On student.
he student
‘He is a student.’ (Russian)

b. Phrase structure rule



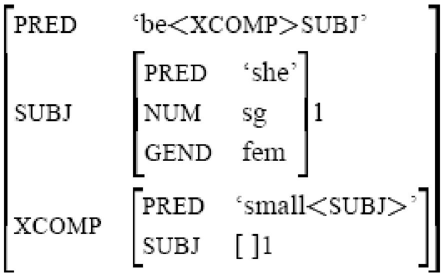
$$\begin{array}{lcccccc} \text{S} \longrightarrow & \text{NP} & \text{VCop} & \vee & \epsilon & & \text{NP} \vee \text{AP} \vee \text{PP} \\ & (\uparrow \text{SUBJ})=\downarrow & \uparrow=\downarrow & & (\uparrow \text{PRED})=\text{‘be<SUBJ,PREDLINK>’} & & (\uparrow \text{PREDLINK})=\downarrow \\ & & & & (\uparrow \text{TENSE})=\text{present} & & \end{array}$$

$$\left[\begin{array}{ll} \text{PRED} & \text{‘null-be<}(\uparrow\text{SUBJ})\ (\uparrow\text{PREDLINK})\text{>’} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{‘pro’} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{masc} \\ \text{PERS} & 3 \end{array}\right] \\ \text{PREDLINK} & \left[\begin{array}{ll} \text{PRED} & \text{‘student’} \\ \text{NUM} & \text{sg} \end{array}\right] \end{array}\right]$$

Figure 51. F-structure of a verbless copula construction

The second variant of the double-tier analysis of the copula construction is the open complement analysis where the structure is subject to functional control. In this analysis the predicate selects for a subject which is controlled by the main subject of the sentence. The French example (227) has the f-structure in Figure 52, both from Dalrymple et al. (2004).

(227) Elle est petite.
she.F.SG is small.F.SG
‘She is small.’ (French)



$$\left[\begin{array}{ll} \text{PRED} & \text{‘be<XCOMP>SUBJ’} \\ \text{SUBJ} & \left[\begin{array}{ll} \text{PRED} & \text{‘she’} \\ \text{NUM} & \text{sg} \\ \text{GEND} & \text{fem} \end{array}\right]1 \\ \text{XCOMP} & \left[\begin{array}{ll} \text{PRED} & \text{‘small<SUBJ>’} \\ \text{SUBJ} & [\ ]1 \end{array}\right] \end{array}\right]$$

Figure 52. Open function analysis of copula constructions

According to Dalrymple et al. (2004), this analysis entails that the lexical entry of the predicate subcategorises for a subject and contains a control equation as shown in (228).

(228) petite (↑ PRED) = ‘small<(↑SUBJ)>’
(↑ SUBJ NUM) =c sg
(↑ SUBJ GEND) =c fem

The conclusions reached by Dalrymple et al. (2004) were not conclusive. They said that the XCOMP analysis is “appropriate for some copular constructions but not for others, even within the same language”. They pointed out that more syntactic tests need to be identified in order to “determine the status of a postcopular constituent both within and across languages”. However, this research left the general perception that the open function is the preferred analysis. XCOMP has effectively replaced PREDLINK in the XLE English grammar and the DCU LFG-based probabilistic parser.

Nordlinger and Sadler (2006), on the other hand, state that the default structure is the single-tier analysis for copula-less languages, while languages which use overt copulas can choose a version of the double-tier analysis. Their focus was on emphasising the flexibility of the LFG framework rather than searching for a unified analysis.

> In the absence of positive evidence to the contrary, the single-tier analysis (which is more economical in assuming less structure) is the default hypothesis for verbless clauses cross-linguistically. (Nordlinger and Sadler, 2006)

5.5.1 Motivation for a Unified Analysis

A critical point in the syntactic analysis of copula constructions in the LFG literature is that it provides more questions than answers. The conclusion Dalrymple et al. (2004) reached is that a unified analysis of copula constructions is not possible either cross-linguistically or inside the same language.

> The fact that different constituents can behave differently in copular constructions means that the full range of copular constructions must be examined within a language in order to analyze it completely. That is, the fact

> that one type of constituent requires a certain analysis of copular constructions does not guarantee that other, superficially similar constructions will be amenable to the same analysis. (Dalrymple et al., 2004, p. 191)

Nevertheless, when talking about Russian, where the copula is null in the present tense but overt in the past and future tenses, they said.

> For such languages, there does not appear to be any evidence that the copula-less constructions have different syntax (or semantics) from the ones with copulas. As such, a unified analysis is desirable. However, a unified analysis is possible for all languages in which the occurrence of the copula is (partially) governed by tense. (Dalrymple et al., 2004, p. 192)

The indeterminacy in the LFG literature regarding copula constructions constitutes practical and theoretical challenges for grammar writing. The practical challenge is that for a new grammar it is hard to make a choice to adopt a representation without clear-cut, well-defined criteria. Instead, a grammar writer is advised to examine the full range of copula constructions and observe the behaviour of different constituents in the predicate position to check whether the copula is overt or non-overt, obligatory or optional, and whether the agreement between subject and predicate is manifested morphologically or not. Nevertheless, these criteria are considered as clues rather than measurable and definite tests. The theoretical challenge is that with three acceptable f-structure representations, generalizations about the predicational syntactic structures are not captured either cross-linguistically or inside the same language. We believe that this divergence is motivated at the c-structure level but not at the f-structure level which is supposed to provide a deeper representation. The presence vs. absence of a copula and the presence vs. absence of morphological features denoting agreement can be considered as parameters of variation across languages. By failing to reach a unified analysis we fail to represent the universal syntactic function of a non-verbal predicate.

Although Nordlinger and Sadler (2006) expressed their conviction that there is no *a priori* reason for copula constructions cross-linguistically to have the same syntactic structure and that it should be left as an empirical issue, they could not help raising the question again after surveying the typological differences in copula constructions:

> The fact that the choice of strategy in a given language can be influenced by superficial matters of grammatical encoding raises the interesting question as to whether the alternative strategies are externally distinct but correspond to the same f-structure. (Nordlinger and Sadler, 2006)

Different types of copula constructions can be considered as merely different strategies or "paradigmatic alternations" (Nordlinger and Sadler, 2006) to express the predicational relationship. This difference should be expressed in the c-structure level, rather than the f-structure level. Every language sets different conditions on the order of the constituents, how to separate between them and how to relate them to each other. By making the f-structure follow the trail of these strategies, we fail to capture the functional generalizations of the predicational construction crosslinguistically as well as inside the same language. We are proposing that f-structure should be grounded, as it is supposed to be, on a functional basis rather than a typological basis. Dyvik (1999) emphasised the idea that f-structures abstract away from constituent order typical of c-structures, and even assumed that f-structures are universal "in the sense that translationally corresponding expressions across languages are assigned the same (or closely similar) f-structures".

We propose that it is preferred to provide a unified analysis of the predication relations cross-linguistically, so that functional parallelism among functionally equivalent constructions can be maintained.

We believe that the source of confusion in arriving at a unified analysis of predicational constructions crosslinguistically is that most analyses are misled by the divergent surface representation and paradigmatic alternations and fail to capture the underlying generalizations.

The concept of parallel levels of representation is a basic assumption in LFG where the c-structure variations do not affect the status of grammatical functions, and that semantic roles are distinct from grammatical functions. For an example, the subject can be expressed in various ways in c-structure, it can be an NP clause, a CP clause, an affix on the verb or a zero-pronoun with no node in the c-structure, yet the grammatical function of SUBJ is assigned to all these variations as the f-structure represents a deeper level of representation. Furthermore the SUBJ can be assigned different semantic roles, as pointed out by the examples in (229) from Lødrup (2006).

(229) a. He ran home (agent SUBJ)
b. He fell down (theme SUBJ)
c. He fantasized (experiencer SUBJ)
d. There is a problem (non-thematic SUBJ)

The distinction between c-structure and f-structure has been maintained, to a great extent, in most syntactic structures, but with the obvious exception of the predicational constructions. Predicational structures are fundamentally similar, crosslinguistically, and yet they receive divergent f-structure analyses in LFG. We need to represent the predicate as a grammatical function that can have various c-structure representations, one grammatical function, and ultimately different semantic roles: predicative, equative, locational, temporal, etc.

Contrary to what is maintained by Dalrymple et al. (2004) that each language can choose either to make its predicates as closed or open complements, or even closed for some and open for the others, we propose that the predicational structures receive a default f-structure analysis that expresses the existence of subject (SUBJ) and predicate (PREDLINK) as primitive grammatical functions and to consider the use of a copula as a parameter of variation across languages. English uses a copula not because adjectives cannot subcategorize for subjects, but because English chooses the “copula +” parameter.

Case marking, word order and agreement features that hold between the subject and the predicate are parameters of variation across languages. It is also a matter of variation among languages to decide how to delimit the subject and predicate, perhaps only by juxtaposing the two elements or by inserting a pronominal or by using a copula verb. In his typological study of copula constructions Curnow (2000) points out that the choice of strategy for encoding the copula construction is conditioned by various factors.

> The choice of construction in these cases depends upon discourse and grammatical factors such as tense and aspect, polarity, the status of the clause as main or subordinate, the person of the Copula subject, and the semantic relation expressed (identification or classification). (Curnow, 2000, p. 2)

Some other syntactic theories have tended to recognize the copula constructions and treat them in a somewhat uniform way. Within the framework of HPSG, Avgustinova and Uszkoreit (2003) identified six types of copula constructions in Russian, only one of them (short adjectives, or adjectives which are lexically predicative) being given a marked analysis, while the rest receive the same representation, regardless of whether the copula is present or not, obligatory or not. The same tendency is expressed in the Minimalist approach by Adger and Ramchand (2003) where they analyzed the various copula constructions in Scottish Gaelic as having the underlying representation of Predicate Phrase (PredP).

The argument we propose for a unified analysis of copula constructions is based on the following premises:

1. The subject-predicate relationship is a universal grammatical relationship that is found cross-linguistically. Typological studies of copula constructions never reported the absence of this clause type in a certain language. Pustet (2003) reported that “serious arguments against the universality of the predicate function have never been proposed.”
2. The distribution of copulas varies crosslinguistically. This is a language-specific variation. Some languages use them along semantic lines, others along morpho-syntactic lines, others along lexical lines, etc.
3. Adjectives have a special affinity to nouns within constructions whether when they are used attributively or predicatively. This affinity does not obliterate their syntactic functions in the predicate position, or allow them to subcategorize for a subject.

Our chosen analysis is the double-tier analysis which uses the closed complement PREDLINK as a specialized grammatical function for the predicate. In defending our chosen analysis we discuss the other alternatives and examine their feasibility, and we also discuss the credibility of the objections raised against our analysis of choice. In doing so, we make use of the mechanisms already available in the framework of LFG without violating any fixed condition or breaking with any established conventions.

5.5.2 Divergent Strategies of Copula Constructions

Many languages have a copula verb that heads a copula construction, yet in many other languages constituents are merely juxtaposed and no copula verb is used. Typological studies (Curnow, 2000, Pustet, 2003) show that between these two poles there is a large spectrum of variation in the strategies used and constraints applied in the use of copula constructions. We will avail ourselves here of the increased attention that has been paid to the copula constructions in LFG and other syntactic theories, as well as typological studies. In this section we study the copula constructions in five selected languages in order to obtain a better understanding of the phenomenon and observe the interesting variety in the choice of strategies used in this relationship.

In this section we show how the interplay of syntax and semantics in the predicational constructions leads to the use of divergent strategies in the formation of copula clauses. Semantic considerations are significantly involved in the choice of the strategies employed in expressing the copula construction in many languages, or as Pustet (2003) puts it, “semantics conditions linguistic form”. This tight relationship between syntax and semantics is also observed by Adger and Ramchand (2003):

> … there is an extremely tight relationship between the syntax and semantics of predication, and that semantic predication always feeds off a syntactic structure containing a predicational head. (Adger and Ramchand, 2003, p. 325)

The languages we choose to analyse are Arabic, Russian (Avgustinova and Uszkoreit, 2003), Irish (Carnie, 1997), Chinese (Tang, 2001), and Scottish

Gaelic (Adger and Ramchand, 2003). These languages use divergent strategies and set various conditions on the construction of copula clauses. The main point we want to make through in this section is that copula constructions use different strategies to encode essentially one and the same grammatical function.

Arabic uses different strategies to express the predicational relationship. The two elements (subject and predicate) can merely be juxtaposed to express predicative and locational relations in the present tense, as in (230). When the predicate is an adjective it agrees with the subject in number and gender, as in (231)–(232).

(230) الرجل في الدار
ar-rağulu fī ad-dāri
the-man in the-house
‘The man is in the house.’

(231) الرجل كريم
ar-rağulu karīmun
the-man.sg.masc generous.sg.masc
‘The man is generous.’

(232) المرأة كريمة
al-marʾatu karīmatun
the-woman.sg.fem generous.sg.fem
‘The woman is generous.’

A pronominal must be inserted between the subject and the predicate in equative relations when both elements are definite, as in (233).

(233) أخي هو الطبيب
ʾaḫ-ī hwa aṭ-ṭabību
brother-my he the-doctor
‘My brother is the doctor.’

A copula verb is used in the past and future tenses, and also in the negated present, as shown in the examples in (234), (235) and (236) respectively.

(234) كان الرجل كريما
kāna ar-rağulu karīman
was the-man generous
‘The man was generous.’

(235) سيكون التقرير جاهزا
sayakūnu at-taqrīru ğāhizan
will-be the-report ready
‘The report will be ready.’

(236) ليس الرجل كريما
laisa ar-rağulu karīman
is-not the-man generous
‘The man is not generous.’

Russian (all examples taken from Avgustinova and Uszkoreit, 2003) also employs various strategies. The following example shows the Russian short adjective. This is the adjective which can only be used predicatively while its attributive use is not allowed. In the present tense the copula is not allowed, as in (237a), but must be used in the past and future tenses, as shown in (237b).

(237) a. On gord rezul’tatami.
he.NOM.SG.M proud.PRD-ADJ.SG.M results.INST.PL
‘He is proud of the results.’

b. On ne byl gord rezul’tatami.
he.NOM.SG.M not was proud.PRD-ADJ.SG.M results.INST.PL
‘He was not proud of the results.’

In the examples in (238) ordinary adjectives and nouns are used in predicative (ascription) constructions. The use of a copula verb in the present is unnatural while a copula must be used in the past and future tenses.

(238) a. On durak | tolstyj
he.NOM.SG.M fool.NOM.SG.M | fat.NOM.SG.M
‘He is a fool | fat.’

b. On byl durak | tolstyj
he.NOM.SG.M was fool.NOM.SG.M | fat.NOM.SG.M
‘He was a fool | fat.’

In equative (identificational) construction, as shown in (239), an overt copula can be used in the present tense. But in the absence of a copula the left periphery must be separated from the right periphery intonationally by a pause and orthographically by a dash. Still the past and future must use overt copulas.

(239) a. On est’ brat Maksima.
he.NOM.SG.M is brother.NOM.SG.M Maxim.GEN
‘He is Maxim’s brother.’

b. On – brat Maksima.
he.NOM.SG.M brother.NOM.SG.M Maxim.GEN
‘He is Maxim’s brother.’

In the localization (locational and temporal), as shown in (240), predicational constructions again the copula is unnatural in the present and is required in the past and future.

(240) Boris na sobranii.
Boris.NOM at meeting.LOC
'Boris is at a meeting.'

In predicational constructions denoting existence and possession, as shown in (241), the use of the copula is optional.

(241) a. Za uglom (est') magazin
behind corner.SG.M.INST (is) store.NOM.SG.M
'There is a store around the corner.'

b. U Kati (est') samovar.
at Katia.GEN (is) samovar.NOM.SG.M
'Katia has a samovar.'

In modern Irish (all examples from Carnie, 1997) there are two types of copula constructions according to whether the relation is predicative or equative. In the predicative construction, as shown in (242a), the copula verb is followed by the predicate which is followed by an optional agreement morpheme, and the subject comes in the final position. In the equative construction, as shown in (242b), the copula is followed by an obligatory agreement morpheme which is followed by the subject and the predicate comes last.

(242) a. Is dochtúir (é) Seamus
COMP doctor (AGR) Seamus
'Seamus is a doctor.'

b. Is é Seamus an captain
COMP AGR Seamus the captain
'Seamus is the captain.'

From the above examples we notice that Irish has two different strategies (word order and the agreement morpheme) in encoding the copula construction according the two different semantic domains. The semantic distinction between equative and predicative gives a straightforward explanation of the differences in word order and obligatory vs. optional presence of the agreement morpheme in Irish.

In Chinese (all examples from Tang, 2001) the copula verb *shi* is optional in predicative sentences, as shown in (243), and obligatory in specificational and equative sentences, as shown in (244).

(243) Zhangsan (shi) Zhongguoren.
Zhangsan be Chinese
'Zhangsan is a Chinese.'

(244) Wo mai de *(shi) zhe duo hua. (specificational)
I buy DE be this Cl flower
'What I bought is this flowers.'

Moreover, predicative copula constructions are constrained by more detailed pragmatic considerations. In the example in (245) the predicate expresses the speaker's opinion or attitude and the clause is grammatical. Contrastively, the example in (246) expresses a fact and, therefore, the clause is considered unnatural or incomplete.

(245) Zhangsan shagua.
Zhangsan fool
'Zhangsan is a fool.'

(246) ??Zhangsan xuesheng.
Zhangsan student
'Zhangsan is a student.'

There are certain conditions that must be realized to make the predicate in (246) more natural. For example the predicate can be modified by an evaluative adjective, as in the example (247), or specified by a noun in compounding construction to make the predicate more complete, as shown in (248).

(247) Zhangsan hao xuesheng.
Zhangsan good student
'Zhangsan is a good student.'

(248) Zhangsan daxue sheng.
Zhangsan university student
'Zhangsan is a university student.'

Scottish Gaelic (all examples from Adger and Ramchand, 2003) shows as well interesting variations. A copula construction is formed from an AP or PP in the predicate position, as shown by the examples in (249) and (250) respectively.

(249) Tha Calum faiceallach.
Be-PRES Calum careful
'Calum is (being) careful.'

(250) Tha Calum anns a'bhùth.
Be-PRES Calum in the shop
'Calum is in the shop.'

However, when an NP is placed in the position of the predicate the construction is ungrammatical, as shown in (251) below, and a preposition is needed, as in (252). The preposition incorporates a pronoun which agrees with the subject. This is explained by Adger and Ramchand by the fact that APs and PPs denote eventuality (stage level), while NPs lack eventuality (individual level). This is why an expletive preposition is needed.

(251) *Tha Calum tidsear.
Be-PRES Calum teacher
'Calum is a teacher.'

(252) Tha Calum 'na thidsear.
Be-PRES Calum in+3sg teacher
'Calum is a teacher.'

In predicative construction Scottish Gaelic can use an inverted structure where the predicate precedes the subject, as in (253).

(253) Is mòr an duine sin.
COP big that man
'That man is big.'

In equative constructions where a DP is used as a predicate, a third person masculine pronoun must be inserted after the copula, as in (254).

(254) 'S e Calum an tidsear
COP 3sg Calum (DP1) the teacher (DP2)
'Calum is the teacher.'

Adger and Ramchand (2003) assumed that the different forms of copula construction have essentially one underlying structure. They attribute the divergence in structure to the particular semantic specification of the predicate.

5.5.3 Adjectives as a Hybrid Category

With regard to the predicational construction, adjectives are receiving more attention in LFG, as well as other theoretical frameworks than any other constituent, to the extent of blurring the predicational relationship itself. The short form predicative adjectives in Russian have been considered as predicators (Avgustinova and Uszkoreit, 2003). They are also considered as the main head of the copula construction in HPSG (Adger and Ramchand, 2003). Similarly Nordlinger and Sadler (2006) draw evidence for the single-tier analysis of copula construction in LFG mainly from the behaviour of adjectives in some languages where they carry verbal morphology such as Abkhaz. Nevertheless they also emphasise that nominal predicates in some languages (such as Bininj Gun-wok) show verbal morphology.

Dalrymple et al. (2004) follow this trend and make a clear dichotomy between adjectives and other constituents in the predicate position by assuming that Japanese adjectives (where a copula is optional) function as the main head and subcategorize for the clausal subjects, whereas nouns (where a copula is always required) function as closed complements. Moreover they used agreement between predicative adjectives and subjects, as in the French examples in (255), as the main argument for the open complement analysis.

(255) Elle est petite.
she.fem.sg is small.fem.sg
'She is small.'

Therefore, we think that a special section on adjectives is motivated to account for the peculiar behaviour of adjectives and to put them in perspective to other constituents.

Syntactic and typological studies have viewed adjectives as a category that falls in the middle between nouns and verbs. Bresnan (1995) proposed a set of tests to distinguish adjective from verbs, and discussed the semantic and syntactic constraints that govern the conversion of verbs into adjectives. Beyssade and Dobrovie-Sorin (2005) on the other hand contrasted adjectives to nouns, stating that nouns denote sets of individuals while adjectives denote properties instantiated in individuals. Pustet (2003) in her typological study of the copula constructions has viewed adjectives as a hybrid category, with both verbal and nominal characteristics.

To put adjectives in perspective, we need to view the relationship between the subject and prototypical predicate as the relationship between a slot and filler, or analogically between a host and a guest. A host (analogous to the subject) can invite many guests (predicates), as illustrated in Table 9.

host/subject	copula	guest/predicate
the idea	is	a shamble good out of date in my head that we need more time affording more money

Table 9. The host-guest relationship between the subject and the predicate

One of the guests (the adjective) shows a special affinity with the host. This affinity is revealed as they have matching qualities (agreement) and they are sometime seen together without an intruder (short adjectives in Russian forbid the use of a copula verb). This, however, neither means that all other guests should be entangled in this affinity nor that the special guest is not a “guest”. This analogy means that the predicational relationship must be viewed across the board. All predicates stand in a functional predicational relationship to the subject as they all say something about the subject.

5.5.4 The single-tier analysis

Now we are going to go into the details of the different approaches to dealing with copula constructions in LFG, and we are going to question their validity one by one. The first approach is the single-tier analysis. In this approach the predicate (or the copula complement) is taken to be the head of the construction that subcategorizes for a SUBJ. Dalrymple et al. (2004) stated that this is the chosen analysis for Japanese adjectives in the predicate position where a copula is optional. In this case the adjective is considered the head whether the copula is overt or non-overt. The examples in (256) both have the same open function f-structure as shown in Figure 53.

(256) a. hon wa akai
book red
'The book is red.'

b. sono hon wa akai desu
this book red is
'This book is red.' (Dalrymple et al., 2004)

$$\begin{bmatrix} \text{PRED} & \text{'red<(}\uparrow\text{ SUBJ)>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'book'} \end{bmatrix} \end{bmatrix}$$

Figure 53. F-structure of a Japanese copula sentence

On the other hand, with Japanese nouns the copula is required and therefore it cannot accept an open function.

Dalrymple et al.'s (2004) argument for this analysis is that, as the copula is optional, the adjective provides the main PRED for the clause. They assumed that an adjective has a subcategorization power comparable to a verb.

> … the adjective is the syntactic head of the predicate phrase. If this is not considered a sufficient criterion for assuming that it subcategorizes for the (prototypical) subject of the sentence, then even the assumption that ordinary verbs subcategorize for subjects may be called into question. (Dalrymple et al., 2004, p. 191)

However, Dalrymple et al.'s (2004) analysis is unconventional, according to Harold Somers (personal communication 18 January 2008). Somers explains that Japanese adjectives belong to two subclasses, one of which (i-adjectives) has all the paradigmatic characteristics of verbs, and is potentially marked morphologically for tense, negation and politeness, while the other (na-adjectives) requires a copula. The adjective *Akai* means 'be red', since the word for 'red' is *aka*. The word *desu* in (256b) is just the polite form of the *-i* ending. The copula is needed in the polite form, not otherwise.

The main argument for the single tier-analysis in the case of Japanese sentences is that if the copula can be omitted then the complement is open, and if the copula cannot be omitted then the complement is closed. However, there are many reasons to counter this argument. First, this hypothesis fails to capture the generalization of the copular structure, and allows c-structure variations to penetrate into f-structure, which is supposed to give a deeper representation of the structure. We believe that it is important to view the syntactic position of the predicate in its totality. This position can be filled by an adjective, noun, preposition, adverb, or complement clause. Some constituents may have certain requirements, but the syntactic function is still the same.

Second, in our view, the presence vs. absence of a copula is not enough to motivate a divergent analysis for the same syntactic function. Copula use is conditioned in many languages according to numerous contexts; even in English the presence of the copula is not required in small clauses, such as the examples in (257).

(257) a. I consider him a monster.
b. I consider him to be a monster.

Predicates require overt/non-overt copulas depending on various criteria, such as the type of the constituent (adjective or noun, Japanese), tense (Arabic, Hebrew and Russian require an overt copula in the past and future), or formality (Japanese polite forms involve a copula). This shows that the requirement of an overt copula is triggered according to different conditions in different languages. So posing different syntactic representation fails to capture the generalization shared across these languages.

Third, while it is true that the adjective is a hybrid category (Pustet, 2003), the verb's power to project onto the sentence structure cannot in any way be rivalled by any other lexical item. Verbs are the "inherent predicators" (Avgustinova and Uszkoreit, 2003), and they are the uncontested predicators in the general case (Bresnan, 1995). Moreover, verbs and adjectives function in basically different relationships. In the subject-predicate clauses the predicate gives information about the subject, while in the verb–subject clauses, the subject is generally the doer of the action which in most cases carries the roles “volitional” and “agentive”.

Fourth, the predicate cannot be the head because it does not operate on the subject nor does it assign case to it. The evidence for this comes from Arabic. In Arabic, the verb assigns the nominative case to the subject and the accusative case to the object, and no other operator can override its power. Similarly, the preposition assigns the genitive case to the object, and no other operator can override its power either. However, in copula constructions the subject and predicate take the default case, i.e. the nominative case, as in (258).

(258) الرجل كريم
ar-rağulu karīmun
the-man.nom generous.nom
‘The man is generous.’

If the sentence is introduced by an affirmative particle, the subject takes the accusative case and the predicate remains unchanged, as in (259).

(259) إن الرجل كريم
ʾinna ar-rağula karīmun
indeed the-man.acc generous.nom
‘The man is indeed generous.’

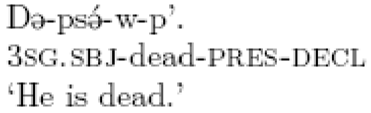
If the sentence is introduced by the copula verb كان *kāna* ‘was’ the predicate takes the accusative case and the subject remains unchanged, as in (260).

(260) كان الرجل كريما
kāna ar-rağulu karīman
was the-man.nom generous.acc
‘The man was generous.’

So, even though the subject and predicate remain adjacent, external operators can change their cases, which is not possible in any other governable relationship.

Nordlinger and Sadler (2006) pose a more powerful motivation for the single-tier analysis, that is the case of predicates which carry verbal morphology. In some languages the predicates carry morphological features (such as tense, mood and aspect) that are normally specifically indicated on verbs, but not on nouns. This is shown by the example from the Abkhaz language in (261) from Nordlinger and Sadler (2006).

(261)



Də-psə́-w-p'.
3SG.SBJ-dead-PRES-DECL
'He is dead.'

Avgustinova and Uszkoreit (2003), in their HPSG analysis of the copula constructions in Russian, present an attitude that is similar to the single-tier analysis in LFG. They assume that Russian short adjectives are "Lexically predicative non-verbal categories" that subcategorize for a subject. Short adjectives are distinct from all other constituents in two ways. First they are exclusively used as predicates, and their attributive use is ungrammatical. Second, an overt copula is not allowed with short adjectives in the present tense. This is shown by the example from Russian in (262) from Avgustinova and Uszkoreit (2003).

(262)	On	gord	rezul'tatami.
	he.NOM.SG.M	proud.PRD-ADJ.SG.M	results.INST.PL
	He is proud of the results.		

Unlike Avgustinova and Uszkoreit (2003) who analysed the predicate as a subcategorizing head in a single case only (short adjectives) while giving a different analysis to all other copula construction, Nordlinger and Sadler (2006) took the existence of a verbal morphology on adjectives and nouns as an evidence of the single-tier analysis in general, without restricting it to certain constituents or conditions.

In principle we need to allow grammatical functions to be expressed differently in different languages and in different contexts where there is a real motivation. For example, objects in one language can be rendered as obliques in another. So the existence of verbal morphology on the predicate may be considered enough in our estimation to trigger a single-tier analysis. In this case we say that the predicate expresses itself in a specific language and in specific conditions as a subcategorizing head, while for the rest of the constituents the relationship is expressed as a subject-predicate binary relationship. As Avgustinova and Uszkoreit (2003) pointed out the predicate position can be filled by various types of constituents to express different semantic roles such as equative, specificational, existential, locative, possessive, etc. So if we assume that a noun or an adjective may subcategorize for the main subject of the clause, how can we account for the subject when the predicate is a prepositional phrase or a complement phrase?

Regarding Avgustinova and Uszkoreit's (2003) analysis of short adjectives, we can counter their analysis with two arguments. First, the justification that short adjectives are used predicatively but not attributively may be motivated by semantic or pure lexical idiosyncrasies. Pustet (2003) points out that in English there are both adjectives that cannot be used attributively, as in (263), as well as adjectives that cannot be used predicatively, as in (264).

(263) a. The man is ready.
 b. * a ready man

(264) a. the former president
 b. * the president is former

In English also there is a whole class of adjectives that are restricted in their use. A participial adjective can serve in the attributive position but not the predicative position, as shown in the examples in (265) and (266). This can be explained as restrictions in the lexical properties of certain adjectives or structural constraints related to adjectival derivation, rather than representing different syntactic functions.

(265) a. an escaped prisoner
 b. * the prisoner is escaped

(266) a. a fallen leaf
 b. * the leaf is fallen

Second, the copula is used with short adjectives in the past and future tenses, as shown in (267) from Avgustinova and Uszkoreit (2003). This means that the short adjective's power as a main predicator is contested.

(267) On byl | budet gord rezul'tatami.
 he.NOM was | will-be proud.PRD-ADJ.SG.M results.INST.PL
 'He was | will be proud of the results.'

The strongest argument against the validity and general applicability of the single-tier analysis is put forward by Nordlinger and Sadler (2006), that is the case of tense stacking in languages such as Tariana, where there are two sets of tense affixes: one marking independent nominal tense, and the other marking propositional tense, as shown in (268).

(268) Pi-ya-dapana-miki-Ri-naka.
2SG-POSS-house-PST-NF-PRES.VIS
‘This is what used to be your house (I can see it).’ (Tariana: Nordlinger and Sadler, 2006 citing Aihkenvald, 2003)

Nordlinger and Sadler (2006) emphasise that a single-tier analysis of such constructions will result in a conflict in the tense feature, and that it must be analysed as a double-tier construction where there are two levels of f-structure: one level stands as the locus of the nominal tense and the other level the locus of the propositional tense.

5.5.5 The double-tier open function analysis

Now we are going to investigate the second approach for analysing the copula constructions to check its validity. The double-tier analysis is different from the single-tier analysis, as noted earlier, in that in the double-tier analysis the predicate is not considered as the clausal head, or main predicator. The predicator is either the copula, when it is present, or a higher structure (dummy predicate) when no copula is used. Nordlinger and Sadler (2006) did not delve into the investigation of the distinction between the two variants of the double-tier approach, i.e. open and closed copula complements, and represented both types simply as GF.

Dalrymple et al. (2004) consider that the open function XCOMP analysis is the chosen representation for languages where the predicate shows agreement with the subject, and cite the French example, reproduced as (269), for which they proposed the f-structure reproduced as Figure 54.

(269) Elle est petite. (Dalrymple et al., 2004)
she.F.SG is small.F.SG
'She is small.'

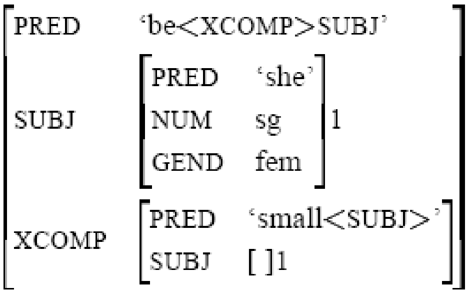


Figure 54. An open complement f-structure of a French copula sentence

Dalrymple et al. argue that the motivation for this analysis is first that "the adjective simply agrees with its own SUBJ, in the same way as verbs do." Second, the XCOMP analysis allows us to write simple and standard control equations, as in (270) on the lexical entry of the adjective to specify the agreement features.

(270) petite (↑ PRED) = 'small <SUBJ> '
(↑ SUBJ NUM) =c sg
(↑ SUBJ GEND) =c fem

They (*ibid.*) maintained that the closed complement PREDLINK analysis, shown in Figure 55, will result in non-standard inside-out control equations, shown in (271).

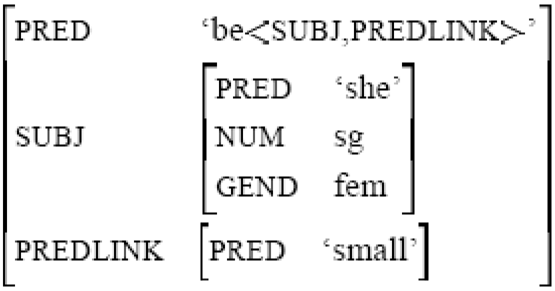


Figure 55. A closed complement f-structure of a French copula sentence

(271) petite (↑ PRED) = 'small'
((PREDLINK ↑) SUBJ NUM) =c sg
((PREDLINK ↑) SUBJ GEND) =c fem

Third, they assumed that “the XCOMP analysis allows for a much simpler analysis and one which is similar to that of other cases of subject-predicate agreement, such as subject–verb agreement.”

Unfortunately, all of these motivations are questionable. First French adjectives do not agree in the same way as verbs. French verbs agree in person with their subjects while adjectives do not. In our view agreement alone is not enough to justify the claim that the predicate subcategorizes for the subject. Agreement is a relation that holds between a verb and subject, and also between a noun and adjective, a noun and relative pronoun, a noun and demonstrative pronoun, etc. Dalrymple et al. themselves questioned the feasibility of agreement alone as a reason for justifying an open function.

> In other languages, however, some considerations may weaken the status of agreement as an argument for assuming an XCOMP analysis. In languages like Norwegian, for example, there is no subject-verb agreement, so that subject-adjective agreement must be treated differently from subject-verb agreement in any case. Another issue is that predicative adjective agreement may be governed by semantic rather than syntactic features. (Dalrymple et al., 2004, p. 196)

It is quite reasonable to maintain that agreement between subject and predicate is governed by the semantics rather than the syntax. This is why the English example in (272b) is ungrammatical while the others are acceptable. This shows that agreement here is not captured merely through grammatical rules.

(272) a. They are doctors.
b. *They are a doctor.

(273) a. They are the cause of our trouble.
b. They are a big problem.

Second, simple standard equations can be written to specify the agreement relation without the inside-out non-standard ones. But the equation need not be written in the lexical entry of the adjective, as it is practically and theoretically implausible to say that the lexical entries of all adjectives and nouns subcategorize for subjects and that they agree with the subject. We adhere to

Rosén's (1996) view that the relation between the subject and predicate is governed by the structure and so the agreement specifications must be written in the phrase structure rules.

> In Maori, the first NP is the predicative complement and the second is the subject. Since this information comes from the syntax and not from the lexicon, it might seem natural to let the phrase structure rule for this sentence type introduce a PRED that could subcategorize for these functions. (Rosén, 1996)

As we adopt a constructional approach to the copula clauses, we believe that the agreement equation should be placed in the phrase structure instead, as in (274).

(274)	S →	NP	VCop	∨	ϵ	NP ∨ AP
		(↑ SUBJ)=↓	↑=↓		(↑ PRED)='be<SUBJ,PREDLINK>' (↑ TENSE)=pres	(↑ PREDLINK)=↓ (↓ GEND)=(↑ SUBJ GEND) (↓ NUM)=(↑ SUBJ NUM)

Third we do not need to analyse copula constructions in the same way as subject–verb constructions as they are syntactically, semantically and typologically different. They use different syntactic structures cross-linguistically to denote different sorts of relationships and semantic roles. We need to formalise the analysis of the predicational constructions instead of making them a subset of the subject–verb constructions. Subject–predicate constructions are fundamentally different from subject–verb constructions in the following ways.

1. They express relations rather than actions or events.
2. They are usually shorter.

> Verbless [copula-less] clauses differ from verbal clauses (apart from the use of the verb) chiefly in the number of constituents used. Verbal clauses often have, beside the verb and its subject, several constituents which modify the verb, and are related to each other only through their relationship to the verb. Verbless clauses are typically composed only of two constituents, which are in some way equated by the structure. (Revell, 1989, p. 1)

3. They use a semantically void copula verb or no verb at all.

It has been the tradition of generative grammar to treat copula verbs as raising verbs (Adger and Ramchand, 2003, Carnie, 1997). However, we believe that raising verbs as in *He seems nice* should be treated as quasi-copulas instead.

The most compelling evidence against the general applicability of the open function comes from Dalrymple et al. (2004) who maintained that a closed complement analysis is mandated when the predicate already has a verb, such as the *that*-clauses, (275a); gerunds, (275b); and infinitival clauses, (275c) (examples from Dalrymple et al., 2004). In these instances the predicate already has a subject distinct from the subject of the main clause.

(275) a. The problem is that they appear.
b. The problem is their appearing.
c. The problem is (for them) to leave before 6.

They show that the XCOMP analysis requires the subject of the main clause to be the subject of the predicate, and this results in a clash, as shown in Figure 56.

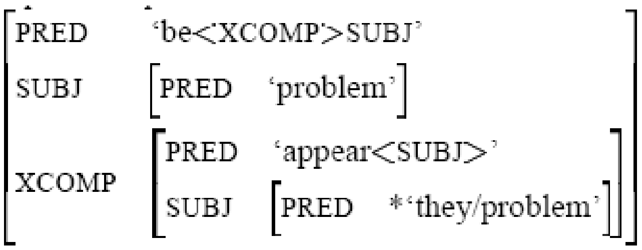


Figure 56. F-structure with a conflicting subject (Dalrymple et al., 2004)

Therefore a closed complement analysis, as shown in Figure 57, is compulsory to avoid this clash.

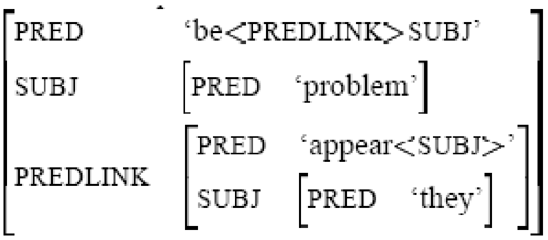


Figure 57. F-structure with a no conflict (Dalrymple et al., 2004)

5.5.6 The double-tier closed function as the chosen analysis

This is the third approach for analyzing the copula constructions in LFG. In our opinion this is the best possible representation as no serious challenges have been given against the general applicability of this analysis. Dalrymple et al. (2004) emphasize that the closed complement analysis is the chosen account for English copula constructions.

> In English … an adjective cannot occur on its own as the syntactic head of a predicate; a copula is always required. This provides a functionally-motivated account of the existence of the copula: it is needed because the adjectives themselves are unable to combine directly with overt SUBJs …
>
> Given this, the copula can be seen as giving to the adjective a needed grammatical prothesis: a SUBJ argument to which to link the adjective's semantic role. This analysis entails that the syntactic head of the predicate is the copula, not the adjective. (Dalrymple et al., 2004, p. 193)

We also maintain that the closed complement analysis is the default syntactic representation for all languages. The presence vs. absence of a copula, presence vs. absence of agreement features on the predicate are all paradigmatic alternations that do not affect the syntactic function. It is also the only account which succeeds in providing a valid representation for all constituent types which take various semantic roles, as shown in Table 10.

Example	**Constituent type of the complement**
He is a doctor.	Noun
He is good.	Adjective
He is here.	Adverb
He is in the garden.	PP
The idea is that we need more time.	CP

Example	**Semantics of the complement**
He is a doctor.	Predicative
He is my father.	Equative
This is what we want.	Specificational
The meeting is tomorrow.	Temporal
He is in the garden.	Location

Table 10. Constituent types and semantic roles of copula complements

Only the closed function analysis allows for a unified account of the predicational phenomenon. Other accounts which assume that the predicative adjective is a head subcategorizing for the subject definitely find it harder to do so with other constituents such as NP and PP. Lødrup (2006) proposes for sentences like (276) to manipulate lexical rules as in (277) to make nouns and prepositions subcategorize for subjects.

(276) a. The pills made him a monster
b. She seems in a bad mood

(277) a. 'monster' => 'be-a-monster<(↑ SUBJ)>'
b. 'in<(↑ OBJ)>' => 'be-in-state-of<(↑ SUBJ) (↑ OBJ)>'

The analysis, however, sounds unnatural and unnecessarily complex. Both Dalrymple et al. (2004) and Rosén (1996) agree on the fact that common nouns should not be considered as taking a subject in their argument structures.

> This [requiring a subject argument] does not seem implausible for adjectives, especially in languages such as French with adjectival agreement, but is less so for PPs and particularly for NPs. That is, it seems unlikely that every NP in a given language, regardless of the syntactic construction in which it appears, requires a subject. (Dalrymple et al., 2004, pp. 197-198)

> And in any case, this analysis [having the PRED of the NCOMP subcategorize for a SUBJ] would mean that all nouns would have to be subcategorized for subjects, which is certainly not desirable. (Rosén, 1996)

The closed complement analysis is also the best representation for verbless constructions. A large number of languages do not use a copula verb to express the predicational relationship.

> The class of languages which contain be-less sentences is widespread; it includes languages from practically every language family and from every continent. (Carnie, 1995, p. 251)

In the analysis of copula-less languages we do not assume that a copula verb is elided, we consider that the relationship is intrinsically expressed merely by juxtaposing the constituents. In Maori a copula verb is never used, but the relationship is expressed by the grammatical construction as a whole (Rosén, 1996). Therefore constituents are not related through a verb, either overt or non-overt, but through the structure of the clause, as further emphasized by Butts (2006) for Aramaic.

> Nexus can be expressed, however, by means other than a finite verb. In Aramaic, the verbless clause, that is, a clause lacking a finite verb as core constituent, is defined as a clause in which nexus is expressed not by a finite verb, but by the syntactical juxtaposition of subject and predicate. (Butts, 2006, p. 56)

In our view, no special treatment of copula-less constructions is considered necessary, as they are semantically and functionally equivalent to constructions where overt copulas are used.

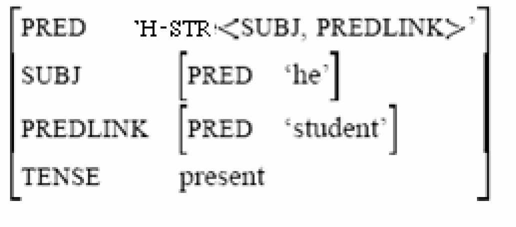
> … verbless constructions … are generally functionally equivalent (or at least, in functional overlap with) with copula constructions in other languages (or even within the same language). (Nordlinger and Sadler, 2006)

The presence or absence of a copula is a parameter of variation. The copula itself is considered semantically redundant. In the typological and syntactic literature the copula verb has been described as “light”, “bleached” and “semantically void”.

We adopt Nordlinger and Sadler’s (2006) account of the copula-less construction as involving a higher structure. So we assume that the main predicator is “H-STR” for “Higher-STRucture” instead of “be” in the LFG literature which entails the assumption that there is an elided *be*-like verb. In many languages the mere juxtaposition of subjects and predicates is enough to express the predicational relationship without assuming elision of the copula verb. Further, it might be questioned why a predicator is needed after all if the clause is composed of two juxtaposed constituents with no elliptical copula. However, we need a predicator not only to satisfy the coherence condition in LFG, but also to state the fact that a grammatical sentence is composed of a subject and a predicate, nothing more, nothing less. A predictor is also needed to convey sentential information such as tense and negation. So for the Arabic example in (278) we have the phrase structure rules in (279) and the f-structure in Figure 58.

(278) هو طالب
hwa ṭālibun
he student
‘He is a student.’

(279)
$$\begin{array}{llllll} \text{S} \rightarrow & \text{NP} & \text{VCop} & \vee & \epsilon & \text{NP} \vee \text{AP} \\ & (\uparrow \text{SUBJ})=\downarrow & \uparrow=\downarrow & & (\uparrow \text{PRED})=\text{'H-STR<SUBJ,PREDLINK>'} & (\uparrow \text{PREDLINK})=\downarrow \\ & & & & (\uparrow \text{TENSE})=\text{pres} & (\downarrow \text{GEND})=(\uparrow \text{SUBJ GEND}) \\ & & & & & (\downarrow \text{NUM})=(\uparrow \text{SUBJ NUM}) \end{array}$$



$$\begin{bmatrix} \text{PRED} & \text{'H-STR<SUBJ, PREDLINK>'} \\ \text{SUBJ} & \begin{bmatrix} \text{PRED} & \text{'he'} \end{bmatrix} \\ \text{PREDLINK} & \begin{bmatrix} \text{PRED} & \text{'student'} \end{bmatrix} \\ \text{TENSE} & \text{present} \end{bmatrix}$$

Figure 58. F-structure of an Arabic copula sentence

We also consider that SUBJ and PREDLINK are primitive grammatical functions that denote the subject and predicate in the universally acknowledged predicational construction.

We conclude that a unified analysis of copula constructions is motivated as all different strategies employed in the predicational structures basically express the same grammatical function. The assumption that the copula complement is closed PREDLINK enables us to account for all constituents that can occupy the predicate position and express cross-linguistic generalizations related to functional use of copula constructions.

6 Syntactic Disambiguation

Ambiguity is a major problem for large-scale computational grammars. A grammar which covers a realistic and representative portion of a natural language produces a large number of possible parses, most of them unintuitive to humans (King et al., 2000).

Kuhn and Rohrer (1997) pointed out that ambiguity poses a problem for grammar writers, for the parsing systems and for the applications based on the parsers. It becomes hard and time consuming for a grammar writer to inspect all the solutions by hand when the grammar produces hundreds and sometimes thousands of parse trees. Ambiguity also causes an efficiency problem to parsers, as the more ambiguities are produced the longer the time spent on processing and the heavier the load on the system memory. The applicability problem arises from the fact that almost all applications need one analysis per sentence, and with the increased number of ambiguities there is a reduced possibility that the first solution will be the best or the most correct solution.

Sometimes there is a correlation between ambiguity and parse time, and sometimes each is not affected by the other. Yet it can be said that the work that aims at reducing parse time falls under the category of ambiguity management. Some ambiguities may be computationally time-consuming and yet they do not surface as valid solutions. This usually happens when the number of subtrees increases dramatically, but they do not make their way up as valid trees.

Ambiguity is a problem faced both by hand-crafted rule-based grammars as well as Probabilistic Context-Free Grammars (PCFGs). While rule interaction is mainly responsible for the ambiguity in rule-based grammars (King et al., 2000), ambiguity in PCFGs also remains after the probability estimates are made. Trees with the least probability scores are discarded while trees bearing the maximum probability are singled out as candidate analyses and the parser has to choose in a non-deterministic way from among them (Infante-Lopez and Rijke, 2004).

Therefore PCFGs have to deal with the ambiguity problem to reduce the size of candidate trees and consequently reduce the level of non-determinism.

Structural ambiguity resolution is a central issue in natural language analysis. A sentence is structurally ambiguous if it can be represented by more than one syntactic structure. Ambiguity appears as a daunting problem especially for large-coverage grammars, where the number of parse trees grows dramatically as the number of rules and lexical readings increases. It is practically impossible to eliminate the ambiguities altogether, yet it always remains the task of grammar writers to try keep the ambiguity rate within a manageable boundary.

Ambiguity is an inherent characteristic of human languages. When we see the words in (280)–(282) in isolation, we cannot determine the intended meaning. It is only by drawing from contextual, probabilistic and real world knowledge clues that we are able to interpret such phrases. Computational analysis of human language is even more complicated, as there are, beside the real ambiguities, system ambiguities that result from the interaction of rules and the competition of constraints.

(280) عالم
ʿālam / ʿālim
'world/scientist'

(281) الديمقراطية
ad-dīmqrāṭīh
'democratic/democracy'

(282) صدام
ṣidām / ṣaddām
'conflict / Saddam'

MacDonald et al. (1994) maintain that disambiguation involves activating one alternative of a given type and inhibiting all others. They view this as a winner-take-all process. They point out that ambiguity is resolved in terms of a competition model, which assumes that languages provide cues that interact (or "compete") with one another during processing in order to select a certain interpretation and inhibit the others. They also believe that there are contingencies among different representations (lexical, syntactic and semantic), and therefore disambiguation must be achieved at all levels of representation.

While working on ambiguity in this section we use a test suite to test the effects of the different methods and techniques in reducing ambiguity. This test suite contains 254 real sentences basically used as a reference for development. The test suite was collected during the various stages of the grammar development from a corpus of news articles from the Al-Jazeera web site. At one stage 79 real sentences of various lengths were selected from four news articles. The shortest sentence was three words and the longest one was 46, and they show a wide variability in the complexity level of syntactic structures. In a subsequent stage we randomly selected 175 sentences ranging between 10 and 15 words to be included in the test suite. The sentences in this category share the same characteristics; they tend to be simpler and to avoid deep embedding.

When testing our grammar against the test suite we found that the average number of optimal solutions (preferred solutions that surface after applying optimality ranking techniques, explained in detail in 6.4.2) is 135 and the average number of suboptimal solutions is 1.45E+04. When testing the English grammar on 44 randomly selected sentences from the BBC news website we found that the average number of optimal solutions is 243, and that the average number of suboptimal solutions is 9.48E+08. Therefore if we take the English grammar as average, we find that our grammar is less than average with regards to the number of ambiguities.

We believe, however, that this comparison is neither indicative nor meaningful. First, the Arabic test suite was used as a reference for development and not randomly selected. This makes the ambiguity rate lower than could be expected. Second, the Arabic test suite includes a number of short sentences (between 10 and 15 words in lengths). Third, comparing ambiguity between different grammars is, in essence, not possible, as John Maxwell pointed out (personal communication (email), 7 June 2007).

> I think that it is difficult to compare ambiguity rates between different languages using the ParGram grammars. This is because the ambiguity rate of a grammar varies over time. When a grammar is first being developed, there are usually just a few ambiguities per sentence. As the grammar matures, there are more and more ambiguities caused by rare constructions. So the number of ambiguities per sentence goes up. If this starts to bother the grammar writer, then he or she will add dispreference marks to eliminate rare constructions if they are not called for. So the ambiguity rate goes down again. Since it is hard to know what state a grammar is in, it is hard to know whether the variation in ambiguity between grammars is caused by the different languages or the different states that the grammars are in.

We start this section by identifying sources of syntactic ambiguities in Arabic. We then move on to explore the full range of tools and mechanism implemented in the XLE/LFG literature in ambiguity management, showing how they were applied to our Arabic grammar.

In this research we deal with ambiguity not as one big problem, but rather as a number of divisible problems spread over all levels of the analysis. The task of handling the ambiguity problem is dealt with in three stages. The pre-parsing stage contains all the processes that feed into the parser whether by splitting a running text into manageable components (tokenization), analyzing words (morphological analyzer) or tagging the text. These processes are at the bottom of the parsing system and their effect on ambiguity is tremendous as they directly influence the number of solutions a parser can produce. The pre-parsing stage covers the topics of tokenization, morphological analysis, MWEs, shallow mark-up, and fully-blown pre-bracketing.

The parsing stage is the process when the syntactic rules and constraints are applied to a text, and the subcategorization frames are specified. The discussion of the parsing phase covers the issues of granularity of phrase structure rules, lexical specifications, application of syntactic constraints, and domain specific adaptation.

The post-parsing stage has no effect on the number of solutions already produced by the parser, but this stage only controls the selection and ranking of these solutions. The post-parsing phase involves packing ambiguities, optimality marks for preferences, using discriminants, and stochastic disambiguation.

6.1 Sources of Syntactic Ambiguity in Arabic

Sources of syntactic ambiguities are identified by King et al. (2000) as rule interactions and alternative definitions of lexical entries. Building on these suggestions we find that structural ambiguities can be boiled down to three areas. The first area is alternative c-structure rule interactions which define word order variations, phrasal attachment and scope of coordination. The second area is disjunctions in f-structure descriptions which specify phrases with alternative feature values and phrases with alternative grammatical functions. The third area is lexical entries which describe alternative parts of speech, alternative subcategorization frames, alternative morphological features, and the choice between MWEs and compositional interpretation.

However, it is not usually possible to point at a certain ambiguity and say definitively that the source of ambiguity is in one or the other domain, as ambiguity in one field usually propagates across the other fields.

Daimi (2001) highlighted the idea that the problem of ambiguity in Arabic had not received enough attention by researchers. This, to a great extent, is still the case today. Although most aspects of the ambiguity problem are shared among human languages, it is still worthwhile to show how the special characteristics of a certain language contribute towards increasing or reducing ambiguities.

Daimi (2001) pointed out that many of the ambiguous English cases discussed in the literature do not necessarily apply to Arabic at all, and cited the example in (283) where the pronoun *her* causes an ambiguity in English as it can be interpreted as either accusative or genitive, but in Arabic the pronoun is not ambiguous as it will either be cliticized to the verb or the noun.

(283) I saw her yesterday ~ I saw her cat

Daimi (2001) further emphasised the idea that ambiguities are not parallel cross-linguistically and that when translating a sentence from a source language to a target language, there are four possibilities:

(a) unambiguous source sentence	→	unambiguous target sentence
(b) unambiguous source sentence	→	ambiguous target sentence
(c) ambiguous source sentence	→	unambiguous target sentence[3]
(d) ambiguous source sentence	→	ambiguous target sentence[4]

This is why the ambiguity problem should be investigated in each language in its own terms. Each language has its own peculiarities and idiosyncrasies, and therefore ambiguities are distributed and resolved differently in each language.

Arabic has its particular weak spots which are prone to produce a great deal of ambiguities, and which must be handled with special attention. In this section we are going to focus specifically on four ambiguity-generating areas in Arabic which, in our estimation, have the greatest impact. These are the pro-drop nature of the language, word order flexibility, lack of diacritics, and the multifunctionality of Arabic nouns.

6.1.1 Pro-drop Ambiguity

A great deal of ambiguity is caused by the pro-drop nature of the Arabic language. The pro-drop theory (Baptista, 1995, Chomsky, 1981) stipulates that a null category (*pro*) 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. In Arabic the subject can be explicitly stated as an NP or implicitly understood as a pro-drop. Arabic has rich agreement morphology. Arabic verbs conjugate for number, gender and person, which enables the missing subject to be reconstructed. A syntactic parser, however, is left with the challenge to decide whether or not there is an omitted pronoun in the subject position (Chalabi,

[3] Presumably this means that each disambiguated reading of the SL sentence is unambiguous in the TL.

[4] In this case the ambiguity in the target language might be “the same” as in the source language or, by coincidence the disambiguated readings could themselves lead to ambiguous TL sentences, as in (b).

2004b). The challenge to decide whether there is a pro-drop or not comes from the fact that many verbs in Arabic can be both transitive and intransitive. In case these verbs are followed by only one NP the ambiguity arises. We can explain this using the example in (284).

(284) قاوم الجندي
qāwama al-ǧundī
resisted.masc.sg the-soldier

In (284) we are not sure whether the NP following the verb is the subject (in this case the meaning is ‘The soldier resisted’) or it is the object and the subject is an elliptic pronoun meaning ‘he’ and understood by the masculine mark on the verb (in which case the meaning will be ‘He resisted the soldier’). This ambiguity is caused by three facts: first there a possibility for a pro-drop subject following Arabic verbs, second the verb *qāwama* ‘resisted’ can be both transitive and intransitive, and third the agreement features on the verb match the post-verbal NP which makes it eligible to be the subject. This ambiguity results in two analyses as shown in Figure 59 and Figure 60. In the pro-drop case, the person, number and gender morphosyntactic features on the verb are used to reconstruct the number, gender and person features for the “pro” subject.

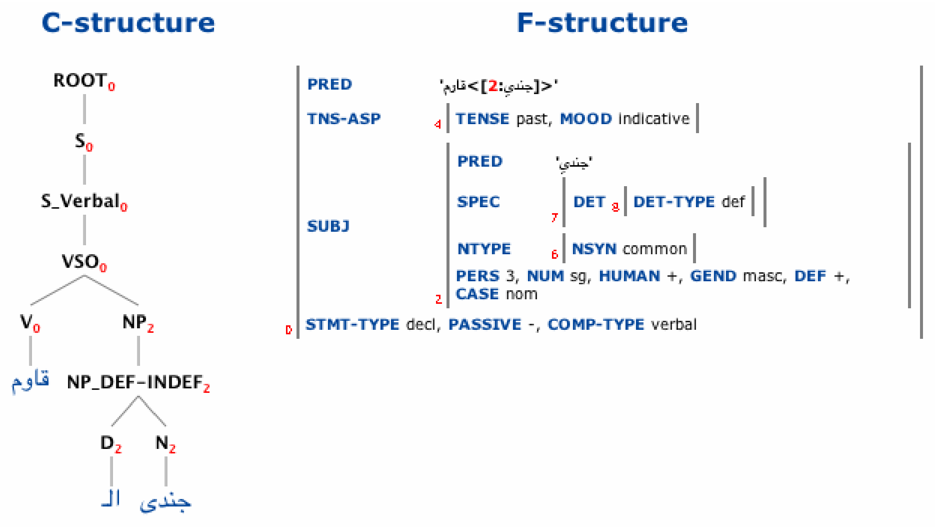

C-structure
ROOT0
S0
S_Verbal0
VSO0
V0
NP2
قاوم
NP_DEF-INDEF2
D2
N2
الـ
جندى
F-structure
PRED 'قاوم<[2:جندي]>'
TNS-ASP 4 TENSE past, MOOD indicative
SUBJ
PRED 'جندي'
SPEC 7 DET 8 DET-TYPE def
NTYPE 6 NSYN common
PERS 3, NUM sg, HUMAN +, GEND masc, DEF +,
2 CASE nom
0 STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal

Figure 59. First analysis of a possible pro-drop sentence

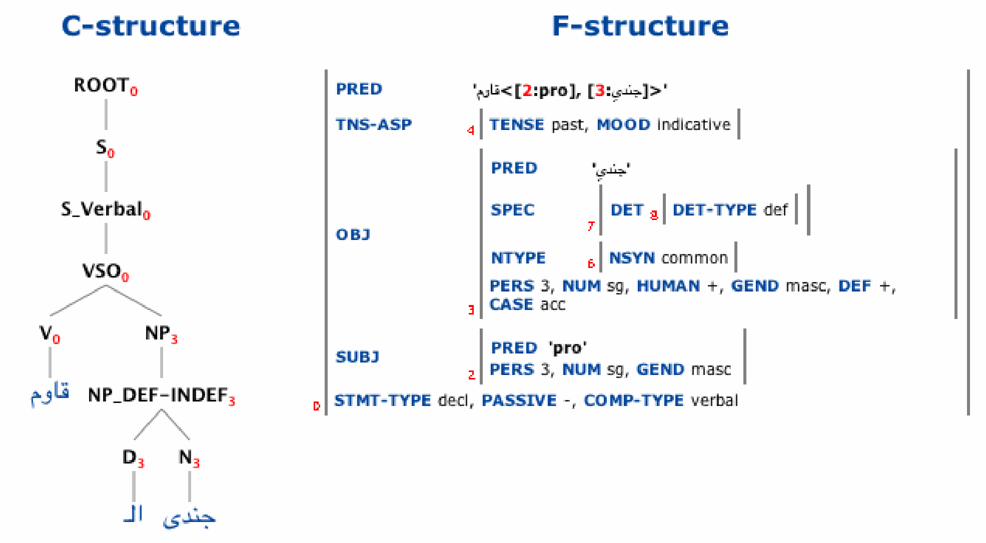

C-structure
F-structure
ROOT0
S0
S_Verbal0
VSO0
V0
NP3
قاوم
NP_DEF-INDEF3
D3
N3
الـ
جندى
PRED 'قاوم<[2:pro], [3:جندي]>'
TNS-ASP 4 TENSE past, MOOD indicative
OBJ
PRED 'جندي'
SPEC 7 DET 8 DET-TYPE def
NTYPE 6 NSYN common
3 PERS 3, NUM sg, HUMAN +, GEND masc, DEF +, CASE acc
SUBJ 2 PRED 'pro' PERS 3, NUM sg, GEND masc
0 STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal

Figure 60. Second analysis of a possible pro-drop sentence

6.1.2 Word Order Ambiguity

A lot of ambiguities are also caused by the relatively free word order in Arabic. Arabic allows VSO, SVO and VOS constructions, as shown in (285), (286), (287) respectively. While SVO is easily detected by the parser and usually does not cause an ambiguity problem, VOS gets mixed up with VSO. The difference between the nominative and accusative cases which normally distinguish the subject and the object is a matter of diacritics, which do not show in the surface forms as they are usually omitted in modern writing. This means that every VSO sentence has a VOS interpretation causing a serious ambiguity problem. In our grammar allowing VOS beside VSO without any constraints almost doubled the number of ambiguities for 15% of the sentences.

(285) أكل الولد التفاحة (VSO sentence)
ʼakala al-waladu at-tuffāḥata
ate the-boy.nom the-apple.acc
'The boy ate the apple.'

(286) الولد أكل التفاحة (SVO sentence)
al-waladu ʼakala at-tuffāḥata
the-boy.nom ate the-apple.acc
'The boy ate the apple.'

(287) أكل التفاحة الولد (VOS sentence)
ʼakala at-tuffāḥata al-waladu
ate the-apple.acc the-boy.nom
'The boy ate the apple.'

The VOS word order, however, is not a frequent construction and the sentence in (287) will sound unusual in Modern Standard Arabic (MSA), while normally perfect in Classical Arabic. Nevertheless, the construction still occurs in MSA but is marked by some constraints. There are certain conditions that allow the object to come before the subject. One of these conditions is when the object is a pronoun, as in (288).

(288) شكرهم الولد
šakara-hum al-waladu
thanked-them the-boy
‘The boy thanked them.’

Ryding (2005) stipulates as a condition for allowing the object to precede the subject that “**the object is substantially shorter** than the subject” (emphasis in original), as in (289) and (290). This condition, however, is not easily stated as a constraint in a computational grammar. A more precise condition might be when the object is definite and the subject is indefinite the object is allowed to precede the subject. When the subject is indefinite it tends to be modified by adjectives or prepositional phrases leading to the apparent length criterion.

(289) كتب التقرير فريق من المختصين
kataba at-taqrīra farīqun min al-muẖtaṣṣīn
wrote the-report team of the-specialists
A team of specialists wrote the report.

(290) غطى أحداثها عشرون ألف صحفي (Ryding, 2005)
ġaṭṭā ʼaḥdāṯu-hā ʿišrūna ʼalfa ṣaḥafī
covered events-its twenty thousand reporter
Twenty thousand reporters covered its events.

Moreover, subjects normally precede oblique objects, as in (291), but this is not always the case. Buckley (2004) pointed out that the prepositional phrase (oblique) precedes an expressed subject under two particular conditions: when the prepositional phrase contains a personal pronoun and when the subject is indefinite. This is shown in (292) and (293). Similarly Badawi et al. (2004) noted that word order is affected by reasons of emphasis as well as the tendency for the heavy (definite) element to precede the light (indefinite) element.

(291) وافق البرلمان على الاقتراح
wāfaqa al-barlamānu ʿalā al-iqtirāḥi
agreed the-parliament on the-motion
'The parliament agreed on the motion.'

(292) دخلت إلى المنزل فتاة (Buckley, 2004)
daḫalat ʾilā al-manzili fatātun
entered into the-house girl
'A girl entered the house.'

(293) ابتسم لها الطبيب (Buckley, 2004)
ʾibtasama la-hā aṭ-ṭabību
smiled to-her the-doctor
'The doctor smiled at her.'

Oblique objects also normally follow other objects, as shown in (294). However, obliques tend to precede objects when obliques contain a pronoun or when the object is indefinite, as shown in (295) and (296).

(294) أنفق مالا كثيرا على الرحلة
ʾanfaqa malan kaṯīran ʿalā ar-riḥlah
spend.past.sg.masc money.sg.masc plentiful.sg.mas on the-trip
'He spent a lot of money on the trip.'

(295) لم أتلق منه أي رد (Buckley, 2004)
lam ʾatalaqqa min-hu ʾayya raddin
not received.1st from-him any reply
'I didn't received from him any reply.'

(296) قلت لمنيرة كل الحقيقة (Buckley, 2004)
qultu li-munīrata kulla al-ḥaqīqati
said.1st to-Muneera all the-truth
'I told Muneera the whole truth.'

This exchangeability in position between obliques and subjects and objects is complicated by the fact that parenthetical phrases (mostly PPs) can appear virtually anywhere in the sentence, as shown in (297)–(299). This allows PPs in many instances to have alternate interpretations, i.e. either as obliques or parenthetical phrases.

(297) أشار الرئيس في كلمته إلى موقفه من الاحتلال
ʾašāra ar-raʾīsu fī kalimati-hi ʾilā mawqifi-hi min al-iḥtilāli
hinted the-president in speech-his to stand-his from the-occupation
'The President hinted, in his speech, at his stance against the occupation.'

(298) وصل المدير إلى الاجتماع في الوقت المناسب
waṣala al-mudīru ʾilā al-iğtimāʿi fī al-waqti al-munāsibi
arrived the-manager to the-meeting in the-time the-suitable
‘The manager arrived at the meeting at a suitable time.’

(299) وفي لندن عبر المسئولون عن قلقهم
wa-fī landan ʿabbara al-masʾūlūna ʿan qalaqi-him
and-in London expressed the-official of worry-them
‘And in London the official expressed their worry.’

Word order flexibility also affects copula constructions. Copula sentences normally follow the order of subject and predicate, as shown in (300), but they can also be inverted allowing the predicate to come before the subject. This, as well, must be properly constrained, otherwise it will lead to an exploding number of ambiguities. Among these constraints are when the subject is a CP (Badawi et al., 2004), as in (301). Another condition is when the subject is indefinite and the predicate is a PP or an existential adverb (Ryding, 2005), as in (302) and (303) respectively.

(300) الفتاة في الدار
al-fatātu fī ad-dāri
the-girl in the-house
‘The girl is in the house.’

(301) على بوش أن يساعد الفلسطينيين
ʿalā būš ʾan yusāʿida al-filisṭīniyyīn
on Bush to help the-Palestinians
‘Bush must help the Palestinians.’

(302) في الدار فتاة
fī ad-dāri fatātun
in the-house girl
‘In the house there is a girl.’

(303) هناك موضوعان مهمان (Ryding, 2005)
hunāka mawḍūʿāni muhimmāni
there topic.dl important.dl
‘There are two important topics.’

6.1.3 Diacritic Ambiguity

Diacritics, or short vowels, are largely omitted in modern texts, the matter that makes morphological and subsequently syntactic analysis difficult and highly ambiguous. Chalabi (2000) assumes that the absence of diacritization in Arabic poses a computational complexity “one order of magnitude bigger than handling Latin-based language counterparts”. In Arabic, in most instances, a word can have different pronunciations without any explicit orthographical effect due to the lack of diacritics. These different pronunciations distinguish between a noun and verb (304), active and passive forms (305), and imperative and declarative forms (306). Some verb forms have the middle letter doubled to make the verbs causative (transitive), but this also does not appear in orthography (307). Some agreement morphemes on the verbs are ambiguous leaving the open the selection between a variety of gender and person features (308).

(304) شرب šrb

شَرِبَ	شُرْبٌ
šariba	šurbun
‘drank’	‘drinking’

(305) أرسل ʼrsl

أَرْسَلَ	أُرْسِلَ
ʼarsala	ʼursila
‘sent’	‘was sent’

(306) قاوم qāwm

قَاوَمَ	قَاوِمْ
qāwama	qāwim
‘resisted’	‘Resist!’

(307) وصل wṣl

وَصَلَ	وَصَّلَ
waṣala	wṣṣala
‘arrived’	‘connect’

(308) كتبت ktbt

كَتَبْتُ	كَتَبْتَ	كَتَبْتِ	كَتَبَتْ
katabtu	katabta	katabti	katabat
wrote.1.sg	wrote.2.masc.sg	wrote.2.fem.sg	wrote.3.fem.sg
‘I wrote’	‘You wrote’	‘You wrote’	‘She wrote’

Frequently a single form can have a combination of the types of ambiguities mentioned above leading to an increased ambiguity level, as shown in Figure 61 for a surface form composed of only three letters but with seven different readings.

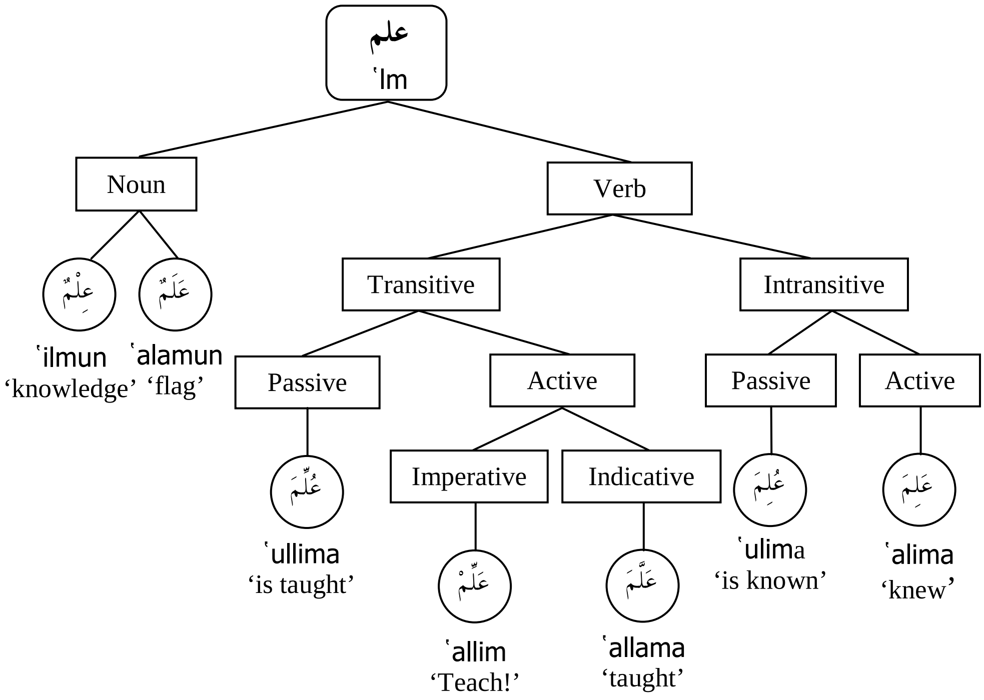


Figure 61. Ambiguity caused by the lack of diacritics

6.1.4 Multifunctionality of Arabic Nouns

Arabic nouns are characterised by their multifunctionality. Arabic nouns are derived from verbs and can take verbal functions in the sentence. Some nouns also can become prepositions, adverbs, adjectives or quantifiers.

Reaching a clear-cut understanding of Arabic word categories has been hindered by a millennium-long underspecification of the parts of speech in Arabic. Sibawaih (late 8th century) (1966) opens his famous book *Al-Kitab* with a classification of the parts of speech in Arabic into nouns, verbs and particles. This classification has remained until this day as a leading principle of Arabic grammar (Suleiman, 1990).

The verb is an uncontested category, and easily identified as an expression that denotes both action and tense. Particles as well are easily distinguished by their non-derivational aspects and by their morphological rigidity. Arabic nouns remain as the most elusive to define as they encompass a wide array of categories.

Wright (1896/2005) uses the term “noun” as an umbrella etymology that encompasses six types: a noun substantive (*nomen substantivum*), adjective (*nomen adjectivum*), numeral adjective (*nomen numerale*), demonstrative pronoun (*nomen demonstrativum*), relative pronoun (*nomen conjuctivum*) and personal pronoun (*pronomen*). Moreover, prepositions are subdivided into two categories: true prepositions such as إلى ʾilā ‘to’, and في fī ‘in’, and prepositions derived from nouns taking the accusative case (considered by traditional Arabic grammarians as adverbs) such as بين baina ‘between’, and تحت taḥta ‘under’. There are also true adverbs such as فقط faqaṭ ‘only’, and هنا hunā ‘here’, and nouns taking the accusative case and functioning as adverbs, such as كثيرا kaṯīran ‘frequently’, and مجانا mğğānan ‘freely’.

Therefore, the tripartite division could be considered as an archetypal classification rather than detailed listing. In a comprehensive morphological and syntactic description it is the detailed listing that is needed. It can be stated that the nature of Arabic derivational morphology (which is based to a great extent on the concept of roots and patterns) influenced the view of the tripartite division of parts of speech. For example, the noun and the adjective undergo the same inflection processes and, therefore, they are considered as one category by many researchers.

Morphologically speaking, adjective are the hardest to separate from nouns. Wright (1896/2005) identified four “nominal” categories that are essentially adjectives. The first is the active and passive participles (*nomina agentis* and *nomina patientis*) such as كاتب kātib ‘writing, a scribe’ and مكتوب maktūb ‘written, a letter’. He noted that these verbal adjectives often become in Arabic, as in other languages, substantives. The second type is semi-participial adjectives (in Arabic terminology, صفات مشبهة بأسماء الفاعل والمفعول ṣifātun mušabbahatun biʾasmāʾi al-fāʿili wa-l-mafʿūl ‘adjectives which are made like, or assimilated to, the participles’) such as سهل sahl ‘easy’ and صعب ṣaʿb ‘difficult’. The third type is the comparative and superlative adjectives (اسم التفضيل ʾismu at-tafḍīli ‘the noun of pre-eminence’), such as أعذب ʾaʿḏab ‘sweeter, sweetest’ and أكبر akbar ‘bigger, biggest’. The fourth type of adjectives is the relative adjectives, (الأسماء المنسوبة al-ʾasmāʾu al-mansūbah ‘relative nouns’), which are formed by adding the suffix يٌّ -

iyy to the nouns, and denote that a person or thing belongs to or is connected with the noun (in respect of origin, family, etc.), such as أرضـي ʾarḍiyy ‘earthly’, from أرض ʾarḍ ‘earth’.

The multifunctionality of Arabic nouns leads to an increased number of alternative possibilities and therefore leads to an increased ambiguity level. The multi-functionality of Arabic nouns can be summarized as follows:

- Arabic verbal nouns are categorically nouns, as shown in (309). They can also act syntactically as verbs heading an embedded clause, as in (310), or an adjunct phrase, as in (311). When verbal nouns function as verbs they inherit the same subcategorization frames from the verbs from which they were derived.

(309) أثمر البحث عن نتائج مبشرة (noun in nominal function)
ʾaṯmara al-bahṯu ʿan natāʾiǧa mubašširatin
brought the-research for results promising
‘The research brought promising results.’

(310) حاول البحث عن حل آخر (verbal noun in embedded clause)
ḥāwala al-bahṯa ʿan ḥallin ʾāẖara
tried the-searching for solution another
‘He tried searching for another solution.’

(311) زار زعماء المعارضة بحثا عن الدعم (Verbal noun in adjunct clause)
zāra zuʿamāʾa al-muʿāraḍati bahṯan ʿan ad-daʿmi
visited leaders the-opposition searching for the-support
‘He visited opposition leaders, searching for support.’

In many instances the two choices for analysing the noun (as nominal and as verbal) are available, leading to increased ambiguity. This is shown by the noun phrase in (312), which has two solutions, as shown in Figure 62 and Figure 63.

(312) لإرسائها
li-ʾirsāʾi-hā
for-establishing-it
‘for establishing it/for its establishment’

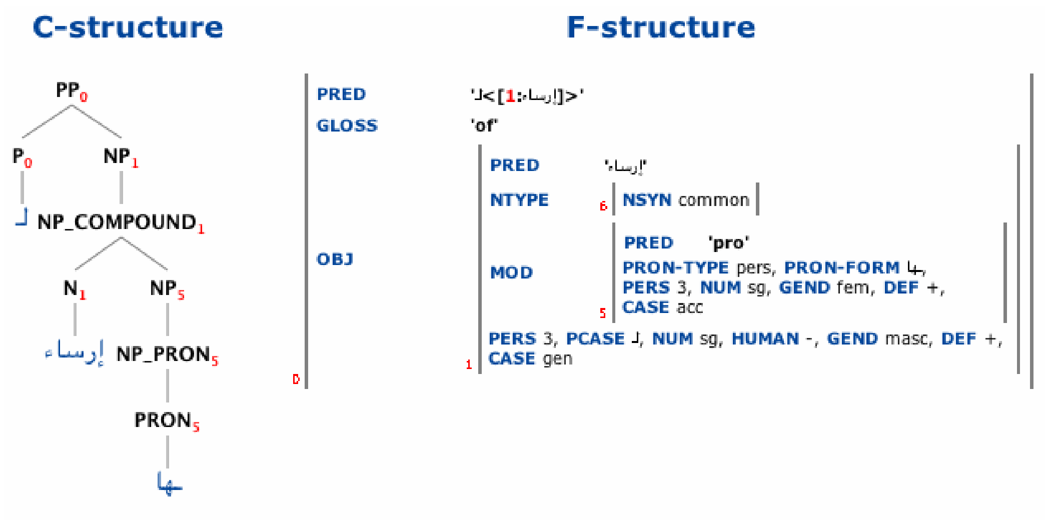


Figure 62. C- and f-structure of a noun with a nominal function

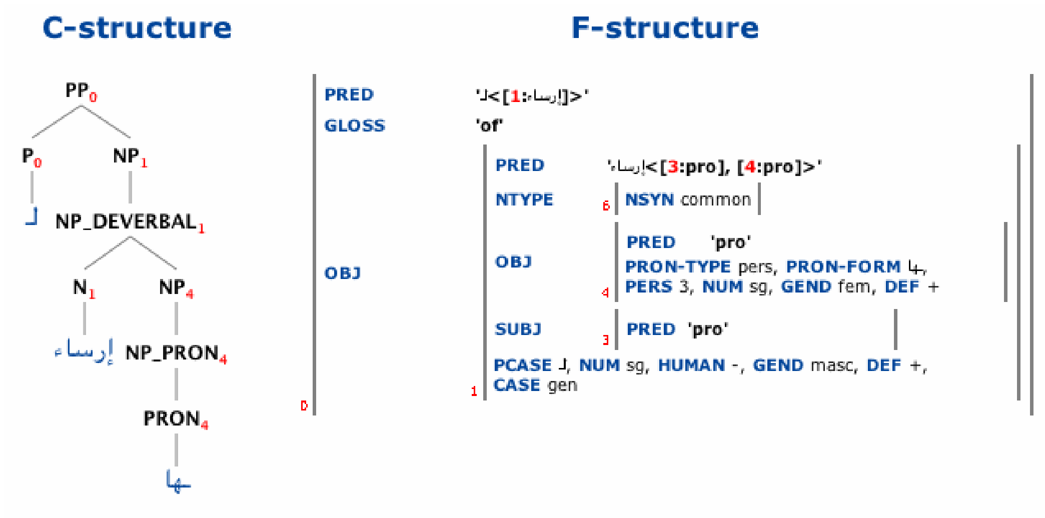


Figure 63. C- and f-structure of a noun with a verbal function

- Active and passive participles are generally adjectives, but they can also act as substantives, and as verbs heading adjunct phrases, as in (313) and (314). When participles function as verbs they inherit the same subcategorization frames as the verbs from which they were derived.

(313) معربا عن أسفه، قدم استقالته (Active participle XADJUNCT)
muʿriban ʿan ʾasafi-hi, qaddama ʾistiqalata-hu
expressing of regret-his, offered resignation-his
Expressing his regret, he offered his resignation.

(314) عاد إلى البيت منهارا (Passive participle XADJUNCT)
ʿāda ʾilā al-baiti munhāran
came to the-home devastated
'He came home devastated.'

- Nouns can also function as prepositions, adverbs and quantifiers. Some nouns can combine with a preposition to form an adverbial prepositional phrase. This is shown in (315)–(318). In all these cases the noun can still be used to perform an ordinary nominal function.

(315) a. وقف خلف أخيه (noun as preposition)
waqafa ḫalfa ʾaḫī-hi
stood behind brother-his
'He stood behind his brother.'

b. وقف في الخلف (noun in nominal function)
waqafa fī al-ḫalfi
stood in the-back
'He stood in the rear.'

(316) a. هذه الطريقة خطأ أساسا (noun as adverb)
hadihi aṭ-ṭarīqatu ḫaṭaʾun ʾasāsan
this the-method wrong basically
'This method is wrong basically.'

b. هذا البنيان يقف على أساس سليم (noun in nominal function)
hadā al-bunyānu yaqifu ʿalā ʾasāsin salīmin
this the-building stand on basis sound
'This building stands on a sound basis.'

(317) a. جرى بسرعة (noun in adverbial PP)
ğarā bi-surʿatin
ran with-speed
'He ran quickly.'

b. السرعة تؤدي إلى المخاطر (noun in nominal function)
as-surʿatu tuʾaddī ʾilā al-maḫāṭir
speed lead to the-dangers
'Speed leads to dangers.'

(318) a. جميع الطلاب حاضرون (noun as quantifier)
ğamīʿu aṭ-ṭullābi ḥāḍirūna
all the-students present
'All students are present.'

b. الجميع حاضرون (noun in nominal function)
al-ğamīʿu ḥāḍirūna
all present
'All are present.'

It is noteworthy that Arabic adjectives, likewise, can function as substantives. Adjectives, such as كريم karīm 'generous', بخيل baḫīl 'miserly', أحمق ʾaḥmaq 'foolish' and حكيم ḥakīm 'wise', can denote an adjective or a person. Although some English adjectives can be used in this sense (such as *the rich* and *the poor*), in Arabic the process is far more systematic and productive. This process can be used with almost any adjective that is able to denote a human entity. In many instances when translating Arabic adjectives which function as nouns into English, the translation is composed of the adjective added to the word *person*, *one*, *man,* or *woman*. It is easy to identify the gender as Arabic adjectives are inflected for gender, as in (319).

(319) يحكى أن غنيا وفقيرا تقابلا
yuḥkā ʾanna ġaniyyan wa-faqīran taqābalā
told that rich and-poor met
‘It is told that a rich man and a poor man met.’

6.2 Disambiguation in the pre-parsing phase

Disambiguation in the pre-parsing phase has the greatest effect on ambiguity reduction. MacDonald et al. (1994) emphasized the bottom-up priority concept and cited Seidenberg et al. (1982) as assuming that the information provided by natural language tends to be useful in deciding between alternatives at a given level of representation but much less effective at preselecting one of the alternatives at a higher level.

This entails that effort spent on managing ambiguity at the morphology level, for instance, can yield more significant results in controlling the overall ambiguities than the effort spent on the higher levels of syntax or semantics. The lexical representation of a word includes information about the word’s morphological features, POS category, argument structures, and semantics. If the word is ambiguous at a lower level of representation the ambiguity cascades exponentially into the other levels.

The pre-parsing stage contains all the processes that feed into the parser whether by splitting a running text into manageable components (tokenization), analyzing words (morphological analyzer) or tagging the text. These processes are at the bottom of the parsing system and their effect on ambiguity is tremendous as they directly influence the number of solutions a parser can produces. The pre-parsing stage covers tokenization, morphological analysis, MWEs, shallow mark-up, and full-blown pre-bracketing.

6.2.1 Tokenization

In an interesting experiment that shows the impact of tokenization on the parsing process, Forst and Kaplan (2006) made some improvements to the German tokenizer and reported that the revised tokenizer increased the coverage of the grammar from 68.3% to 73.4% when tested on 2000 sentences of the TiGer Corpus.

The Arabic tokenizer has been discussed in depth in Chapter 3, but here we are going to evaluate the tokenization effect on parse time and the ambiguity level.

We mentioned earlier that we have a deterministic and a non-deterministic tokenizer. Testing a randomly selected 16 word sentence using the deterministic tokenizer yielded 4 solutions while the non-deterministic tokenizer yielded 1280 solutions. However, using the non-deterministic tokenizer in Arabic does not generally affect the number of parses or parse time. It is obvious that the XLE system has an efficient mechanism in dealing with large finite state automata.

Some tokenization readings are genuine, yet highly infrequent and undesired in real-life data. These undesired readings create spurious ambiguities, as they are confused with more common and more acceptable forms. For example the Arabic preposition بعد ba’da ‘after’ has a possible remote reading if split into two tokens بـ@عد, which is made of two elements: بـ bi ‘with’, and عد ʿaddi ‘counting’, meaning ‘by counting’. The same problem occurs with MWEs. The solution to this problem is to mark the undesired readings. This is implemented by developing a filter, a finite state transducer that contains all possible undesired tokenization possibilities and attaches the “+undesired” tag to each one of them.

Using the tokenization filter to discard the compositional analyses of MWEs reduced the number of parses to less than half for 29% of the sentences in the test suite (254 sentences), and the overall parse time was reduced by 14%.

Among the functions of a tokenizer is to separate clitics from stems. Some clitics, however, when separated, become ambiguous with other clitics and also with other free forms. For example the word كتابهم kitābahum has only one morphological reading (meaning ‘their book’), but after tokenization كتاب@هم there are three different readings, as the second token هم can either be a clitic genitive pronoun, ‘their’, which is the intended reading, or a free pronoun, ‘they’, or even a noun meaning ‘worry’. This problem is solved by inserting a *kashida* that precedes enclitics and follows proclitics to distinguish them from each other as well as from free forms. Using the *kashida* for clitics reduced the number of parses to less than half for 5% of the sentences in the test suite (254 sentences).

6.2.2 Morphological analysis

The Arabic morphology component has been discussed in depth in Chapter 2, but there is still a need to explore the impact of the morphology on the syntax and how it contributes to resolving the problem of syntactic ambiguities and the efficiency of the parser in terms of parse time.

The morphology component feeds the parser with information on the morpho-syntactic features, such as number (singular, dual or plural), gender (masculine or feminine), person (first, second or third), tense (past, present, future), mood (declarative or imperative) and voice (active or passive). Therefore any ambiguity on the morphology level will propagate exponentially into the higher levels.

A classical well-known problem with Arabic morphology is the lack of diacritics, or short vowels, which reflect the pronunciation. The lack of diacritics results in two morphologically different words having orthographically identical forms. A widespread example of this ambiguity is the class of verbs which do not contain a weak letter in their formative radicals. These verbs typically have ambiguous person, mood and voice. This is shown in (٣٢٠), and the problem is graphically illustrated by the feature grid in Figure 64.

(٣٢٠) شكرت škrt

شَكَرَتْ	شَكَرْتَ	شَكَرْتِ
šakarat	šakarta	šakarti
thank.3.past.fem.sg	thank.2.past.masc.sg	thank.2.past.fem.sg
‘she thanked’	‘you thanked’	‘you thanked’
شُكِرَتْ	شُكِرْتَ	شُكِرْتِ
šukirat	šukirta	šukirti
thank.3.past.pass.fem.sg	thank.2.past.pass.masc.sg	thank.2.past.pass.fem.sg
‘She was thanked’	‘You were thanked’	‘You were thanked’
شَكَرْتُ	شُكِرْتُ	
šakartu	šukirtu	
thank.1.past.sg	thank.1.past.pass.sg	
‘I thanked’	‘I was thanked’	

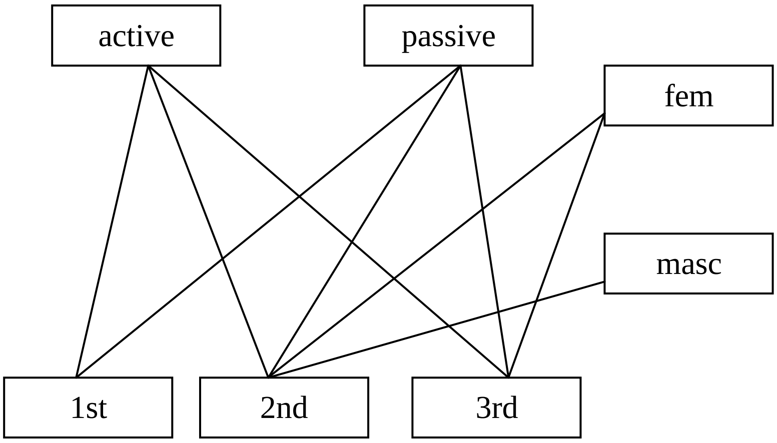


Figure 64. Feature grid of an ambiguous form

Every time we succeed in eliminating the “passive” possibility, we are effectively eliminating a good deal of ambiguities for each verb. We can eliminate the “passive” option for verbs depending on the verb’s nature (transitive or intransitive) or relying on personal judgment of plausibility.

In our morphology we specified which verbs can have the passive forms, and which verbs cannot. Out of 1532 verbs, only 36% are allowed to have passive forms (504 transitive verbs, and 43 intransitive verbs). The imperative form in Arabic is also mostly marked with diacritics which are not used in modern writing. Therefore, constraining the number of possible imperative forms will also help in reducing ambiguity. In our morphology we specified which verbs can have imperative forms, and which verbs cannot. Out of 1532 verbs, only 484 verbs (32%) are allowed to have an imperative form (324 transitive verbs, 160 intransitive verbs).

We tried to test the impact of the restrictions of the passive and the imperative forms in the morphology on the syntax through reverse testing. So we removed the relevant flag diacritics from our finite state morphology which is responsible for indicating which verbs can have passive and imperative forms and which cannot. Then we ran the grammar against the test suite. We found that this significantly increased the number of possible parses for 17% of the sentences and increased the parse time by 9%.

We would like also to discuss another morpho-syntactic feature, namely the "humanness" feature. English grammar distinguishes whether a noun is human or non-human in order to provide the appropriate relative pronoun (*who* or *which*). Arabic, however, does so more often to make proper agreement in number and gender between nouns and adjectives, nouns and relative pronouns, nouns and demonstrative pronouns, and also between the subject and its predicate in copula constructions. When the noun is human the adjective for example agrees with it in number. Yet when the noun is non-human, then if it is plural the adjective is singular, otherwise (i.e. if it is singular or dual), the adjective agrees with the noun in number. This is why a new feature ±human is added to all nouns in the morphology. This specification helps to constrain the Arabic grammar and to specify the correct agreement relations, reducing the number of ambiguities in many instances by almost half.

6.2.3 MWEs

MWEs have been discussed in depth in Chapter 4, but here we are going to evaluate the effect of MWEs on parse time and the ambiguity level.

MWEs encompass a wide range of linguistically related phenomena that share the criteria of being composed of two words or more, whether adjacent or separate. Filtering out the compositional analyses of MWEs in an early state of the analysis, i.e. tokenization, reduced the number of parses to less than half for 29% of the sentences in the test suite, and the parse time is reduced by 14%.

In the early stages of analysing MWEs we tried to allow compositional readings along with the MWE readings and to give a positive OT preference mark to MWEs. However, we found in some instances that the interaction of preference marks can lead the compositional readings to surface as optimal solutions and MWE readings to be suppressed as suboptimal, as in the example (321) where the MWE expression has the adverbial reading ‘quickly’ while the compositional reading has the PP reading ‘with speed’.

(321) جرى بسرعة
ǧarā bi-surʿah
ran.masc.sg.3 with-speed
MWE reading: ‘He ran quickly.’
Compositional reading: ‘He ran with speed.’

Figure 65 shows the compositional reading which passes as the optional solution, while Figure 66 shows the correct MWE which is suppressed by the parser as a suboptimal solution.

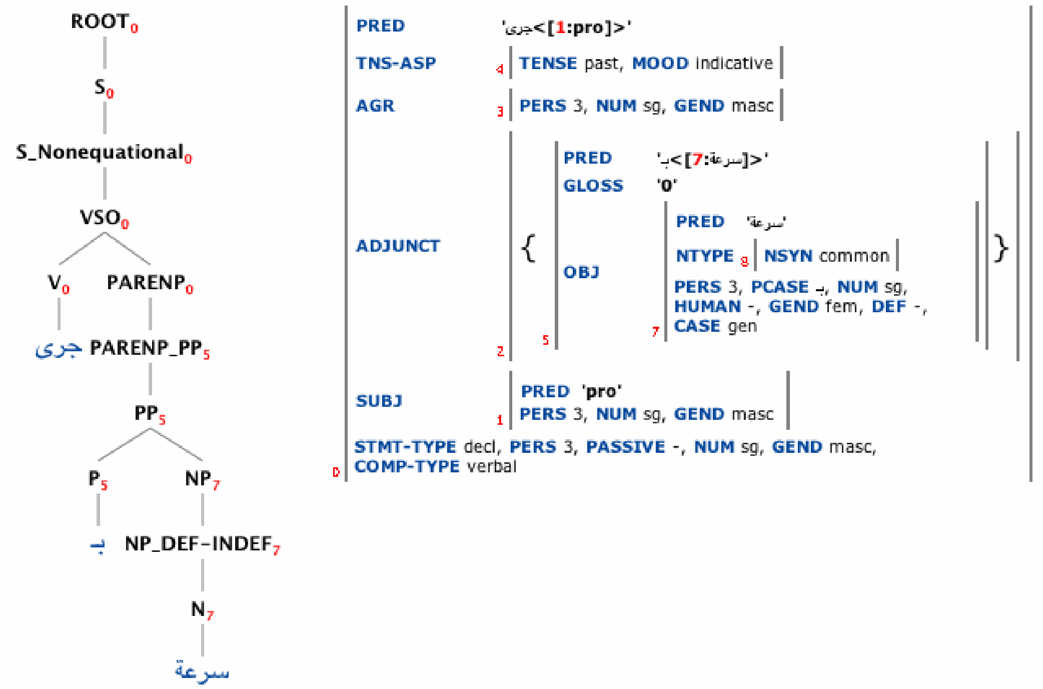

ROOT0
S0
S_Nonequational0
VSO0
V0
PARENP0
جرى
PARENP_PP5
PP5
P5
NP7
بـ
NP_DEF-INDEF7
N7
سرعة
PRED 'جرى<[1:pro]>'
TNS-ASP 4 TENSE past, MOOD indicative
AGR 3 PERS 3, NUM sg, GEND masc
ADJUNCT
PRED 'بـ<[7:سرعة]>'
GLOSS '0'
OBJ
PRED 'سرعة'
NTYPE 8 NSYN common
PERS 3, PCASE بـ, NUM sg, HUMAN -, GEND fem, DEF -, CASE gen
7
5
2
SUBJ
PRED 'pro'
PERS 3, NUM sg, GEND masc
1
STMT-TYPE decl, PERS 3, PASSIVE -, NUM sg, GEND masc, COMP-TYPE verbal
0

Figure 65. A composition reading of a MWE surfacing as the optimal solution

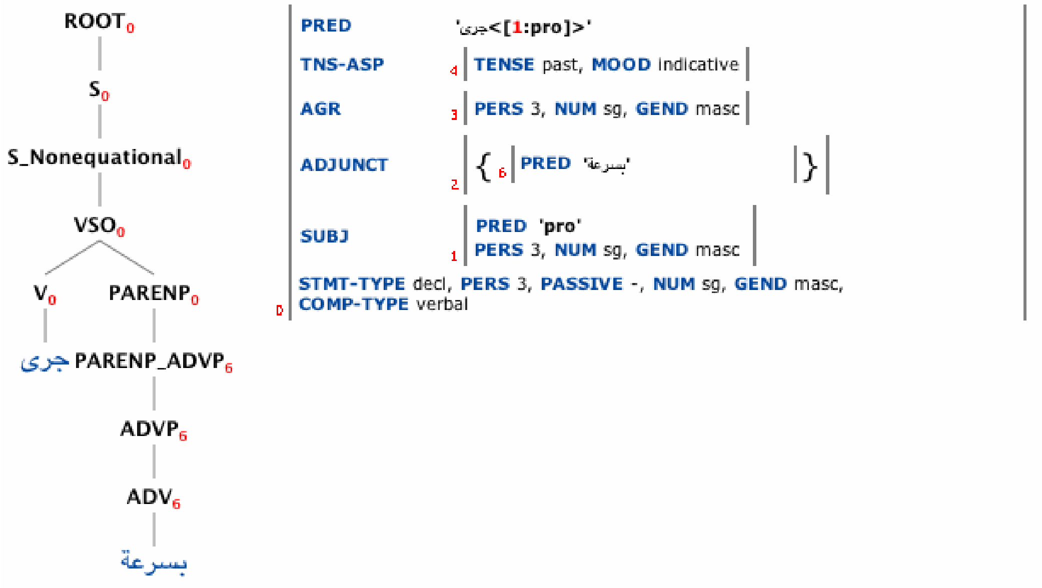

ROOT0
S0
S_Nonequational0
VSO0
V0
PARENP0
جرى
PARENP_ADVP6
ADVP6
ADV6
بسرعة
PRED 'جرى<[1:pro]>'
TNS-ASP 4 TENSE past, MOOD indicative
AGR 3 PERS 3, NUM sg, GEND masc
ADJUNCT 2 { 6 PRED 'بسرعة' }
SUBJ 1 PRED 'pro' PERS 3, NUM sg, GEND masc
0 STMT-TYPE decl, PERS 3, PASSIVE -, NUM sg, GEND masc, COMP-TYPE verbal

Figure 66. A MWE reading suppressed as a suboptimal solution

Moreover, the compositional readings cause an efficiency problem by increasing the number of solutions and parse time. Therefore we opted for pruning the compositional readings in the early stage of tokenization. However, this remains as an empirical decision, and if evidence shows that plausible parses are lost in certain cases, we can either handle these cases individually or make the tokenization allow these the solutions and mark them with a certain tag, so that they can be incorporated in the optimality hierarchy.

The MWE transducer is now part and parcel of the system. Our system cover 2818 MWEs. When they are removed 34% of the sentences in the test suite are affected, either by failing to find a parse (20 sentences) or having parses with almost double the number of ambiguities.

Examples in (322)–(327) show MWEs that caused the system to fail to find a parse, as a compositional analysis is not available in the core morphology. These expressions vary in their grammatical category.

(322) على الفور (Adverb)
ʿalā al-fawri
on the-immediate
'immediately'

(323) أبو حليقة (Proper name)
ʾabū ḥalīqah
'Abu Haliqah'

(324) ميؤوس منها (Adjective)
maiʾūsun min-hā
despaired of-it
‘hopeless’

(325) بعيدا عن (Preposition)
baʿīdān ʿan
far from

(326) غير أن (Subordinating conjunction)
ġaira ʾanna
but that
‘but, however’

(327) شهود عيان (Compound noun)
šuhūdu ʿayānin
witnesses seeing
‘eye witnesses’

Similarly, examples (328)–(331) show MWEs that caused the system to have an increased number of ambiguities, while the correct parse is not provided. A compositional analysis is already available in the core morphology, but the problem is that a compositional analysis does not provide the correct parse. These expressions also vary in their grammatical category.

(328) الشرق الأوسط (Named entity)
aš-šarq al-ʾawsaṭ
the-east the-middle
‘the Middle East’

(329) لأن (Subordinating conjunction)
li-ʾanna
for-that
‘because’

(330) رام الله (Proper noun – place)
rām allah
‘Ramallah’

(331) بشكل عشوائي (Adverb)
bi-šaklin ʿašwāʾiyyin
with-way random
‘randomly’

6.2.4 Shallow markup (tagging)

This technique has not been applied in the Arabic grammar but we would like to mention some experiments that report on its efficiency in managing ambiguity in other languages, in order to point out its feasibility.

Kaplan and King (2003) integrated three types of shallow mark-up (POS tagging, named entities, and labelled bracketing) into the ParGram LFG English grammar. Labelled bracketing is when a constituent is labelled both with the phrase type as well as the grammatical function. For example, *the boy* in *He saw the boy,* will be bracketed [NP-OBJ the boy]. They observed that named-entity mark-up improves both speed and accuracy and labelled brackets also can be beneficial, but that POS tags are not particularly useful. This confirms the earlier findings that MWEs have a great effect in reducing ambiguity and increasing efficiency. A large portion of our list of MWEs include names of countries, institutions and organizations.

Dalrymple (2006) showed that if a perfect POS tagger were available, a reduction in ambiguity of about 50% would be attained. The problem is that a perfect POS tagger does not exist, as creating a perfect POS tagger needs a perfect parser and perfect world knowledge to be integrated into the system; neither is currently available for any language.

6.2.5 Full-blown pre-bracketing using a probabilistic parser

This technique, as well, has not been applied in the Arabic grammar but we would like to mention some experiments that report on its efficiency in managing ambiguity in other languages, in order to point out how feasible it is for future work.

In a recent experiment Cahill et al. (2007) tried to increase the speed of the English hand-crafted rule-based grammar which produces deep linguistic analysis by pruning the search space at an earlier stage of the parsing process. They retrained a state-of-the-art probabilistic parser and used it to pre-bracket the sentences before inputting them to the XLE English parser, in an attempt to constrain the valid c-structure space for each sentence. The job of the XLE parser then was limited to drawing deep f-structure representations from the available c-structures. Their evaluation shows that this strategy decreases the time taken to parse by about 18% while maintaining accuracy.

This technique, however, reduces the usability of the rule-based parser as it makes it dependent on the probabilistic parser with its advantages and limitations.

6.3 Disambiguation in the Parsing Phase

We mean by the parsing phase everything that is related to the actual rule writing, starting with how fine grained the rules are, how the projection of lexical entries onto the grammar is specified, how the grammatical constraints are used, and how domain-specific adaptation can help the grammar to be more focused.

Before we started working on the performance and ambiguity management issues, it came to a point when further development of the grammar became extremely difficult because of the highly increasing time the grammar took to parse our test suite. The grammar took 141 minutes (CPU time) to parse a test suite of 229 sentences. This meant that when writing a new rule, we had to wait about two and half hours to see the result of the change.

Moreover, the number of valid analyses for long sentences ran astronomically into several millions, a level of ambiguity that is not conceived to be motivated by any linguistic complexity of the language.

Our aim during working with the performance of the grammar was to reduce both parse time and spurious ambiguities, and to keep them within a manageable boundary. After a series of refining, fine-tuning and corrections, there is a 95% reduction in parse time. It takes the grammar now 7 minutes (CPU time) to parse the test suite. Ambiguity is also significantly reduced. The average number of optimal solutions was 767 and the average number of suboptimal solutions was

4.37E+07. Now the average number of optimal solutions is 135 (a reduction of 82%) and the average number of suboptimal solutions is 1.45E+04. Table 11 shows a comparison of the number of valid parses for some selected sentences. Numbers in the final two columns show totals of optimal and suboptimal solutions offered for each sentence.

#	Sentence	Before fine-tuning	After fine-tuning
1	بوش يطالب بتسريع خطى الديمقراطية بالشرق الأوسط būš yuṭālibu bi-tasrīʿi ḫuṭā Bush calls for-speeding steps ad-dīmuqrāṭiyyah bi-š-šarqi al-ʾawsṭi the-democracy in-the-eat the-middle ‘Bush calls for speeding up the steps of democracy in the Middle East.’	1+523	3+16
2	أكد الرئيس الأميركي جورج بوش ضرورة تسريع الإصلاحات الديمقراطية في الشرق الأوسط بعد الحرب على العراق لإنهاء عقود من الحرمان والكبت ʾakkada ar-raʾīsu al-ʾamīrkī confirmed the- president the-American ǧūrǧ būš ḍarūrata tasrīʿI al-ʾiṣlāḥāti George Bush necessity speeding the-reforms ad-dīmuqrāṭiyyati fī aš-šrqi al-ʾawsaṭi the-democratic in the-east the-middle baʿda al-ḥarbi ʿlā al-ʿirāqi li-ʾinhāʾi after the-war on Iraq for-ending ʿuqūdin min al-ḥirmāni decades of the-depravation wa-l-kabti and-the-oppression ‘The American President George Bush confirmed the necessity of speeding up the democratic reforms in the Middle East after the war on Iraq to end decades of depravation and oppression.’	27+8116335	2+12
3	وأضاف أن موجة من الديمقراطية بدأت تجتاح المنطقة بعد الإطاحة بالرئيس العراقي صدام حسين. wa-ʾaḍāfa ʾanna mawǧatan mina and-added that wave of ad-dīmuqrāṭīti badaʾat taǧtāḥu al-manṭiqati the-democracy started sweep the-region baʿda al-ʾiṭāḥati bi-l-raʾīsi after the-deposing of-the-president al-ʿirāqiyy ṣaddām ḥusain the-Iraqi Saddam Hussein ‘And he added that a wave of democracy started to sweep the region after deposing the Iraqi President Saddam Hussein.’	33+5259	2+5

Table 11. Ambiguity comparison for some sentences before and after fine-tuning

In this section we will explain the avenues we explored to reduce the ambiguity level and improve the efficiency and performance of the parser. This phase covers the issues of granularity of phrase structure rules, lexical specifications, application of syntactic constraints, and domain specific adaptation.

6.3.1 Granularity of Phrase Structure Rules

Nagata (1992) studied the effect of phrase structure granularity on the efficiency and performance of a unification-based HPSG parser, and concluded that using "medium grained" phrase structure rules will make a unification based grammar fast, efficient and maintainable. He suggested that *medium-grained* phrase structure rules reduce the computational loads of unification without intractably increasing the number of rules.

According to Nagata (1992), a coarse-grained grammar is one which uses very few phrase structure rules, and which relies heavily on disjunctions and strong constraints on features. A medium-grained grammar is one which consists of atomic phrase structure rules and medium constraints on features. A fine-grained grammar represents most constraints in phrase structure rules, and the number of rules can reach several thousands. Nagata (1992) pointed out that an example of a coarse-grained HPSG-based Japanese grammar has about 20 generalized phrase structure rules, while a medium-grained grammar has about 200 phrase structure rules.

In the LFG literature, Maxwell and Kaplan (2003) agreed with Nagata's finding that a medium-grain phrase structure grammar performs better than either a coarse-grain or fine-grain grammar. They conducted experiments that proved that processing all of the phrasal constraints first using a chart, and then using the results to decide which functional constraints to process is more efficient than interleaving phrasal and functional constraints. This is because the phrasal constraints can be processed in cubic time, whereas the functional constraints may run in exponential time in the worst case. They suggested that the global well-formedness of phrasal constraints can serve as a polynomial filter for the computation of functional constraints.

The Arabic grammar can basically be described as composed of coarse-grained phrase structure rules. This is well indicated by the small number of grammar rules in the grammar, 57 rules. The reason the Arabic grammar is so coarse is, we believe, due to the nature of the language. Arabic relies on the morphology more than on the configurational structure, to decide the sentence type of imperative, interrogative, passive, negative and declarative. Yet we believe that it is still possible to make the grammar more fine-grained.

The grammar previously had 25 rules, so we attempted to make some rules more fine-grained. Splitting the rule for non-equational sentences in (332) into three rules: VSO, SVO and VOS, as shown in (333), led to a 10% reduction in parse time with no effect on the number of valid solutions. The number of subtrees was generally increased (sometimes decreased) by a fraction.

```
(332)  S_Nonequational -->
       { "The VSO word order"
                 V: ^=! ;
                 NP: (^SUBJ)=! (! CASE) = nom
                 (NP: (^OBJ)=!  (! CASE) = acc)
       | "The SVO word order"
                 NP: (^ SUBJ)=! (! CASE) = nom;
                  V: ^=!
                 (NP: (^OBJ)=!  (! CASE)=acc)
       | "The VOS word order"
                 V: ^=!
                 NP: (^ OBJ)=! (! PRON-TYPE)=c pers (! CASE)=acc;
                 NP: (^ SUBJ)=! (! CASE)=nom}.

(333)  S_Nonequational -->
       { VSO
       | SVO
       | VOS}.

       VSO --> V: ^=! ;
                 NP: (^SUBJ)=! (! CASE) = nom
                 (NP: (^OBJ)=!  (! CASE) = acc).

       SVO --> NP: (^ SUBJ)=! (! CASE) = nom;
                  V: ^=!
                 (NP: (^OBJ)=!  (! CASE)=acc).

       VOS -->           V: ^=!
                 NP: (^ OBJ)=! (! PRON-TYPE)=c pers (! CASE)=acc;
                 NP: (^ SUBJ)=! (! CASE)=nom.
```

Splitting the sentence types between equational and non-equational led to a 3% reduction in parse time with no effect on the number of valid solutions. The number of subtrees was generally increased.

We expanded the NP rule into 12 subtypes: NP_Compound, NP_Demonstrative, NP_Proper, NP_Pronoun, NP_Deverbal, NP_Number, NP_Date, NP_Adjective, NP_Superlative, NP_Partitive, NP_Relative, and NP_Definite-Indefinite. This change led to no change in the number of valid solutions. Parse time was not affected. There was an increase of approximately 5% in the number of subtrees for most sentences.

In general, the granularity of phrase structure rules affects the speed and performance of the grammar. The fewer the rules in a grammar, the greater the number of disjunctions. Previously, with 25 rules the grammar had 3341 disjunctions. Now the grammar has 57 rules and 2564 disjunctions, and has much greater coverage. The resolution of disjunctions is computationally expensive in terms of memory resources and parse-time, and the fewer disjunctions a grammar has, the better it is expected to perform.

6.3.2 Exhaustive Lexical Description

In the field of psycholinguistics MacDonald et al. (1994) challenged the general view that lexical and syntactic ambiguities are dichotomous, involving different types of knowledge representations. They also criticized the blind application of Fodor’s (1983) concept of modularity in language processing, which assumes the existence of a number of autonomous and encapsulated modules responsible for analyzing different types of information. They gave an alternative account in which both lexical and syntactic ambiguities are resolved by the same processing mechanisms.

MacDonald et al. (1994) maintained that languages are structured at multiple levels simultaneously, including lexical, morphological, syntactic, and discourse levels. They went on to show that these levels are entangled in such a way that ambiguity at any given level will propagate into other levels. For example, the word *watch* has two meanings (i.e., ‘time piece’ and ‘observe’), and it is also ambiguous in its grammatical category (noun or verb), and the verb can have different syntactic structures, including transitive and intransitive.

MacDonald et al. (1994) believed that syntactic ambiguities are caused by ambiguities associated with lexical items and that syntactic ambiguity resolution is guided by lexical information. However, we believe that while it is true that many instances of syntactic ambiguities originate from different lexical interpretations, there are many instances where syntactic ambiguities are not triggered by any lexical interpretations, such as PP attachment, scope of coordination, and the interaction of rules.

MacDonald et al. (1994) made a strong claim that both lexical and syntactic ambiguities are governed by the same processing mechanisms. Both the lexical and syntactic domains are managed by frequency information and contextual constraints. They argued in favour of the linguistic theories that eliminate the strong distinction between accessing a meaning and constructing a syntactic representation. They suggested that the parallel between these domains derives from the fact that syntactic ambiguities are based on ambiguities at the lexical level. This is compatible with their assumption that the comprehension of a given sentence is the process of concurrently deriving a number of linked representations at three major levels: lexical, syntactic, and discourse.

In the LFG framework the role of the lexicon in the sentence structure is emphasized significantly. Although syntactic structures are represented independently in phrase structure trees, such representations are constrained by properties of lexical items. Many pieces of information required by the syntax will be stored in lexical entries, and lexical entries project certain structures onto the syntax. This is how syntax and the lexicon are interlinked and the boundaries between the two systems are greatly blurred.

Argument structures, or subcategorization frames, are one type of information associated with words, and they play an important role in causing or resolving syntactic ambiguities. Argument structures dictate the kind of phrases that optionally or obligatorily occur with a lexical item and the relationships between these phrases.

Subcategorization frames can become a source of structural ambiguity as many words can be associated with several different argument structures. A typical example is the transitive vs. intransitive argument structures, as shown in (334) for the verb أكل ’akala ‘eat’.

(334) أكل V
{ (↑ PRED)='أكل<(↑ SUBJ)(↑ OBJ)>'
| (↑ PRED)='أكل<(↑ SUBJ)>'}.

The English lexicon contains 9,652 verb stems and 23,525 subcategorization frames (Riezler et al., 2002), meaning that each verb has an average of 2.4 subcategorization frames.

There are 1507 verbs in the Arabic grammar with 1660 subcategorization frames, with an average of 1.1 subcategorizations per verb. This could be because the subcategorizations frames in our grammar are underspecified, or that our grammar being limited to the news domain has the advantage of cutting down the number of ambiguities, or it could just be a difference between English and Arabic. Furthermore there are 1327 verbal nouns with generally a single subcategorization frame for each noun.

The subcategorization frames for Arabic verbs and verbal nouns were entered manually into the LFG grammar lexicon. There are a total of 1507 verbs classified into 17 subcategorization frames, as shown in Table 12.

#	Arguments	Examples
1	Subject	أتى ʾatā ‘come’
2	Subject-Complement	أثبت ʾaṯbata ‘prove that’
3	Subject-Object	أعد ʾaʿadda ‘prepare’
4	Subject-Object-Complement	أبلغ ʾablaġa ‘inform sb that’
5	Subject-Object-Secondary Object	أعطى ʾaʿṭā ‘give sb sth’
6	Subject-Object-Oblique	أخفى ʾaḫfā ‘hide sth from’
7	Subject-Oblique	أخفق ʾaḫfaqa ‘fail in’
8	Subject-Oblique-Complement	أكد ʾakkada ‘confirm to sb that’
9	Subject-Oblique1-Oblique2	اتفق ʾittafaqa ‘agree with sb on’
10	Subject-Object-Oblique1-Oblique2	اشترى ʾištarā ‘buy sth from sb for’
11	Subject-Xcomp	أراد ʾarāda ‘want’
12	Subject-Object-Xcomp	اعتبر ʾiʿtabara ‘consider’
13	Subject-Oblique-XComp	طلب ṭalaba ‘request from sb to do sth’
14	Subject-BetweenAnd	تنقل tanaqqala ‘move between … and …’
15	Subject-FromTo	أقلع ʾaqlaʿa ‘fly from … to …’
16	Subject-Object-BetweenAnd	نسق nassaqa ‘coordinate sth between … and …’
17	Subject-Object-FromTo	ترجم tarǧama ‘translate … from … to …’

Table 12. Subcategorization frames for Arabic verbs

We would like here to explain what we mean by “verbal nouns”. In Arabic there is a class of nominals derived from verbs. They are assumed to inherit some or all of the verb’s argument structure. Verbal nouns and verbs share the same root, so morphological analyzers that take the root as the base form can easily relate them together. In our implementation, however, the stem, not the root, is used as the base form, and so verbal nouns have to be entered separately into the lexicon.

The derivation process in Arabic uses non-concatenative morphotactics: unlike English -*ing*, or -*ed* suffixes. There is no way to distinguish verbal nouns from nominal nouns as they have the same form.

In our application, there are a total of 1327 verbal nouns having subcategorization frames as shown in Table 13.

#	Arguments	Examples
1	Subject-Object	إتمام ʾitmām ‘completing’
2	Subject-XComp	محاولة muḥāwalah ‘trying to’
3	Subject-Object-Oblique	إخفاء ʾiḫfāʾ ‘hiding sth from’
4	Subject-Object-XComp	إجبار ʾiğbār ‘forcing sb to’
5	Subject-Oblique	إحجام ʾiḥğām ‘refraining from’
6	Subject-Oblique1-Oblique2	اتفاق ʾittifāq ‘agreeing with … on …’
7	Subject-Complement	إثبات ʾiṯbāt ‘proving that’
8	Subject-Object-Secondary Object	إعطاء ʾiʿṭāʾ ‘giving sb sth’
9	Subject-Object-Complement	طمأنة ṭamʾanah ‘comforting sb that’
10	Subject-Oblique-Complement	مستشهد mustašhid ‘citing from … that’
11	Subject-Oblique-XComp	التماس ʾiltimās ‘requesting from sb to’
12	Subject-Object-FromTo	تحويل taḥwīl ‘transferring sth from … to …’
13	Subject-FromTo	انتقال ʾintiqāl ‘moving from … to …’
14	BetweenAnd	اقتتال ʾiqtitāl ‘fight between … and …’
15	Subject-BetweenAnd	تنقل tanaqqul ‘moving between … and …’
16	Oblique	حرب ḥarb ‘war on’

Table 13. Subcategorization frames for Arabic verbal nouns

Perhaps the most effective way to reduce ambiguities is to write more accurate and more constrained subcategorization frames. So for every oblique object in the argument structure the form of the preposition is explicitly specified, as shown in (335)–(337).

(335) أسفر [*result*] V XLE @(V-Subj-Obl %stem عن [*in*]).

(336) أسهم [*contribute*] V XLE @(V-Subj-Obl %stem في [*to*]).

(337) أشار [*point*] V XLE @(V-Subj-Obl %stem إلى [*to*]).

These lexical entries call a template with two arguments: the first is the lexical entry of the verb and the second is the form of the preposition required. The template as shown in (338) requires the oblique object to be headed with a preposition of the same form.

(338) V-Subj-Obl(P_ PF_)=
(^ PRED)='P_<(^ SUBJ)(^ OBL)>' (^ OBL OBJ PCASE)=c PF_.

The same sort of specification has been done with verbal nouns. When a verbal noun takes an oblique object, the lexicon must specify the lexical form of the preposition as shown in (339)–(341).

(339) إبلاغ [*informing*] N XLE @(Subj-Obj-Obl %stem بـ [*of*]).

(340) إجبار [*forcing*] N XLE @(Subj-Obj-Obl %stem على [*on*]).

(341) إجلاء [*evacuating*] N XLE @(Subj-Obj-Obl %stem من [*from*]).

To test how effective this condition is in reducing the ambiguity level, we removed the constraining equation from the templates allowing verbs and verbal nouns to have obliques but without specifying the lexical form of the preposition. This change affected 49% of the test suite and increased the average number of optimal solutions from 135 to 159 and the average number of suboptimal solutions almost tripled from 1.45E+04 to 3.38E+04.

The lexical specification is not limited to subcategorization frames, but it contains also any structurally relevant information. For every verb there is a specification of whether the verb is a main verb or copula verb. For equi verbs the control relationship is stipulated, as shown in (342) for the equi verb حاول ḥāwala 'try'.

(342) حاول V (↑ PRED)= 'حاول<(↑ SUBJ) (↑ COMP)>'
(↑ COMP SUBJ NUM) = (↑ SUBJ NUM)
(↑ COMP SUBJ GEND) = (↑ SUBJ GEND)
(↑ COMP SUBJ PERS) = (↑ SUBJ PERS).

6.3.3 Application of Syntactic Constraints

MacDonald et al. (1994) maintain that grammatical knowledge plays an important role in constraining the potential interpretations of a sentence. They cite MacWhinney and Bates (1989) as providing an account of the effect of context in resolving ambiguity in terms of a competition model, which assumes that languages provide cues that interact (or "compete") with one another during processing in order to select a certain interpretation and inhibit the others.

The LFG theory relies heavily on constraint satisfaction mechanism in managing syntactic ambiguities. Here we will explore two of the most notorious hotspots of ambiguity, i.e. coordination and PP attachment, and see how they are handled through constraints in the grammar. However, it must be noted that all constraints are subject to empirical testing, as constraints can be modified or removed in the light of new evidence.

6.3.3.1 Coordination

Coordination is a well-known hotspot of ambiguity, especially when the boundaries of the coordinated phrases can be defined in two or more different ways. This is known as an ambiguity in the scope of conjunction.

There are two types of coordination: constituent and non-constituent coordination (Kaplan and Maxwell, 1995). In constituent coordination two phrases of the same category are coordinated, e.g. *John and Mary went to London*. In non-constituent coordination the coordinated elements are fragments of phrases, e.g. *John went to London and Mary to Paris*. Only constituent coordination is covered in our grammar until now.

In the LFG framework coordinated constituents are treated as sets. The phrase structure notation for creating a set function for the coordinated constituents is presented by Kaplan and Maxwell (1995) as in (343) which means that the two NPs on the right-hand side are members of the set NP on the left-hand side.

$$\text{(343)}\quad NP \rightarrow \underset{\uparrow \in \downarrow}{NP} \quad CONJ \quad \underset{\uparrow \in \downarrow}{NP}$$

For the coordinated sentences in (٣٤٤), Figure 67 shows how the two sentences are represented as a set containing the f-structures that correspond to sentences.

(٣٤٤) ذهب الولد ونامت البنت
ḏahaba al-waladu wa-nāmati al-bintu
went the-boy and-slept the-girl
'The boy went and the girl slept.'

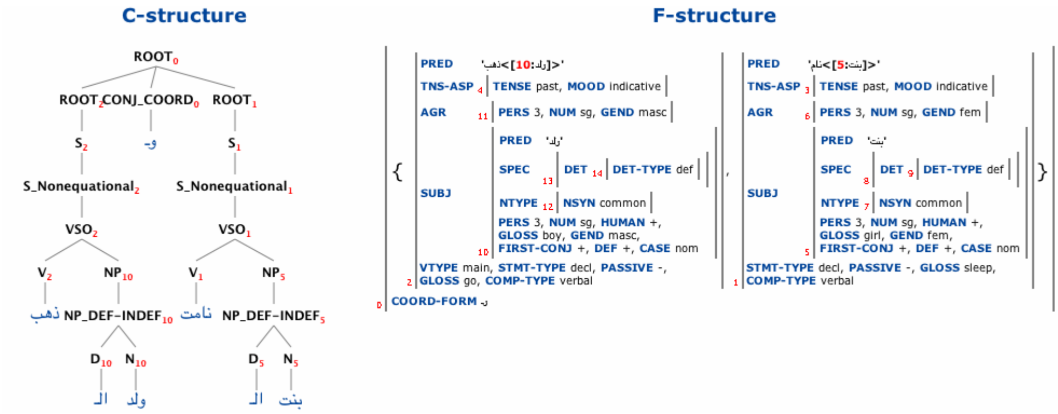


Figure 67. Constituent coordination represented as a set

Some features however are distributive and other features are not. In Arabic NPs, the features of number, gender, person, and humanness are non-distributive and are controlled through special conditions.

In Arabic, if the subject is a coordinate NP occurring in the post-verbal position, the verb exhibits what is termed “first conjunct agreement” by many researchers, e.g. Sadler (2003) and Hoyt (2004), i.e. the verb agrees only with the first conjunct of a coordinate subject. Alternatively if the subject occurs in the pre-verbal position, verbs exhibit agreement with the whole set, after the features of the coordinate NP are resolved according to specific conditions.

The first conjunct agreement is handled in our grammar through the phrase structure rules, as shown in (٣٤٥). The NP in the subject position which occurs in the post-verbal position is given a check feature of FIRST-CONJ which takes the value of ‘+’.

(٣٤٥) S → V NP

$$S \rightarrow \underset{\uparrow=\downarrow}{V} \quad \underset{\substack{(\uparrow \text{ SUBJ})=\downarrow \\ (\downarrow \text{ FIRST-CONJ})=+}}{NP}$$

Then the NP coordination template (the set of rules responsible for resolving the agreement features on coordinate noun) checks for the feature “FIRST-CONJ”. If it is found the whole conjunction is given the same features for number, gender and person as the first conjunct. The example in (٣٤٦) and the corresponding representation in Figure 68 show how first conjunct agreement is treated in our grammar.

(٣٤٦) ذهبت البنت والولد
ḏahabat al-bintu wa-al-waladu
went.fem.sg the-girl and-the-boy
‘The girl and the boy went.’

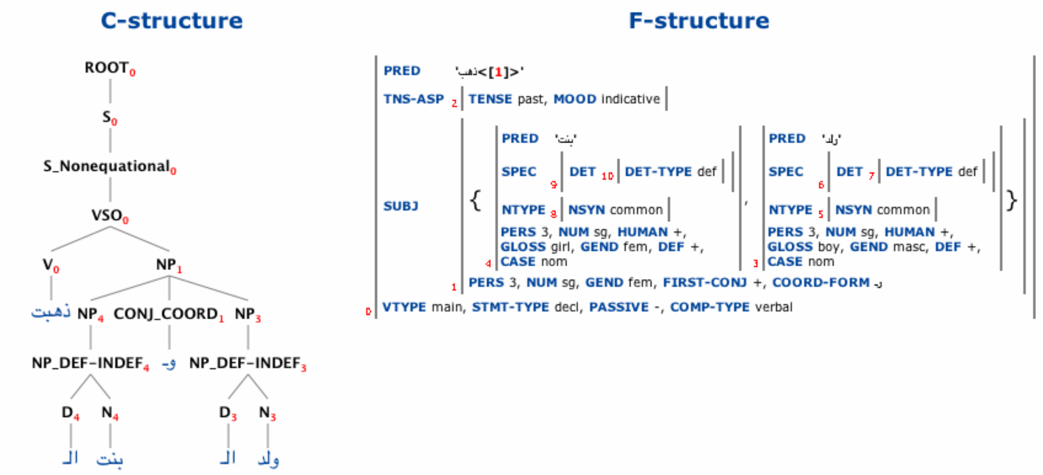

C-structure
ROOT
S
S_Nonequational
VSO
V
NP
ذهبت
NP
CONJ_COORD
NP
NP_DEF-INDEF
و-
NP_DEF-INDEF
D
N
D
N
الـ
بنت
الـ
ولد
F-structure
PRED 'ذهب<[1]>'
TNS-ASP TENSE past, MOOD indicative
SUBJ
PRED 'بنت'
SPEC DET DET-TYPE def
NTYPE NSYN common
PERS 3, NUM sg, HUMAN +, GLOSS girl, GEND fem, DEF +, CASE nom
PRED 'ولد'
SPEC DET DET-TYPE def
NTYPE NSYN common
PERS 3, NUM sg, HUMAN +, GLOSS boy, GEND masc, DEF +, CASE nom
PERS 3, NUM sg, GEND fem, FIRST-CONJ +, COORD-FORM و
VTYPE main, STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal

Figure 68. First conjunct agreement

Kuhn and Sadler (2007) studied different type of Single Conjunct Agreement, including First Conjunct Agreement, and proposed an interesting solution. They first criticized the traditional representation of the f-structure for coordinate NPs as unordered sets. They appealed to this meta principle and argued that some of the mathematical properties of sets turned out to be less adequate. They suggested using a slightly different formal device, which they called “local f-structure sequences”, and assumed that this new device would cater more readily for the typological differences between languages regarding agreement and the phenomenon of Single Conjunct Agreement.

If the agreement does not follow the first conjunct agreement condition, the resolution of the features in conjoined subjects follows these rules:

- Gender: The gender of the whole NP is masculine unless all conjuncts are feminine nouns.
- Person: The resolution of the person feature follows this priority order. The person of the whole conjunction is 1st if any NP is in the 1st person.

The person of the whole conjunction is 2nd if any NP is in the 2nd person. Otherwise the person is 3rd.

- Number: the number of the whole NP will be plural unless there are only two conjuncts and both are singular, in which case the whole NP is dual.

The example in (٣٤٧) and the corresponding representation in Figure 69 show how the agreement features are resolved.

(٣٤٧) البنت والولد ذهبا
al-bintu wa-al-waladu   ḏahabā
the-girl and-the-boy   went.dual.masc
'The girl and the boy went.'

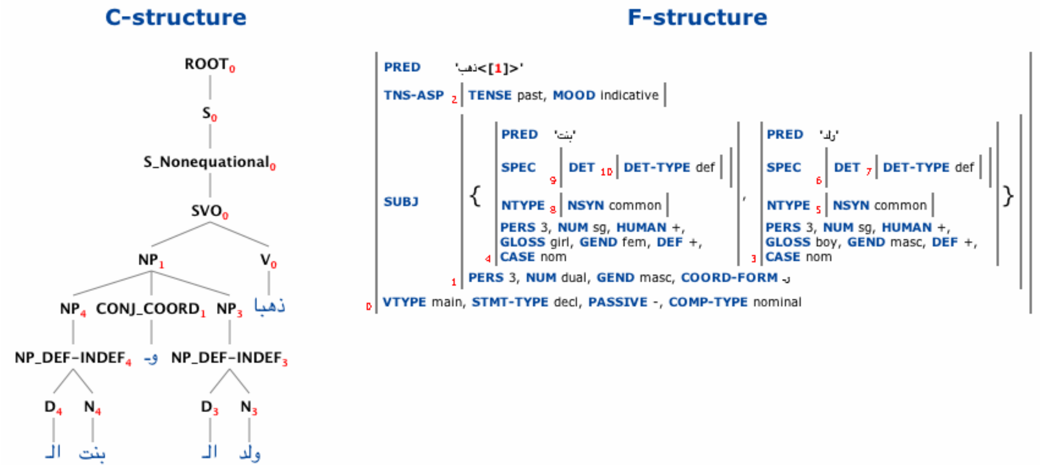


Figure 69. resolution of the agreement features in conjoined NPs

In our grammar we found that reducing the number of possible scopes of coordinated constituents helps greatly in reducing the ambiguity level for sentences that contain coordination. We did this by specifying which nodes in the phrase structure trees are allowed to undergo coordination and which are not. For example we only allow NPs to undergo coordination and forbid sub-categories under NPs (such as NP_Demonstrative and NP_Compound) from undergoing coordination. We also found that stating a condition that forbids conjoined NPs from having an embedded coordinated structure was also very effective in reducing ambiguity. It is possible that this constraint might rule out some valid analyses, but we did not see any good solutions pruned in the data we reviewed. However, this remains as an empirical question and the constraint can be modified in the light of new evidence. As for the agreement features, we found it hard to quantify how the correct specification of the agreement features of the conjoined NP is effective in reducing ambiguity as this has more to do with the well-formedness of the grammar.

6.3.3.2 PP attachment

Ambiguity in PP attachment arises when a PP can either modify the preceding verb (verb attachment) or the noun (noun attachment). The PP attachment problem in Arabic is magnified by the fact that not only the object follows the verb, but the subject as well. In the example in (٣٤٨) we see that the PP, ‘in the Middle East’ can modify the object noun, ‘democracy’, or the verb, ‘protect’.

(٣٤٨) يحمي الديمقراطية في الشرق الأوسط
yaḥmī ad-dīmuqrāṭiyyata fī aš-šarqi al-ʾawsaṭi
protect.sg.masc the-democracy in the-east the-middle
‘He protects democracy in the Middle East.’

There are examples when the PP has a sure attachment to the verb when the preceding noun is pronominal or proper noun, but in most other cases an ambiguity is created. Each additional PP increases the number of possible attachment solutions leading to increased ambiguity.

While there are cases of PP attachment ambiguity that certainly need some deep knowledge, simple superficial knowledge can be used as possible contextual cues to predict proper attachment. Due to the difficulty of modelling semantics and world knowledge in order to resolve the PP attachment ambiguity, researchers have considered word co-occurrence statistics in annotated corpora.

Hindle and Rooth (1993) are the pioneering researchers in looking for a solution for the PP attachment ambiguity problem using probabilistic methods. They proposed that PP attachment can be resolved on the basis of the lexical preference (or what they termed “lexical association”) by weighing the relative strength of association of the preposition with the preceding noun and verb, estimated on the basis of word distribution in a large corpus. They used human judges to decide the PP attachment for 880 test sentences on the basis of verb, noun and preposition alone, i.e. without seeing the rest of the sentence. The human judges had an average accuracy of 86%, while the lexical association procedure based on the co-occurrence frequency had an accuracy of 78%. This proves that the task of judging the PP attachment is neither easy nor error-free, even for human judges.

There have been several attempts to extend and improve Hindle and Rooth's (1993) model to try to achieve better results. Brill and Resnik (1994) applied their Error-Driven Transformation-Based Learning algorithm. First, unannotated text is passed through the initial-state annotator that assigns a default structure (right association). The text is then compared to a manually annotated corpus and transformations are learned. The learning is based on 4-tuples of v, n1, p, and n2, where n1 is the object and n2 is the object of preposition. They reported that this model yielded an accuracy of 81%.

Zavrel et al. (1997) used another statistical method, Memory-Based Learning, in trying to improve the performance of PP attachment resolution. They also took account of the object of preposition in their statistical analysis and achieved 84% accuracy.

In the field of psycholinguistics MacDonald et al. (1994) pointed out that "the relative plausibility of the alternatives suggests the preferred interpretation". They suggest that both nouns and verbs have different preferences about which thematic role is likely to be assigned by the preposition. They assume that there are three potentially highly constraining sources of biasing lexical frequency information for interpreting the PP attachment ambiguity:

- **The verb.** Action verbs tend to occur with modifying PPs (conveying instrument or manner roles) more than perception or psychological (mental state) verbs.
- **The noun.** Nouns which have argument structure representations tend to occur with modifying PPs, for example, nouns related to communication (*mail*, *message*, etc.) occur often with theme (*mail about the parking situation*).

- **The PP.** Location PPs (*in the room, next to the nightstand*) are neutral in their attachment preferences, while temporal PPs, such as *in three minutes*, tend to have verb attachment preferences. Prepositions themselves provide highly constraining information. The preposition *of*, for example, nearly always attaches to a preceding noun and assigns an attribute (*book of poems*) or theme role (*destruction of the city*), whereas prepositions such as *into*, *onto,* and *to* nearly always assign a goal role (*took the dog into the house*). The preposition *with*, on the other hand, is ambiguous as it assigns a broader range of roles, including manner, instrument, attribute, and location.

In our application, drawing on all of the above mentioned ideas, we set a wide range of constraints to decide whether the PP functions as an oblique object to a verb, a modifier of a noun or an adjunct to the sentence. These constraints take into account the nature of the verb, noun, preposition and prepositional object. The major weakness of our work with PP attachment resolution is that the constraints are based mainly on intuition and observation of a small subset of data, and not on any corpus-driven statistics. Although these constraints help in keeping the PP attachment within a reasonable boundary, PP attachment is still a hot spot of ambiguity in our grammar.

In this section we are going to show hard-coded constraints used in the grammar to constrain obliques in subcategorization frames, limit the effect of word order flexibility, constrain sentential PP adjuncts, constrain copula PP complement, and constrain PPs modifying nouns.

Constraining obliques in subcategorization frames

In section 6.3.2 on Lexical Specification we saw how verbs and verbal nouns which subcategorize for obliques are constrained by specifying the lexical form of the preposition. We showed that removing this specification affected 49% of the test suite and increased the average number of optimal solutions from 135 to 159 and the average number of suboptimal solutions almost tripled from 1.45E+04 to 3.38E+04.

It must be noted here that it is not only verbs and verbal nouns that can subcategorize for obliques. Common nouns and adjectives can subcategorize for PPs as well, and the lexical form of the preposition must be clearly stated, as shown in (٣٤٩) and (٣٥٠) which link to the template in (٣٥١). The template uses the constraining equation to define the lexical form of the preposition that must be used with these nouns and adjectives. The specification of obliques for some nouns and adjectives led to a reduced level of ambiguity for further 4% of the test suite.

(٣٤٩) a. حرب [*war*] N XLE @(TakesOblPP %stem على [*on*]).
b. طريق [*road*] N XLE @(TakesOblPP %stem لـ [*to*]).

(٣٥٠) a. مملوء [*full*] ADJ XLE @(TakesOblPP %stem بـ [*of*]).
b. ضروري [*necessary*] ADJ XLE @(TakesOblPP %stem لـ [*for*]).

(٣٥١) TakesOblPP(P_ Prep_) = (^ PRED)='P_<(^ OBL)>' (^ OBL OBJ PCASE)=c Prep_.

Constraining word order flexibility

In Arabic the oblique generally follows the subject. In a few cases the oblique precedes the subject, as in (352).

(352) وافقت عليها جميع الفصائل الفلسطينية
wāfaqat ʿalai-hā ǧamīʿu al-faṣāʾil al-filisṭīniyyah
agreed on-it all the-factions the-Palestinian
‘All Palestinian factions agreed on it.’

Such cases, i.e. where the oblique precedes the subject, must be explicitly constrained. The conditions in (353) help to constrain in the grammar in three ways. First, the subject must not be a ‘pro’, which means that it can be neither a pronoun nor a pro-drop. Second, the object of the oblique must be a pronoun. Third, the oblique is allowed to precede the subject only with certain verbs. This constraint helped in reducing ambiguity for 6% of the test suite.

(353) (PP: (^ OBL)=!
(^ SUBJ PRED)~='pro'
(^ OBL OBJ)='pro'
{(^ PRED)=c 'سقط' |(^ PRED)=c 'شعر' |(^ PRED)=c 'وافق' })
NP: (^ SUBJ)=! (! CASE)=nom)

Similarly, obliques generally follow objects, but in a few cases an oblique may precede the object, as in (354).

(354) يملي على الدول كيفية إدارة شئونها
yumlī ʿalā ad-duwali kaifiyyati ʾidārati šuʾūni-hā
dictate on the-countries how managing affairs-its
‘He dictates to the countries how to manage their affairs.’

In our grammar this is constrained only by the lexical form of the verb. This is shown by the code in (355) which stipulates that the oblique is allowed to precede the object only with certain verbs. This constraint helped reduce ambiguity for 5% of the test suite.

```
(355)    (PP: (^ OBL)=!
              {(^ PRED)=c 'أملى'  |(^ PRED)=c 'ألحق'  |(^ PRED)=c 'أطلق' })
          NP: (^ OBJ)=! (! CASE)=acc)
```

Constraining sentential adjuncts

Perhaps the constraint with the greatest impact in our grammar is that concerning the PP attachment as a sentential adjunct. A sentential adjunct PP (or the parenthetical phrase) can occur anywhere in the sentence, yet it usually prefers the final position, as shown in (356).

(356) عززت الشرطة الإجراءات الأمنية بعد الهجمات
ʿazzazat aš-šurṭah al-ʾiğrāʾāti al-ʾamniyyati baʿda al-hağamāt
reinforced the-police the-measures the-security after the-attacks
‘The police reinforced the security measures after the attacks.’

The code in (357) shows three sorts of constraints. First, it is constrained by the lexical form of the preposition alone. Second, it is constrained by the lexical form of the preposition along with the lexical form of the object of preposition. Third it is constrained by permitting the morpho-syntactic class of the prepositionals (ADVPREP) in this position. Prepositionals are prepositions derived from nouns taking the accusative case (considered by traditional Arabic grammarians as adverbs) such as بين baina ‘between’, and تحت taḥta ‘under’. These constraints led to a reduced level of ambiguity in 56% of the test suite.

```
(357)   PARENP_PP --> PP:
                  {(^ OBJ PCASE)=c في
                  | (^ OBJ PCASE)=c بـ
                  | (^ OBJ PCASE)=c من (^ OBJ PRED)=c 'جانب'
                  | (^ OBJ PCASE)=c لدى
                  | (^ OBJ PCASE)=c منذ
                  | (^ OBJ PCASE)=c على
```

```
                    {(^ OBJ PRED)=c 'حد'
                     |(^ OBJ PRED)=c 'سبيل'
                     |(^ OBJ PRED)=c 'خلفية' }
          | (^ OBJ PCASE)=c لـ
                    {(^ OBJ PRED)=c 'سنة'
                     |(^ OBJ PRED)=c 'فترة'
                     |(^ OBJ SUBJ)}
          | (^ OBJ PCASE)=c إلى (^ OBJ COMP-FORM)=c أن
          | (^ ADVPREP)=c +}.
```

Constraining Copula Complement PP

The copula complement can be realized as a prepositional phrase, as in (358).

(358) كان الانفجار في العاصمة
kāna al-infiǧāru fī al-ʿāṣimah
was the-explosion in the-capital
‘The explosion was in the capital.’

The PP in the copula complement position can also be constrained. The condition in (359) shows that not all prepositions are allowed to head a copula complement, but only a limited number of prepositions can occupy this position. This constraint led to reducing ambiguity in 12% of the test suite.

```
(359)  PP: (^ PREDLINK)=!
        {(! OBJ PCASE)=c في
         |(! OBJ PCASE)=c على
         |(! OBJ PCASE)=c لدى
         |(! OBJ PCASE)=c فوق
         |(! OBJ PCASE)=c تحت
         |(! OBJ PCASE)=c عند
         |(! OBJ PCASE)=c لـ }
```

Constraining PPs Modifying Nouns

The PP can be attached to a noun and function as an adjunct or modifier to this noun, as in (360).

(360) الإصلاحات الديمقراطية في الشرق الأوسط
al-ʾiṣlāḥātu ad-dīmuqrāṭiyyati fī aš-šarqi al-ʾawsaṭi
the-reforms the-democratic in the-east the-middle
‘the democratic reforms in the Middle East’

The code in (361) shows four different types of constraints. First, the noun must not be a proper name. Second the object of the preposition cannot be a verbal noun. Third, there is a constraint on the lexical form of the preposition. Fourth, there is a constraint on the lexical form of the preposition and the object of preposition. This set of constraints led to a reduction in ambiguity in 33% of the sentences in the test suite.

```
(361)   PP-NounAdjunct -->  PP: ! $ (^ ADJUNCT)
        (^ NTYPE NSYN)~=proper
        ~(! OBJ SUBJ)
        {(! OBJ PCASE)=c في
          | (! OBJ PCASE)=c بـ (! OBJ PRED)~='سبب'
          | (! OBJ PCASE)=c على
          | (! OBJ PCASE)=c من
          | (! OBJ PCASE)=c إلى (! OBJ PRED)=c 'جانب'
          | (! OBJ PCASE)=c لـ
          | (! OBJ PCASE)=c بين
          | (! OBJ PCASE)=c حيال
          | (! OBJ PCASE)=c عن}.
```

6.3.3.3 Mending Non-exclusive Disjunctions

The resolution of disjunctive feature constraints is computationally expensive, and may be exponential in the worst case (Frank, 1999). Disjunctions are the alternative paths that a rule can take. If these are not clearly defined in order to be mutually exclusive, they inevitably lead to an overflow in the number of generated solutions. Reviewing these rules leads to removing a considerable amount of spurious ambiguities.

As a hypothetical problem to show the power of non-exclusive disjunctions in generating a large number of ambiguities, we changed the condition in (362) which states the that the case of the subject NP must be nominative into the condition in (363) which contains a non-exclusive disjunction and which states that the case of the subject NP is either nominative or not accusative or not genitive. This single change affected 72% of the sentences in the test suite, and the affected sentences had a three-fold increase in the number of possible solutions.

```
(362)   NP: (^ SUBJ)=! (! CASE)=nom.
(363)   NP: (^ SUBJ)=! {(! CASE)=nom | (! CASE)~=acc | (! CASE)~=gen}.
```

In real-life situations non-exclusive disjunctions are the most intricate and hardest to discover and fix. This is why the XLE platform comes with a built-in utility “check-grammar-disjuncts” for spotting non-exclusive disjunctions.

Back in the history of our grammar development, changing the way a rule was written to avoid a non-exclusive disjunction led to a huge reduction in parse time by 68%. The number of subtrees was reduced then by approximately 10%.

6.3.4 Domain Specific Adaptation

Concentrating on a domain reduces ambiguity by allowing us to focus on selective access to lexical entries and syntactic structures and to avoid needless details in both levels.

MacDonald et al. (1994) maintained that contextual information can result in the activation of a single meaning of an ambiguous word, and cited Duffy et al. (1988) as assuming that the context can “reorder” access to meanings by promoting or demoting interpretations. Though they meant context in the narrow sense of adjacent words, we can use context here in a broader sense of the domain or field in which the discourse is used.

We focused on the news domain in our research. We assume that the imperative mood and the interrogative constructions are not expected to occur in news articles with significant frequency. As we have shown earlier in section 6.2.2 on morphology, removing the constraints on forming the passive voice affected 1058 verbs out of 1640, and removing the constraints on imperative mood affected 1110 verbs. When the new version of the morphology (after removing the constraints on forming the imperative and passive forms) was integrated into the grammar, the number of possible parses was significantly increased for 17% of the sentences in the test suite, and the parse time was increased by 9%. When interrogative construction was commented out from the grammar, this led to a further 3% reduction in parse time.

The morphology is restricted to the domain of MSA, and, therefore, Classical Arabic forms were avoided. Some proper names are associated with classical senses that are no longer used in the language, such as those listed in (364). Some classical entries are totally no longer in use, such as those in (365). All these forms are homographic with other forms that are in contemporary usage and their inclusion would only complicate the ambiguity problem.

(364) a. حسام ḥusām ‘Husam / sword’
b. حنيفة ḥanīfah ‘Hanifah / orthodox’

(365) a. قف qaffa ‘to dry’
b. أبد ’abada ‘be untamed’
c. أب ’abba ‘desire’

The grammar is restricted to MSA constructions as well, and, therefore obsolete constructions, such as the OVS word order, are forbidden. Infrequent constructions, such as the VOS word order, are highly constrained. We have shown in section 6.1.2 on word order ambiguity how the VOS word order converges with VSO to create an ambiguity issue. The nominative and accusative cases which distinguish the subject and the object, and which are normally marked by diacritics, do not show in the surface forms. This leads every VSO sentence to have a VOS interpretation causing an ambiguity problem. In our grammar allowing the VOS beside VSO without any constraints almost doubled the number of ambiguities for 15% of the sentences.

Such domain-oriented adaptation relieves the load on the system and makes the parser faster and more efficient.

6.4 Disambiguation in post-parsing phase

The post-parsing stage has no effect on the number of solutions already produced by the parser, as it neither increases nor decreases the number of parses. This stage is primarily responsible for controlling the presentation, analysis, and reordering the ranking of these solutions.

The post-parsing phase covers the issues of packing ambiguities, optimality marks for preferences, using discriminants, and stochastic disambiguation. These are considered as add-on utilities that complement the core parsing system.

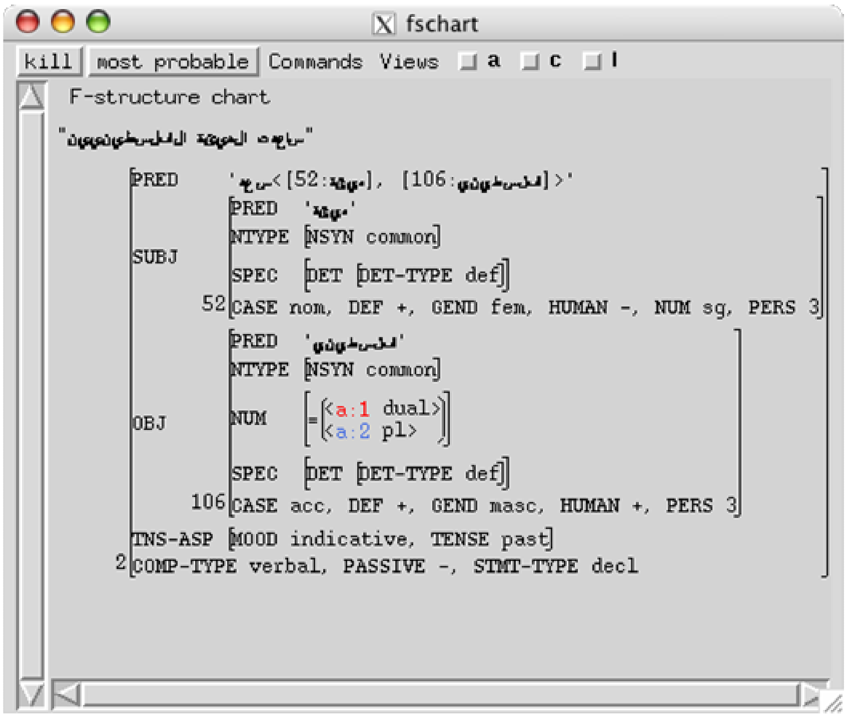
6.4.1 Packing ambiguities

The need for packed representation stems from the fact that developers usually need to determine the source of the multiple solutions the parser produces (King et al., 2000). Some parses are legitimate, while others are spurious and need to be eliminated. Searching through the parse forest by examining one solution after the other can be tedious, time consuming and impractical if not impossible. Grouping the solutions in packed representations can effectively speed up the process of detection and revision.

This is why XLE comes with a built-in facility for showing packed representations of the alternative solutions (King et al., 2000). When a sentence is parsed, XLE displays four windows: c-structure, f-structure window, f-structure chart window which shows a packed representation of all analyses, and a chart-choices window. The example in (366) shows a real ambiguity which arises from the fact that the noun in the object position is morphologically ambiguous and can be interpreted as either plural or dual.

(366) ساعدت الهيئة الفلسطينيين
sāʿadat al-haiʾatu al-filisṭīniyyīn/ al-filisṭīniyyain
helped the-agency the-Palestinian.pl/ the-Palestinian.dual
‘The agency helped the Palestinians/ the two Palestinians.’

The f-structure chart window in Figure 70 provides a list of choices that are caused by alternative solutions. Please note that the Arabic letters are distorted in the XLE chart, due to the fact that the Mac OS shell does not link the Arabic letters properly.



```
F-structure chart
"ساعدت الحكومة الفلسطينيين"

     PRED     'ساعد<[52:حكومة], [106:فلسطيني]>'
                PRED   'حكومة'
                NTYPE  [NSYN common]
     SUBJ
                SPEC   [DET [DET-TYPE def]]
             52 CASE nom, DEF +, GEND fem, HUMAN -, NUM sg, PERS 3
                PRED   'فلسطيني'
                NTYPE  [NSYN common]
     OBJ        NUM    [=(<a:1 dual>
                          <a:2 pl>  )]
                SPEC   [DET [DET-TYPE def]]
            106 CASE acc, DEF +, GEND masc, HUMAN +, PERS 3
     TNS-ASP  [MOOD indicative, TENSE past]
   2 COMP-TYPE verbal, PASSIVE -, STMT-TYPE decl
```

Figure 70. F-structure chart for packed ambiguities in XLE

Figure 70 shows two alternative analyses of the sentence in (366). These analyses are identical in all respects except for the value of the feature NUM for the NP in the object position, which may be dual or plural. In the f-structure chart window, the two values are labelled as *a:1* and *a:2*. These labels are active. If the user clicks on a choice the corresponding solution is displayed in the c-structure and f-structure windows. This facility is useful in grammar debugging and development.

6.4.2 Optimality marks for preferences

The Optimality Theory (OT) was first developed by Prince and Smolensky (1993) for phonology, but later it was extended to other fields such as syntax and semantics. The model used in LFG for ranking preferences and constraints (sometimes referred to as OT-LFG) is inspired by the OT, but does not strictly comply with the principles of the original theory. The major difference between OT and it application in LFG is that in the original OT there are no rules and no hard inviolable constraints. All constraints are ranked in a hierarchy and they are all violable. Choosing a form is based on resolving the conflict between competing constraints, maintaining that violations of high ranking constraints is more serious than violations of low ranking ones (Kager, 2000, Wunderlich, 2005). In LFG, however, the idea is different. There are hard-coded rules, constraints, and disjunctions (or options). The disjunctions are ranked so that preferred solutions can be filtered from dispreferred ones.

Optimality ranking in LFG also functions as a weighting approach that gives the grammar writer control over the means of expression to filter implausible readings (Kuhn and Rohrer, 1997).

In LFG, OT is a projection (o-projection or o-structure) used on top of grammar constraints to rank alternative paths of the phrase structure or alternative features in the f-structure (Frank et al., 2001). It takes the form in (367) which states that the Arabic sentence, S, is expanded into a V followed either by an explicit NP subject or a pro-drop subject, and each choice is marked with an o-projection mark for preference ranking.

(367)
$$\mathrm{S} \rightarrow \underset{\uparrow=\downarrow}{\mathrm{V}} \left\{ \begin{array}{c} \mathrm{NP} \\ (\uparrow \mathrm{SUBJ})=\downarrow \\ \mathrm{MARK1} \in \mathrm{o}^* \\ \\ \varepsilon \\ (\downarrow \mathrm{PRED})= \text{'pro'} \\ (\uparrow \mathrm{SUBJ})=\downarrow \\ \mathrm{MARK2} \in \mathrm{o}^* \end{array} \right\}$$

Optimality marks have proved to be the most effective utility in ambiguity management in the post parsing phase by limiting the number of possibilities according to a predefined set of criteria for preferences and dispreferences. Table 14 shows how optimality ranking is very effective especially when the number of possible solutions grows dramatically. The number before the ‘+’ sign in the table is the number of optimal solutions and the number after the sign is the number of suboptimal solutions. Frank et al. (2001) maintained that OT marking is an effective mechanism in filtering syntactic ambiguity, even though the preference constraints are constantly faced with exceptions and counterexamples.

#	Sentence	Solutions
1	وأوضح أن كثيرا من الدول العربية منها البحرين والأردن والمغرب بدأت تسعى إلى تحقيق الديمقراطية، مضيفا أن المستقبل سيكون أفضل بمشاركة المرأة بشكل كامل في المجتمع. wa-ʾawḍaḥa ʾanna kaṯīran mina ad-duwali al-ʿarabiyyati and-clarified that many of the-countries the-Arab min-hā al-baḥrain wa-l-ʾurdun wa-l-maġrib badaʾat from-them Bahrain and-Jordan and-Morocco started tasʿā ʾilā taḥqīqi ad-dīmuqrāṭiyyat, muḍīfan ʾann try to achieving the-democracy, adding that al-mustaqbala sayakūnu ʾafḍala bi-mušārakati the-future will-be better with-participation al-marʾati bi-šaklin kāmilin fī al-muğtamaʿi the-woman in-way complete in the-society ‘And he pointed out that many Arab countries including Bahrain, Jordan and Morocco started to try to apply democracy, adding that the future will be better with the full participation of women in the society.’	4+260
2	وأوضحت المذكرة أن قدرة السلطات الأفغانية على الحفاظ على النظام وضمان أمن المواطنين والزوار محدودة. wa-ʾawḍaḥati al-muḏakkiratu ʾanna qudrata as-suluṭāti and-clarified the-memorandum that ability the-authorities al-ʾafġāniyyata ʿalā al-ḥifāẓi ʿalā an-niẓāmi the-Afghani on the-preserving on the-order wa-ḍamāni ʾamni al-mwāṭinīna wa-l-zuwwāra maḥdūdah and-safeguarding safety the-citizens and-the-visitors limited ‘The memorandum pointed out that the ability of the Afghani authorities to keep order and safeguard the safety of citizens and visitors is limited.’	30+266
3	ويعد خطاب بوش أحدث محاولة من جانبه ليبرر الحرب على العراق بأنها كانت ضرورية لتبني الديمقراطية في المنطقة في وقت يجد فيه نفسه يتعرض لانتقادات wa-yuʿaddu ḫiṭāba būš ʾaḥdaṯa muḥāwalatin min and-considered speech Bush most-recent attempt on ğānibi-hi li-yubarrira al-ḥarba ʿalā al-ʿirāqi bi-ʾanna-hā kānat part-his to-justify the-war on Iraq by-that-it was ḍarūriyyatan li-tabannī ad-dīmuqrāṭiyyati fī al-minṭaqati fī waqtin necessary to-adopt the-democracy in the-region in time yağidu fī-hi nafsahu yataʿarraḍu li-intiqādātin find in-it himself being-subject to-criticisms ‘And Bush’s speech is considered as the most recent attempt on his part to justify the war on Iraq that it was necessary to spread democracy in the region at a time that he finds himself subject to criticisms.’	32+4304

Table 14. Effect of optimality in reducing the number of possibilities

Optimality marks in LFG are a means to express a dispreference for infrequent readings, without having to rule them out, as in a different context these readings may be the most plausible, or even the only possible, analysis (Frank et al., 2001). OT can also be very valuable in domain specific applications where certain constructions and choices need to be eliminated in specific domains. The OT mechanism can also increase the robustness of a grammar by adding low-ranked fallback rules, which allows for parsing common grammatical mistakes and marginal constructions (Frank et al., 2001).

There are different types of Optimality Marks for selecting plausible solutions and performing various other purposes. These marks are explained as follows (mainly from Frank et al., 2001).

1. **Preference Marks**. Preference marks are prefixed with a plus sign and are used when one choice is preferred. In the Arabic grammar preference marks are used for function words. They are also used to give preference to obliques over adjuncts.

2. **Dispreference Marks**. Dispreference marks are used for rare constructions. The marks make sure that these constructions surface only when no other, more plausible, analysis is possible. In our grammar these are used to mark the first and second person (as they are not typically found in the news domain), and to mark the passive readings. It is also used with demonstrative pronouns when they function as NPs, as this is an unlikely possibility in the grammar.

3. **STOPPOINT Marks**. According to the XLE manual[5] the STOPPOINT marks are used to allow XLE to process the input in multiple passes, using larger and larger versions of the grammar. STOPPOINTS are processed in order from right to left, so that the first STOPPOINT considered is the rightmost. Constructions marked with STOPPOINTs are tried only when the system fails to find optimal or suboptimal parses. STOPPOINTs are useful for speeding up the parser by only considering rare constructions when no other analyses are available. In our grammar STOPPOINTS are used for uncommon morphological forms and

[5] http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html

uncommon grammatical constructions. They are also used to mark the case when adjectives function as NPs.

4. **UNGRAMMATICAL Marks**. These are used to increase the robustness of the grammar by allowing it to parse ungrammatical constructions which do not conform to the well-formedness constraints, such as relaxing the subject–verb agreement constraint in English. We did not use this mark in our grammar as robustness is not a primary objective in the current stage.

5. **NOGOOD Marks**. These marks indicate that the analysis is always bad, even if there are no other analyses. They are used to comment some rules out of the grammar. Kuhn and Rohrer (1997) used NOGOOD marks to create different grammar versions with switched off parts of rules. This mark is used in our grammar to exclude the imperative and interrogative constructions, and to comment out some subcategorization frames.

OT-LFG has been extended by allowing the system to learn preference ranking from a corpus, and also the preferences are ranked on a continuous scale of numbers (Forst et al., 2005, Kuhn, 2002).

6.4.3 Using discriminants

The parser usually produces tens, hundreds and sometimes even thousands of solutions. In this case, reviewing the solutions by hand to select the correct one becomes a tedious, impractical and even impossible task. To deal with this problem, the research group at Bergen University utilized a smart and efficient disambiguation process based on the use of *discriminants* in the TREPIL project (Norwegian treebank pilot project 2004-2008) which is mainly aimed at the construction of a Norwegian parsed corpus (Oepen and Lønning, 2006, Rosén et al., 2005a, Rosén et al., 2005b, Rosén et al., 2006).

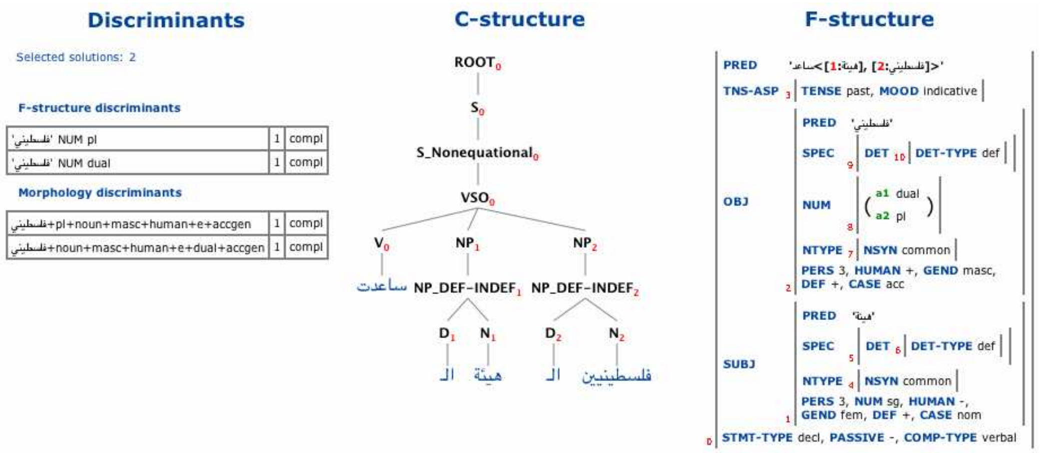
Rosén et al. (2005a) believed that treebank construction on the basis of automatic parsing is more desirable than manual annotation, as manual annotation is costly and prone to errors and inconsistencies. Treebank construction can be an immediate benefit of the grammar, as treebanks are highly in demand for statistical research that will ultimately benefit the grammar by allowing it to expand and incorporate accurate information on the frequency of structures and subcategorization frames (Rohrer and Forst, 2006). Stochastic disambiguation techniques require the existence of a treebank to extract the frequency information about parse choices (Riezler et al., 2002). Moreover a complete probabilistic parser can be constructed through methods of induction from a treebank (O'Donovan et al., 2004).

As part of the work in the TREPIL project, the XLE Web Interface (XLE-Web) is developed as a web-based tool for parsing with XLE and viewing c-structures and f-structures along with the discriminants for disambiguation. In our work with the Arabic grammar we found that the XLE Web Interface was of immediate benefit as it allowed us to view our parse results in a better platform that renders the Arabic characters correctly.

Discriminants are defined by Carter (1997) (as cited by Rosén et al., 2005a) as small independent choices which interact to create dozens of analyses. It is maintained that disambiguation can be done more quickly and efficiently if it is based on these elementary linguistic properties, or discriminants, than if it is based on the solutions themselves. Rosén et al. (2005a) defined a discriminant in LFG terms as “any local property of a c-structure or f-structure that not all analyses share.” They classified discriminants into three types: c-structure discriminants which deal with node branching, f-structure discriminants, which deal with the feature–value matrices in the f-structure, and morphology discriminants which deal with the different tags received from the morphological processor.

For the example in (366), repeated here for convenience as (368), Figure 71 shows the XLE-Web interface with the f-structure, c-structure and list of discriminants. This example has one morphology discriminant that is reflected in the f-structure, which also shows one discriminant. The discriminants are the different values of the NUM feature. The discriminants are active links so that a human disambiguator can choose a discriminant by clicking on it, or reject it by clicking on *compl*, i.e. complement.

(368) ساعدت الهيئة الفلسطينيين
sāʿadat al-haiʾatu al-filisṭīniyyīn/ al-filisṭīniyyain
helped the-agency the-Palestinian.pl/ the-Palestinian.dual
'The agency helped the Palestinians/ the two Palestinians.'



Selected solutions: 2

F-structure discriminants

'فلسطيني' NUM pl	1	compl
'فلسطيني' NUM dual	1	compl

Morphology discriminants

فلسطيني+pl+noun+masc+human+e+accgen	1	compl
فلسطيني+noun+masc+human+e+dual+accgen	1	compl

```
PRED       'ساعد<[1:هيئة], [2:فلسطيني]>'
TNS-ASP  3 TENSE past, MOOD indicative
OBJ        PRED   'فلسطيني'
           SPEC 9 DET 10 DET-TYPE def
           NUM  8 (a1 dual
                   a2 pl)
           NTYPE 7 NSYN common
         2 PERS 3, HUMAN +, GEND masc,
           DEF +, CASE acc
SUBJ       PRED   'هيئة'
           SPEC 5 DET 6 DET-TYPE def
           NTYPE 4 NSYN common
         1 PERS 3, NUM sg, HUMAN -,
           GEND fem, DEF +, CASE nom
0 STMT-TYPE decl, PASSIVE -, COMP-TYPE verbal
```

Figure 71. XLE-Web interface showing discriminants

6.4.4 Stochastic disambiguation

MacDonald et al. (1994) believed that frequency affects the order in which the meanings of ambiguous words are accessed. They, however, pointed out that most theoretical linguists in the past have shown little interest in issues concerning statistical properties of language, as linguistics at the time was influenced by Chomsky's argument that the concept of well-formedness of syntactic structures cannot be accounted for by using statistics. MacDonald et al. (1994) said that the Chomskian example, *Colorless green ideas sleep furiously*, used to be quoted to show how nonsensical sentences with a low frequency can still be considered grammatical. They, however, maintained that this view has changed as frequency information has come be acknowledged as relevant to sentence comprehension.

MacDonald et al. (1994) emphasised that even when the grammar admits multiple alternative interpretations at a given level of representation, they often differ substantially in frequency and thus have different probability. These probabilistic sources of information interact to allow the system to settle on a certain choice and discard the others.

Manning and Schütze (1999) maintained that hand-crafted rules are ineffective in resolving the ambiguity problems and that a statistical NLP model, which learns lexical and structural preferences from corpora offers a better solution to the ambiguity problem as “statistical models are robust, generalize well, and behave gracefully in the presence of errors and new data.” They also pointed out that the parameters of statistical disambiguation can be learned from corpora, reducing the need for human effort in producing NLP systems.

In the LFG literature, stochastic disambiguation is used as an automatic probability-based disambiguation component. It relies on an already annotated corpus to compute the probability of alternative parses and assign a score to each alternative. This approach however has to face a classical problem, that is the quality and size of the treebanks used. If the functional annotations in the treebanks are rudimentary and the size of treebanks is small the application of statistical estimation will be hindered (Riezler et al., 2002).

In our situation, we did not use this utility, since it requires costly data preparation to obtain labelled trees. Even if a treebank does exist, manual work is required to make sure that the annotation is consistent with LFG formalisms. There is an Arabic treebank available in the LDC, but due to time and scope limitations we could not acquire this treebank to explore the feasibility of using it in probabilistic disambiguation.

7 Grammar Development, Testing and Evaluation

The syntactic parser for Arabic is developed within the framework of LFG (Lexical Functional Grammar) (Bresnan, 2001, Kaplan and Bresnan, 1982). In this parser a cascade of finite state transducers are used to cover the preprocessing phases: normalization, tokenization, morphological transduction, and multiword expressions transduction. Beside core transducers there are backup transducers to provide estimates when exact analyses are not possible. These backup guessers are the non-deterministic tokenizer and the morphological guesser. Tools for processing the corpus by breaking a running text into sentences and providing frequency statistics on lexical entries are developed in Visual Basic.

The Arabic grammar at the current stage has 57 grammar rules and 13,600 morphological entries. The corpus we use contains 5.3 million words in 17,958 articles comprising 209,949 sentences of news articles (Al-Jazeera news articles for the year 2003 and the first half of 2004). The average sentence length in this corpus is 25 words. The Arabic grammar parser on its own in this stage provides 33% coverage (complete parses) for short sentences (10 to 15 words) of an unseen subset of the data. The coverage is extended to 92% using robustness techniques, such as morphological guessers and a fragment grammar. Although the grammar does not target longer sentences at this stage, just to have a rough idea about the grammar coverage regarding longer sentences, it was found that the grammar provided 16% coverage (complete parses) for sentences ranging between 16 and 25 words in length.

This chapter starts by discussing the development of hand-crafted rule-based grammars, showing that it is not usually a fast process, but it takes years of building and investigation. We then explain the stages of Arabic grammar development and the tools used for processing the corpus for the purpose of testing and developing the grammar. We then report on an evaluation experiment conducted on unseen data to show how much coverage the grammar has achieved at the current stage. We also apply a set of robustness tools (guessers and fragment grammar) and show how these utilities are effective in increasing the coverage, and providing useful pieces of information, when complete parses are not possible.

7.1 How Fast Can a Grammar Be Developed?

In the available publications within the ParGram project we found that the German grammar is the best documented with regards to the coverage at different stages and at different points in the development history. So we decided to trace German to see how fast the grammar has developed since its inception until the present day.

Work started on the German LFG-based grammar sometime before 1999 and was reported by Butt et al. (1999b). In an evaluation experiment in 2000, the grammar covered 35% of free newspaper text (Dipper, 2003). In 2003 the grammar achieved 50% coverage (Rohrer and Forst, 2006). Forst and Kaplan (2006) reported that “The revised tokenizer increases the coverage of the grammar in terms of full parses from 68.3% to 73.4% on sentences 8,001 through 10,000 of the TiGer Corpus.” In 2006, Rohrer and Forst (2006) reported that “In parsing the complete treebank, 86.44% of the sentences receive full parses.”

From this data we can draw a timeline for the growth in coverage in the German grammar as shown in Figure 72. Yet it must be noted that the testing and evaluation experiments mentioned above are not homogeneous as some of them were conducted against free newspaper texts while others against the TiGer treebank.

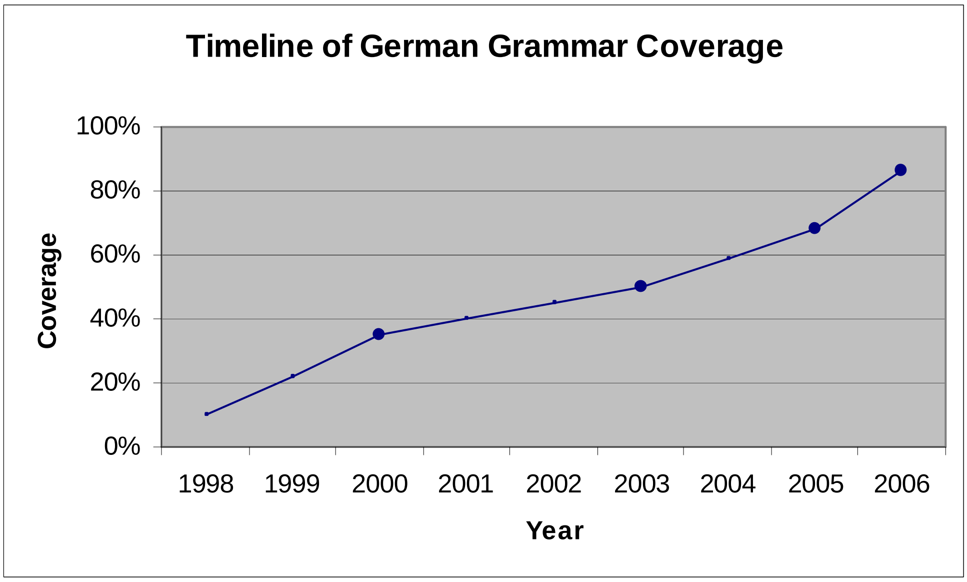


Figure 72. Timeline of German grammar coverage

Regarding the English grammar, work started well before 1999 (Butt et al., 1999b). We managed to find only two coverage evaluation experiments.

Riezler et al. (2002) reported that the grammar provided 74.7% coverage as full parses for section 23 of the Wall Street Journal corpus. Kaplan et al. (2004) reported that the XLE grammar achieved 79% coverage as full parses of section 23 of the Wall Street Journal corpus. From these two experiments we can draw an (admittedly more tentative) indicative timeline of the growth of English grammar coverage, as shown in Figure 73.

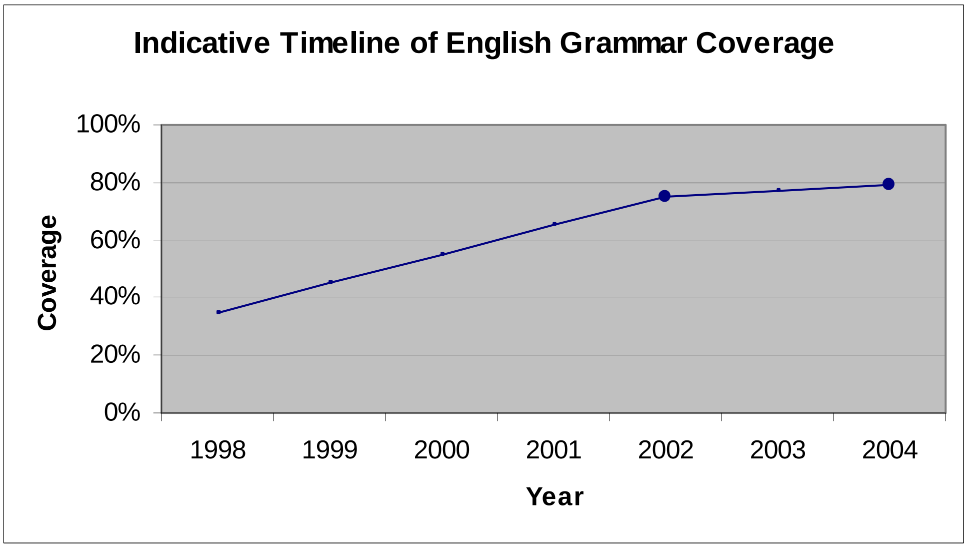


Figure 73. Indicative timeline of English grammar coverage

The data presented for German and English grammars above shows that the development of a hand-crafted rule-based grammar is not usually a fast process, but it takes years of building and investigation. The development is usually hampered by the need to study and analyse subtle and complex constructions.

7.2 Stages of Arabic Grammar Development

Among the ParGram community, there is no agreed-upon set of methodologies and strategies for developing a grammar. The seminal work in the project started with parsing a tractor manual in different parallel languages (Butt et al., 1999b). The aim of the project was to explore the different syntactic structures in different languages within LFG and to ensure that maximal parallelism is maintained in their representations.

In our grammar development we wanted to achieve large-scale coverage of the Arabic news corpus. The development process passed through different stages of maturity and complexity starting with a "toy" stage and ending with a more focused perspective towards the gradable complexity levels.

7.2.1 Stage One

In the first stage of the development process, a test suite of 175 made-up sentences was created to aid the grammar in providing coverage of the basic Arabic sentence structures. The test suite included various possible word orders (VSO, SVO, VOS), copula-less constructions, gender and number variations, transitive and intransitive verb constructions, sentential and nominal modifications, coordination, questions, negations, demonstrative and relative clauses, complement phrases, compounding and sentences with multiword expressions. The development process in this stage is straightforward. The test suite has short sentences each representing one construction. There is no lexical variety or deep embedding, and usually there is no ambiguity or complexity characteristic of real-life data. Therefore the grammar at this stage was considered as a "toy" grammar.

7.2.2 Stage Two

In the second stage of development four articles were chosen from the corpus to be used as a reference for further development. We expanded our grammar rules and lexicons to accommodate the complexity and variability of the real life data. Eventually 79 real sentences of various lengths were parsed successfully. The shortest sentence is three words and the longest one is 46 words. Before we could analyse large sentences they had to be broken into smaller and more manageable chunks in order to narrow down a problem or focus on a certain structure which cannot be easily traced in the sentence as a whole.

This phase allowed the grammar to mature considerably, as it made the grammar see a sense of real-life data, and deal with high levels of complexity and variations. However, this strategy has its limitations. Four articles cannot be representative of a dataset consisting of 17,958 articles. Some constructions contained in the four articles may be too complex or rare, some subcategorization frames may be infrequent, and there were some typing or grammatical errors that we would not like to handle at this stage. During this stage the development process was slow as it was in many instances stuck with complex constructions, and we did not manage to yield reasonable coverage.

7.2.3 Stage Three

Therefore we decided to move to new criteria for selecting a reference set from the data. These criteria are based on sentence length. The concept of sentence length is a useful concept in both grammar development and grammar evaluation. This concept has been manipulated by many researchers working in the field of grammar development, whether hand-crafted or probabilistic. In the probabilistic paradigm, Charniak (1996) excluded in his experiment all sentences that exceeded 40 words in length on the grounds that their frequency is low and that the average sentence length in English is 22 words. In this regard, Arabic is somewhat similar to English. Based on a corpus of 5.3 million words comprising over 200,000 sentences of news articles we found that the average sentence length is 25 words. The frequency of sentences above 40 words is 6%. In the paradigm of rule-based grammar development, Maxwell and Kaplan (1996) show that there is a correlation between sentence length, parsing algorithm and parse time as shown in Table 15 reproduced from (Maxwell and Kaplan, 1996).

Order	10	20	30	40	50
O(n)	.01 sec.	.02 sec.	.03 sec.	.04 sec.	.05 sec.
$O(n^2)$	.1 sec.	.4 sec.	.9 sec.	1.6 sec.	2.5 sec.
$O(n^3)$	1 sec.	8 sec.	27 sec.	64 sec.	125 sec.
$O(n^6)$	17 min.	18 hours	8 days	47 days	180 days
2^n	1 sec.	17 min.	12 days	35 years	357 centuries

Table 15. Parsing time of different algorithms for sentences with different lengths

Table 15 above shows that as the sentence grows in length it also grows in complexity, and that beside the algorithm used in parsing, the parse time is also affected by the sentence length.

Realizing the importance of sentence length as a factor in complexity, we investigated the distribution of sentence length in the corpus and found out that sentences which are between 10 and 15 words represent 12% of the whole corpus, as shown in Figure 74. We found that sentences less than 10 words in length are most likely to include fragments (e.g. headings and captions), and not complete sentences. Sentences exceeding 50 words are more likely to include more than one sentence separated by a comma instead of a full stop.

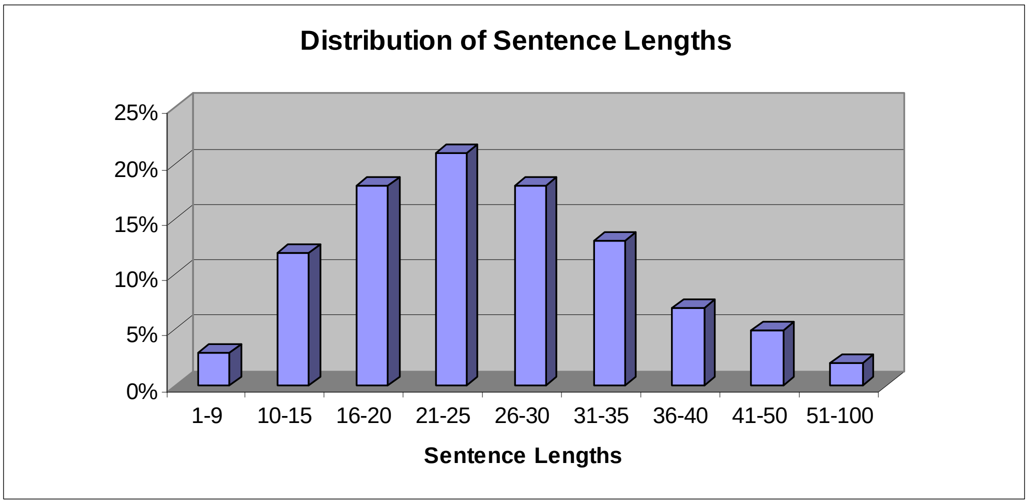


Figure 74. Distribution of sentence lengths in the Arabic corpus

Therefore, we randomly selected 175 sentences ranging between 10 and 15 words to be used as reference data. We found that the sentences in this category share the same characteristics. Sentences tend to be simpler and to avoid deep embedding. We found that a good strategy of grammar development could be to move from one grade of complexity to the next, and this can be done, to a great extent, by moving from one range of sentence length to the other.

Another possible venue to explore to enhance and expand the grammar coverage is using a treebank. Rohrer and Forst (2006) relied on a treebank to see where the grammar was incomplete and to determine the frequency of constructions.

Extending the coverage of the Arabic grammar may be possible by relying on the Penn Arabic Treebank (Maamouri et al., 2003). A treebank could be very helpful as it contains a lot of useful information on word categories and sentence structures. From a treebank we can extract statistical information on the distribution of syntactic structures: which structures are frequent and which are rare. Relying on a treebank could also help the grammar writers to base their judgment on realistic information instead of using personal judgments and intuitions. For example, with the aid of a treebank we can have material evidence that some sentence structures are no longer used in modern writing (such as the OVS word order), and therefore they can be eliminated from the grammar.

7.3 Corpus Tools for Grammar Development and Testing

The corpus is collected from articles published on the Al-Jazeera website[6] in the news domain between January 2003 and June 2004. It includes 17,958 articles, containing 5,300,481 words, and 209,949 sentences.

The 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 the 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

[6] http://www.aljazeera.net

minds." Al-Jazeera employs presenters and reporters from across the spectrum of the Arabic-speaking countries.

In order to collect data and process it, a number of text processing tools were developed. News articles were downloaded from the Al-Jazeera website. These articles were in HTML format and included a lot of tags that related to the presentation of the text but would not be relevant to further processing. On the contrary these tags were misleading to statistics on the number and frequency of words. Therefore, a tool was developed in Visual Basic to remove HTML tags from the files and to put all articles in a database, where they could be sorted according to subject or date.

Tools for segmenting running text into sentences were developed in Visual Basic, as well. We relied on the period as a mark for demarcating the sentence boundary. However, the reliance on the period as a sentence delimiter is contested because the use of punctuation marks in Arabic, in general, is not systematic. To move from one idea to the next an Arab author might use a period, a comma or even a conjunction. In our extended corpus of Al-Jazeera articles of all subjects (news, arts, sports, interviews, etc.) from 2001 until June 2004 we found 23,102 sentences out of 1,180,643 sentences (26,640,519 words) had 100 words or more, with the largest sentence reaching 803 words. This means that a whole article might be expressed in one sentence, or more accurately in sentences demarcated by means other than period. We found that sentences below 50 words in length are more likely to be long sentences, and sentences of 50 words or more are more likely to be two or more sentences joined together by a comma or a coordinating or resumptive conjunction (a feature in Arabic discourse which is not known in Indo-European languages).

In the news domain things are not as difficult since only 15 sentences in our section of the news corpus (January 2003 until June 2004) reach or go beyond the 100 words threshold. Maybe the influence of translations from European news agencies influenced a strong tendency towards a more systematic use of punctuation in Arabic. In our news domain it has been found that the period can function satisfactorily as a marker of sentence boundaries. Table 16 shows the statistics of the Al-Jazeera news corpus.

Number of articles	17,958
average number of words per article	295
Average number of sentences per article	11
Average number of words per sentence	25
The following statistics exclude tags, headings and captions	
Total number of words	5,300,481
Total number of sentences	209,949
Sentences less than 10 words	6,352
Sentences between 10 and 15 words	26,098
Sentences between 16 and 20 words	38,419
Sentences between 21 and 25 words	44,587
Sentences between 26 and 30 words	38,206
Sentences between 31 and 35 words	26,902
Sentences between 36 and 40 words	15,167
Sentences between 41 and 50 words	11,042
Sentences with more that 51 words	3,176

Table 16. Corpus Statistics

However, the system cannot take the period blindly as a marker of sentence boundary, as it needs to pay attention to some facts. Numbers use a period as a decimal point, such as 1.5. A period is used in Latin acronyms quoted in Arabic texts, such as *B.B.C.* A period is also used in file names and web site addresses, such as *www.aljazeera.net*. A period is used as well in abbreviations such as أ. 'Mr.', د. 'Dr.', etc. The rules we followed to make allowance for these issues, and to avoid the incorrect use of the period as sentence delimiter, are:

- Preserve any dot between two digits. This solves the decimal point problem in examples such as *1.5*.
- Preserve any dot between two Latin characters. This solves the problem of acronyms, such as *B.B.C.* and URL addresses, such as *www.aljazeera.net*.
- Preserve any dot that follows a single letter. This solves the problem of abbreviations, such as أ. 'Mr.' and د. 'Dr.'
- Preserve dots with a list of hard-coded abbreviations. This list includes the transliteration of Latin alphabet such as آر. 'R.' and a single original Arabic abbreviated word, that is إلخ. 'etc.'

Using these criteria, a Visual Basic tool successfully extracts sentences and puts them in a database. The sentence database structure, as shown in Table 17, provides information on word count as well as subject and date. This allows us to easily obtain statistics on the corpus.

Field Name	Data Type	Field Type/Size
ID	AutoNumber	Long Integer
TopicID	Text	250
SentenceNo	Number	Integer
Sentence	Memo	
WordCount	Number	Integer
Date	Yes/No	
Heading	Yes/No	
LikelyCaption	Yes/No	
Subject	Text	150
Year	Number	Integer
Month	Number	Byte
Day	Number	Byte

Table 17. Design of the Al-Jazeera Sentence Database

From this database we also define scopes for testing and references for development. From the field “WordCount” we can choose sentences of any length we like, from the field “Subject” we can limit our scope to the “news” section alone, and from the fields “Year”, “Month”, and “Day” we can extract sentences in any date we want. The field “Date” is a string of text that comes at the start of each article to give the date in Gregorian and Hijri calendars, “Heading” is the title of the article, and “LikelyCaption” is the caption that comments on photos in the text. These three fields are given a binary Yes/No data type. This allows us to include or exclude them from our data and statistics as they have their peculiar characteristics which are significantly different from ordinary sentences.

A third tool was developed in Visual Basic to produce statistics on frequency information of word forms. These word frequency statistics are mainly helpful in lexicon building as words with the highest frequencies are given priority to be included in the lexicon. For instance, it was found that in news section in 2003 there were 95,182 unique words, with the longest word consisting of 22 characters and the shortest words two characters in length.

7.4 Evaluating the Grammar Coverage

The Arabic grammar parser at the current stage provides 33% coverage (complete parses) for short sentences (10 to 15 words) when tested on an unseen subset of data. The coverage is raised to 92% when using a set of robustness techniques: non-deterministic tokenizer, morphological guesser and fragment grammar. The grammar does not target long sentences at this stage, but just to have a rough idea about the grammar coverage regarding longer sentences, it was found that the grammar provided 16% coverage (complete parses) for sentences ranging between 16 and 25 words in length.

We tried to ensure that the data used in evaluation is different from the data used in development. The reference set used in development was collected from the first ten days in August 2003, while the evaluation data was collected from the first five days of January 2004. The gap between the reference and test data is 5 months which we think is enough to ensure that no articles or sections of articles are repeated in the two sets.

In our evaluation experiment we evaluate both coverage and accuracy. The coverage evaluation shows 33% coverage (complete parses) for short sentences (10 to 15 words). The evaluation experiment was conducted against 207 test sentences. Of them, 69 sentences found a complete parse, and 138 sentences could not be completely parsed using the grammar alone.

As for the accuracy evaluation, we found that accuracy evaluation experiments with the English grammar are usually conducted automatically against a gold standard of the PARC 700 dependency bank (Kaplan et al., 2004). This automatic measurement was not possible in Arabic because such gold standards are not available. Instead we conducted a manual evaluation of the grammar accuracy and reviewed all 69 sentences by hand to check the c-structure and f-structure of the analyses and provide a score according to the number and type of errors found. In our experiment we classified errors into minor errors and serious errors.

Minor errors include one of the following instances:

- PP attachment
- Active/passive variation
- Pronominal reference
- Scope of coordination
- Best solution is not first solution, but among the first 10
- Wrong phrase structure in a small embedded clause (Not CP clause)

Whereas, a serious error includes one of the following instances

- Wrong phrase structure in the main clause. This happens when the system builds the wrong tree because it assigns a POS or a subcategorization frame that is wrong in the context.
- Three or more minor errors

We based our criteria for accuracy evaluation, shown in Table 18, on assuming a scoring scheme based on the error type and the number of errors:

1 - No serious or minor errors

2 - One minor error

3 - Two minor errors

4 - One or more serious errors

Score	Total	%	Subtotal	Error Type
1	20	29		
2	33	49	7	Correct analysis is not first analysis
			2	Error in embedded clause
			1	Passive/active error
			3	POS in embedded clause
			12	PP attachment
			4	Scope of coordination
			3	Wrong phrase structure in embedded NP
			2	Wrong subcategorization in embedded NP
3	6	9	1	Best analysis is not number 1, POS error in embedded NP
			1	POS in embedded NP, PP attachment
			4	Two PP attachments
4	9	13	1	Three errors: two PP attachments and one POS in embedded clause
			8	Wrong phrase structure

Table 18. Accuracy scores for sentences with complete parses

We also assume that analyses scoring 1, 2, or 3 are acceptably accurate while analyses scoring 4 are not acceptable. The result of the evaluation experiment is shown in Table 18.

This evaluation experiment shows that 87% of the sentences that received a complete parse passed the acceptable accuracy threshold, while 13% were marked with serious faults that rendered the analysis unacceptable. This result is somehow comparable to the 90% accuracy results reported by the English grammar for parsing section 23 of the Wall Street Journal text (Cahill et al., 2007).

7.4.1 Robustness Techniques for Increased Coverage

In some applications it is desirable to have any sort of output, even with low accuracy, for every input. This is why XLE has been provided with a “FRAGMENT” grammar, which is a partial parsing technique. When a complete parse is not found in the standard grammar, the FRAGMENT grammar allows the sentence to be analyzed as a sequence of well-formed chunks which have both c-structures and f-structures corresponding to them (Riezler et al., 2002).

The FRAGMENT grammar is a robustness or fall-back technique that allows the system to give a partial parse in case a full parse cannot be attained. Using this robustness technique English is assumed to achieve 100% grammar coverage on unseen data (Riezler et al., 2002).

In our grammar we used a set of robustness techniques for increasing the grammar coverage. First, using a non-deterministic tokenizer (see 3.2.2) and a morphological guesser as a fail safe strategy improves the coverage from 33% to 57%. This variance could be used as an indication of how much coverage could be achieved by expanding the morphology.

Second, adding a fragment grammar on top of the morphological guesser and the non-deterministic tokenizer raised the coverage to 92%. A fragment grammar builds well-formed chunks from input sentences for which no correct analysis could be found. It also ensures that the least number of chunks is produced.

We also found out that a fragment grammar is very useful as it conveys, in many instances, most of the structure and meaning of a sentence. For example, the sentence in (369) is parsed by the system and assigned the fragment c-structure representation in Figure 75. We find that the “S” chunk under “Fragment 0” in Figure 75 carries most of the information. The chunk means ‘But he did not mention any more information’ which is both meaningful and informative, and no important content is wasted.

(369) لكنه لم يذكر مزيدا من المعلومات أو الإيضاحات عن ذلك
lākinn-hu lam yaḏkur mazīdan min al-maʿlūmāti ʾaw
but-he not mention more of the-information or
al-ʾīaḍāḥāti ʿan ḏalika
the-clarifications about that
‘But he did not mention any more information or clarifications on that.’

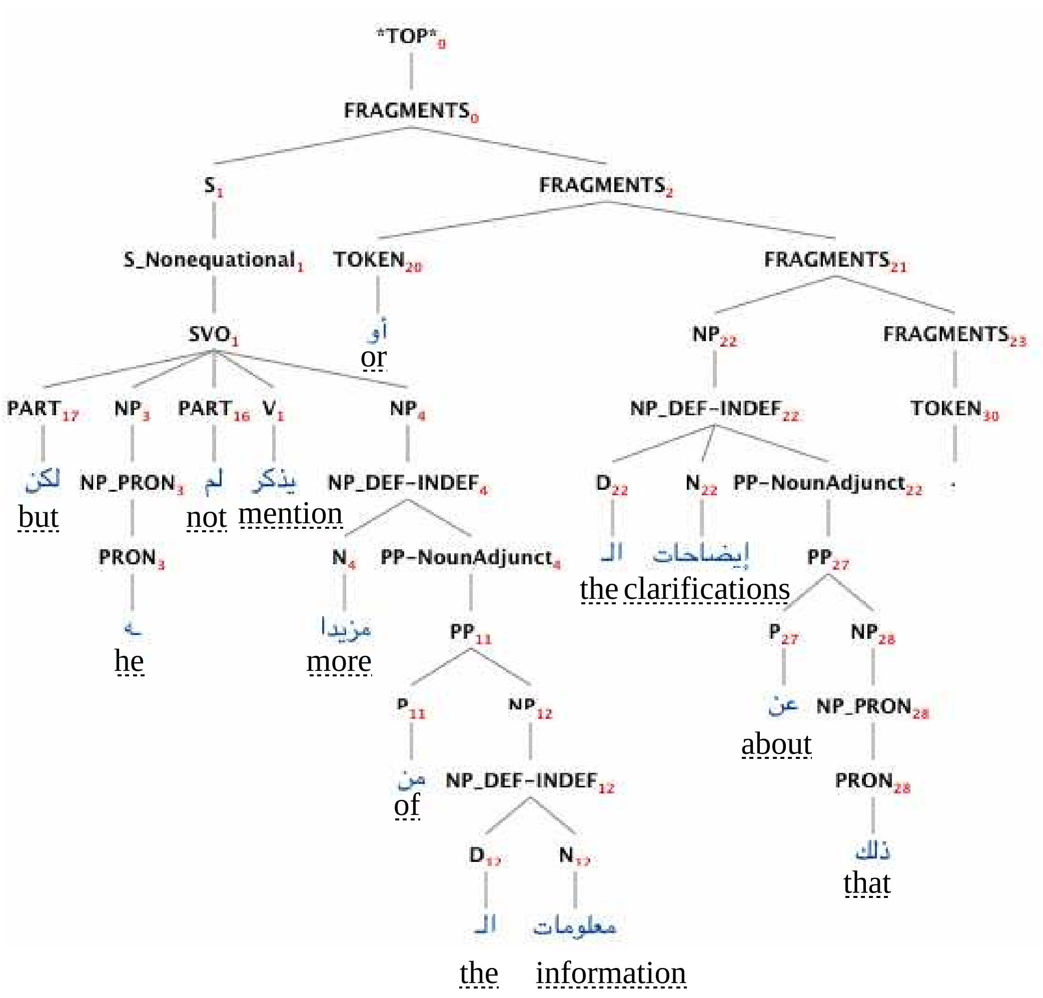


Figure 75. A fragment analysis of an Arabic sentence

Similarly, for the example in (370) the system produces the fragment c-structure representation in Figure 76. However, the chunks have two sentences: the first is ill-formed while the second is a well-formed informative piece of structure. The second chunk sentence means ‘The head of the news chamber said that the office will abide by the decision’ which still carries a great deal of the information contained in the sentence.

(370) ماجد خضر مدير مكتب بغداد مؤقتا ورئيس غرفة الأخبار قال إن المكتب سيلتزم بالقرار
māǧid ḫiḍr mudīru maktabi baġdāda muʾaqqtan wa- raʾīsu ġurfati
Maged Khidr manager office Baghdad interim and-head chamber
al-ʾaḫbāri qala ʾinna al-maktaba s-yaltazimu bi-l-qarāri
the-news said that the-office will-abide by-the-decision
‘Maged Khidr, the interim manager of Baghdad office and head of the news chamber, said that the office will abide by the decision.’

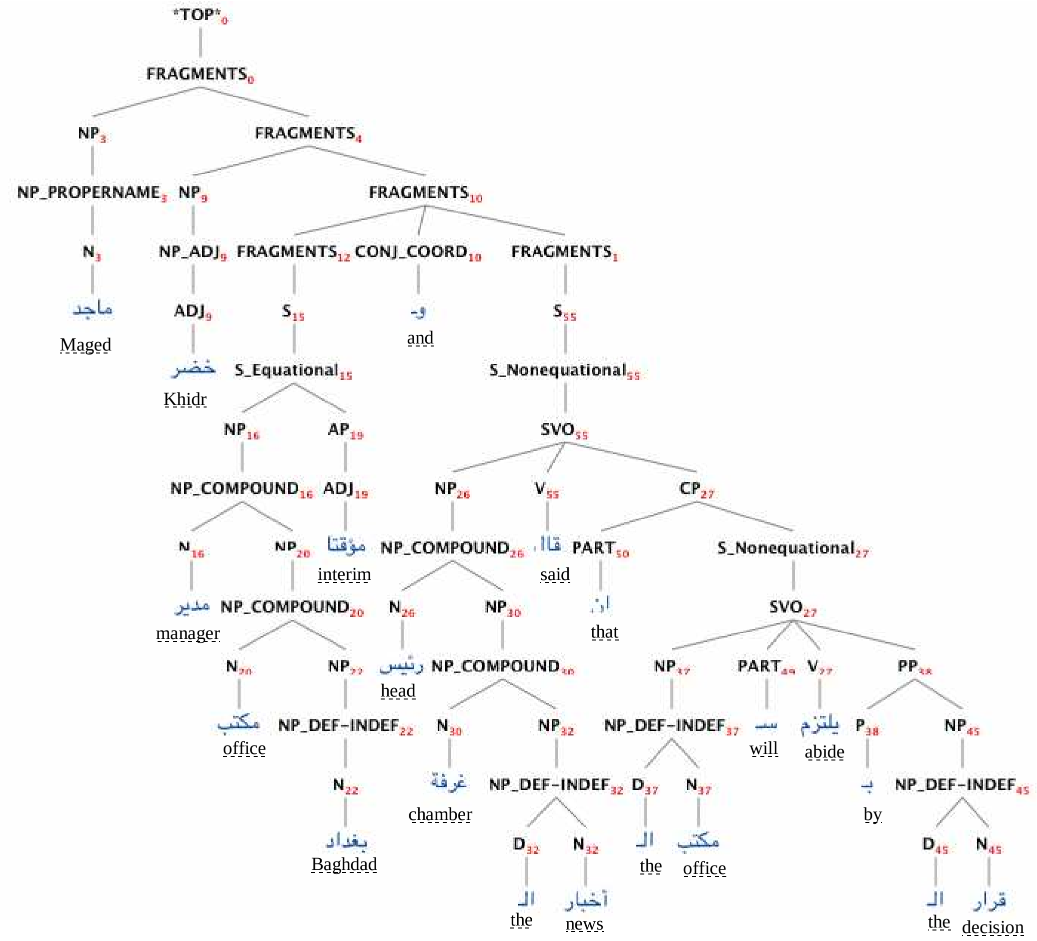


Figure 76. A fragment analysis of an Arabic sentence

In the conclusion of this chapter we would like to emphasise that the quality of the fragment grammar depends on the quality and coverage of the core parser and also on the quality of the morphological guessers. The more coverage the core parser has, the less non-determinism the system has to cope with in the fragment stage.

8 Towards Machine Translation

This chapter concludes the thesis by recapitulating the prospect of Machine Translation (MT) within the ParGram project. We first define what is meant by, and what could be expected from, MT. We give a short explanation of the rule-based transfer approach. We then demonstrate the MT component in the ParGram project. We apply simple transfer rules to translate a small sentence from Arabic into English, and point out what needs to be done in order to produce a fully-fledged MT system. We also show what possible extensions can be implemented in the system, as a whole, in the future.

8.1 What Is MT?

MT is defined as “the automatic translation of text or speech from one language to another” (Manning and Schütze, 1999). It involves making the computer acquire and use the kind of knowledge that translators need to perform their work. However, the endeavour is not an easy one. To successfully undertake a translation task, human translators needs to have four types of knowledge (Eynde, 1993):

1) Knowledge of the source language (SL) (lexicon, morphology, syntax, and semantics) in order to understand the meaning of the source text.

2) Knowledge of the target language (TL) (lexicon, morphology, syntax, and semantics) in order to produce a comprehensible, acceptable, and well-formed text.

3) Knowledge of the relation between SL and TL in order to be able to transfer lexical items and syntactic structures of the SL to the nearest matches in the TL.

4) Knowledge of the subject matter. This enables the translator to understand the specific and contextual usage of terminology.

Ultimately, the translation process is not considered successful unless the output text has the same meaning as the input text (Catford, 1965). Therefore, the transfer of lexical items and syntactic structures is not considered successful translation if the overall meaning is not conveyed.

In addition to the types of knowledge mentioned above, translators must have a special skill in their craft. To a great extent, translation “is an intelligent activity, requiring creative problem-solving in novel textual, social and cultural conditions” (Robinson, 1997). Not only does the translation depend on linguistics, but it also “draws on anthropology, psychology, literary theory, philosophy, cultural studies and various bodies of knowledge, as well as on its own techniques and methodologies” (Trujillo, 1999).

It is not so easy for the computer to translate as to conduct a mathematical operation. In order for the computer to translate, it must “to some degree ‘understand’ the input” (Willis, 1992). However, this understanding is not easily available because there are many factors that cloud the meaning. The meaning of a human utterance is “open to doubt, depending on such things as knowledge, context, association and background” (Boulton, 1960).

After computer engineers and linguists met with many failures in the beginning of MT application, they now understand the intricacy of the task. Many researchers today are directing their efforts towards MT fully aware of the elusiveness of the colossal task. MT has become a “testing ground for many ideas in Computer Science, Artificial Intelligence and linguistics” (Arnold et al., 1994).

Once a far-away dream, MT today has become a reality. Against all odds many advances have been made, many successes have been achieved and many translation applications have now hit the market. However, this reality is not as big as people initially hoped. Commenting on the capacity and prospect of MT, Hutchins and Somers said that there are no MT systems which can produce a perfect translation at the touch of a button, and that this is “an ideal for the distant future, if it is even achievable in principle” (Hutchins and Somers, 1992). Though these words are said a decade and a half ago, they are still expressive of the state of the art of MT today. The translation process is so complicated for the machine to handle. The machine cannot deal with all types of texts in all fields. No MT manufacturer dare claim that their applications can produce a hundred per cent accurate and comprehensible output.

Despite the progressive reality of MT today, some people still argue that studies in MT are useless because the machine can never translate great literary works like those of Shakespeare or Dickens. However, translating literary works is not within the scope of MT, because “translating literature requires special literary skill” (Arnold et al., 1994) and creativity from the translator. It is usually a poet or a man of letters (not an ordinary translator) who attempts this sort of texts.

The machine cannot and will not replace translators entirely, but it complements them and helps them in a variety of ways. MT can handle the huge routine tasks. Technical manuals and periodicals, for example, are a perfect material for MT. They use no figurative or flowery language. They have specific subject fields and restricted styles, terminology, structures, and vocabularies. MT can also provide raw translation which can be revised or ‘post-edited’ to give a high quality translation in a shorter time.

Different strategies have been adopted by different research groups at different times. Strategy choice reflects both the depth of linguistic manipulation and the breadth of ambition. At the early stages of MT research and development, little was understood about linguistic complexities. A simple methodology was followed by replacing SL words with their equivalents in the TL with a few rules for local reordering. As MT research grew, scientists concentrated more on the analysis of SL with higher levels of abstractness. In this section I will give a brief account of the transfer strategy, as it is the strategy upon which the translation component (XTE) in the ParGram project is based.

8.1.1 Transfer

The transfer method is a middle course between two other approaches: direct and interlingua MT strategies. The difference between the three strategies can be captured in Figure 77 (from Vauquois, 1978).

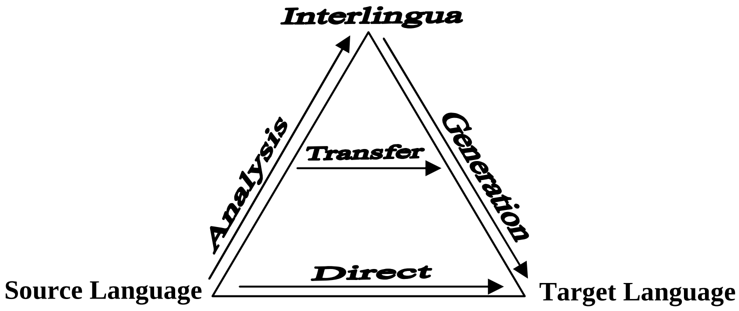


Figure 77. Difference between direct, transfer, and interlingua MT methods

As can be seen in Figure 77, the direct method has no modules for SL analysis or TL generation but applies a set of rules for direct translation. In the interlingua method the SL is fully analyzed into a language-independent representation from which the TL is generated. The transfer method is a middle course between the two approaches. Both the interlingua and the transfer methods utilize abstract representations, but they place different demands on these representations (Bennett, 2003). The transfer strategy can be viewed as “a practical compromise between the efficient use of resources of interlingua systems, and the ease of implementation of direct systems” (Trujillo, 1999). The SL is analyzed into a language-dependent representation which carries features of the SL. Then a set of transfer rules are applied to transform this representation into a representation that carries features of the TL. At the end the generation module is used to produce the target output.

Compared to the interlingua method, there are two advantages of the transfer method that make it appealing for many researchers. The first advantage is the applicability of the transfer system. While it is difficult to reach the level of abstractness required in interlingua systems, the level of analysis in transfer models is attainable. The second advantage is the ease of implementation. Developing a transfer MT system requires less time and effort than interlingua. This is why many operational transfer systems have appeared in the market.

One clear disadvantage of the transfer method is that it is costly when translation between many languages is required. The transfer method “involves a (usually substantial) bilingual component, i.e., a component tailored for a specific SL-TL pair” (Tucher, 1987). This entails significant effort and time for each new language added to the system. Mathematically speaking, the number of transfer modules for *n* languages is "*n* × (*n* – 1)" in addition to *n* analysis and *n* generation modules (Hutchins and Somers, 1992).

A range of Arabic-English MT products are produced by ATA Software[7]. The company presumably uses the transfer method. Sakhr Software has also developed its Arabic English MT solution[8] within the transfer paradigm.

It must be noted, however, that with the advances in computer science, Statistical Natural Language processing has taken centre-stage in Computational Linguistics research. Statistical-Based Machine Translation (SBMT) systems do not use any hard-coded linguistic information. Instead they rely on corpora to conduct probability statistics based on the frequency of occurrence. The best-known Arabic-English MT system built in the SBMT paradigm is the Google free online text translation[9].

Yet the recent few years have witnessed some revival of interest in rule-based MT systems. Statistical methods are criticised for their reliance on relatively shallow input, and their value has been doubted in the long run. It has been maintained that a semantic analysis is necessary to preserve the semantic content of the input and a rule-based generator is needed to secure the well-formedness of the output (Flickinger et al., 2005).

8.2 Using XLE to Do MT

There is a growing interest in MT systems that support some degree of ambiguity preservation to alleviate the tedious task of ambiguity handling during parsing and transfer (Dymetman and Tendeau, 2000). These systems rely on packed structures which factorize ambiguities in a compact representation. Emele and Dorna (1998) cite the well-known PP attachment ambiguities as a good example of preservable syntactic ambiguities. These ambiguities can be transferred from the SL into the TL without requiring costly disambiguation, and

[7] http://www.atasoft.com
[8] http://tarjim.sakhr.com
[9] http://translate.google.com/translate_t?langpair=ar|en

they are even termed “free rides” (Hutchins and Somers, 1992). The transfer module in XLE is an ambiguity preserving translation tool that circumvents the need for disambiguation by generating a target sentence that has exactly the same ambiguities as the source (Wedekind and Kaplan, 1996). The translation approach in XLE is based on the idea of transferring ambiguous LFG f-structure representations based on packed f-structure representations (Emele and Dorna, 1998).

Machine Translation in XLE came as an offshoot of the ParGram LFG grammar development project. It uses the transfer strategy, as transfer rules need to be written to transform f-structures from the source language to the target language. This translation strategy is frequently referred to as *Chart Translation* (Kay, 1999), because the idea is to translate f-structure charts from SL to TL. After the application of the transfer rules to create new charts in the TL, the parser in the TL is then used for generation. Frank (1999) criticized the conventional translation architectures, where ambiguity filters are applied early to reduce the size of complexity, yet risking discarding correct solutions too early on the basis of poor evidence. She adhered to the idea that ambiguities should be propagated forward within the translation chains, and that the translation system must not take decisions which it is not well prepared to take. Therefore she advocated preserving ambiguity, and proposed that the ambiguity could be solved later by drawing clues from the translation output and through human interaction with the system. The decisions taken by the human disambiguator can then be used in memory-based learning techniques to propagate the human decisions for similar ambiguity problems.

Researchers in the XLE translation project (Frank et al., 2001, Kay, 1999, Wedekind and Kaplan, 1996) emphasize that the transfer system does not attempt to resolve ambiguities, but it transforms the packed representation (or packed ambiguities) from the SL to the TL. They consider this as an advantage as it avoids taking decisions about ambiguity handling at the wrong time. They believe that it is not the job of the transfer component to handle ambiguities, but the problem should be handled in the stages before or after transfer.

MT in XLE is facilitated by the fact that the ParGram project ensures that isomorphism is maintained cross-linguistically. The same grammar model is used for different languages, and common guidelines are provided to sustain consistency in feature notations and to keep the divergence to a minimum. The aim of the ParGram project is to create grammars in different languages with the fewest possible divergences. The grammar writers working in the ParGram project meet semi-annually to make sure that parallelism is observed in naming conventions, notations, formalisms, and what new features are to be added to or removed from the common inventory. They also make sure that divergence is linguistically motivated and well-justified. These parallel grammars are believed to provide the cornerstone for an MT project (Frank, 1999).

Moreover, the LFG formalism itself provides a favourable background for translation (Butt et al., 1999a). LFG has two main levels of representation. The first is the c-structure level which is a phrase structure tree that encodes consistency (dominance) and surface order (precedence). The second is the f-structure level which is more abstract and which provides information on morphosyntactic features (such as number, gender and person) and grammatical functions (such as subject, object and oblique). The greatest variability and divergence among languages appear in the c-structure, while more convergence and parallelism appear in the f-structure. Therefore f-structures are better suited as a base for MT.

Harold Somers maintained that using LFG for translation captured the interest of researchers inside and outside of the ParGram community. The main idea is that f-structures are deep enough to transcend superficial surface structure differences between languages, but not so deep as to invite the difficulties of a true interlingua approach.

The idea of using LFG's concept of structural correspondences for the purpose of MT first appeared in Kaplan et al. (1989). The main concept was to introduce two levels of correspondence: one to map between the f-structures, and the other to map between the semantic structures of the two languages. We can even trace the attempt to use LFG in MT to an earlier date. Hutchins (1988) reported on an English-Japanese experimental MT system (NTRAN) at UMIST, Manchester, UK. It was an interactive system, written in Prolog, which produced LFG-type f-structures; from these were derived the s-structure interface representations which were then converted into equivalent Japanese interfaces; these s-structures were then used to generate the Japanese f-structures and surface strings.

Interestingly enough, LFG has been appealing even to researchers working on Statistical NLP. Way (1999) proposed a hybrid for MT based on both LFG and Data-Oriented Parsing (DOP) to improve upon Data-Oriented Translation (DOT). Owczarzak et al. (2007) proposed using f-structures for MT evaluation. The Dublin City University have been working on grammar induction based on automatic f-structure annotation algorithm for the Penn Treebanks (O'Donovan et al., 2004).

It must be noted, however, that the work in the transfer component within the ParGram project has not evolved into a full-fledged MT system. The translation module has not been used in a large-scale implementation, but it is considered merely as a first step, experimental prototype (Frank, 1999). In Frank’s experiment 99 sentences were translated from French into English. No operational system has yet been implemented to translate between any language pairs.

Apart from Kay (1999) and Frank (1999) there is relatively little published on the XTE. Yet the tool has been actively used as a rule-rewriting facility for text summarization and sentence condensation (Crouch, 2004), as confirmed by Tracy Holloway King (personal communication, email, 11 March 2008). In Powerset, a question-answering search engine, XTE is also used to go from f-structures to semantics.

In recent years, interest in rule-based MT systems saw a resurgence with the LOGON system. It is an experimentation LFG-based MT system for translation from Norwegian into English (Flickinger et al., 2005). It is based on semantic transfer using Minimal Recursion Semantics (MRS), instead of the traditional f-structure transfer model. The Norwegian LOGON system uses the HPSG transfer component instead of XTE. The basic concept, however, is still the same. The system uses LFG grammar for parsing and producing MRSs, uses the

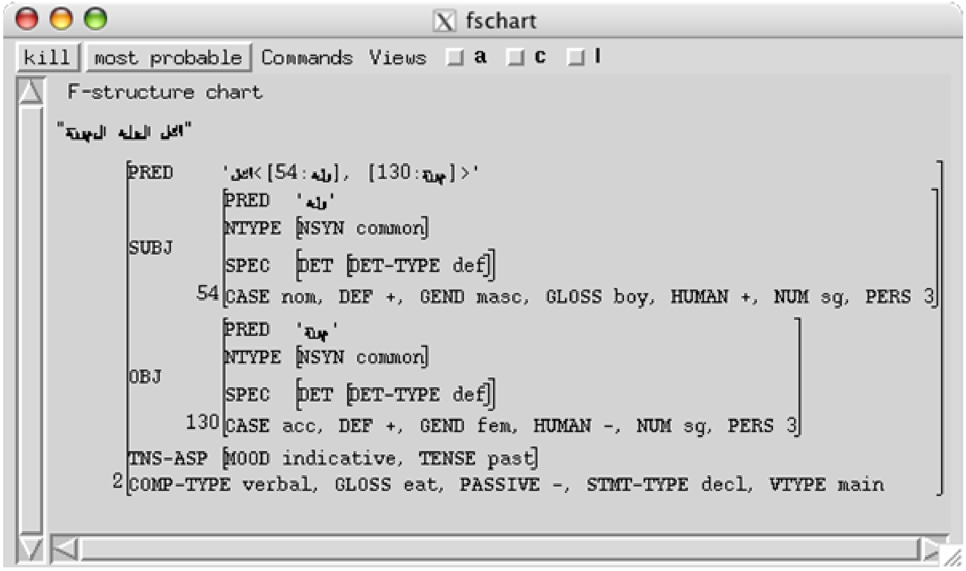
HPSG transfer to transfer the structures into a deep representation of the target language, then generation is made using the HPSG English grammar. This information was also provided by Tracy Holloway King (personal communication, email, 11 March 2008).

8.2.1 The Transfer Module

The transfer module in XLE is called XTE (the Xerox Translation Environment) (Kay 1999). It is a re-write system that works on f-structures to convert them from one notation describing one language to another notation describing another language.

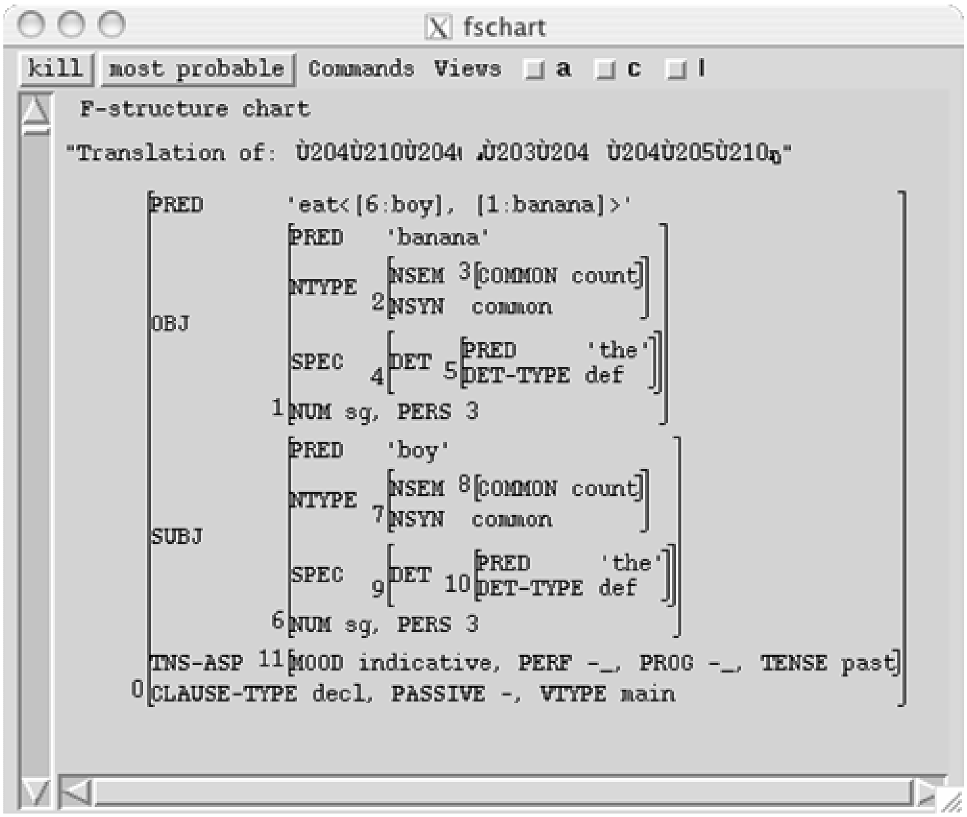
Here we will describe some basic rule writing to convert Arabic f-structures into English f-structures from which the generator seamlessly produces the English translation, as discussed below. We will show how the rules are applied to translate the Arabic sentence in (371) by converting the Arabic f-structure in Figure 78 into the English f-structure in Figure 79.

(371)	الولد أكل الموزة		
	al-waladu	ʾakala	al-mūzata
	the-boy.nom	ate	the-banana.acc
	‘The boy ate the banana.’		



```
"اكل الولد الموزة"

        PRED      'اكل<[54:ولد], [130:موزة]>'
                   PRED    'ولد'
                   NTYPE   NSYN common
        SUBJ       SPEC    DET DET-TYPE def
                54 CASE nom, DEF +, GEND masc, GLOSS boy, HUMAN +, NUM sg, PERS 3
                   PRED    'موزة'
                   NTYPE   NSYN common
        OBJ        SPEC    DET DET-TYPE def
               130 CASE acc, DEF +, GEND fem, HUMAN -, NUM sg, PERS 3
        TNS-ASP   MOOD indicative, TENSE past
      2 COMP-TYPE verbal, GLOSS eat, PASSIVE -, STMT-TYPE decl, VTYPE main
```

Figure 78. F-Structure in Arabic before transfer into English



```
 F-structure chart
"Translation of: Ù204Ù210Ù204ι ,Ù203Ù204 Ù204Ù205Ù210ŋ"

   PRED        'eat<[6:boy], [1:banana]>'
                PRED    'banana'
                NTYPE  2NSEM 3[COMMON count]
                        NSYN  common
   OBJ
                SPEC   4DET 5PRED      'the'
                             DET-TYPE def
              1 NUM sg, PERS 3
                PRED    'boy'
                NTYPE  7NSEM 8[COMMON count]
                        NSYN  common
   SUBJ
                SPEC   9DET 10PRED      'the'
                              DET-TYPE def
              6 NUM sg, PERS 3
   TNS-ASP 11[MOOD indicative, PERF -_, PROG -_, TENSE past]
 0 CLAUSE-TYPE decl, PASSIVE -, VTYPE main
```

Figure 79. F-Structure in English after transfer from Arabic

First we need to state that the order in which the transfer rules are written is important as each rule works on the output of the previous one (Frank, 1999). To translate the sentence we first translate the Arabic nouns into English through the re-write rules[10] such as those in (372).

(372) PRED(%X, ولد), +NTYPE(%X, %%) ==> PRED(%X, boy).
PRED(%X, موزة), +NTYPE(%X, %%) ==> PRED(%X, banana).

Then we need to translate the verb along with the explicit statement of subcategorization frames from Arabic into its equivalent in English, as in (373).

(373) PRED(%X, أكل), SUBJ(%X, %Subj), , OBJ(%X, %Obj) ==> PRED(%X, eat), SUBJ(%X, %Subj), , OBJ(%X, %Obj).

Within this formalism we can also add, delete or change features. The definite article in Arabic which does not have a PRED feature must be realised as *the* in English. This is achieved by the rule in (374).

(374) +DET-TYPE(%X, def) ==> PRED(%X, the).

[10] See the transfer documentation on: http://www2.parc.com/isl/groups/nltt/xle/doc/transfer-manual.html

For English nouns the features of humanness, gender, definiteness and case are irrelevant and therefore they are discarded through the deletion rule in (375).

(375) HUMAN(%%, %%) ==> 0.
+NTYPE(%X, %%), GEND(%X, %%) ==> 0.
+NTYPE(%X, %%), CASE(%X, %%) ==> 0.
DEF(%%, %%) ==> 0.

Although the Arabic sentence follows the VSO word order and the English sentence needs to follow the SVO word order, the transfer component does not need to include any statements about word order. It transfers only features and grammatical functions. The best surface structure will be rendered by the TL generator.

The XTE transfer component as mentioned above is not designed to resolve ambiguities. Yet we can use transfer rules to truncate implausible readings. If we look at the sentence in (376), we see that there is ambiguity in the subject position.

(376) قرأ الرجل الكتاب
qaraʾa ar-rağulu/ar-riğlu al-kitāba
read the-man/the-foot the-book.
'The man/foot read the book.'

The ambiguity in (376) stems from the fact that diacritics are omitted in modern writing and this is why الرجل becomes ambiguous between two readings, ar-rağulu/ar-riğlu 'the-man/the-foot'. Yet in the example it is quite obvious that the intended reading is *the man* and not *the foot*, as the subject of *read* must be a human entity. All nouns in our morphology are already assigned a feature of ±human. Therefore, we can write a transfer rule to disallow all non-human entities from becoming the subject of *read*, as in (377).

(377) @verb_subj(%%, قرأ, %Subj), HUMAN(%Subj, -) ==> stop.

In our future work with the transfer component we would like to see how MWEs are handled and how we can account for head-switching, as in (378), and conflational divergence, as in (379)–(380), among other instances of structural divergence.

(378) كاد الولد أن ينام
kāda al-waladu ʼan yanāma
was-nearly the-boy to sleep
‘The boy nearly slept.’

(379) انتحر الولد
ʼintaḥara al-waladu
committed_suicide the-boy
‘The boy committed suicide.’

(380) عبأ اللبن في زجاجة
abbaʼa al-labana fi zuǧāǧah.
stored the-milk in bottle
‘He bottled the milk.’

According to the XLE documentation,[11] the generator in XLE is the inverse of a parser. While a parser takes a string as input and produces f-structures as output, a generator takes an f-structure as input and produces all of the surface strings that could have that f-structure as output. The generation grammar can be made slightly different from the parsing grammar by changing the set of optimality marks and by changing the set of transducers used.

The XTE provides a wide range of notations to express whether a rule is optional or obligatory and to state different conditions on the application of rules. It also provides the facility of using templates and macros to speed up the development process and to state generalizations.

Of course the transfer grammar is not as simple as might be conceived from the demo example. There are a great number of structural divergences between Arabic and English that must be taken care of, such as the functional control relations, agreement conditions, divergent argument structures, copula constructions, etc. Sadler and Thompson (1991) emphasized the structural non-correspondence in translation. The volume of the work required in the transfer phase cannot be possibly determined at this stage as no work has been conducted on real sentences on a large scale. Transfer under XTE, however, proves to be considerably convenient as it relieves the grammar writer from worrying about the word order and surface structure in the TL. This confirms the common belief

[11] http://www2.parc.com/isl/groups/nltt/xle/doc/xle_toc.html

that structural parallelism achieved at the f-structure level facilitates the translation process.

We believe, however, that the work done in the transfer component has not yet matured enough to produce an MT system. The transfer model in XTE in the current stage is a direct mapper between the f-structures of the SL and TL languages, with no specification of the actual translation work. The MT system implemented in XLE is still within the confines of experimentation. There are certain areas that are still underspecified and need more research and engineering to make it a fully-rounded application. Among the drawbacks we notice in the system are:

1. It is not clear how a bilingual lexicon fits in the system. The bilingual lexicon is expected to fit in the transfer component. Yet our initial conception was that it would fit in the LFG lexicon along with the specification of subcategorization frames. The subcategorization frames in many instances provide a viable context for specifying the meaning. If the bilingual lexicon is placed in the transfer component, such context information will have to be re-stated, leading to duplication of the lexicon, thus complicating the work of a lexicographer. When a new verb, for instance, is added to the morphology, it needs to be added to the grammar lexicon to stipulate its subcategorization frames and special constraints and idiosyncrasies. Then again the verb will need to be added to the bilingual lexicon in the transfer component to type in the meanings in the other language, taking into consideration the necessity to stipulate again the various subcategorization frames and idiosyncrasies which inevitably affect the meaning. A good suggestion given by Mary Dalrymple (personal communication, 7 May 2008) is that it would be practical to have a single arch-lexicon which could be automatically processed to produce morphological, syntactic, and transfer lexicons.
2. Chart translation cannot alone produce an MT system as more work needs to be done in the semantic level regarding thematic roles and word-sense disambiguation. Word sense depends on the context, and the transfer component is not equipped to analyze the context.

3. Although there are clear and viable justifications for not attempting to resolve ambiguities during transfer as this may lead to pruning good solutions too early, yet an MT system still needs to output one solution if it wants to have any advantage to a user.
4. The system does not provide a user-friendly interface for inputting a sentence in the SL and viewing the translation in the TL. It is still a system for engineers and developers to test and experiment, moving the output by hand from one phase and inputting it into the next.
5. The quality and coverage of the MT facility in the XTE system depends on the quality and coverage of the parser. As most of the ParGram languages are still struggling to achieve reasonable quality and coverage, fewer efforts have been put into the implementation of an MT system.
6. The transfer grammar is not robust. It must be exhaustive and comprehensive to make sure that the f-structures in the SL are converted to well-formed f-structures in the TL. The new f-structures must follow the rules of the TL to the smallest details, as even a single additional feature could cause the generator to fail.
7. It is not obvious whether the transfer component is reversible or not. The impression we get is that the transfer system is not reversible and that transfer rules must be written for each translation direction.

8.2.2 Possible Extensions to the System

- More work in the morphology is needed to increase the coverage. The coverage of the parser is, to a great extent, correlated with the coverage of the morphological analyser.
- There are currently two main implementations of the tokenizer: one that depends on the morphology and fails to handle unknown words, and one that handles any text but at the cost of a high level of non-determinism. The tokenizer could be a lot more intelligent if it works on a core list of words and guesses only unknown words. We have already experimented with some ideas to achieve this goal. One of these ideas will require the investigation of how Arabic words are formed from letters and syllables to be able to generate possible words.

- MWEs are highly valuable in the system as they decrease non-determinism and increase efficiency. We would like to explore ways of the automatic extraction of MWEs and named entities from annotated/unannotated texts.
- LFG proves to be a strong and flexible theoretical framework. Therefore we would like to dedicate more work to the Arabic XLE parser to increase the coverage and explore how more underspecified Arabic syntactic structures can best be described.
- Treebanks are invaluable in Computational Linguistics research nowadays. It could be very interesting to build a treebank using the Arabic parser and the Norwegian XLE Web Interface and the Discriminants tools.
- We would like to acquire the LDC Arabic treebank to see how it can be used for grammar extension and for stochastic disambiguation.
- Work with the transfer component is an interesting way to see how much divergence there is between Arabic and English and how this divergence can be handled.

In the end I would like to conclude with the articulate words of Beesley and Karttunen (2003, p. 259).

> In practice, linguists are imperfect, and natural languages are somewhat fuzzy and moving targets – we must content ourselves with closer and closer approximation.

Appendix: Demo of System Processing

In this demo we will show the processing sequence for two examples that show the basic sentence structure in Arabic. The purpose of this section is to put all the pieces of processing steps together for convenience, as these steps are only discussed at length in separate chapters in the thesis. We made a small version of the phrase structure rules in our grammar so that they are as simple and understandable as possible. We removed all the unnecessary details from our rules for the purpose of the demo. For example we removed the details related to particles, coordination, subordinating conjunctions, obliques, and parenthetical clauses. Then we tested the small-version grammar to make sure that all the rules are working and contain no errors.

The first example shows the equational (copula) construction. The sentence has an overt copula, but we also show how the variant with non-overt copula is handled using the same phrase structure rules. The second example will show the non-equational (verbal) construction that follows the default word order in Arabic, i.e. VSO, and we also show how the other variant, SVO, is handled by the phrase structure rules. While the first sentence is a straightforward example with no ambiguity. The second will have a simple kind of ambiguity.

SENTENCE ONE:

كانت الشمس مشرقة
kānat aš-šamsu mušriqatun
was the-sun.sg.fem bright.sg.fem
‘The sun was bright.’

TOKENIZATION OUTPUT

```
كانت@الـ@شمس@مشرقة@
```

MORPHOLOGY OUTPUT

كانت	+verb+past+activeكان+3pers+sg+fem
الـ	+defArtالـ
شمس	+noun+nonhumanشمس+fem+sg
مشرقة	+adjمشرق+fem+sg

LEXICAL ENTRIES

Lexical entries are responsible for assigning PRED values for lexical items as well as subcategorization frames for verbs. Lexical rules stating the functional control equations for raising and equi verbs are also written here. They are also used for assigning default values for the features of number and gender on nouns and adjectives. Lexical entries are also used to interpret the morphological features that accompany words. These features come from the morphological analyser and they are usually related to tense, person, number and gender.

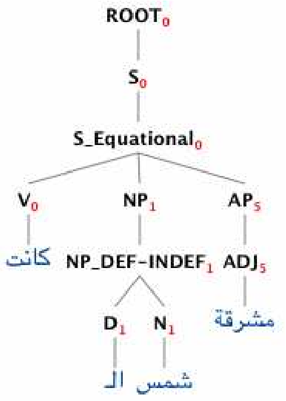
```
كان      V XLE (^ GLOSS)=be
         "It has two subcategorization frames: as a copula verb and as a raising verb"
         {(^ PRED)='%stem<(^ SUBJ)(^ PREDLINK)>'
                  (^ VTYPE)=copular (^ PREDLINK CASE)=acc
         |(^ PRED)='%stem<(^ SUBJ)(^ VCOMP)>' (^ VCOMP SUBJ)=(^ SUBJ) }.

شمس      N XLE (^ GLOSS)=sun (^ PRED)='%stem' (^ PERS)=3
          { (^ NUM) (^ NUM) ~= sg | (^ NUM) = sg } "the default number is singular".

مشرق     ADJ XLE (^ PRED)='%stem' (^ GLOSS) = 'bright'
          { (^ ATYPE)=c predicative | (^ ATYPE)= attributive}.

+past    V_SFX XLE (^ TNS-ASP TENSE) = past.

+active  V_SFX XLE (^ PASSIVE) = -.

+3pers   V_SFX XLE (^ AGR PERS) = 3;
         PRON_SFX_PERS XLE (^ PERS) = 3.

+sg      N_SFX_NUM XLE  (^ NUM) = sg;
         V_SFX_NUM XLE (^ AGR NUM) = sg;
         ADJ_SFX_NUM XLE (^ NUM) = sg;
         PRON_SFX_NUM XLE (^ NUM) = sg.

+fem     N_SFX_GEND XLE (^ GEND) = fem;
         V_SFX_GEND XLE (^ AGR GEND) = fem.
         ADJ_SFX_GEND XLE (^ GEND) = fem;
         PRON_SFX_GEND XLE (^ GEND) = fem.

+defArt  D_SFX XLE (^ SPEC DET DET-TYPE) = def.

+nonhuman         N_SFX XLE (^ HUMAN) = -.
```

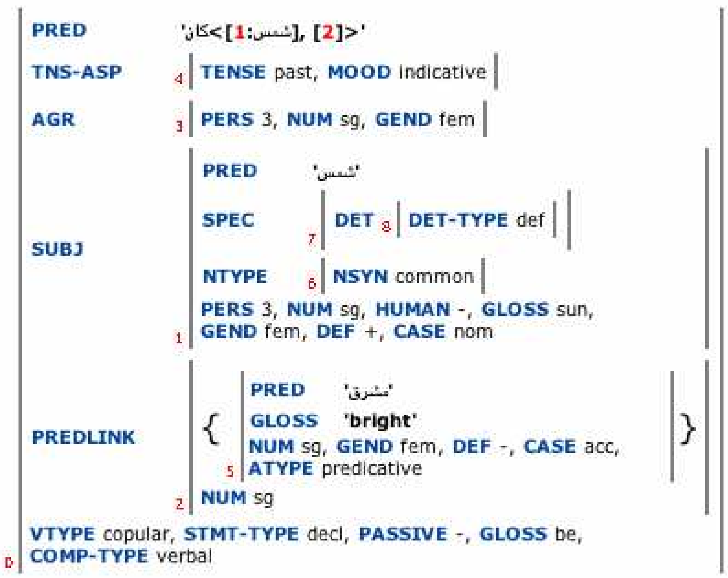
PHRASE STRUCTURE RULES

To see the grammatical notations used in XLE, you can see the online documentation on: http://www2.parc.com/isl/groups/nltt/xle/doc/notations.html. The explanation of some rules is enclosed in double quotes which is the XLE way of writing comments.

```
MT ARABIC RULES (1.0)

 S --> "A sentence can either be equational or nonequational"
            { S_Equational "the class of copular sentences"
            | S_Nonequational}. "sentences composed of main verbs"

S_Equational --> "In a copular construction, the copula verb can be overt or non-overt. Then comes
the subject NP and the predicate AP"
            {V: (^ VTYPE)=c copular @DefSTense (^ COMP-TYPE)=verbal
               (^ AGR NUM)=sg "the verb is invariably singular if it comes before the subject"
               (^ AGR GEND)=(^ SUBJ GEND)  (^ AGR PERS)=(^ SUBJ PERS)
            | e: (^ PRED) = 'H-STR<(^SUBJ)(^PREDLINK)>'(^ VTYPE)=copular
                 (^ COMP-TYPE)=nominal
            @DefSTense}
             NP: (^ SUBJ)=! (! DEF)=c +
                         { (! CASE) (! CASE) ~= nom | (! CASE) = nom } "the default case is nominal";
            AP: ! $ (^ PREDLINK)
                          { (! CASE) (! CASE) ~= nom | (! CASE) = nom } "the default case is nominal"
                          (! ATYPE)=predicative
                         {~(^ SUBJ HUMAN)
                         | {(^ SUBJ HUMAN)=c + (^ SUBJ NUM)=(^ PREDLINK NUM)
                           | (^ SUBJ HUMAN)=c -
                             {(^ SUBJ NUM)=pl (^ PREDLINK NUM)=sg
                             | (^ SUBJ NUM)~=pl (^ SUBJ NUM)=(^ PREDLINK NUM)}}}
                         (! GEND) = (^ SUBJ GEND).

 NP --> {
            NP_DEMONSTRATIVE | NP_DEF-INDEF | NP_PARTITIVE | NP_COMPOUND
            | NP_PROPERNAME | NP_PRON | NP_DEVERBAL |  NP_RELATIVE
            | NP_NUM | NP_SUPERLATIVE | NP_DATE
         }.


NP_DEF-INDEF --> "A common noun is composed of an optional determiner, a noun, and an optional
AP or PP"
             (D: (^ SPEC DET DET-TYPE)=c def (^ DEF)=+)
            N: @(DEFAULT (^ DEF) -) (^ NSEM PROPER PROPER-TYPE)~= name;
             (AP-NounAdjunct)
            [PP-NounAdjunct]*
            (PP-NounObl).

----
MT ARABIC TEMPLATES (1.0)

DefSTense = "This template states the tense, aspect, mood and sentence type"
               {(^ STMT-TYPE) (^ STMT-TYPE)~= decl | (^ STMT-TYPE)=decl}
                 {(^ TNS-ASP MOOD) (^ TNS-ASP MOOD)~= indicative
                  | (^ TNS-ASP MOOD)=indicative}.
```

C-STRUCTURE & F-STRUCTURE



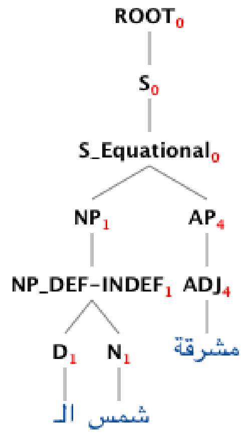


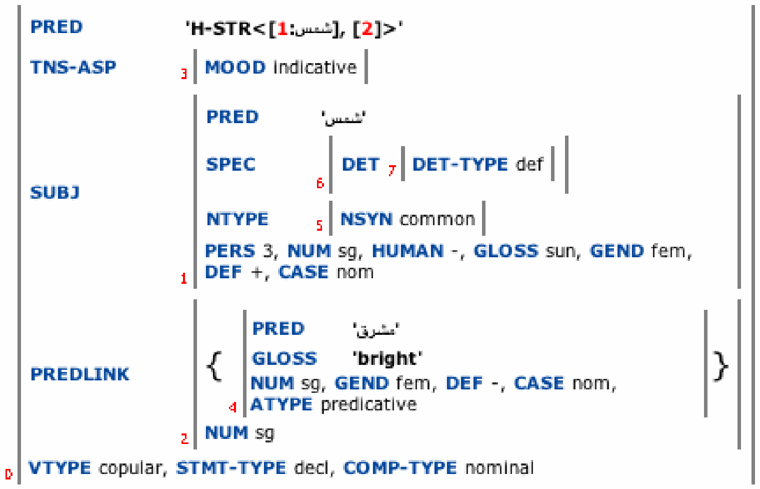
```
PRED        'كان<[1:شمس], [2]>'
TNS-ASP    4| TENSE past, MOOD indicative |
AGR        3| PERS 3, NUM sg, GEND fem |
              PRED      'شمس'
              SPEC     7| DET 8| DET-TYPE def | |
SUBJ          NTYPE    6| NSYN common |
              PERS 3, NUM sg, HUMAN -, GLOSS sun,
           1| GEND fem, DEF +, CASE nom
                 PRED   'مشرق'
                 GLOSS  'bright'
PREDLINK      {  NUM sg, GEND fem, DEF -, CASE acc,  }
               5| ATYPE predicative
           2| NUM sg
   VTYPE copular, STMT-TYPE decl, PASSIVE -, GLOSS be,
0| COMP-TYPE verbal
```

We note that the phrase structure will allow us to parse sentences where the copula is non-overt, such as:

الشمس مشرقة	
aš-šamsu	mušriqatun
the-sun.sg.fem	bright.sg.fem
‘The sun is bright.’	

C-STRUCTURE & F-STRUCTURE





```
PRED        'H-STR<[1:شمس], [2]>'
TNS-ASP    3| MOOD indicative |
              PRED      'شمس'
              SPEC     6| DET 7| DET-TYPE def | |
SUBJ          NTYPE    5| NSYN common |
              PERS 3, NUM sg, HUMAN -, GLOSS sun, GEND fem,
           1| DEF +, CASE nom
                 PRED   'مشرق'
                 GLOSS  'bright'
PREDLINK      {  NUM sg, GEND fem, DEF -, CASE nom,  }
               4| ATYPE predicative
           2| NUM sg
0| VTYPE copular, STMT-TYPE decl, COMP-TYPE nominal
```

SENTENCE TWO:

ساعدت الهيئة الفلسطينيين
sāʿadat al-haiʾatu   al-filisṭīniyyīn/   al-filisṭīniyyain
helped the-agency the-Palestinian.pl/ the-Palestinian.dual
‘The agency helped the Palestinians/ the two Palestinians.’

TOKENIZATION OUTPUT

```
@ساعدت@الـ@هيئة@الـ@فلسطينيين
```

MORPHOLOGY OUTPUT

```
ساعدت       +verb+past+activeساعد+1pers
            +verb+past+activeساعد+3pers+sg+fem
            +verb+past+activeساعد+2pers+sg+fem
            +verb+past+activeساعد+2pers+sg+masc

الـ          +defArtالـ

هيئة        +noun+nonhumanهيئة+fem+sg

فلسطينيين   +adjفلسطيني+masc+dual+accgen
            +adjفلسطيني+masc+pl+accgen
            +noun+humanفلسطيني+masc+dual+accgen
            +noun+humanفلسطيني+masc+pl+accgen
```

LEXICAL ENTRIES

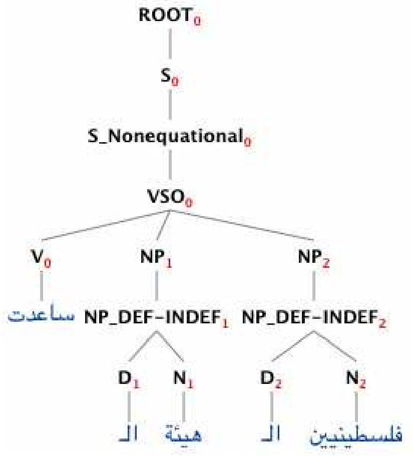
```
ساعد        V XLE (^ GLOSS)=help "This verb has three different subcat frames"
            {  (^ PRED)='%stem<(^ SUBJ)(^ OBJ)(^ COMP)>'
              (^ COMP COMP-FORM)=c أن (^ COMP COMP-TYPE)=c verbal
            |  (^ PRED)='%stem<(^ SUBJ)(^ OBJ)(^ OBL)>' (^ OBL OBJ PCASE)=c على
            |  (^ PRED)='%stem<(^ SUBJ)(^ OBJ)>'}.

هيئة        N XLE (^ GLOSS)=agency (^ PRED)='%stem' (^ PERS)=3
              { (^ NUM) (^ NUM) ~= sg | (^ NUM) = sg } "the default number is singular".

فلسطيني     N XLE (^ GLOSS)=Palestinian (^ PRED)='%stem' (^ PERS)=3
              { (^ NUM) (^ NUM) ~= sg | (^ NUM) = sg } "the default number is singular";
            ADJ XLE (^ PRED)='%stem' (^ GLOSS) = 'Palestinian'
             { (^ ATYPE)=c predicative | (^ ATYPE)= attributive}.
+1pers      V_SFX XLE (^ AGR PERS) = 1;
            PRON_SFX_PERS XLE (^ PERS) = 1.

+2pers      V_SFX XLE (^ AGR PERS) = 2;
            PRON_SFX_PERS XLE (^ PERS) = 2.

+masc       N_SFX_GEND XLE (^ GEND) = masc;
            V_SFX_GEND XLE (^ AGR GEND) = masc;
            ADJ_SFX_GEND XLE (^ GEND) = masc;
            PRON_SFX_GEND XLE (^ GEND) = masc.
```

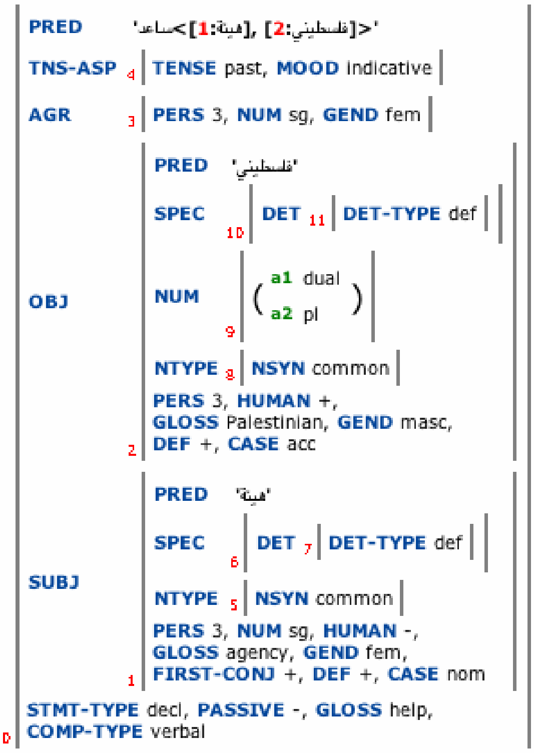
```
+pl       N_SFX_NUM XLE  (^ NUM) = pl;
          V_SFX_NUM XLE (^ AGR NUM) = pl;
          ADJ_SFX_NUM XLE (^ NUM) = pl;
          PRON_SFX_NUM XLE (^ NUM) = pl.

+dual     N_SFX_NUM XLE  (^ NUM) = dual;
          V_SFX_NUM XLE (^ AGR NUM) = dual;
          ADJ_SFX_NUM XLE (^ NUM) = dual;
          PRON_SFX_NUM XLE (^ NUM) = dual.

+accgen   N_SFX_CASE XLE (^ CASE)~= nom
                  { (^ CASE) (^ CASE) ~= acc | (^ CASE) = acc } "defaults to acc";
          ADJ_SFX_CASE XLE (^ CASE)~= nom
                  { (^ CASE) (^ CASE) ~= acc | (^ CASE) = acc } "defaults to acc".

+human    N_SFX XLE (^ HUMAN) = +.
```

PHRASE STRUCTURE RULES

```
MT ARABIC RULES (1.0)

S_Nonequational --> "There are three word orders permitted in Arabic: VSO, SVO and VOS"
          { VSO
          | SVO
          | VOS}.

VSO -->   V: ^=! @DefSTense (^ VTYPE)~= copular (^ COMP-TYPE)=verbal
             {(^ SUBJ PRED)=c 'pro' (^ SUBJ NUM) = (^ AGR NUM)
              | (^ SUBJ PRED)~= 'pro' (^ AGR NUM)=sg)}
             (^ AGR GEND)=(^ SUBJ GEND)  (^ AGR PERS)=(^ SUBJ PERS);
          {NP: (^SUBJ)=! (! FIRST-CONJ)=+
                    (! CASE)=nom (! PRON-TYPE) ~=pers
          | e: (^ SUBJ PRED)='pro' "ProDrop"
                    (^ AGR PERS)= (! PERS) (^ AGR NUM)= (! NUM) (^ AGR GEND)= (! GEND) }
          (NP: (^OBJ)=!  (! CASE)=acc).

SVO -->   NP: (^ SUBJ)=! { (! CASE) (! CASE) ~= nom | (! CASE) = nom } "the default case is nominal" ;
           V: @DefSTense (^ VTYPE)~= copular (^ COMP-TYPE)=nominal
          {(^ SUBJ HUMAN)=c - {(^ SUBJ NUM)=pl (^ AGR NUM)=sg
                               | (^ SUBJ NUM)~=pl (^ AGR NUM)=(^ SUBJ NUM)}
                    | (^ SUBJ HUMAN)~=- (^ AGR NUM)=(^ SUBJ NUM)}
             (^ AGR GEND) = (^ SUBJ GEND) (^ AGR PERS) = (^ SUBJ PERS);
          (NP: (^OBJ)=!  (! CASE)=acc).

VOS -->   V: @DefSTense (^ VTYPE)~= copular (^ COMP-TYPE)=verbal
             (^ AGR NUM)=sg (^ AGR GEND)=(^ SUBJ GEND) (^ AGR PERS)=(^ SUBJ PERS);
          NP: (^ OBJ)=! (! PRON-TYPE)=c pers (! CASE)=acc;
          NP: (^ SUBJ)=! (! CASE)=nom.
```

C-STRUCTURE & F-STRUCTURE WITH PACKED AMBIGUITY



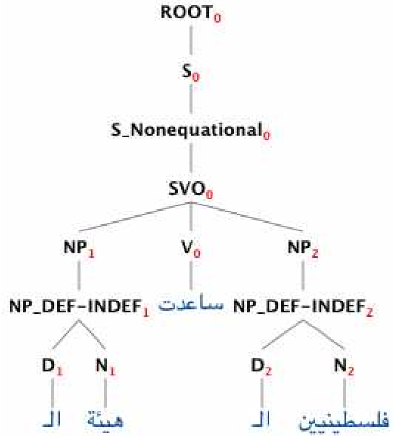


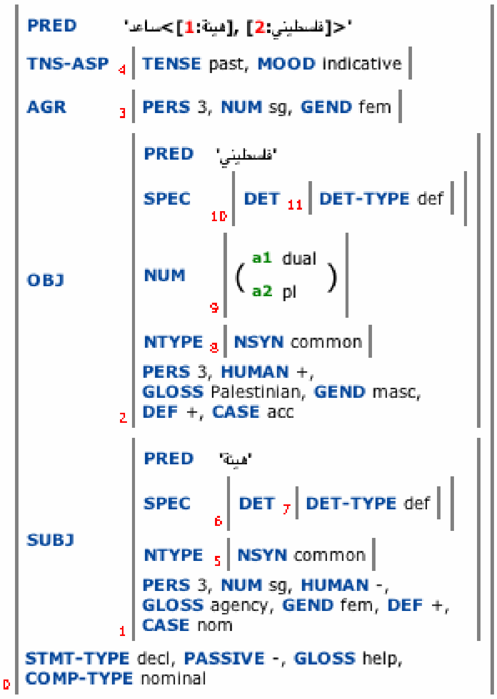
```
PRED      'ساعد<[1:هيئة], [2:فلسطيني]>'
TNS-ASP 4 [TENSE past, MOOD indicative]
AGR     3 [PERS 3, NUM sg, GEND fem]
            PRED   'فلسطيني'
            SPEC   [DET 11 [DET-TYPE def]]
                10
            NUM    ( a1 dual )
                 9 ( a2 pl   )
OBJ         NTYPE 8 [NSYN common]
            PERS 3, HUMAN +,
            GLOSS Palestinian, GEND masc,
          2 DEF +, CASE acc
            PRED   'هيئة'
            SPEC   [DET 7 [DET-TYPE def]]
                 6
SUBJ        NTYPE 5 [NSYN common]
            PERS 3, NUM sg, HUMAN -,
            GLOSS agency, GEND fem,
          1 FIRST-CONJ +, DEF +, CASE nom
STMT-TYPE decl, PASSIVE -, GLOSS help,
0 COMP-TYPE verbal
```

We note that the phrase structure will allow us to parse sentences with SVO word order, such as:

> الهيئة ساعدت الفلسطينيين
> al-haiʾatu sāʿadat al-filisṭīniyyīn/ al-filisṭīniyyain
> the-agency helped the-Palestinian.pl/ the-Palestinian.dual
> ‘The agency helped the Palestinians/ the two Palestinians.’

C-STRUCTURE & F-STRUCTURE WITH PACKED AMBIGUITY




```
PRED      'ساعد<[1:هيئة], [2:فلسطيني]>'
TNS-ASP  4 [TENSE past, MOOD indicative]
AGR      3 [PERS 3, NUM sg, GEND fem]
OBJ          [PRED   'فلسطيني'
              SPEC 10 [DET 11 [DET-TYPE def]]
              NUM   9 ( a1 dual
                        a2 pl )
              NTYPE 8 [NSYN common]
              PERS 3, HUMAN +,
              GLOSS Palestinian, GEND masc,
            2 DEF +, CASE acc]
SUBJ         [PRED   'هيئة'
              SPEC  6 [DET 7 [DET-TYPE def]]
              NTYPE 5 [NSYN common]
              PERS 3, NUM sg, HUMAN -,
              GLOSS agency, GEND fem, DEF +,
            1 CASE nom]
  STMT-TYPE decl, PASSIVE -, GLOSS help,
0 COMP-TYPE nominal
```

References

Abbès, Ramzi, Dichy, Joseph, and Hassoun, Mohamed. 2004. The Architecture of a Standard Arabic lexical database: some figures, ratios and categories from the DIINAR.1 source program. In *The Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004*, Geneva, Switzerland, pp. 15-22.

Žabokrtský, Zdenek, and Smrž, Otakar. 2003. Arabic syntactic trees: from constituency to dependency. In *The 10th Conference of the European Chapter of the Association for Computational Linguistics*, Budapest, Hungary, pp. 183-186.

Adger, David, and Ramchand, Gillian. 2003. Predication and Equation. *Linguistic Inquiry* 34:325-359.

Alegria, Iñaki, Ansa, Olatz, Ezeiza, Xabier Artola Nerea, Gojenola, Koldo, and Urizar, Ruben. 2004. Representation and Treatment of Multiword Expressions in Basque. In *The ACL workshop on Multiword Expressions*, Barcelona, pp. 48-55.

Al-Haq, Fawwaz Al-Abed. 1992. Functional or Anaphoric Control in Jordanian Arabic? *Language Sciences* 14:199-214.

Al-Raheb, Y., Akrout, A., Genabith, J. van, and Dichy, Joseph. 2006. DCU 250 Arabic Dependency Bank: An LFG Gold Standard Resource for the Arabic Penn Treebank. In *Challenge of Arabic for NLP/MT*, London, UK, pp. 105-117.

Androutsopoulou, Antonia. 2001. D-raising and Asymmetric Agreement in French. In *The Role of Agreement in Natural Language. The Fifth Annual Texas Linguistics Society Conference*, Austin, Texas, pp. 35-46.

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

Arnold, Doug, Balkan, Lorna, Meijer, Siety, Humphreys, R. Lee, and Sadler, Louisa. 1994. *Machine Translation: An Introductory Guide*. Manchester: NCC Blackwell.

Asudeh, Ash. 2000. Functional Identity and Resource-Sensitivity in Control. In *The LFG '00 Conference*, Stanford, CA, pp. 19-39.

Asudeh, Ash. 2005. Control and Semantic Resource Sensitivity. *Journal of Linguistics* 41:465-511.

Attia, Mohammed A. 2005. Developing a Robust Arabic Morphological Transducer Using Finite State Technology. In *The 8th Annual CLUK Research Colloquium*, Manchester, UK, pp. 9-18.

Attia, Mohammed A. 2006a. An Ambiguity-Controlled Morphological Analyzer for Modern Standard Arabic Modelling Finite State Networks. In *Challenge of Arabic for NLP/MT Conference*, The British Computer Society, London, UK, pp. 48-67.

Attia, Mohammed A. 2006b. Accommodating Multiword Expressions in an Arabic LFG Grammar. In Salakoski, Tapio, Ginter, Filip, Pahikkala, Tapio, Pyysalo, Tampo: *Lecture Notes in Computer Science: Advances in Natural Language Processing*, 5th International Conference, FinTAL-

2006. Turku, Finland. Vol. 4139, pp. 87-98. Springer-Verlag Berlin Heidelberg.

Attia, Mohammed A. 2007. Arabic Tokenization System. In *ACL-Workshop on Computational Approaches to Semitic Languages: Common Issues and Resources*, Prague, pp. 65-72.

Austin, Peter K. 2001. Lexical functional grammar. In *International Encyclopedia of the Social and Behavioral Sciences*, eds. Neil J. Smelser and Paul Baltes, 8748-8754. Oxford, UK: Elsevier Science Ltd.

Avgustinova, Tania, and Uszkoreit, Hans. 2003. Reconsidering the relations in constructions with non-verbal predicates. In *Investigations into Formal Slavic Linguistics*, ed. Peter Kosta, 461-482. Frankfurt, Germany: Peter Lang.

Azmi, Moinuddin. 1988. *Arabic Morphology: A Study in the System of Conjugation*. Hyderabad: Hasan Publishers.

Badawi, Elsaid, Carter, M. G., and Gully, Adrian. 2004. *Modern Written Arabic, A Comprehensive Grammar*. London and New York: Routledge.

Baldwin, Timothy, Bannard, Colin, Tanaka, Takaaki, and Widdows, Dominic. 2003. An Empirical Model of Multiword Expression Decomposability. In *The ACL-2003 Workshop on Multiword Expressions: Analysis, Acquisition and Treatment*, Sapporo, Japan, pp. 89-96.

Baldwin, Timothy. 2004. Multiword Expressions, an Advanced Course. In *The Australasian Language Technology Summer School (ALTSS 2004)*, Sydney, Australia.

Baptista, Marlyse. 1995. On the Nature of Pro-drop in Capeverdean Creole. *Harvard Working Papers in Linguistics* 5:3-17.

Agarwal, Ashwini, Ray, Biswajit, Choudhury, Monojit, Basu, Anupam, and Sarkar, Sudeshna. 2004. Automatic Extraction of Multiword Expressions in Bengali: An Approach for Miserly Resource Scenarios. In *The International Conference On Natural Language Processing (ICON 2004)*, Hyderabad, India, pp. 165-172.

Beesley, Kenneth R. 1996. Arabic Finite-State Morphological Analysis and Generation. In *COLING 1996: The 16th International Conference on Computational Linguistics*, Copenhagen, Denmark, pp. 89-94.

Beesley, Kenneth R. 1998a. Arabic Morphological Analysis on the Internet. In *The 6th International Conference and Exhibition on Multilingual Computing*, Cambridge, UK.

Beesley, Kenneth R. 1998b. Arabic Morphology Using Only Finite-State Operations. In *The Workshop on Computational Approaches to Semitic languages*, Montreal, Quebec, pp. 50-57.

Beesley, Kenneth R. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. In *The ACL 2001 Workshop on Arabic Language Processing: Status and Prospects*, Toulouse, France.

Beesley, Kenneth R., and Karttunen, Lauri. 2003. *Finite State Morphology*: CSLI studies in computational linguistics. Stanford, Calif.: Csli.

Bennett, Paul. 2003. The relevance of linguistics to machine translation. In *Computers and Translation*, ed. Harold L. Somers, 143-160. Amsterdam: Benjamins.

Bentivogli, L., and Pianta, E. 2003. Beyond Lexical Units: Enriching WordNets with Phrasets. In *The 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, Budapest, Hungary, pp. 67-70.

Beyssade, Claire, and Dobrovie-Sorin, Carmen. 2005. A Syntax-Based Analysis of Predication. In *Semantics and Linguistic Theory (SALT 15)*, UCLA, Los Angeles.

Boulton, Marjorie. 1960. *The Anatomy of Language: Saying what we Mean*. London: Routledge & Kegan Paul Limited.

Bresnan, Joan. 1995. Lexicality and Argument Structure. Invited paper presented at *Paris Syntax and Semantics Conference*, Paris, October 12-14, 1995.

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

Brill, Eric, and Resnik, Philip. 1994. A Rule-Based Approach to Prepositional Phrase Attachment Disambiguation. In *COLING 1994: The 15th International Conference on Computational Linguistics*, Kyoto, Japan, pp. 1198-1204.

Brun, Caroline. 1998. Terminology Finite-state Preprocessing for Computational LFG. In *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics*, Montreal, Quebec, Canada, pp. 196-200.

Buckley, Ronald. 2004. *Modern Literary Arabic - A Reference Grammar*. Beirut: Librairie du Liban.

Buckwalter, Tim. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. In *Linguistic Data Consortium. Catalog number LDC2002L49, and ISBN 1-58563-257-0*.

Buckwalter, Tim. 2004. Issues in Arabic Orthography and Morphology Analysis. In *The Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004*, Geneva, pp. 31-34.

Butt, Miriam, Dipper, Stefanie, Frank, Anette, and King, Tracy Holloway. 1999a. Writing Large-Scale Parallel Grammars for English, French, and German. In *The LFG '99 Conference*, Manchester, UK.

Butt, Miriam, King, Tracy Holloway, Niño, Maria-Eugenia, and Segond, Frédérique. 1999b. *A Grammar Writer's Cookbook*. Stanford, CA: CSLI Publications.

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

Butts, Aaron Michael. 2006. Observations on the Verbless Clause in the Language of Neophyti I. *Aramaic Studies* 4:53-66.

Cahill, Aoife, King, Tracy Holloway, and Maxwell, John T. 2007. Pruning the Search Space of a Hand-Crafted Parsing System with a Probabilistic Parser. In *The ACL 2007 Workshop on Deep Linguistic Processing*, Prague, Czech Republic, pp. 65-72.

Calzolari, N., Lenci, A., and Quochi, V. 2002. Towards Multiword and Multilingual Lexicons: between Theory and Practice. In *Linguistics and Phonetics 2002 (LP2002)*, Urayasu, Japan.

Cantarino, Vicente. 1974. *The Syntax of Modern Arabic Prose*. Bloomington: Indiana University Press.

Carnie, Andrew. 1995. Non-verbal predication and head-movement, MIT, Cambridge, Mass.: Doctoral dissertation.

Carnie, Andrew. 1997. Two Types of Non-verbal Predication in Modern Irish. *The Canadian Journal of Linguistics* 42.1&2:57-73.

Carter, David. 1997. The TreeBanker: A tool for supervised training of parsed corpora. In *The Fourteenth National Conference on Artificial Intelligence (AAAI-97)*, Providence, Rhode Island.

Catford, J. C. 1965. *A Linguistic Theory of Translation*. Oxford: Oxford University Press.

Chalabi, Achraf. 2000. MT-Based Transparent Arabization of the Internet TARJIM.COM. In *Envisioning Machine Translation in the Information Future, 4th Conference of the Association for machine Translation in the Americas, AMTA 2000, Cuernavaca, Mexico*, ed. John S. White, 189-191. Berlin: Springer.

Chalabi, Achraf. 2004a. Sakhr Arabic Lexicon. In *NEMLAR International Conference on Arabic Language Resources and Tools*, Cairo, Egypt, pp 21-24.

Chalabi, Achraf. 2004b. Elliptic Personal Pronoun and MT in Arabic. In *JEP-2004-TALN 2004 Special Session on Arabic Language Processing-Text and Speech*, Fez, Morocco.

Chanod, Jean-Pierre, and Tapanainen, Pasi. 1996. A Non-Deterministic Tokenizer for Finite-State Parsing. In *The European Conference on Artificial Intelligence 1996 Workshop on Extended Finite State Models of Language (ECAI'96)*, Budapest, Hungary, pp. 10-12.

Charniak, Eugene. 1996. Tree-bank grammars. *AAAI/IAAI* 2:1031-1036.

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

Corbett, Greville. 2001. Agreement: Terms and boundaries. In William Griffin (ed.), *The Role of Agreement in Natural Language, Proceedings of the 2001 Texas Linguistic Society Conference*, Austin, Texas.

Crouch, Richard, King, Tracy Holloway, Maxwell, John T., Riezler, Stefan, and Zaenen, Annie. 2004. Exploiting F-structure Input for Sentence Condensation. In *The LFG 04 Conference*, Christchurch, New Zealand, 167-187.

Crystal, David. 1980. *A First Dictionary of Linguistics and Phonetics*. London: Deutsch.

Curnow, Timothy. 2000. Towards a cross-linguistic typology of copula constructions. In *the 1999 Conference of the Australian Linguistic Society*, ed. John Henderson, 1-9: Australian Linguistic Society.

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

Dalrymple, Mary. 2001. *Lexical Functional Grammar*. New York: Academic Press.

Dalrymple, Mary, Dyvik, Helge, and King, Tracy Holloway. 2004. Copular Complements: Closed or Open? In *The LFG 04 Conference*, Christchurch, New Zealand, pp. 188-198.

Dalrymple, Mary. 2006. How much can part-of-speech tagging help parsing? *Natural Language Engineering* 12:373-389.

Darwish, Kareem. 2002. Building a Shallow Morphological Analyzer in One Day. In *The ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, PA, USA.

Deane, Paul. 2005. A Nonparametric Method for Extraction of Candidate Phrasal Terms. In *The 43rd Annual Meeting of the Association for Computational Linguistics (ACL 2005)*, Ann Arbor, Michigan, pp. 605-613.

Diab, Mona, Hacioglu, Kadri, and Jurafsky, Daniel. 2004. Automatic Tagging of Arabic Text: From Raw Text to Base Phrase Chunks. In *HLT-NAACL 2004: Short Papers*, Boston, Massachusetts, pp. 149-152.

Dichy, Joseph, and Hassoun, Mohammed. 1998. Some aspects of the DIINAR-MBC research programme. In *The 6th ICEMCO, International Conference and Exhibition on Multi-lingual Computing*, Cambridge, England.

Dichy, Joseph. 2001. On lemmatization in Arabic. A formal definition of the Arabic entries of multilingual lexical databases. In *The ACL 2001 Workshop on Arabic Language Processing: Status and Prospects*, Toulouse, France.

Dichy, Joseph, and Fargaly, Ali. 2003. Roots & Patterns vs. Stems plus Grammar-Lexis Specifications: on what basis should a multilingual lexical database centred on Arabic be built? In *The MT-Summit IX workshop on Machine Translation for Semitic Languages,* New Orleans, USA.

Dipper, Stefanie. 2003. Implementing and Documenting Large-Scale Grammars - German LFG, Stuttgart University: Doctoral Dissertation.

Ditters, Everhard. 2001. A Formal Grammar for the Description of Sentence Structure in Modern Standard Arabic. In *Workshop on Arabic Processing: Status and Prospects at ACL/EACL,* Toulouse, France.

Duffy, S. A., Morris, R. K., and Rayner, K. 1988. Lexical ambiguity and fixation times in reading. *Journal of Memory and Language* 27:429-446.

Dymetman, Marc, and Tendeau, Frédérique. 2000. Context-Free Grammar Rewriting and the Transfer of Packed Linguistic Representations. In *The 18th International Conference on Computational Linguistics: COLING 2000,* Saarbrücken, Germany, pp. 1016-1020.

Dyvik, Helge. 1999. The universality of f-structure: discovery or stipulation? The case of modals. In *The LFG '99 Conference,* Manchester, UK.

Emele, Martin C., and Dorna, Michael. 1998. Ambiguity preserving machine translation using packed representations. In *COLING-ACL '98: 36th Annual Meeting of the Association for Computational Linguistics and*

17th International Conference on Computational Linguistics, Montreal, Quebec, Canada, pp. 365-371.

Eynde, Frank Van ed. 1993. *Linguistic Issues in Machine Translation*. London: Pinter Publishers.

Falk, Yehuda N. 2001. *Lexical-Functional Grammar: An Introduction to Parallel Constraint-Based Syntax*. Stanford, Calif.: CSLI Publications.

Falk, Yehuda N. 2002. Resumptive Pronouns in LFG. In *The LFG 02 Conference*, Athens, pp. 154-173.

Falk, Yehuda N. 2004. The Hebrew Present-Tense Copula as a Mixed Category. In *The LFG 04 Conference*, Christchurch, New Zealand, pp. 226-246.

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

Fellbaum, Christine ed. 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

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

Flickinger, Dan, Lonning, Jan Tore, Dyvik, Helge, Oepen, Stephan, and Bond, Francis. 2005. SEM-I rational MT. Enriching deep grammars with a semantic interface for scalable machine translation. In *The 10th Machine Translation Summit*, Phuket, Thailand, pp. 165-172.

Fodor, J. A. 1983. *Modularity of Mind*. Cambridge, MA: MIT Press.

Forst, Martin, Kuhn, Jonas, and Rohrer, Christian. 2005. Corpus-based Learning of OT Constraint Rankings for Large-scale LFG Grammars. In *The LFG 05 Conference*, Bergen, Norway, pp. 154-165.

Forst, Martin, and Kaplan, Ronald M. 2006. The importance of precise tokenizing for deep grammars. In *The 5th Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, pp. 369-372.

Frank, Anette. 1999. From Parallel Grammar Development towards Machine Translation. A Project Overview. In *Machine Translation Summit VII*, Singapore, pp. 134-142.

Frank, Anette, King, Tracy Holloway, Kuhn, Jonas, and Maxwell, John. 2001. Optimality Theory Style Constraint Ranking in Large-scale LFG Grammars. In *Formal and Empirical Issues in Optimality-theoretic Syntax*, ed. Peter Sells, 367-397. Stanford: CSLI Publications.

Freeman, Andrew. 2001. Brill's POS tagger and a Morphology parser for Arabic. In *The ACL 2001 Workshop on Arabic Language Processing: Status and Prospects*, Toulouse, France.

Georgopoulos, Carol. 1991. *Syntactic Variables: Resumptive Pronouns and A' Binding in Palauan*. Dordrecht: Kluwer Academic Publishers.

Guenthner, Frantz, and Blanco, Xavier. 2004. Multi-Lexemic Expressions: an overview. In *Lexique, Syntaxe et Lexique-Grammaire*, eds. Christian Leclère, Éric Laporte, Mireille Piot and Max Silberztein. Philadelphia PA, USA: John Benjamins, pp. 239-252.

Habash, Nizar, and Rambow, Owen. 2005. Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In *The*

43rd Annual Meeting of the Association for Computational Linguistics (ACL-2005), Ann Arbor, Michigan, pp. 573-580.

Hajic, Jan, Smrž, Otakar, Buckwalter, Tim, and Jin, Hubert. 2005. Feature-Based Tagger of Approximations of Functional Arabic Morphology. In *The 4th Workshop on Treebanks and Linguistic Theories (TLT 2005)*, Barcelona, Spain, pp. 53-64.

Hindle, Donald, and Rooth, Mats. 1993. Structural Ambiguity and Lexical Relations. *Computational Linguistics* 19:103-120.

Hoyt, Frederick. 2004. Subject-Verb Agreement in Modern Standard Arabic: An LFG Implementation in the Xerox Language Engineering Environment. Unpublished MS, University of Texas at Austin.

Hoyt, Frederick. 2006. Arabic Nominal Clauses. In *The Encyclopedia of Arabic Language and Linguistics*. Leiden: Brill.

Hutchins, W. John. 1988. Recent developments in Machine Translation a review of the last five years In Maxwell, Dan, Schubert, Klaus and Witkam, Toon (eds.). *New Directions in Machine Translation*, Dordrecht, Foris, pp. 7-62.

Hutchins, W. J., and Somers, Harold L. 1992. *An Introduction to Machine Translation*. London: Academic Press.

Ibrahim, Khalil. 2002. *Al-Murshid fi Qawa'id Al-Nahw wa Al-Sarf [The Guide in Syntax and Morphology Rules]*. Amman, Jordan: Al-Ahliyyah for Publishing and Distribution.

Infante-Lopez, Gabriel, and Rijke, Maarten de. 2004. Comparing the Ambiguity Reduction Abilities of Probabilistic Context-Free Grammars. In *The 4th International Conference on Language Resources and Evaluation (LREC 2004)*, Lisbon, Portugal.

Jouitteau, Mélanie, and Rezac, Milan. 2006. Deriving the Complementarity Effect: Relativized Minimality in Breton agreement. *Lingua* 116:1915-1945.

Kager, René. 2000. Optimality Theory. *Computational Linguistics* 26:286-290.

Kamir, Dror, Soreq, Naama, and Neeman, Yoni. 2002. A Comprehensive NLP System for Modern Standard Arabic and Modern Hebrew. In *The ACL-02 Workshop on Computational Approaches to Semitic Languages*, Philadelphia, PA, USA.

Kaplan, Ronald M., and Bresnan, Joan. 1982. Lexical-Functional Grammar: A formal system for grammatical representation. In *The Mental Representation of Grammatical Relations*, ed. Joan Bresnan, 173-281. Cambridge, MA: The MIT Press.

Kaplan, Ronald M., Netter, Klaus, Wedekind, Jürgen and Zaenen, Annie. 1989. Translation by Structural Correspondences. In *The 4th Conference of the European Chapter of the Association for Computational Linguistics (EACL-89)*, Manchester, UK, pp. 272-281.

Kaplan, Ronald M., and Maxwell, John T. 1995. Constituent Coordination in Lexical-Functional Grammar. In *Formal Issues in Lexical-Functional Grammar*, eds. Mary Dalrymple, Ronald M. Kaplan, John Maxwell and Annie Zaenan. Stanford, CA: CSLI Publications.

Kaplan, Ronald M., and King, Tracy Holloway. 2003. Low-Level Mark-Up and Large-scale LFG Grammar Processing. In *The LFG 03 Conference*, Saratoga Springs, NY, pp. 238-249.

Kaplan, Ronald M., Riezler, Stefan, King, Tracy Holloway, Maxwell, John T., Vasserman, Alexander, and Crouch, Richard. 2004. Speed and Accuracy in Shallow and Deep Stochastic Parsing. In *The Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2004)*, Boston, MA, pp. 97-104.

Karttunen, Lauri, Chanod, Jean-Pierre, Grefenstette, G., and Schiller, A. 1996. Regular expressions for language engineering. *Natural Language Engineering* 2:305-328.

Kay, Martin. 1999. Chart translation. In *Machine Translation Summit VII*, Kent Ridge Digital Labs, Singapore, pp. 9-14.

King, Tracy Holloway, Dipper, Stefanie, Frank, Anette, Kuhn, Jonas, and Maxwell, John. 2000. Ambiguity Management in Grammar Writing. In *The Workshop on Linguistic Theory and Grammar Implementation (ESSLLI-2000)*, Birmingham; UK, pp. 5-19.

Kiraz, George Anton. 1998. Arabic Computational Morphology in the West. In *The 6th International Conference and Exhibition on Multi-lingual Computing (ICEMCO)*, Cambridge, UK.

Kuhn, Jonas, and Rohrer, Christian. 1997. Approaching ambiguity in real-life sentences - the application of an Optimality Theory-inspired constraint ranking in a large-scale LFG grammar. In *Beiträge zur 6. Fachtagung der Sektion Computerlinguistik der Deutschen Gesellschaft für Sprachwissenschaft*, Heidelberg.

Kuhn, Jonas and Sadler, Louisa. 2007. Single Conjunct Agreement and the Formal Treatment of Coordination In LFG. In *The LFG 07 Conference*, , Stanford, CA., pp. 302-322.

Kuhn, Jonas. 2002. Corpus-based Learning in Stochastic OT-LFG - Experiments with a Bidirectional: Bootstrapping Approach. In *The LFG 02 Conference*, Athens, pp. 239-257.

Larkey, Leah S., and Connell, Margaret E. 2002. Arabic Information Retrieval at UMass. In *NIST Special Publication 500-250: the 10th Text REtrieval Conference (TREC 2001)*, Gaithersburg, MD: NIST, pp. 562-570.

Levin, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*: University of Chicago Press.

Li, W., Zhang, X., Niu, C., Jiang, Y., and Srihari, R. K. 2003. An Expert Lexicon Approach to Identifying English Phrasal Verbs. In *The 41st Annual Meeting of the Association for Computational Linguistics (ACL-2003)*, Sapporo, Japan, pp. 513-520.

Lødrup, Helge. 2006. Functional Structure. In *Non-transformational Syntax: A Guide to Current Models (to appear)*, eds. Robert D. Borsley and Kersti Börjars. Oxford: Blackwells.

Maamouri, Mohamed, Bies, Ann, Jin, Hubert, and Buckwalter, Tim. 2003. Penn Arabic Treebank, Catalog No.: LDC2003T2006: Linguistic Data Consortium (LDC).

MacDonald, Maryellen C., Pearlmutter, Neal J., and Seidenberg, Mark S. 1994. Lexical Nature of Syntactic Ambiguity Resolution. *Psychological Review* 101:676-703.

MacWhinney, B., and Bates, E. 1989. *The Crosslinguistic Study of Sentence Processing*. Cambridge, England: Cambridge University Press.

Manning, C. D., and Schütze, H. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press.

Marshad, Hassan A., and Suleiman, Saleh M. 1991. A Comparative Study of Swahili ni and Arabic kana as Copulative Elements. *Language Sciences* 3:21-37.

Maxwell, John T., and Kaplan, Ronald M. 1993. The Interface between Phrasal and Functional Constraints. *Computational Linguistics*. 19(4): 571-590.

Maxwell, John T., and Kaplan, Ronald M. 1996. Unification-based Parsers that Automatically Take Advantage of Context Freeness. In *The LFG 96 Conference*, Grenoble, France.

McCarthy, John J. 1985. *Formal Problems in Semitic Phonology and Morphology*: Outstanding dissertations in linguistics. New York; London: Garland.

Meral, Hasan Mesud. 2004. Resumptive Pronouns in Turkish, Bogaziçi University, Istanbul.: Master's Thesis.

Nagata, Masaaki. 1992. An Empirical Study on Rule Granularity and Unification Interleaving Toward an Efficient Unification-Based Parsing System. In *The 15th International Conference on Computational Linguistics (COLING-92),* Nantes, France, pp. 177-183.

Nelken, Rani, and Shieber, Stuart M. 2005. Arabic Diacritization Using Weighted Finite-State Transducers. In *The ACL 2005 Workshop on Computational Approaches to Semitic Languages,* Ann Arbor, Michigan, pp. 79-86.

Nerima, Luka, Seretan, Violeta, and Wehrli, Eric. 2003. Creating a Multilingual Collocations Dictionary from Large Text Corpora. In *The 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL03),* Budapest, Hungary, pp. 131-134.

Nordlinger, Rachel, and Sadler, Louisa. 2006. Verbless clauses: revealing the structure within. In *Architectures, Rules and Preferences: A Festschrift for Joan Bresnan* (to appear), eds. Jane Grimshaw, Joan Maling, Chris Manning, Jane Simpson and Annie Zaenan. Stanford: CSLI Publications.

O'Donovan, Ruth, Cahill, Aoife, Way, Andy, Burke, Michael, and Genabith, Josef van. 2004. Large-Scale Induction and Evaluation of Lexical Resources from the Penn-II Treebank. In *The 42nd Annual Conference of the Association for Computational Linguistics (ACL-04),* Barcelona, Spain, pp. 367-374.

Oepen, Stephan, and Lønning, Jan Tore. 2006. Discriminant-based MRS banking. In *The 4th International Conference on Language Resources and Evaluation*, Genoa, Italy, pp. 1250-1255.

Oflazer, Kemal, Uglu, Özlem Çetino, and Say, Bilge. 2004. Integrating Morphology with Multi-word Expression Processing in Turkish. In *The*

2nd ACL Workshop on Multiword Expressions: Integrating Processing, Spain, pp. 64-71.

Othman, Eman, Shaalan, Khaled, and Rafea, Ahmed. 2003. A Chart Parser for Analyzing Modern Standard Arabic Sentence. In *The MT Summit IX Workshop on Machine Translation for Semitic Languages: Issues and Approaches*, New Orleans, USA.

Owczarzak, Karolina, Graham, Yvette, Genabith, Josef van and Way, Andy. 2007. Using F-structures in Machine Translation Evaluation. In *The LFG 07 Conference*, Stanford, CA., pp. 383-396.

Platzack, Christer. 2003. Agreement and Null Subjects. In *The 19th Scandinavian Conference of Linguistics*, University of Tromsø, Norway, pp. 326-355.

Prince, Alan, and Smolensky, Paul. 1993. Optimality Theory: Constraint interaction in generative grammar. In *Technical Report No. 2*. Cambridge, MA: University Center for Cognitive Science.

Pustet, Regina. 2003. *Copulas. Universals in the Categorization of the Lexicon*. New York: Oxford University Press.

Ramsay, Allan, and Mansour, Hanady. 2007. Towards including prosody in a text-to-speech system for modern standard Arabic. *Computer Speech and Language* 22:84–103.

Ratcliffe, Robert R. 1998. *The Broken Plural Problem in Arabic and Comparative Semitic : Allomorphy and Analogy in Non-concatenative Morphology*: Amsterdam studies in the theory and history of linguistic science. Series IV, Current issues in linguistic theory ; v. 168. Amsterdam ; Philadelphia: J. Benjamins.

Revell, E. J. 1989. The Conditioning of Word Order in Verbless Clauses in Biblical Hebrew. *Journal of Semitic Studies* 34:1-24.

Riezler, Stefan, King, Tracy H., Kaplan, Ronald M., Crouch, Richard, Maxwell, John T., and Johnson, Mark. 2002. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *The 40th Annual Meeting of the Association for Computational Linguistics (ACL-2002)*, Philadelphia, PA, pp. 271-278.

Robinson, Douglas. 1997. *Becoming a Translator: An Accelerated Course*. London: Routledge.

Rohrer, Christian, and Forst, Martin. 2006. Improving coverage and parsing quality of a large-scale LFG for German. In *The 5th Conference on Language Resources and Evaluation (LREC 2006)*, Genoa, Italy, pp. 2206-2211.

Rosén, Victoria. 1996. The LFG Architecture and "Verbless" Syntactic Constructions. In *The LFG 96 Conference*, Grenoble, France.

Rosén, Victoria, Meurer, Paul, and Smedt, Koenraad de. 2005a. Constructing a Parsed Corpus with a Large LFG Grammar. In *The LFG 05 Conference*, Bergen, Norway, pp. 371-387.

Rosén, Victoria, Smedt, Koenraad de, Dyvik, Helge, and Meurer, Paul. 2005b. TREPIL: Developing Methods and Tools for Multilevel Treebank Construction. In *The 4th Workshop on Treebanks and Linguistic Theories*, Barcelona, Spain, pp. 161-172.

Rosén, Victoria, Smedt, Koenraad De, and Meurer, Paul. 2006. Towards a toolkit linking treebanking and grammar development. In *The 5th Workshop on Treebanks and Linguistic Theories*, Prague, Czech Republic, pp. 55-66.

Ryding, Karin C. 2005. *A Reference Grammar of Modern Standard Arabic*. Cambridge: Cambridge University Press.

Sadler, Louisa, and Thompson, Henry S. 1991. Structural Non-Correspondence in Translation. In *The Fifth Conference of the European Chapter of the Association for Computational Linguistics (EACL-1991)*, Berlin, Germany, pp. 293-298.

Sadler, Louisa. 2003. Coordination and Asymmetric Agreement in Welsh. In *Nominals: Inside and Out*, eds. Tracy H. King and Miriam Butt, 85-118: CSLI Publications.

Sag, Ivan A., Baldwin, Timothy, Bond, Francis, Copestake, Ann, and Flickinger, Dan. 2002. Multi-word Expressions: A Pain in the Neck for NLP. In *Computational Linguistics and Intelligent Text Processing*, ed. A. Gelbukh, 1-15: Springer.

Seidenberg, M. S., Tanenhaus, M. K., Leiman, J. L., and Bienkowski, M. 1982. Automatic access of the meanings of ambiguous words in contexts: Some limitations of knowledge-based processing. *Cognitive Psychology* 14:489-537.

Sells, Peter. 1985. *Lectures on Contemporary Syntactic Theories*: CSLI Lecture Notes. Stanford, CA: CSLI.

Sibawaihi, Abu Bishr 'Amr. 1966. *Al-Kitab*. Cairo, Egypt: Dar al-Qalam.

Smadja, Frank, McKeown, Kathleen R., and Hatzivassiloglou, Vasileios. 1996. Translating collocations for bilingual lexicons: a statistical approach. *Computational Linguistics* 22:1-38.

Suleiman, M. Y. 1990. Sibawaihi's 'Parts Of Speech' According To Zajjaji: A New Interpretation. *Journal of Semitic Studies* XXXV/2:245-263.

Suleiman, Saleh M. 1989. On the Pragmatics of Subject-Object Preposing in Standard Arabic. *Language Sciences* 11:215-235.

Tang, Sze-Wing. 2001. Nominal predication and focus anchoring. In *ZAS Papers in Linguistics*, eds. Gerhard Jäger, Anatoli Strigin, Chris Wilder and Niina Zhang, 159-172. Berlin: ZAS.

Trujillo, Arturo. 1999. *Translation Engines: Techniques for Machine Translation*. London: Springer.

Tucher, Allen B. 1987. Current strategies in machine translation research and development. In *Machine Translation: Theoretical and methodological issues*, ed. Sergei Nirenburg, 23-24. Cambridge: Cambridge University Press.

Vaillette, Nathan. 2002. Irish gaps and resumptive pronouns in HPSG. In *The 8th International Conference on Head-Driven Phrase Structure Grammar*, Stanford, pp. 284-299.

Vauquois, Bernard. 1978. L'evolution des logiciels et des modeles linguistiques pour la traduction automatisee. *T.A. Informations* 1:1-21.

Venkatapathy, Sriram. 2004. Overview of my work on Multi-word expressions and Semantic Role Labeling. Technical Report. International Institute of

Information Technology. Hyderabad, India. (http://www.cse.iitk.ac.in/users/iriss05/v_sriram.pdf)

Volk, Martin. 1998. The Automatic Translation of Idioms. Machine Translation vs. Translation Memory Systems. In *Machine Translation: Theory, Applications, and Evaluation. An assessment of the state-of-the-art*, ed. Nico Weber. St. Augustin: Gardez!-Verlag

Way, Andy 1999. A Hybrid Translation Model using LFG-DOP. In *The 10th Irish Conference on Artificial Intelligence and Cognitive Science*, Cork, Ireland, pp.130-136.

Wedekind, Jurgen, and Kaplan, Ronald M. 1996. Ambiguity-preserving Generation with LFG and PATR-style Grammars. *Computational Linguistics* 22:555-568.

Wehr, Hans. 1979. *A Dictionary of Modern Written Arabic*. Ithaca, NY: Spoken Language Services, Inc.

Willis, David. 2000. On the distribution of resumptive pronouns and wh-trace in Welsh. *Journal of Linguistics* 36:531–573.

Willis, Tim. 1992. Processing Natural Language. In *Computing in Linguistics and Phonetics: Introductory Readings*, ed. Peter Roach. San Diego: Academic Press.

Wright, W. 1896/2005. *A Grammar of the Arabic Language*. Cambridge: Cambridge University Press.

Wunderlich, Dieter. 2005. Optimality theory in morphology and syntax. In *Encyclopaedia of Language and Linguistics*. Oxford: Elsevier, 2nd edition, vol. 12, pp. 408-418.

Yona, Shlomo, and Wintner, Shuly. 2005. A finite-state morphological grammar of Hebrew. In *The ACL 2005 Workshop on Computational Approaches to Semitic Languages*, Ann Arbor, Michigan, pp. 9-16.

Zavrel, Jakub, Daelemans, Walter, and Veenstra, Jorn. 1997. Resolving PP attachment Ambiguities with Memory-Based Learning. In *The workshop on Computational Natural Language Learning (CoNLL'97)*, Madrid, pp. 136-144.