NEMLAR

Specifications of the Arabic Written Corpus

Muhammad Atiyya (RDI)
Khalid Choukri (ELDA)
Mustafa Yaseen (AU)

September 29th, 2005

CONTENTS

The NEMLAR project is supported by the INCO-MED programme

http://www.nemlar.org

1. Introduction

Following the work carried out within the other work packages of NEMLAR, the consortium agreed to focus on three main resources:

- Written Corpus
- Speech corpus for TTS applications
- Broadcast News

This document describes the Nemlar Written Corpus (WC).
The data consists of about 500K words of Arabic text from 13 different categories. The text is provided in 4 different versions:

- Raw text
- Fully vowelized text
- Text with Arabic lexical analysis
- Text with Arabic POS-tags

The database was produced and annotated by RDI, Egypt for the Nemlar Consortium. The Nemlar Arabic Written Corpus is owned and copyrighted by the Nemlar Consortium.

The database is distributed on 1 ISO 9660 CD-ROM volume.

The contents of the disk include the documentation files and label files.
The list of the distribution disks and directories are contained in the README.TXT file. Further details regarding the database contents, files and directories are provided in the documentation files in the DOC directory.

File types are identified with the following extensions:
*.DOC - Microsoft Word V6.0 document
*.TBL - DOS text file with Unicode symbols
*.TXT - DOS text file with Unicode symbols
*.PS - Adobe Postscript file
*.PDF - Adobe PDF file

The CD-ROM has the following directory structure:

\:
COPYRIGH.TXT - copyright notice
DISK.ID - UNIX volume ID file
README.TXT - describes files in the distribution media
NMWRC7AR\DATA- data directory
NMWRC7AR\REFS - reference directory
NMWRC7AR\DOC:
DESIGN.DOC - database documentation file
VALREP.DOC - validation report of the database

1.1 Directory structure

The databases should comply with the following directory structure:

```
/<database>/Data/<subcorpus>
```

Where:

<database>	Defined as <dbName><#><language code> where <dbName> is NMWRC, <#> is 7 for NEMLAR, <language code> is <ISO-639 code>: FOR OUR DATABASE: NMWRC7AR
<subcorpus>	Defined as: • RawText • TextwithArabicDiacritization • TextwithArabicLexicalAnalysis • TextwithArabicPOS_tags

Table 1.1 – Directory structure

The directory structure is very simple that; under the root directory reside the LinksMap.txt file, this document references each text file to its source and category), and the following four directories:

RawText,
TextWithArabicLexicalAnalysis,
TextWithArabicPOS_Tags
TextWithArabicDiacritization

with each of the 4 directories respectively contains; raw text files, the same files lexically analyzed, the same files POS tagged, and the same files fully diacritized concatenatively.

In addition to the previous structures the following directories are used to store the other files:

/	Root directory
/<database>/DOC	Documentation
/<database>/REFS	Referenced articles that are listed in this document

Table 1.2 – Documentation files hierarchy

The root directory contains three mandatory files:

- COPYRIGH.TXT: a copyright text in ASCII format,
- DISK.ID: an 11-character string with the volume name (required for systems that cannot read the physical volume label),
- README.TXT: an ASCII text file that lists all files of the database, except for signal and label files which can be indicated by their name template.

1.2 File naming conventions

DD_YYMMDD_SSS_LL.<ext>

where:

DD	Database identification code (00-ZZ) For NEMLAR: N7
YYMM DD	Year, month, day of recording
SSS	Source of recording (three characters, e.g. 'RTM')

LL	Two letter ISO 639 language code, e.g. 'AR'
<ext>	File type code, i.e. .WAV = speech file .SAM = SAM annotation file .TRS = transcription in XML format

Table 1.3 - Filename convention

1.3 Label files

All delivered files are plain Unicode text files for maximum portability.

2. Database design and collection

2.1 Sampling parameters

Sampling parameters taken into considerations are:

- TimeSpan, (mostly recent; i.e. late 1990's till now 2005)
- Standard Arabic only is considered as it is the most commonly accepted variant through out the native Arabic speakers, and due to its regularity that can be consistently modeled by our tools.
- Multiple miscellaneous domains are considered; political, scientific, general news, … (see the table just below)

The corpus design of the resource is a function of two criteria: a sampling strategy and a definition of the size of the resource and of each "session/genre" if we consider this as important feature).

The criteria must ensure:

- To produce corpora that in principle offer a representation for the variety of syntactic, semantic, pragmatic features of modern Arabic
- To produce corpora that are a comparable "characteristic" to other BLARK resources for other languages (not to mention BNC , PAROLE, etc.)

Given the restriction on the size of the resource due to the cost of the resources, the general size of the Written Corpus is 500 K words. The rationals behind the choice of the corpora:

- things that exist.
- things that would lead to a good balanced corpus, comparable to others.
- what we can afford within our budget.

This makes the domain categories size distribution as follows:

Written Corpus	Written Corpus size
political news	48,000 words
political debate	30,000 words
Islamic text (Preaching and others)	29,000 words
Phrases of common words (for TTS speakers DB LR)	8,500 words
Text taken from Broad Cast News for TTS speakers DB LR.	5,500 words

Business	20,000 words
Arabic literature	30,000 words
General news	100,000 words
Interviews	56,000 words
Scientific press	50,000 words
Sports press	50,000 words
Dictionary entries explanation	52,000 words
Legal domain text	21,000 words
Total size:	**500,000 words**

2.2 Written Corpus processing procedure

The process goes briefly as follows; Arabic linguists run RDI's tool Fassieh© where they initiate a book project and feed to it with their plain text files where each file is considered as a logical page. They then press a button to run the automatic analysis engine on the text in some page (no manual intervention at this phase; just machine time). The accuracy of that automatic analysis is typically around 95%.

To reach about 99%+ (as is the case of this LR)accuracy rate, the linguists uses the visual revision mode of Fassieh© where all the possible analyses of each word are elegantly ranked and visualized and the linguist has to either approve the 1st most likely analysis (most of the time) or select another one (in the 4% minority of the cases) by clicking its column head.

The productivity of a well trained linguist is at least 1.5K words per work-day. The team of RDI has worked on that tool Fassieh© and also RDI has afford this tool with the necessary training for the team of AU to do their part of the LR.

Given the process above is done, Fassieh© is asked to produce the following kinds of output files:

Lexical analysis files: Each file of this kind is simply the same as the input file with each Arabic word (a string of Arabic alphabets delimited by non Arabic alphabets) is replaced by its lexical analysis in the following notation expressed by the following BNF rule:

{Wv;(Tn) T: (Pn) P, (Rn) R, (Fn) F, (Sn) S}

where;

Wv is the vowelized mnemonic of the vowelized full form word.
Tn is the mnemonic of word type.
T is the ID of the word type.
Pn is the mnemonic of word prefix.
P is the ID of the word prefix.
Rn is the mnemonic of word root.
R is the ID of the root prefix.
Fn is the mnemonic of word pattern (or form).
F is the ID of the word pattern.
Sn is the mnemonic of word suffix.
S is the ID of the word suffix.

For the Arabic lexical structure model of RDI see Appenidx 1 of this document.

POS tagging files: Each file of this kind is simply the same as the input file with each Arabic word (a string of Arabic alphabets delimited by non Arabic alphabets) is replaced by its POS tags vector in the following notation expressed by the following BNF rules:

{[Wv];Tp}

where;

Wv is the vowelized mnemonic of the vowelized full form word.
Tp is the POS tags vector of the full form word and is defined as (T)#; where

T is a POS tag mnemonic.

For the Arabic POS tagging model of RDI see Appenidx 2 of this document.

Vowelization files: Each file of this kind is simply the same as the input file with each Arabic word (a string of Arabic alphabets delimited by non Arabic alphabets) is fully vowelized (ponetically written).

There're 4 more vowelization marks that we added to the standard set of Arabic diacritics in order to get a fully vowelized Arabic text that is ready to be 1-to-1 mapped into any phonetic transcription notation (IPA, SAMPA, ..):

1- **@** *means long vowel,*
2- **×** *means unpronounced char.,*
3- **^** *unwritten long vowel ALIF,*
4- **~** *means Alif Layyina (long vowel ALIF but written as YAA).*

All the tools are proprietary one or RDI. The corner stone one is *Fassieh*©.

3. Deviations from Nemlar specifications

None

4. List of references

For more details on the underlying theory of the produced analyses, the following references can be consulted:

1- **[Attia, 2005]** Attia, M., Theory and Implementation of a Large-Scale Arabic Phonetic Transcriptor, and Applications, and Applications. PhD thesis, Dept. of Electronics and Electrical Communications, Cairo University, 2005.
2- **[Attia, Rashwan, Khallaaf, 2004]** Attia, M., Rashwan, M., Khallaaf, G., A Formalism of Arabic Phonetic Grammar and Application on the Automatic Arabic Phonetic Transcription of Transliterated Words, NEMLAR int'l conference in Cairo, Sept. 2004.
3- **[Attia, Rashwan, 2004]** Attia, M., Rashwan, M., A Large-Scale Arabic POS Tagger Based on a Compact Arabic POS Tags Set, and Application on the Statistical Inference of Syntactic Diacritics of Arabic Text Words, NEMLAR int'l conference in Cairo, Sept. 2004.
4- **[Attia, Rashwan, Khallaaf, 2002]** Attia, M., Rashwan, M., Khallaaf, G., On Stochastic Models, Statistical Disambiguation, and Applications on Arabic NLP Problems, The Proceedings of the 3rd Conference on Language Engineering; CLE'2002, the Egyptian Society of Language Engineering (ESLE).
5- **[Attia, 2000]** Attia, M., A Large-Scale Computational Processor of The Arabic Morphology, and Applications, MSc. thesis, Dept. of Computer Engineering, Faculty of Engineering, Cairo University, 2000.

Electronic versions of these references are downloadable at the link

http://www.rdi-eg.com/rdi/Technologies/paper.htm

Appendix 1 - RDI's Arabic lexical analysis model

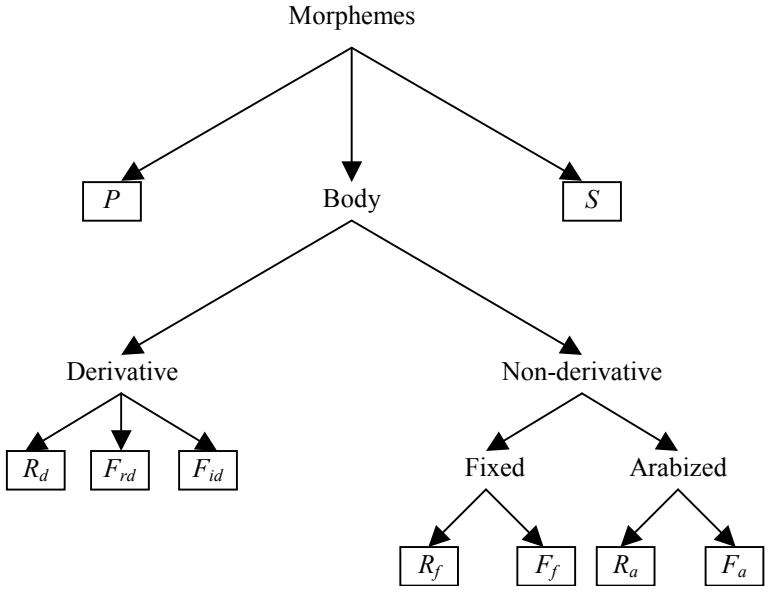
This section quickly gives a brief idea about the lexical processing component; namely RDI's ArabMorpho© (ArabMorpho©, 2000), deployed to extract Arabic lexical diacritization. For a fully detailed description, the reader is referred to (Attia, 2000).

Due to the highly derivative and inflective nature of the Arabic language, it is much more comprehensive, effective, and economic to deal with its compact set of basic building entities; i.e. morphemes, than its unmanageably huge generable vocabulary. Following that morpheme-based approach, the canonical lexical structure of any Arabic word **w** according to ArabMorpho© has been formulated as a quadruple;

$$\boldsymbol{w} \longrightarrow \underline{q} = (t : p, r, f, s)$$

where t is Type Code (with possible types are Regular Derivative, Irregular Derivative, Fixed, Arabized), p is Prefix Code, r is Root Code, f is Pattern Code, and s is Suffix Code.

These kinds of morphemes in the Arabic lexicon of ArabMorpho© are clearly classified in the figure below.



Classifying the 9 types of morphemes in the Arabic lexicon of ArabMorpho©.

With a dynamic coverage ratio exceeding 99.8% – without counting the Arabic transliterated foreign words - the knowledge base (i.e. the lexicon) of ArabMorpho© based on this model are composed from only about 7,700 morphemes with the fully agent-oriented linguistic description of each[1]. The sizes of each kind of morphemes in the figure above are as follows:

[1] *Building the lexicon of* ArabMorpho© *has taken about 4 working years of RDI team surveying, filtering, and transforming the classic material in dozens of classical Arabic linguistic resources to the computational format of our lexical model. The following resources were among the most useful ones to our team; (Al-Faraaby, 1974)) ,Ibn Faaris,*

1- P: About 260 Arabic prefixes.
2- R_d: About 4,600 Arabic derivative roots.
3- F_{rd}: About 1,000 Arabic regular derivative patterns.
4- F_{id}: About 300 Arabic irregularly derived words.
5- R_f: About 250 Roots of Arabic fixed words.
6- F_f: About 300 Arabic fixed words.
7- R_a: About 240 Roots of Arabized words.
8- F_a: About 290 Arabized words.
9- S: About 550 Arabic suffixes.

Without attempting to explain the lexical analysis and synthesis process where the reader is referred again to (Attia, 2000), the table below shows this model in application on few sample Arabic words.

Sample word	**Type**	**Prefix &** Prefix Code	**Root &** Root Code	**Pattern &** Pattern Code	**Suffix &** Suffix Code
فَمَا	Fixed	فَـ **2**	اَلَّذِي **87**	مَا **48**	- **0**
تَتَنَاوَلَه	Regular Derivative	تـ **86**	ن و ل **4077**	تَفَاعَلَ **176**	ـه **8**
اَلْكِتَابَات	Regular Derivative	الـ **9**	ك ت ب **3354**	فِعَال **684**	ـات **27**
اَلْعِلْمِيَّة	Regular Derivative	الـ **9**	ع ل م **2754**	فِعْل **842**	ـيَّة **28**
مِنْ	Fixed	- **0**	مِنْ **63**	مِنْ **118**	- **0**
مَوَاضِيع	Regular Derivative	- **0**	و ض ع **4339**	مَفَاعِيل **93**	- **0**

ArabMorpho©'s canonical lexical structure of sample words

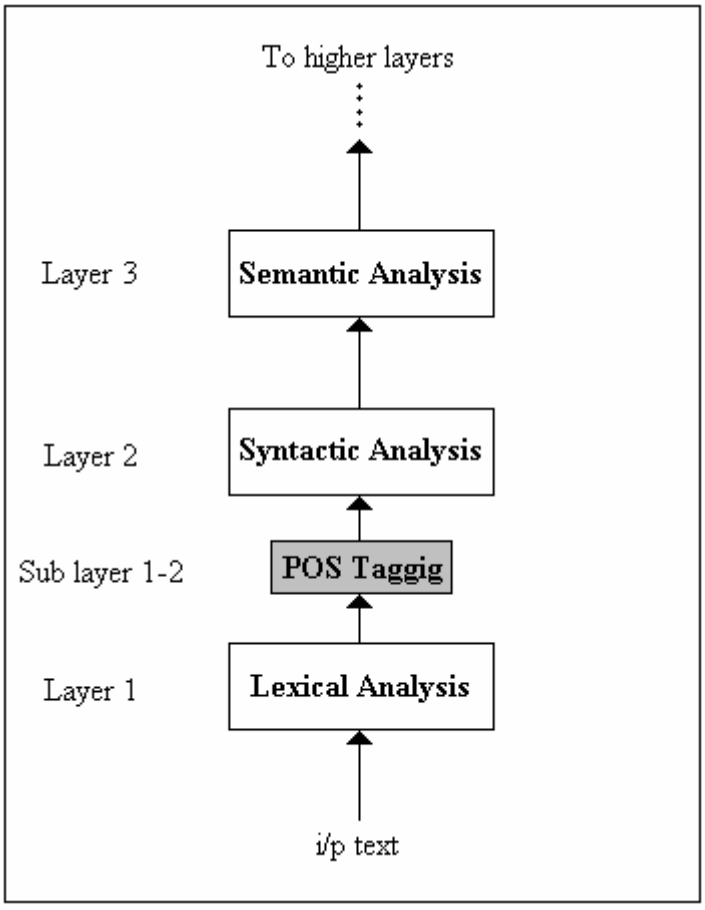
1969), (Addahdaah, 1990), (Abdul Fattaah, 1983), (Fayyaadh, M., 1990), (Arragehy, 1993), (Al-Waseett Diction, 1985), (Holy Quraan Vocabulary Diction, 1990), (Al-Kabeer Diction, 1992), (Arraazy), (Udhayma).

Appendix 2 – Defining and positioning Arabic POS tagging

Part-Of-Speech (POS) tagging is a fundamental linguistic analysis process where POS tags that convey the basic context-free syntactic features of input surface text words are extracted.

Among several linguistic processing tasks for whom POS tagging may be quite useful – including our two problems tackled here - POS tags are the most essential input features for all kinds of natural language computational syntax parsers which are in turn one step higher in the ladder towards language understanding and machine translation as well.

Based on that definition the position of POS tagging is obviously a middle sub layer between the two fundamental lexical and syntactic ones on the NLP ladder as shown in the figure below.



The position of POS tagging on the NLP ladder.

Just before delving to the next few sections that anatomize our large-scale Arabic POS tagger; RDI's ArabTagger© (ArabTagger©, 2004), it may be useful to remind the reader with the underlying ArabMorpho©'s lexical canonical structure – presented earlier in chapter (3) - of any Arabic word w as a quadruple of morphemes;

$$w \rightarrow Q = (t : p, r, f, s)$$

where t is Type Code (with possible types are Regular Derivative, Irregular Derivative, Fixed,

Arabized), p is Prefix Code, r is Root Code, f is Pattern Code, and s is Suffix Code.

With a dynamic coverage ratio exceeding 99.8% coverage ratio; the knowledge base (i.e. lexicon) of our lexical analyzer ArabMorpho© based on this model comprises only about 7,700 morphemes along with the full agent-oriented linguistic description of them. (ArabMorpho©, 2000), (Attia, 2000)

Compact Arabic POS Tags Set

Composing an Arabic POS tags set necessitates scanning the lexico-syntactic features of each possible word of the Arabic vocabulary which is apparently infeasible. Instead, thanks for the morpheme-based approach, the features of each morpheme in the relatively compact ArabMorpho© knowledge base have been scanned, then digested through several iterations of decimation into a non redundant compact Arabic POS tags set.

During that scanning process the following criteria has been adhered to:

6- All the existing lexico-syntactic features must be named and registered, which aims to the completeness of the resulting POS tags set.

7- All the named and registered features must be atomic, which aims to compactness and avoids redundancy in the resulting tags set. This in turn is vital for the effectiveness of the based upon POS tagging process - which is essentially an abstraction process - and all higher processing layers as well.

8- All the named and registered features can be ensured upon the POS labeling of the morphemes in our Arabic lexical knowledge base. (More on this point in the next section below)

The table below displayed below shows our Arabic POS tags set along with the meaning of each tag verbalized in both English and Arabic. Moreover, the 62 tags in the set are functionally categorized in order to maximize clarity.

While some tags in the following table may have corresponding ones in other languages; e.g. English, others do not have such counterparts and are specific to the Arabic language.

Cat.	Mnemonic	Meaning in English	Meaning in Arabic
Start of word marker	SOW	Start-Of-Word marker	
Pad-ding strin g	Padding	Padding string	
Features of noun and verb prefixes	NullPrefix	Null prefix	
	Conj	Conjunctive	
	Confirm	Confirmation by Laam	
	Interrog	Interrogation by Hamza	
Features of noun and verb suffixes	NullSuffix	Null suffix	
	ObjPossPro	Object or possession pronoun	
Verb and noun syntactic cases	MARF	1st Arabic syntactic case	
	MANSS	2nd Arabic syntactic case	
Features of noun-only prefixes	Definit	Definitive article	
Features of noun-only stems	Noun	Nominal	
	NounInfinit	Nouns made of infinitives	
	NounInfinitLike	"NounInfinit" like	
	SubjNoun	Subject noun	
	ExaggAdj	Exaggeration adjective	
	ObjNoun	Object noun	
	TimeLocNoun	Noun of time or location	
	NoSARF	An Arabic feature of a specific class of nouns	
Features of noun-only suffixes	PossessPro	Possessive pronoun	
	RelAdj	Relative adjectives maker	
	Femin	Feminine	
	Masc	Masculine	
	Single	Singular	
	Binary	Binary	
	Plural	Plural	
	Adjunct	Adjunct	
	NonAdjunct	NonAdjunct	
	MANSS_MAGR	2nd or 3rd Arabic syntactic case	
	MAGR	3rd Arabic syntactic case	
Features of verb-only prefixes	Present	Present tense	
	Future	Future tense	
Features of verb-only stems	Active	Active sound	(
	Passive	Passive sound	(
	Imperative	Imperative	
	Verb	Verb	
	Intransitive	Intransitive verb	
	MAJZ	4th Arabic syntactic case	
	Past	Past tense	
	PresImperat	Present tense, or imperative	
Features of verb-only suffixes	SubjPro	Subject form pronoun	
	ObjPro	Object form pronoun	
	MANS_MAJZ	2nd or 4th Arabic syntactic case	
Features of: mostly functional fixed words, and scarcely affixes	Prepos	Preposition	
	Interj	Interjection	
	PrepPronComp	Preposition-Pronoun Compound	
	RelPro	Relative pronoun	
	DemoPro	Demonstrative pronoun	
	InterrogArticle	Interrogation article	
	JAAZIMA	For specific articles that make the consequent verb in the 4th Arabic syntactic case	
	CondJAAZIMA	Feature of a class of Arabic conditionals	
	CondNot JAAZIMA	Feature of a class of Arabic conditionals	
	LAA	Arabic specific article	
	LAATA	Arabic specific article	
	Except	Article of exception	
	NoSyntaEffect	A class of articles that have no syntactic effect	
	DZARF	Feature for certain kind of Arabic adverbs	
	ParticleNAASIKH	A class of particles that make the subject of the consequent nominal sentence in 2nd Arabic syntactic case	
	VerbNAASIKH	A class of auxiliary verbs that make the predicate of the consequent verbal sentence in 2nd Arabic syntactic case	
	ParticleNAASSIB	Arabic specific class of particles that make the consequent verb in 2nd Arabic syntactic case	
	MASSDARIYYA	Arabic specific article	
For words beyond our morphological model	Translit	Transliterated Arabic string	

Arabic POS Labeling

Having the Arabic POS tags set been designed, labeling the morphemes of the lexical knowledge base comes as the next job which is a straightforward one given that the following three main points are carefully considered:

1- For morphologically analyzed words; the f part of the quadruples gives the Arabic POS tagging of stems, while the p and s parts give the Arabic POS tagging of affixes. Hence, the root morphemes of all kinds which do not participate to tagging are not Arabic POS labeled.

2- Due to the atomicity of the tags in the Arabic POS tags (see section 4.2) and in same time the compound nature of Arabic morphemes in general, POS labels of Arabic morphemes are vectors not simple scalars.

3- Only ensured Arabic POS tags are considered in the Arabic POS labeling of morphemes. i.e. When an Arabic POS tag is a possible - or even a highly probable – but not an ensured feature of a given morpheme, it is not included in its Arabic POS label vector.

The following few morpheme labeling examples (ArabMorpho©) are listed below in order to concretely illustrate the process:

Morpheme type and code	***Morpheme shown as Arabic string***	***Arabic POS vector label***
Prefix; 9		*[Definit]*
Prefix; 125		*[Future,Present,Active]*
Regular derivative pattern; 482		*[Noun,SubjNoun]*
Regular derivative pattern; 67		*[Noun,NounInfinit]*
Irregular derivative pattern; 29		*[Noun,NoSARF,Plural]*
Fixed pattern; 8		*[Noun,SubjPro]*
Fixed pattern; 39		*[Noun,Masc,Single,Adjunct,MARF]*
Suffix; 27		*[Femin,Plural]*
Suffix; 427		*[Present,MARF,SubjPro,ObjPro]*
Suffix; 195		*[RelAdj,Femin,Binary,NonAdjunct,MARF]*

Sample Arabic POS labels from RDI's ArabMorpho© knowledge base.

Arabic POS Tagging

The Arabic POS tagging process is implemented in the following steps:

1- The Arabic strings sequence to be POS tagged are morphologically analyzed and combinatorially disambiguated using RDI's ArabMorpho© as explained in chapter 3. (ArabMorpho©, 2000), (Atiyya, 2000), (Atiyya, et al, 2002) This results in a disambiguated quadruples sequence where each string is substituted by either one quadruple or a mark of Transliterated string.

2- For the prefix, pattern, and suffix morphemes of each quadruple in the sequence, the Arabic POS labels; APOS(p) APOS(t:f) APOS(s) are retrieved from the Arabic lexicon of ArabMorpho©.

3- The Arabic POS tags vector of each word in the sequence is then composed using the formula:

$$APOS(w) = Concat(APOS(p), APOS(t:f), APOS(s))$$

where the Concat function simply concatenates the POS sub vectors of the constituting morphemes after eliminating any mutual redundancy among their tags.

The resulting Arabic POS tags vectors by RDI's ArabTagger© (ArabTagger©, 2004) of the words in a real-life phrase are shown in the table below:

Phrase words		*Most likely Arabic POS tags vectors*
وقد	**وَقدْ**	[SOW,Conj,NoSyntaEffect,NullSuffix]
صرحت	**صَرَّحَتْ**	[SOW,NullPrefix,Verb,Past,Single,Femin]
رئيسة	**رَئِيسَة**	[SOW,NullPrefix,Noun,ExaggAdj,Single,Femin]
الوزراء	**اَلْوُزَرَاء**	[SOW,Definit,Noun,Plural,NoSARF,NullSuffix]
في	**فِي**	[SOW,NullPrefix,Prepos,NullSuffix]
نيوزيلندا	**نيوزيلندا**	[SOW,Translit]

The resulting Arabic POS tagging of a real-life phrase using ArabTagger©.