Arabic Natural Language Processing

Nizar Habash

Columbia University

Center for Computational Learning Systems

With modifications

Road Map

- Introduction
- Orthography
  - Arabic Script
  - MSA Phonology and Spelling
  - Encoding Issues
- Morphology
- Syntax

Introduction

- What is ‘Arabic’?
  - – A Semitic language
  - – Arabic Script
    - With or without diacritics
    - More ambiguity!
  - – *Arabic Language*
    - *Modern Standard Arabic (MSA)*
    - *Arabic Dialects*

Road Map

- Introduction
- Orthography
  - Arabic Script
  - MSA Phonology and Spelling
  - Encoding Issues
- Morphology
- Syntax

Arabic Script

Arabic script is an alphabet with allographic variants, optional zero-width diacritics and common ligatures.

الخَطُ العَرَبِي

Arabic script is used to write many languages: Arabic, Persian, Kurdish, Urdu, Pashto, etc.

PDF Files problem

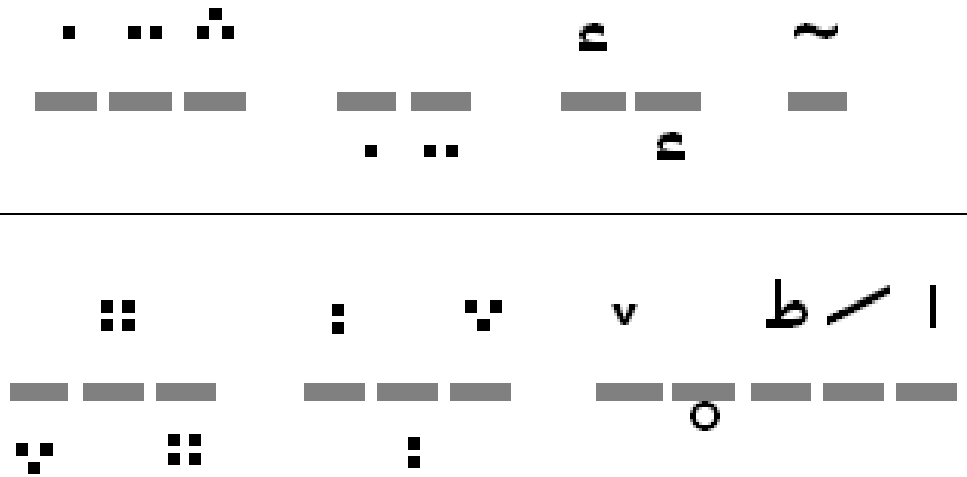
Arabic Script

Alphabet

- letter forms

ا ٮ ح د ر س ص ط ع
ڡ ک ل م ں ه و ى ء

- letter marks
  - Arabic only



  - Other languages
    - Persian, Kurdish, Ur
      Pashto, etc.

- *OCR output ambiguity*

Arabic Script

Alphabet (MSA)

- letters (form+mark)
  - Distinctive

ش	س	ث	ت	ب
/ʃ/	/s/	/θ/	/t/	/b/

- Non-distinctive

ا أ إ آ ئ ؤ ء

/ʔ/

glottal stop aka hamza

Arabic Script

Letter Shapes

- No distinction between print and handwriting
- No capitalization
- Right-to-left
- Ambiguous shapes
- Connective letters
- Disconnective letters
- OCR problems

ز	د	ا	ن	ب	ك	م	ش	غ	**Stand alone**
			ﻧ	ﺑ	ﻛ	ﻣ	ﺷ	ﻏ	**initial**
ﺰ	ﺪ	ﺎ	ﻨ	ﺒ	ﻜ	ﻤ	ﺸ	ﻐ	**medial**
			ﻦ	ﺐ	ﻚ	ﻢ	ﺶ	ﻎ	**final**

Arabic Script

Letter shaping

ك ت ب ← كتب = كتب

/katab/ b t k

to write

ك ت ا ب ← كتاب = كتاب

/kitāb/ b ā t k

book

Arabic Script

Diacritics

- Zero-width characters
- Used for short vowels

  كَتَب /katab/ *to write*

- Nunation is used for nominal indefinite marker in MSA

  كِتَابٌ /kitābun/ *a book*

Nunation
بً /ban/
بٌ /bun/
بٍ /bin/

Vowel
بَ /ba/
بُ /bu/
بِ /bi/

Arabic Script

Diacritics

- No-vowel marker (*sukun*)

  مَكْتَب /maktab/ *office*

- Double consonant marker (*shadda*)

  كَتَّب /kattab/ *to dictate*

- Combinable

بُّ	بٍّ	بًّ
/bbu/	/bbin/	/bban/

No Vowel
بْ /b/

Double Consonant
بّ /bb/

Arabic Script

Putting it together

Simple combination

Arab /ʕarab/ عرب = عَرَب ← عَ رَ ب

West /ʁarb/ غرب = غَرْب ← غَ رْ ب

Ligatures

Peace /salām/ سلام سلام ✗ ← س ل ا م

Arabic Script

Tatweel

- ‘elongation’
- aka kashida
- used for text highlight and justification

حقوق الانسان

حقـوق الانسـان

حقــوق الانســان

حقـــوق الانســـان

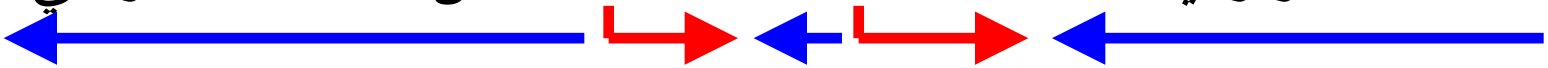
human rights /ħuqūq alʔinsān/

Arabic Script

“Arabic” Numerals

- Decimal system
- Numbers written left-to-right in right-to-left text : dual directions

استقلت الجزائر في سنة 1962 بعد 132 عاما من الاحتلال الفرنسي.



Algeria achieved its independence in 1962 after 132 years of French occupation.

- Three systems of enumeration symbols that vary by region

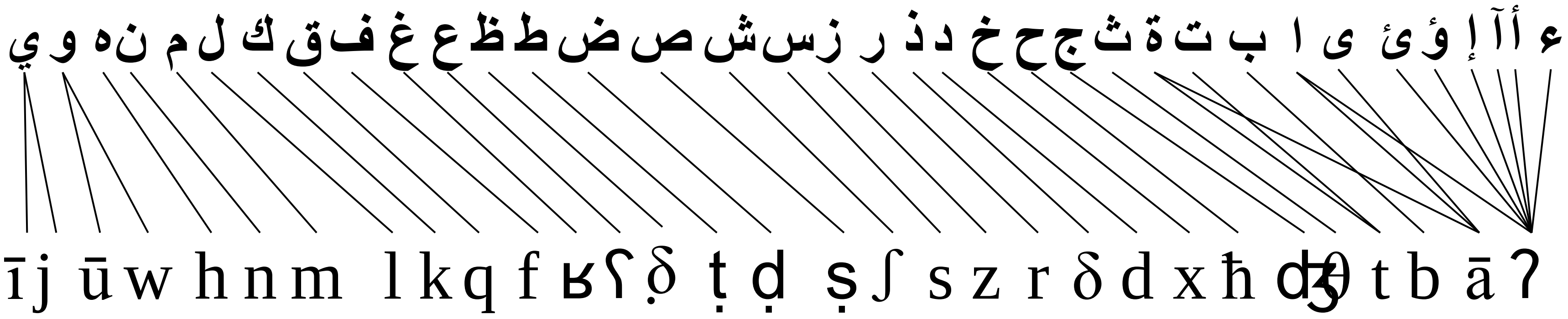
Western Arabic *Tunisia, Morocco, etc.*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Indo-Arabic *Middle East*	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Eastern Indo-Arabic *Iran, Pakistan, etc.*	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹

Road Map

- Introduction
- Orthography
  - Arabic Script
  - MSA Phonology and Spelling
  - Encoding Issues
- Morphology
- Syntax

MSA Phonology and Spelling

- Phonological profile of Standard Arabic
  - – 28 Consonants
  - – 3 short vowels, 3 long vowels, 2 diphthongs
- Arabic spelling is mostly phonemic …
  - – Letter-sound correspondence


ء أ آ إ ؤ ئ ى ا ب ت ة ث ج ح خ د ذ ر ز س ش ص ض ط ظ ع غ ف ق ك ل م ن ه و ي
ī j ū w h n m l k q f ʁ ʕ δ̣ ṭ ḍ ṣ ʃ s z r δ d x ħ ʤ θ t b ā ʔ

MSA Phonology and Spelling

- Arabic spelling is mostly phonemic …

Except for

- Medial short vowels can only appear as diacritics
- Diacritics are optional in most written text
  - Except in holy scripture
  - Occasionally appear in newspapers to mark less common readings, resolve certain ambiguities
    - كتب /katab/ to write كُتب /kutib/ to be written
- Dual use of ا, و, ي as consonant and long vowel
  - ا (/‘/,/ā/) و (/w/,/ū/) ي (/j/,/ī/)

MSA Phonology and Spelling

- Arabic spelling is mostly phonemic …

Except for (continued)

- Morphophonemic characters
  - Feminine marker ة (*ta marbuta*)
    - كبير /kabīr/ (big ♂) كبيرة /kabīra/ (big ♀)
  - Derivation marker
    - /ʕaṣa/ (to disobey عصى) (a stick عصا)
- Hamza variants (6 characters for one phoneme!)
  - (ء أآإؤئ) بهاءه بهاؤه بهائه /baha’/ + 3MascSing (his glory)

MSA Phonology and Spelling

- Arabic spelling can be ambiguous
- But how ambiguous? Really?
- Classic example

  ths s wht n rbc txt lks lk wth n vwls

  this is what an Arabic text looks like with no vowels
- Not exactly true
  - Long vowels are always written
  - Initial vowels are represented by an ا ‘alef’
  - Some final short vowels are represented

  ths is wht an Arbc txt lks lik wth no vwls

Will revisit ambiguity in more detail again under morphology discussion

Humans can read, machines may be troubled!

Road Map

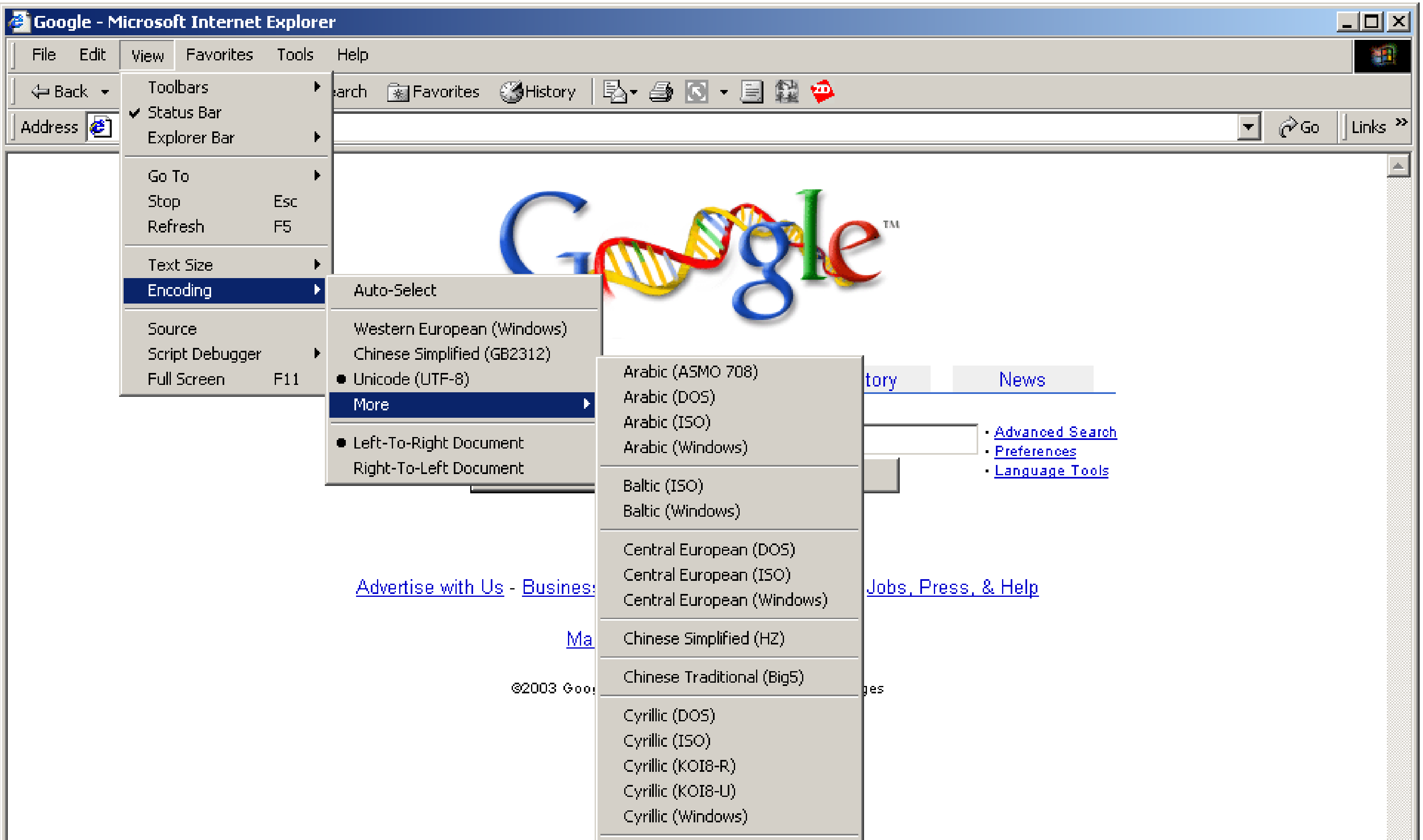
- Introduction
- Orthography
  - Arabic Script
  - MSA Phonology and Spelling
  - Encoding Issues
- Morphology
- Syntax

Encoding Issues

- Encoding Arabic
  - Data entry, storage, and display
  - Ease of use for *Arabic-*
  - Multi-script support
  - Multilingual support (extended Arabic characters)
- Types of Encoding
  - Machine character sets
    - Graphemic (shape insensitive, logical order)
    - Allographic (shape/direction sensitive) [obsolete]
  - Human accessible
    - Transliteration
    - Phonetic spelling (IPA)
    - Romanization

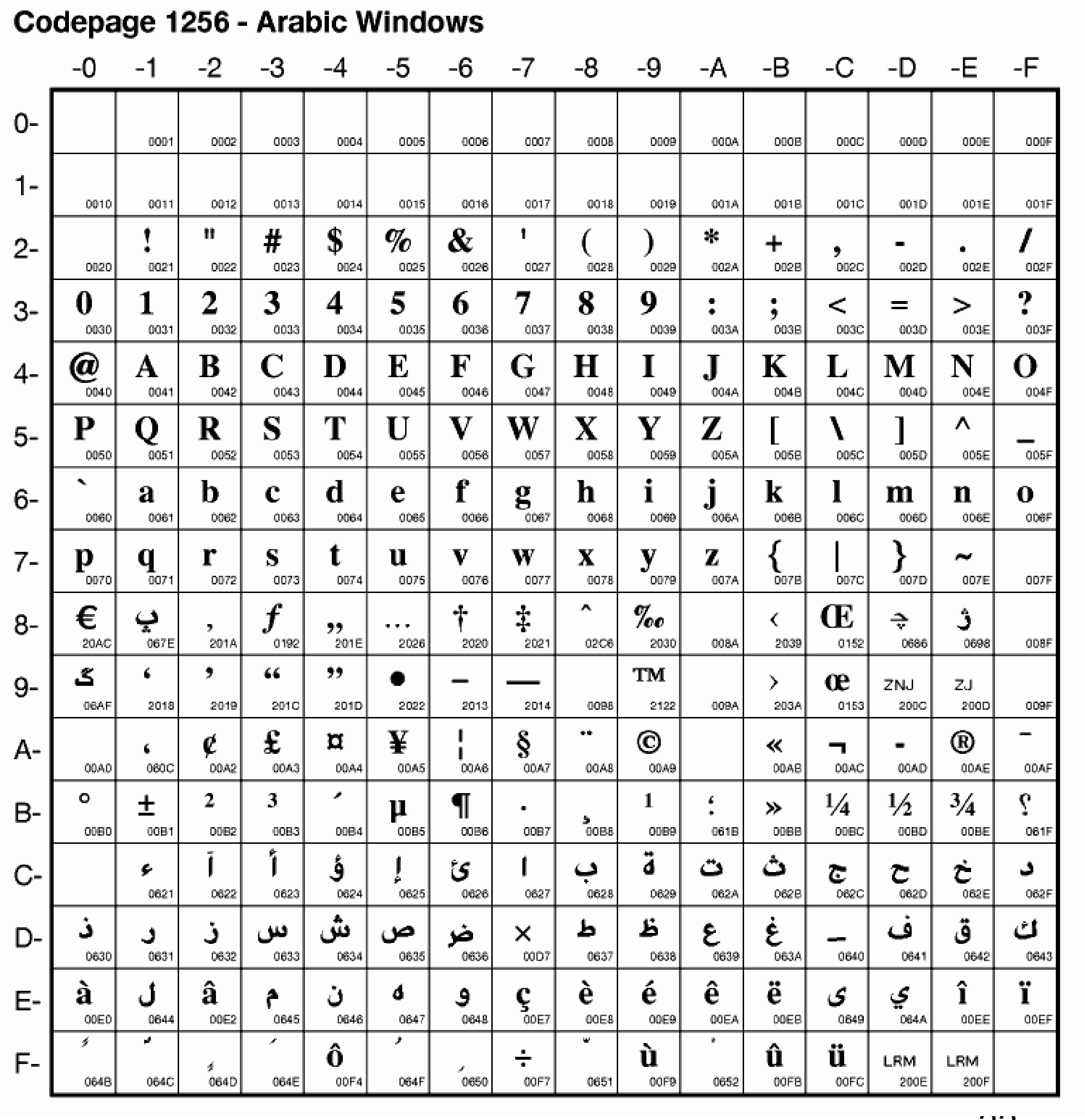
Encoding Issues

- Many Conflicting Character Sets for Arabic



Encodings

- CP-1256
  - Commonly used
  - 1-byte characters
  - Widely supported input/display
  - Minimal support for extended Arabic characters
  - bi-script support (Roman/Arabic)
  - Tri-lingual support: Arabic, French, English (ala ANSI)

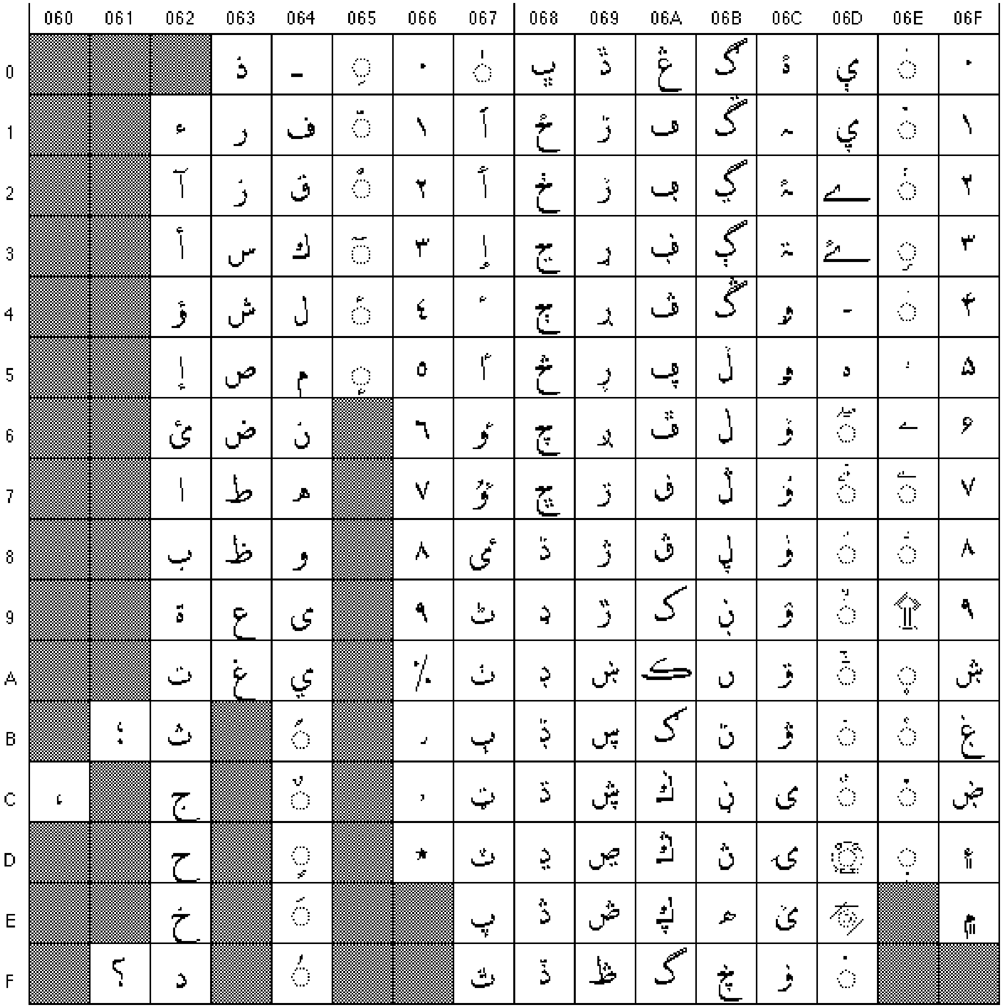


Codepage 1256 - Arabic Windows

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-		0001	0002	0003	0004	0005	0006	0007	0008	0009	000A	000B	000C	000D	000E	000F
1-	0010	0011	0012	0013	0014	0015	0016	0017	0018	0019	001A	001B	001C	001D	001E	001F
2-	0020	! 0021	" 0022	# 0023	$ 0024	% 0025	& 0026	' 0027	(0028	) 0029	* 002A	+ 002B	, 002C	- 002D	. 002E	/ 002F
3-	0 0030	1 0031	2 0032	3 0033	4 0034	5 0035	6 0036	7 0037	8 0038	9 0039	: 003A	; 003B	< 003C	= 003D	> 003E	? 003F
4-	@ 0040	A 0041	B 0042	C 0043	D 0044	E 0045	F 0046	G 0047	H 0048	I 0049	J 004A	K 004B	L 004C	M 004D	N 004E	O 004F
5-	P 0050	Q 0051	R 0052	S 0053	T 0054	U 0055	V 0056	W 0057	X 0058	Y 0059	Z 005A	[005B	\ 005C	] 005D	^ 005E	_ 005F
6-	` 0060	a 0061	b 0062	c 0063	d 0064	e 0065	f 0066	g 0067	h 0068	i 0069	j 006A	k 006B	l 006C	m 006D	n 006E	o 006F
7-	p 0070	q 0071	r 0072	s 0073	t 0074	u 0075	v 0076	w 0077	x 0078	y 0079	z 007A	{ 007B	| 007C	} 007D	~ 007E	007F
8-	€ 20AC	پ 067E	‚ 201A	ƒ 0192	„ 201E	… 2026	† 2020	‡ 2021	ˆ 02C6	‰ 2030	008A	‹ 2039	Œ 0152	چ 0686	ژ 0698	008F
9-	گ 06AF	‘ 2018	’ 2019	“ 201C	” 201D	• 2022	– 2013	— 2014	0098	™ 2122	009A	› 203A	œ 0153	ZNJ 200C	ZJ 200D	009F
A-	00A0	، 060C	¢ 00A2	£ 00A3	¤ 00A4	¥ 00A5	¦ 00A6	§ 00A7	¨ 00A8	© 00A9		« 00AB	¬ 00AC	- 00AD	® 00AE	¯ 00AF
B-	° 00B0	± 00B1	² 00B2	³ 00B3	´ 00B4	µ 00B5	¶ 00B6	· 00B7	¸ 00B8	¹ 00B9	؛ 061B	» 00BB	¼ 00BC	½ 00BD	¾ 00BE	؟ 061F
C-		ء 0621	آ 0622	أ 0623	ؤ 0624	إ 0625	ئ 0626	ا 0627	ب 0628	ة 0629	ت 062A	ث 062B	ج 062C	ح 062D	خ 062E	د 062F
D-	ذ 0630	ر 0631	ز 0632	س 0633	ش 0634	ص 0635	ض 0636	× 00D7	ط 0637	ظ 0638	ع 0639	غ 063A	ـ 0640	ف 0641	ق 0642	ك 0643
E-	à 00E0	ل 0644	â 00E2	م 0645	ن 0646	ه 0647	و 0648	ç 00E7	è 00E8	é 00E9	ê 00EA	ë 00EB	ى 0649	ي 064A	î 00EE	ï 00EF
F-	ً 064B	ٌ 064C	ٍ 064D	َ 064E	ô 00F4	ُ 064F	ِ 0650	÷ 00F7	ّ 0651	ù 00F9	ْ 0652	û 00FB	ü 00FC	LRM 200E	LRM 200F	

Encodings

- Unicode
  - Becoming the standard more and more
  - 2-byte characters
  - Widely supported input/display
  - Supports extended Arabic characters
  - Multi-script representation



	060	061	062	063	064	065	066	067	068	069	06A	06B	06C	06D	06E	06F
0				ذ	ـ	◌ِ	٠	◌ٰ	ڀ	ڐ	ڠ	ڰ	ۀ	ې	◌۠	۰
1			ء	ر	ف	◌ّ	١	ٱ	ځ	ڑ	ڡ	ڱ	ہ	ۑ	◌ۡ	۱
2			آ	ز	ق	◌ْ	٢	ٲ	ڂ	ڒ	ڢ	ڲ	ۂ	ے	◌ۢ	۲
3			أ	س	ك	◌ٓ	٣	ٳ	ڃ	ړ	ڣ	ڳ	ۃ	ۓ	◌ۣ	۳
4			ؤ	ش	ل	◌ٔ	٤	ٴ	ڄ	ڔ	ڤ	ڴ	ۄ	۔	◌ۤ	۴
5			إ	ص	م	◌ٕ	٥	ٵ	څ	ڕ	ڥ	ڵ	ۅ	ە	ۥ	۵
6			ئ	ض	ن		٦	ٶ	چ	ږ	ڦ	ڶ	ۆ	◌ۖ	ۦ	۶
7			ا	ط	ه		٧	ٷ	ڇ	ڗ	ڧ	ڷ	ۇ	◌ۗ	◌ۧ	۷
8			ب	ظ	و		٨	ٸ	ڈ	ژ	ڨ	ڸ	ۈ	◌ۘ	◌ۨ	۸
9			ة	ع	ى		٩	ٹ	ډ	ڙ	ک	ڹ	ۉ	◌ۙ	۩	۹
A			ت	غ	ي		٪	ٺ	ڊ	ښ	ڪ	ں	ۊ	◌ۚ	◌۪	ۺ
B		؛	ث		◌ً		٫	ٻ	ڋ	ڛ	ګ	ڻ	ۋ	◌ۛ	◌۫	ۻ
C	،		ج		◌ٌ		٬	ټ	ڌ	ڜ	ڬ	ڼ	ی	◌ۜ	◌۬	ۼ
D			ح		◌ٍ		٭	ٽ	ڍ	ڝ	ڭ	ڽ	ۍ	۝	◌ۭ	۽
E			خ		◌َ			پ	ڎ	ڞ	ڮ	ھ	ێ	۞		۾
F		؟	د		◌ُ			ٿ	ڏ	ڟ	گ	ڿ	ۏ	◌۟		

Encoding Issues
Arabic Display

- Memory (logical order) →

ÔÇÑßÊ ÝáÓØíä (Palestine) Ýí ÇæáãÈíÇÏ (Olympics) 2000 æ 2004.

شاركت فلسطين (Palestine) في اولمبياد (Olympics) 2000 و 2004.

or this way for those with direction-bias

←

.4002 æ 0002)scipmylO(ÏÇíÈãáæÇ íÝ)enitselaP(äíØÓáÝ ÊßÑÇÔ

.4002 و 0002)scipmylO(اولمبياد في)enitselaP(فلسطين شاركت

Encoding Issues
Arabic Display

- Memory (logical order)

ÔÇÑßÊ ÝáÓØíä (Palestine) Ýí ÇæáãÈíÇÏ (Olympics) 2000 æ 2004.

شاركت فلسطين (Palestine) في اولمبياد (Olympics) 2000 و 2004.

- Display (visual order)
  - Bidirectional (BiDi) support
    - Numbers and Roman script

شاركت فلسطين (Palestine) في اولمبياد (Olympics) 2000 و 2004.

  - Letter and ligature shaping

شاركت فلسطين (Palestine) في اولمبياد (Olympics) 2000 و 2004.

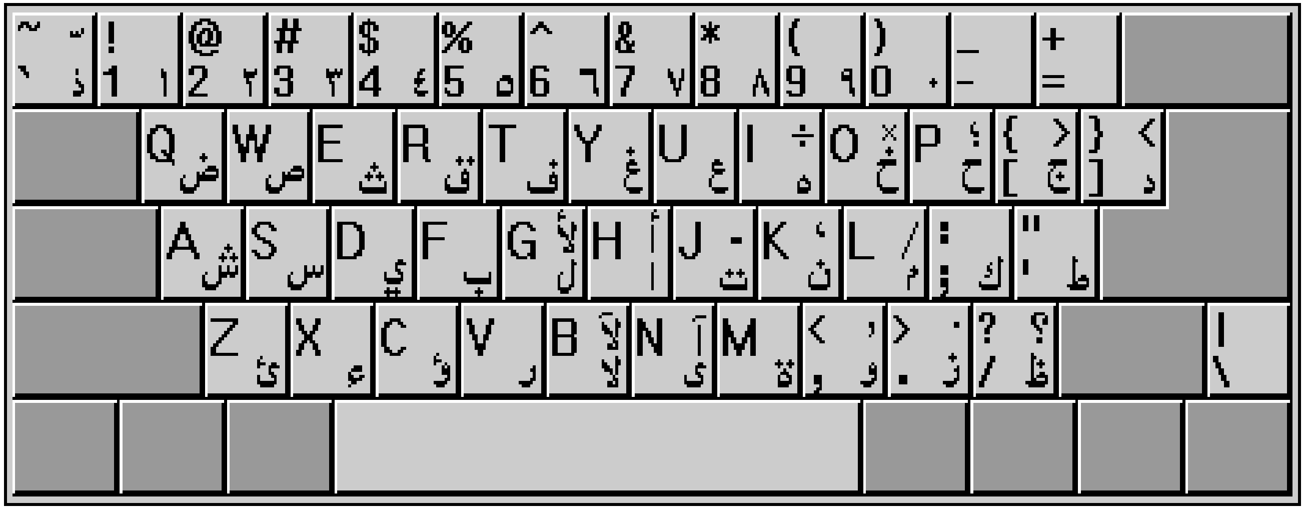
Display Problems

		Display Encoding			
		CP-1256	ISO-8859	Unicode	Western
Actual Encoding	CP-1256	تدشين منطقة حرة في دبي للتجارة الالكترونية	ٍ□ ة حرة□تدشِل كلظ ترنلِة□دبٍ ففتجارة افاف	Y□蹼祃ģgش ψ□ՠŎgāā□A餚	ÊÏÔíä ãäØÞÉ ÍÑÉ Ýí ÏÈí ááÊÌÇÑÉ ÇáÇáßÊÑæäíÉ
	ISO-8859	ة حرة â×و هوêتدش ننتجارة êدب êل ةêوèانانمتر	تدشين منطقة حرة في دبي للتجارة الالكترونية	Y□柒既 ģg□ψ□甁ŎgGG ㊑親g	ÊÏÔêæ åæ×âÉ ÍÑÉ áê ÏÈê ääÊÌÇÑÉ ÇäÇäãÊÑèæêÉ
	Unicode	ï»؟طھط¯ط´ظٹظ† ظ…ظ†ط·ظ‚ط© طط±ط© ظپظٹ ط¯ط¨ظٹ ظ„ظ„طھط¬ط§ط±ط© ط§ظ„ط§ظ„ظƒطھط±ظˆظ†ظٹط©	ظ؟؛ُ□ظ□ظ□عع□ ععظ□عظ□ ظ–ظ□ظ□ عع□ ظ□ظ□ع□ ععظ□ظ،ظ□ظ□ظ□ ظ□عظ□ععظ□ظ□عع□ عظ□	تدشين منطقة حرة في دبي للتجارة الالكترونية	ï»¿Øª Ø¯Ø´ÙŠÙ† Ù…Ù†Ø·Ù‚Ø© ØØ±Ø© ÙÙŠ Ø¯Ø¨ÙŠ Ù„Ù„Ø ªØ¬Ø§Ø±Ø© Ø§Ù„Ø§Ù„ÙƒØªØ±Ù ˆÙ†ÙŠØ©

- Wrong encoding
- Partial support problems

Encoding Issues
Arabic Input

- Standard graphemic keyboard
- Logical order input


~ ` ّ ذ ! 1 ١ @ 2 ٢ # 3 ٣ $ 4 ٤ % 5 ٥ ^ 6 ٦ & 7 ٧ * 8 ٨ (9 ٩) 0 ٠ _ - + =
Q ض W ص E ث R ق T ف Y غ U ع I ÷ ه O × خ P ؛ ح { < [ج } >] د
A ش S س D ي F ب G لأ ل H أ ا J ـ ت K ، ن L / م : ; ك " ' ط
Z ئ X ء C ؤ V ر B لآ لا N آ ى M ة < , ’ و > . ز ? / ؟ ظ | \

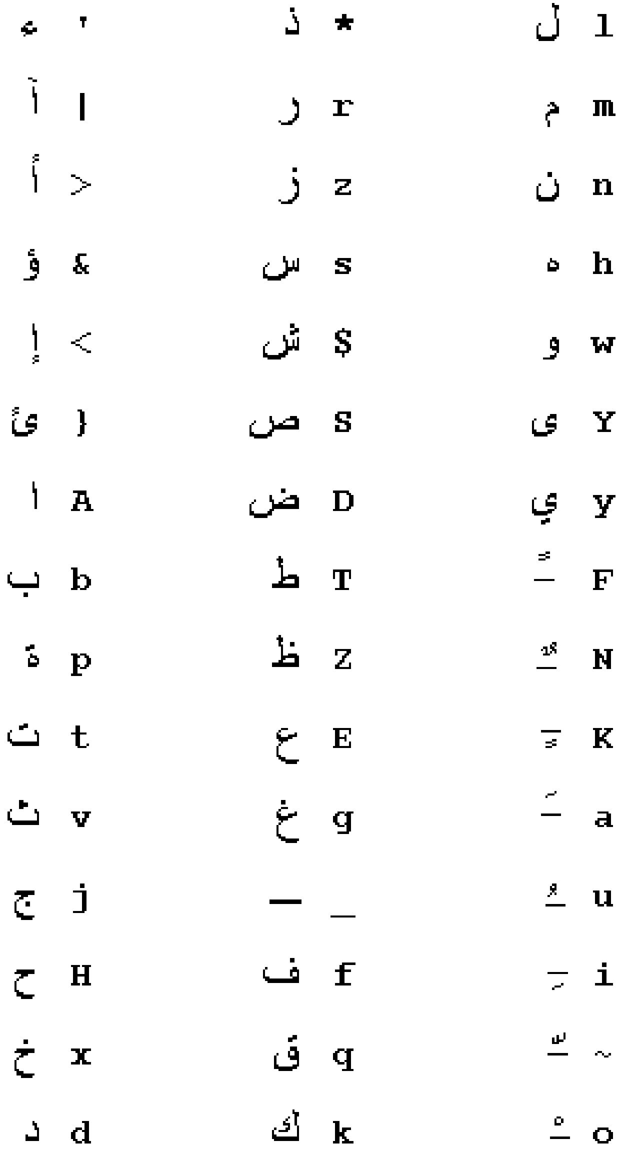
م ا ل س ⇨ سلام

http://www.cyrillic.com/kbd/btc.html

Encodings

Buckwalter Encoding

- Romanization
  - One-to-one mapping to Arabic script spelling
  - Left-to-right
  - Easy to learn/use
  - Human & machine compatible
- Penn Arabic Tree Bank
- Some characters can be modified to allow use with XML and regular expressions
- Roman input/display
- Monolingual encoding (can’t do English and Arabic)
- Minimal support for extended Arabic characters



ء	'	ذ	*	ل	l
آ	|	ر	r	م	m
أ	>	ز	z	ن	n
ؤ	&	س	s	ه	h
إ	<	ش	$	و	w
ئ	}	ص	S	ى	Y
ا	A	ض	D	ي	y
ب	b	ط	T	ـً	F
ة	p	ظ	Z	ـٌ	N
ت	t	ع	E	ـٍ	K
ث	v	غ	g	ـَ	a
ج	j	ـ	_	ـُ	u
ح	H	ف	f	ـِ	i
خ	x	ق	q	ـّ	~
د	d	ك	k	ـْ	o

Road Map

- Introduction
- Orthography
- Morphology
  - Derivational Morphology
  - Inflectional Morphology
  - Morphological Ambiguity
  - Arabic Computational Morphology
- Syntax

Morphology

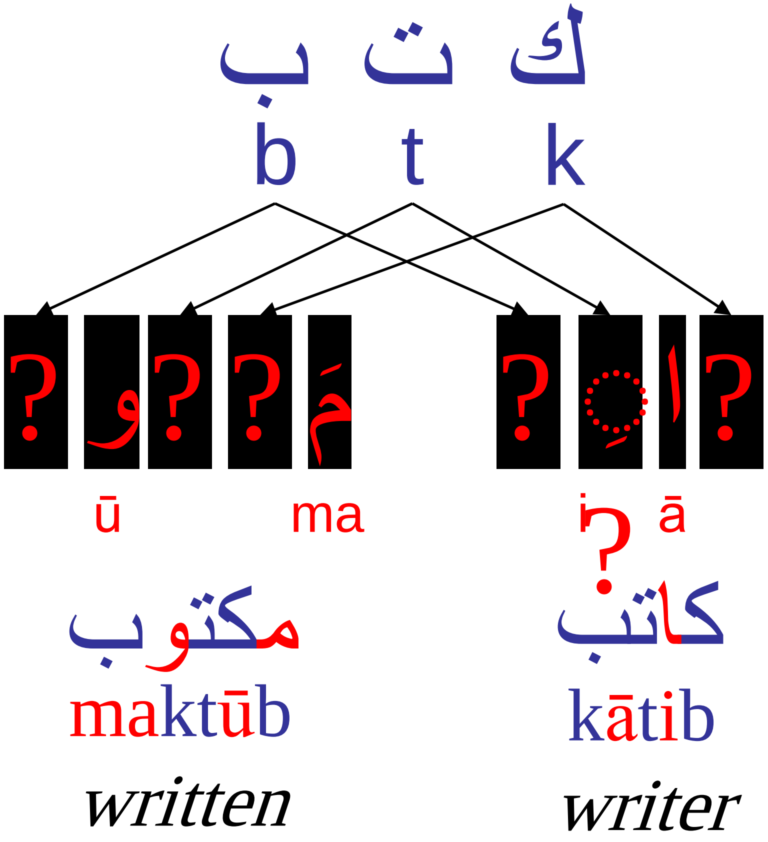
- Type
  - Concatenative: prefix, suffix, infix
  - Templatic: root+pattern
- Function
  - Derivational
    - Creating new words
    - *Mostly templatic*
  - Inflectional
    - Modifying features of words
      - Tense, number, person, mood, aspect
    - Mostly concatenative

Road Map

- Introduction
- Orthography
- Morphology
  - Derivational Morphology
  - Inflectional Morphology
  - Morphological Ambiguity
  - Arabic Computational Morphology
- Syntax

Derivational Morphology

- Templatic Morphology
  - Root
  - Pattern
  - Lexeme

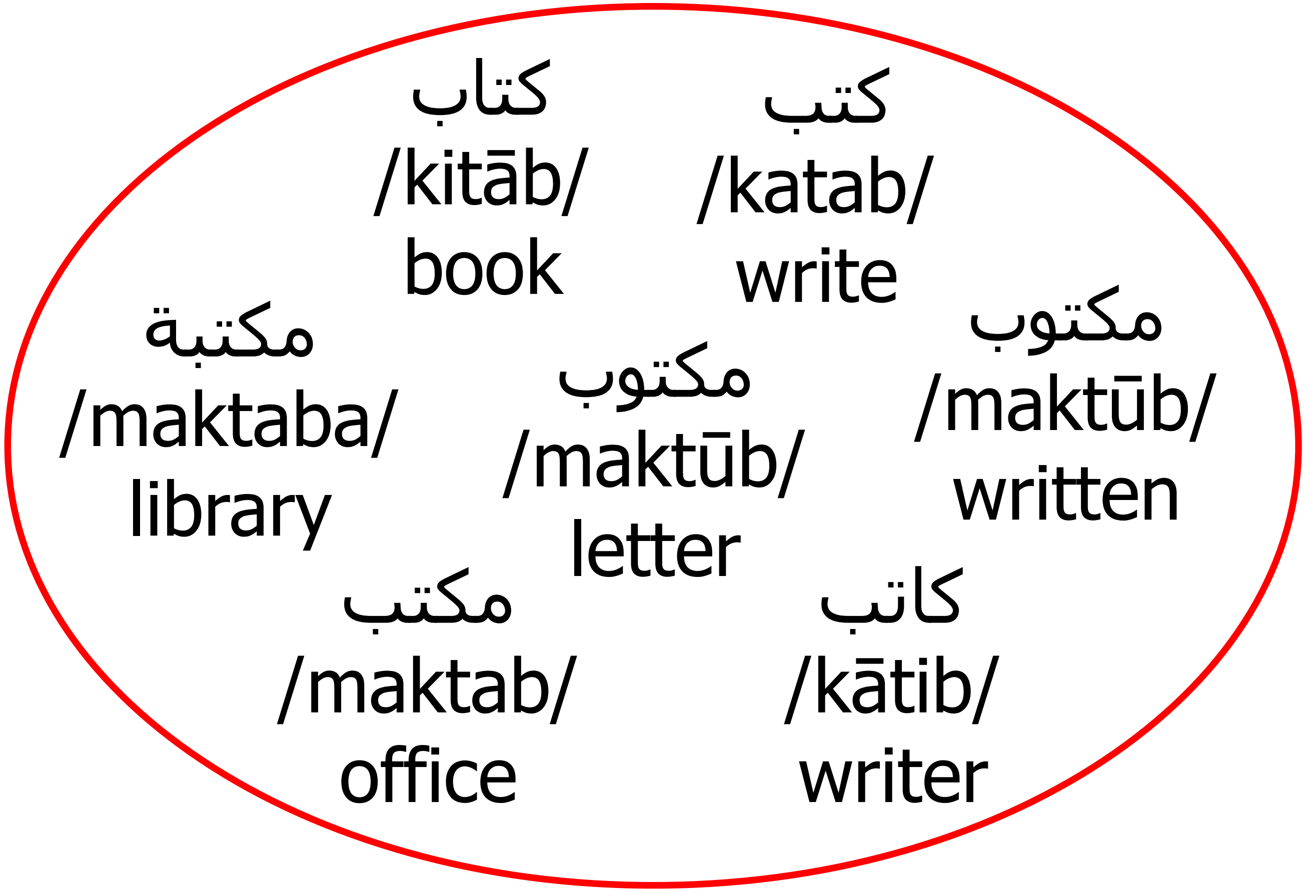


*Lexeme.Meaning = (Root.Meaning+Pattern.Meaning)*Idiosyncrasy.Random*

Derivational Morphology

Root Meaning

- ك ت ب KTB = notion of *“writing”*



كتاب
/kitāb/
book

كتب
/katab/
write

مكتبة
/maktaba/
library

مكتوب
/maktūb/
letter

مكتوب
/maktūb/
written

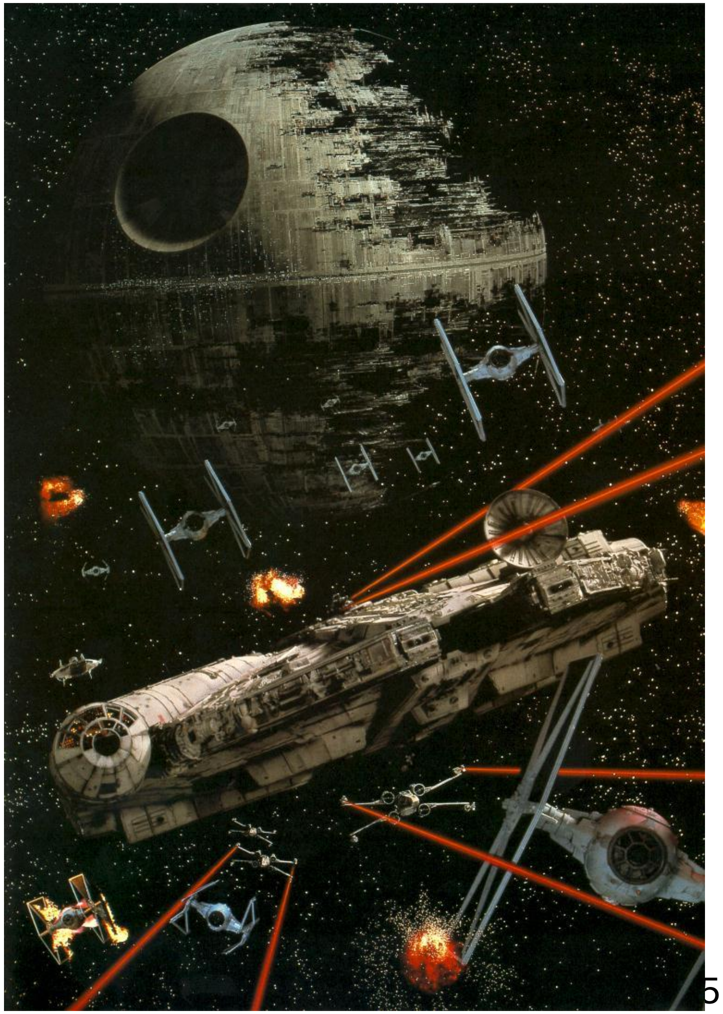
مكتب
/maktab/
office

كاتب
/kātib/
writer

Derivational Morphology

Root Meaning

- LHM-2
- Notion of “battle”
  - – ملحمة /malħama/
    - Fierce battle
    - Massacre
    - Epic



Road Map

- Introduction
- Orthography
- Morphology
  - Derivational Morphology
  - Inflectional Morphology
  - Morphological Ambiguity
  - Arabic Computational Morphology
- Syntax

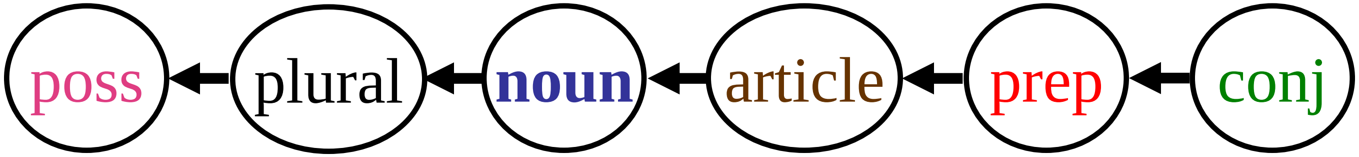
Inflectional Morphology

- Derivational Morphology
  - Lexeme ≈ Root + Pattern
- Inflectional Morphology
  - Word = Lexeme + Features
- Features
  - Part-of-speech
    - *Traditional*: Noun, Verb, Particle
    - *Computational*: N, PN, V, Adj, Adv, P, Pron, Num, Conj, Det, Aux, Pun, IJ, and others
  - Noun-specific
    - Number: singular, dual, plural, collective
    - Gender: masculine, feminine, Neutral
    - Definiteness: definite, indefinite
    - Case: nominative, accusative, genitive
    - Possessive clitic

Inflectional Morphology

- Features (continued)
  - Verb-specific
    - Aspect: perfective, imperfective, imperative
    - Voice: active, passive
    - Tense: past, present, future
    - Mood: indicative, subjunctive, jussive
    - Subject (Person, Number, Gender)
    - Object clitic
  - Others
    - Single-letter conjunctions
    - Single-letter prepositions

Inflectional Morphology
Nouns



وكبيوتنا

/wakabiyūtinā/

و + ك + بيوت + نا

wa+ka+biyūt+nā

and+like+houses+our

And like our houses

وللمكتبات

/walilmaktabāt/

و+ل+ال+مكتبة+ات

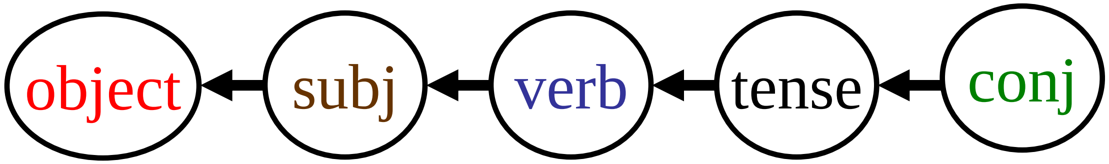
wa+li+al+maktaba+āt

and+for+the+library+plural

And for the libraries

- Morphotactics (e.g. ل+ال → لل)
- Arabic *Broken Plurals* (templatic)

Inflectional Morphology
Verbs



object ← subj ← verb ← tense ← conj

فقلناها

/faqulnāhā/

ف+ قال+ نا+ ها

fa+qul+na+hā

so+said+we+it

So we said it.

وسنقولها

/wasanaqūluhā/

و+ س+ ن+ قول + ها

wa+sa+na+qūl+u+hā

and+will+we+say+it

And we will say it

- Morphotactics
- Subject conjugation (suffix or circumfix)

Inflectional Morphology

- Perfect verb subject conjugation (*suffixes only*)

	Singular	Dual	Plural
1	كتبتُ katabtu	كتبنا katabnā	
2	كتبتَ katabta	كتبتما katabtumā	كتبتم katabtum
3	كتبَ kataba	كتبا katabā	كتبوا katabtū

- Imperfect verb subject conjugation (*prefix+suffix*)

	Singular	Dual	Plural
1	اكتبُ aktubu	نكتبُ naktubu	
2	تكتبُ taktubu	تكتبان taktubān	تكتبون taktubūn
3	يكتبُ yaktubu	يكتبان yaktubān	يتكتبون yaktubūn

Feminine form and other verb moods not shown

Road Map

- Introduction
- Orthography
- Morphology
  - Derivational Morphology
  - Inflectional Morphology
  - Morphological Ambiguity
  - Arabic Computational Morphology
- Syntax

Morphological Ambiguity

- Derivational ambiguity
  - قاعدة: basis/principle/rule, military base, Qa'ida/Qaeda/Qaida
- Inflectional ambiguity
  - تكتب: you write, she writes
  - Segmentation ambiguity
    - وجد: he found; و+جد: and+grandfather
    - للغة: ل+لغة: for a language; ل+اللغة: for the language
- Spelling ambiguity
  - Optional diacritics
    - كاتب: /kātib/ writer , /kātab/ to correspond
  - Suboptimal spelling
    - Hamza dropping: أ, إ → ا
    - Undotted ta-marbuta: ة → ه
    - Undotted final ya: ي → ى

Morphological Ambiguity

- Multiple sources of ambiguity

بين

/bayyana/	Verb	*he declared/demonstrated*
/bayyanna/	Verb	*they [feminine] declared/demonstrated*
/bayyin/	Adj	*clear/evident/explicit*
/bayna/	Prep	*between/among*
/biyin/	Proper Noun	*in Yen*
/biyn/	Proper Noun	*Ben*

- Hard to measure specific causes of ambiguity
  - Derivational ambiguity* (diacritized tokens)
    - 1.09 entries/token
    - 1.01 entries/token (within same part-of-speech)
  - Spelling ambiguity* (undiacritized tokens)
    - 1.28 entries/token
    - 1.08 entries/token (within same part-of-speech)

** in Buckwalter’s Lexicon (~40,000 lexemes)*

Road Map

- Introduction
- Orthography
- Morphology
  - Derivational Morphology
  - Inflectional Morphology
  - Morphological Ambiguity
  - Arabic Computational Morphology
- Syntax
- Machine Translation Issues
- Syntax

Arabic Computational Morphology

- Representation units
  - Natural token وللـمـكـتـبـــات
    - White space separated strings (as is)
    - Can include extra characters (e.g. tatweel/kashida)
  - Word وللمكتبات
  - Segmented word
    - Can include any degree of morphological analysis
    - Pure segmentation: و ل لمكتبات
    - Arabic Treebank tokens (with recovery of some deleted/modified letters): و ل المكتبات

Arabic Computational Morphology

- Representation units (continued)
  - Prefix + Stem + Suffix
    - ولل+مكتب+ات–
    - Can create more ambiguity
  - Lexeme + Features
    - مكتبة[+Plural +Def +و+ ل]
  - Root + Pattern + Features
    - كتب + ةa3a21aم + [+Plural +Def +ل +و]
    - Very abstract
  - Root + Pattern + Vocalism + Features
    - كتب + ة321م + a.a.a + [+Plural +Def +ل +و]
    - Very very abstract

Arabic Computational Morphology

- Approaches
  - Finite state machines (Beesely,2001) (Kiraz,2001) (Habash et al, 2005b)
  - Concatenative analysis/generation (Buckwlater,2002) (Cavalli-Sforza et al, 2000)
  - Lexeme+Feature analysis/generation (Habash, 2004)
  - Shallow stemming (Darwish,2002) (Aljlayl and Frieder 2002)
  - Machine learning (Diab et al,2004) (Lee et al,2003) (Rogati et al, 2003)
- Issues
  - Appropriateness of system representation for an application
    - Machine Translation vs. Information Retrieval
    - Arabic spelling vs. phonetic spelling
  - System coverage
  - System extendibility
  - Availability to researchers
  - Use for analysis and generation

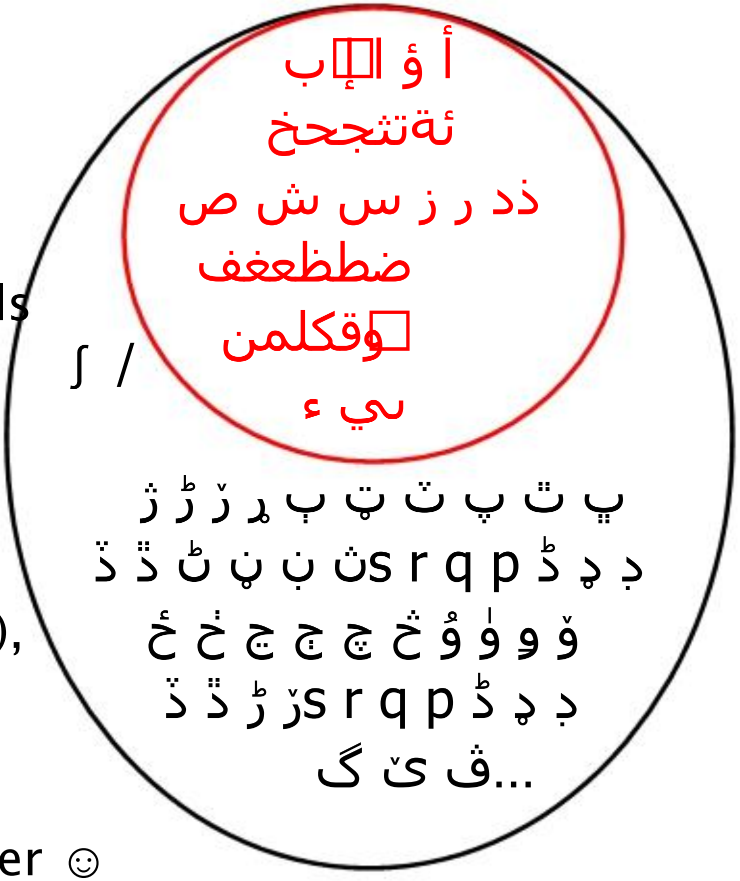
Road Map

- Introduction
- Orthography
  - Arabic Script
  - MSA Phonology and Spelling
  - Recognizing Arabic vs. Persian/Urdu/Pashto/Kurdish/Sindhi/…
  - Encoding Issues
- Morphology
- Syntax

Arabic Script
Other languages

Arabic

- No more than 3 dots
- Dots either above or below
- Marks are 1/2/3 dots, hamza (ء) or madda (~) only
- Rare borrowing for foreign words
- پ/p/, ڤ /v/, چ گ ڤ /g/, چ /t ʃ /
- regionally variable



Not Arabic

- Extra marks: haft (v), ring (o), taa (ط), four dots (::), vertical dots (:)
- Some Numerals (i,h,g)

Once you learn the alphabet, it is easier ☺

HOW TO DETECT ARABIC??

Morphology and Syntax

- Rich morphology crosses into syntax
  - Pro-drop / Subject conjugation
  - Verb sub-categorization and object clitics
    - $\text{Verb}_{\text{transitive}}$+subject+object
    - $\text{Verb}_{\text{intransitive}}$+subject but not $\text{Verb}_{\text{intransitive}}$+subject+object
    - $\text{Verb}_{\text{passive}}$+subject but not $\text{Verb}_{\text{passive}}$+subject+object
- Morphological interactions with syntax
  - Agreement
    - Full: e.g. Noun-Adjective on number, gender, and definiteness (for persons)
    - Partial: e.g. Verb-Subject on gender (in VSO order)
  - Definiteness
    - Noun compound formation, copular sentences, etc.
    - Nouns+DefiniteArticle, Proper Nouns, Pronouns, etc.

Morphology and Syntax

- Morphological interactions with syntax (continued)
  - Case
    - MSA is case marking: nominative, accusative, genitive
    - Almost-free word order
    - Case is often marked with optionally written short vowels
      - This effectively limits the word-order freedom in published text
- Agglutination
  - Attached prepositions create words that cross phrase boundaries

ت& Wl+ل	li+Almaktabāt
for the-libraries	[PP li [NP Almaktabāt]]

- Some morphological analysis (minimally segmentation) is necessary even for statistical approaches to parsing

Road Map

- Introduction
- Orthography
- Morphology
- Syntax
  - Morphology and Syntax
  - Sentence Structure
  - Phrase Structure
  - Computational Resources

Sentence Structure

Two types of Arabic Sentences

- Verbal sentences

  - [Verb Subject Object] (VSO)

  - كتب الأولاد الأشعار

    Wrote the-boys the-poems

    The boys wrote the poems

- Copular sentences (aka nominal sentences)

  - [Topic Complement]

  الأولاد شعراء

    the-boys poets

    The boys are poets

Sentence Structure

- Verbal sentences
  - Verb agreement with gender only
    - Default singular number
    - كتب الولد-الأولاد wrote$_{\text{3MascSing}}$ the-boy/the-boys

      كتبت البنت-البنات wrote$_{\text{3FemSing}}$ the-girl/the-girls
  - Pronominal subjects are conjugated
    - كتبت wrote-you$_{\text{MascSing}}$
    - كتبتم wrote-you$_{\text{MascPlur}}$
    - كتبوا ا wrote-they$_{\text{MascPlur}}$
  - Passive verbs
    - Same structure: Verb$_{\text{passive}}$ Subject$_{\text{underlyingObject}}$
    - Agreement with surface subject

Road Map

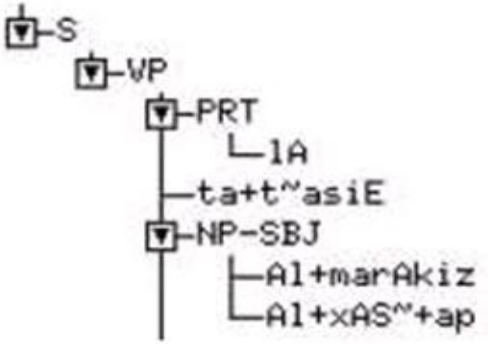
- Introduction
- Orthography
- Morphology
- Syntax
  - Morphology and Syntax
  - Sentence Structure
  - Phrase Structure
  - Computational Resources

Computational Resources

- Monolingual corpora for building language models
  - Arabic Gigaword
    - Agence France Presse
    - AlHayat News Agency
    - AnNahar News Agency
    - Xinhua News Agency
  - Arabic Newswire
  - United Nations Corpus (parallel with other UN languages)
  - Ummah Corpus (parallel with English)
- Distributors
  - Linguistic Data Consortium (LDC)
  - Evaluations and Language resources Distribution Agency (ELDA)

Computational Resources

- Penn Arabic Treebank (PATB)
  - Started in 2001
  - Goal is 1 Million words
  - Currently 650K words
    - Agence France Presse , AlHayat newspaper, AnNahar newspaper
- POS tags
  - Buckwalter analyzer
  - Arabic-tailored POS list
- PATB constituency representation
  - Some modifications of Penn English Treebank
    - (e.g. Verb-phrase internal subjects)



Efficiency Enhancement Tools for Arabic Search Engines: A Statistical Approach

by

Adnan Yahya

joint work with

Ali Salhi
Birzeit University, Palestine

Presented at AI Class, November 3, 2011

Talk Outline

1. Introduction and Motivation
2. Our Arabic NLP Tools and Methods
3. Utility of employed Structures.
4. Search of Arabic PDF Files.
5. Next Steps

Introduction

- The World-Wide Web is rapidly changing and expanding.
- Estimates of the size of the World-Wide Web vary from 15 to 30 billion pages.
- The share of the Arabic is around 1% (for 5% of population).
- Usually one is interested in web searches that return exactly the needed info.
  Returning too little : not sufficient
  Returning too much : not useful.

Introduction

- Sometimes you want to search for more by (expanding the query or correcting it and
- Sometimes you want to search for less by eliminating certain query words
- Issues like document structuring, user profiling and using NLP tools are part of the solution
- The problems are present in other languages but may be more complicated for Arabic

Arabic NLP Tools and Methods

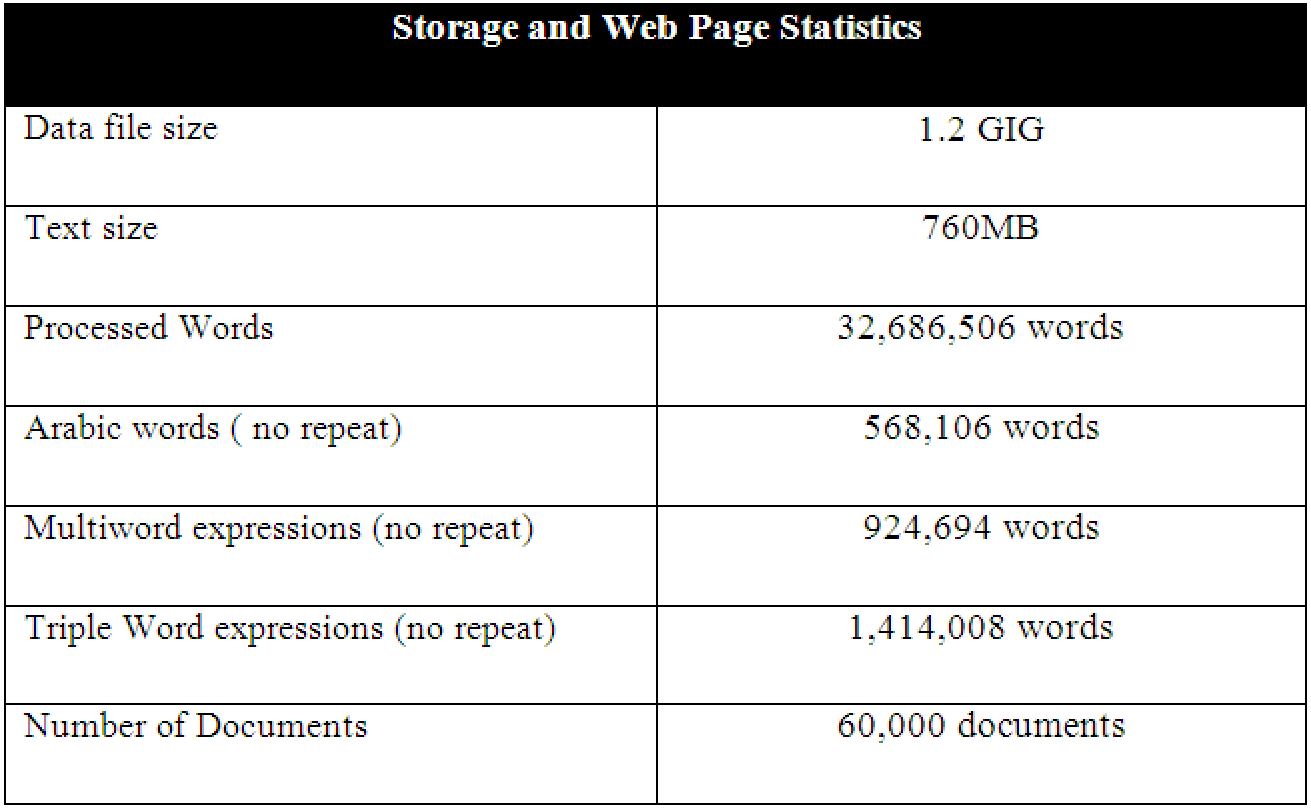
1. It was decided to look for practical solutions
2. Developing NLP tools and methods based on the availability of:

- Arabic Corpus,
- Arabic Word Database,
- Stop Word List

Arabic Corpus Construction

- For developing NLP tools we employed a statistical/Corpus based approach.
- We started with contemporary data we obtained initially from Al-Sharq Alawsat newspaper.
- Then we expended it with available sources (not without major problems)

Some Initial Statistics



Storage and Web Page Statistics	
Data file size	1.2 GIG
Text size	760MB
Processed Words	32,686,506 words
Arabic words (no repeat)	568,106 words
Multiword expressions (no repeat)	924,694 words
Triple Word expressions (no repeat)	1,414,008 words
Number of Documents	60,000 documents

Expansion Directions

- Topical Corpora: Topics sub-topics for Categorization.
- Multisource: books, journals,..
- Multiregional: dialects
- Translation tools usage
- Structures for efficient storage-manipulation.

Arabic Word Database (AWD)

- A database with all the Arabic words on the web with their frequencies
- The starting database was (568,106) *words*.
- Still expanding as we find more documents and better ways to deal with old ones
- We retain the substructures of data even after aggregating it: so we retain the ability to work with subsets of data as well
- Efforts at cleaning: infrequent words +

Construction of Stop Word List

- We created an Arabic stop words list consisting of
- the Arabic prepositions,
- pronouns, interrogatives, particles,
- English stop words list (translated).
- The list has (1065) words
- Much larger that English due to morphology
- One or multiple lists?: function dependant

Utility of the Built Structures

- Arabic Language Detection.
- Arabic Automatic Categorizer
- Arabic Query Live Suggestion
- Query modification by Word Elimination/Expansion
- Arabic Automatic Categorizer
- Person Names unification-translation

Arabic Language Detection

1. Determine whether the language of the document is Arabic or just a language using Arabic alphabet, say *Persian* or *Urdu* in order to crawl and index it.
2. Based on spell checking of words in the text against Arabic words and looking for a threshold that enables accurate but fast decisions.
3. A side tool: recover wrong language data entries.

Arabic Automatic Categorizer

- Web documents are assigned to one of predefined categories.
- Can be used to **refine** the search results. Search for query "السلام" under the category *religions* to avoid *political* results. (or الزراعة)
- We have been playing with the Granularity of categorization
- Have implications to the corpora properties
- Multi-stage categorization, user assisted categorization

Arabic Automatic Categorizer

Steps in Building the automatic categorizer :

1. Defined a set of topics: *Politics, Religion, Science, Economics and Commerce, Sports, Entertainment, Arts, Social Sciences, Engineering/technology*.
2. Word-based vector for each topic was prepared.
3. Calculate the similarities for each topic and select the most similar: How!

Arabic Query Live Suggestion

- When the user types in the search box, the system queries the suggestions tables to retrieve a list of possible suggestions.
- Large suggestion tables.
- Attempting the use of two and three words from the text corpus to reduce the suggestion space.

Word Elimination

- Eliminate non-discriminating words (stop words) in queries
- May speed search process.

Example: when a user search for (ما هي التعددية), the search engine has to make three runs to find match, while looking for the last word التعددية is enough to find relevant pages.

Query Expansion

- Looking for relevant words or derivatives.
- In Arabic many words can be derived from a single root.
- we ran the root extractor tool on the AWD words.
- For each root we now have almost all the Arabic words that can be derived from it.

Arabic Query Expansion

The Query expansion is then carried out as follows:

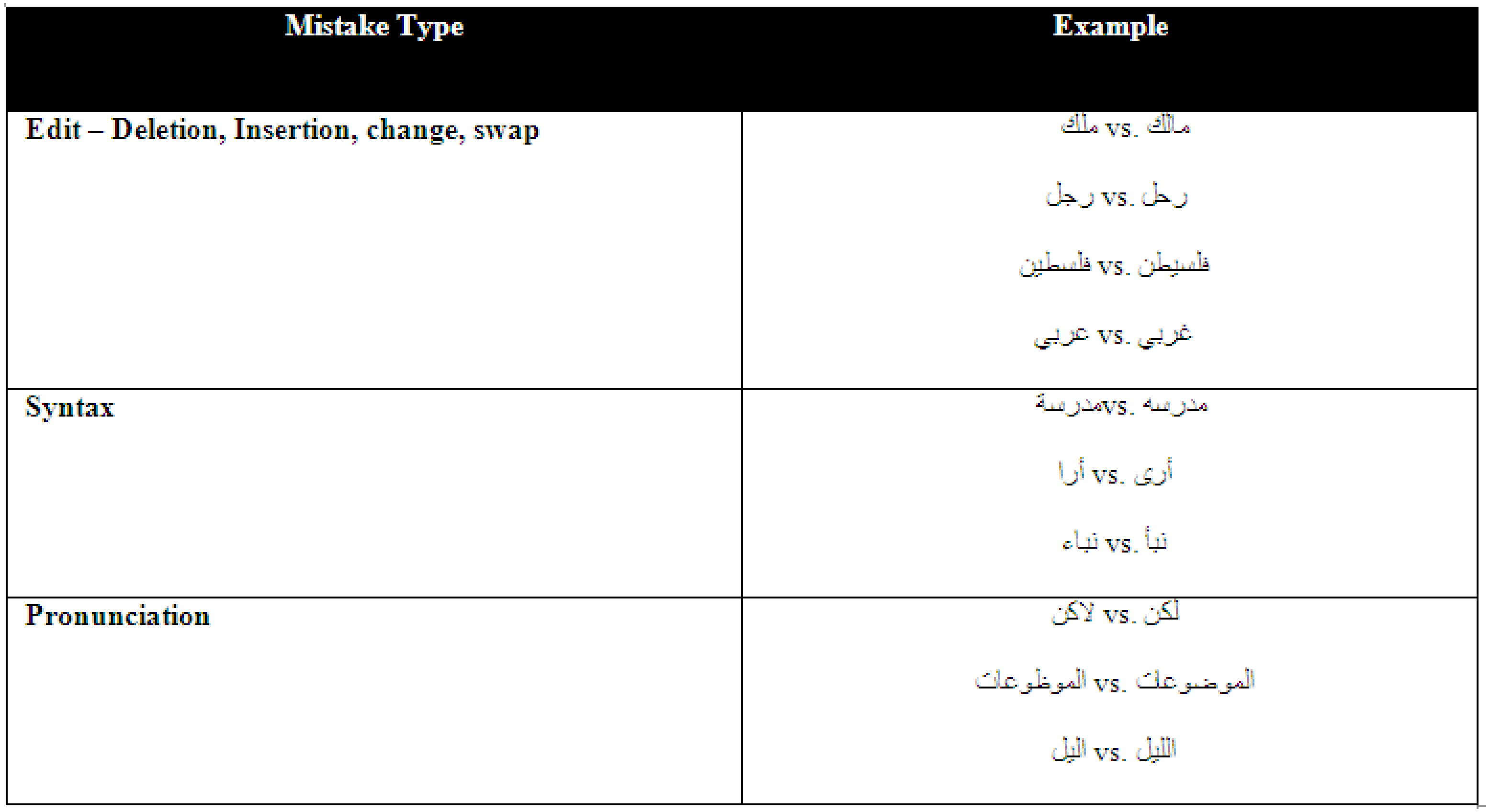
1. Each nonstop word is returned to its root, then a list of words that share the same root are retrieved from the AWD.
2. A weighing system is used to assign weight to the expanded queries say by frequencies

The Query Correction and Suggestion System

- The query correction main function is to correct user entered queries.
- For each query we examine the query correctness by comparing its words with words in the AWD. If there is a match, then the query is considered correct. Otherwise, it is misspelled.
- Why do we need such dictionary? Aren’t the normal dictionaries enough?

The Query Correction and Suggestion System

Common Arabic Spelling Errors



Mistake Type	Example
Edit – Deletion, Insertion, change, swap	ملك vs. مالك رجل vs. رحل فلسطين vs. فلسيطن عربي vs. غربي
Syntax	مدرسةvs. مدرسه أرا vs. أرى نباء vs. نبأ
Pronunciation	لاكن vs. لكن الموظوعات vs. الموضوعك اليل vs. الليل

Search Arabic PDF files

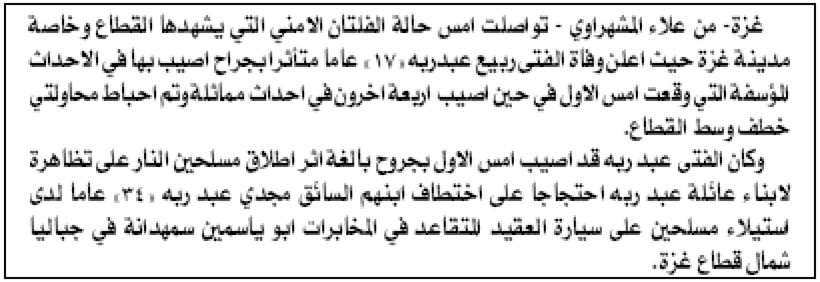
- One of the topics that concerns us through this work is Arabic PDF files because of prevalence;
- How we can index them so as to make them searchable by our search engine
- The concern is with *legacy* files

Search Arabic PDF files

- **cross platform font encoding problem in Arabic PDF.**

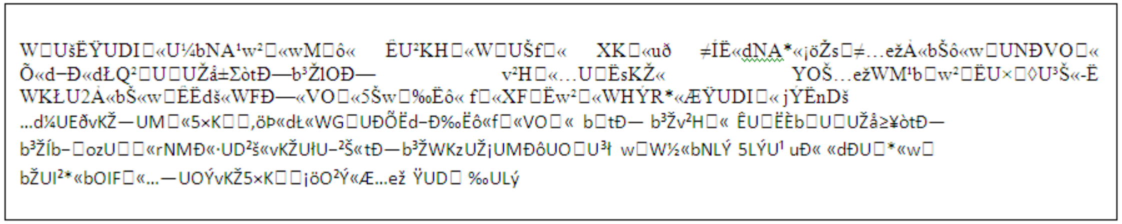
This problem comes from the fact that windows and MACs encode characters in different ways once we leave the ASCII characters. Thus, when we move Arabic characters from MAC to PC we can notice that the PC identifies them as unreadable: we call this *legacy* pdf files

Search Arabic PDF files



غزة- من علاء المشهراوي - تواصلت امس حالة الفلتان الامني التي يشهدها القطاع وخاصة
مدينة غزة حيث اعلن وفاة الفتى ربيع عبدربه ،١٧، عاما متأثرا بجراح اصيب بها في الاحداث
للؤسفة التي وقعت امس الاول في حين اصيب اربعة اخرون في احداث مماثلة وتم احباط محاولتي
خطف وسط القطاع.
وكان الفتى عبد ربه قد اصيب امس الاول بجروح بالغة اثر اطلاق مسلحين النار على تظاهرة
لابناء عائلة عبد ربه احتجاجا على اختطاف ابنهم السائق مجدي عبد ربه ،٣٤، عاما لدى
استيلاء مسلحين على سيارة العقيد للتقاعد في المخابرات ابو ياسمين سمهدانة في جباليا
شمال قطاع غزة.

Text As Created In InDesign Under MAC



```
W□UšËŸUDI□«U¼bNA¹w²□«wM□ô«    ÊU²KH□«W□UŠf□«    XK□«uð    ≠ÍË«dNA*«¡öŽs□≠…ežÀ«bŠô«w□UNÐVO□«
Õ«d−Ð«dŁQ²□U□UŽå±ΣòtÐ—b³Ž1OÐ—                    v²H□«…U□ËsKŽ«                    YOŠ…ežWM¹b□w²□ËU×□◊U³Š«-Ë
WKŁU2À«bŠ«w□ÊËdš«WFÐ—«VO□«5Šw□‰Ëô« f□«XF□Ëw²□«WHÝR*«ÆŸUDI□« jÝËnDš
…d¼UEðvKŽ—UM□«5×K□□,öÞ«dŁ«WG□UÐÕËd−Ð‰Ëô«f□«VO□«  b□tÐ— b³Žv²H□«  ÊU□ËÈb□U□UŽå≥¥òtÐ—
b³ŽÍb−□ozU□□«rNMÐ«·UD²š«vKŽUłU−²Š«tÐ—b³ŽWKzUŽ¡UMÐôUO□U³ł  w□W½«bNLÝ 5LÝU¹ uÐ« «dÐU□*«w□
bŽUI²*«bOIF□«…—UOÝvKŽ5×K□□¡öO²Ý«Æ…ež ŸUD□ ‰ULý
```

The Same Text Imported By Notepad under PC

Search of Arabic PDF files

Crawling the web for PDF files.

- Crawling the web for PDF files.
- Character mapping (partial for legacy files).
- Search using approximate match rather than exact.
- Needs extra work to increase the hit ratio.

Next Steps

- Continue to expand and clean the word collection: diverse topics
- Observe the behavior and contrast with other work
- Fine tune the lists and monitor scalability
- We have reported on work in progress

Thanks

Incremental Search

- *I*ncremental search enable users to find NEW data on the web.
- Not an Arabic feature but we thought it may be interesting
- User search for query A for the first time the search engine will provide the user with available results.
- Next time the user searches for A, the search engine will provide the new results introduced since his last search for A.

The Query Correction and Suggestion System

Correction is carried out in four stages:

1-Checking stage.

2-Processing Stage.

3-Filtering stage.

4-Weighting stage.

Stemming and Root Extraction

- Indexing based on the roots which are far more abstract than stems will improve the retrieval effectiveness over stems and words.

- Arabic words are formed by adding prefixes (letters and vowels at the start), infixes (vowels) and suffixes (letters and vowels at the end) to the root.

- The form of an Arabic word is usually determined by its **gender**, **number**, **grammatical case**, **whether it is definitive or not**, and finally **if there is a preposition attached to it.**

Stemming and Root Extraction

Stemming is carried out in the following steps.

- There are seven levels of processing (L3,L4, L5, L6, L7, L8, Ln) that the query may pass through the process of stemming.
- At each level of processing all the possible combinations of (prefix core suffix) is examined.
- The combinations where the core is a correct Arabic (corpus/dictionary) word, prefix and suffix are extracted.
- Level four processes infixes.

Stemming and Root Extraction

- Example: Stemming "وسيأخذونهما”
- LN: وسيأخذونهما,   و+سيأخذون+هما
- L7: سيأخذون,   سي+أخذ+ ون
- L3: أخذ,   root