JHU Summer School on Human Language Technology

Baltimore, June 27, 2006

Introduction to Arabic Natural Language Processing

Nizar Habash

Columbia University

Center for Computational Learning Systems

- Focus of this tutorial
  - Phenomena
  - Concepts
  - Approaches & Resources
- What is ‘Arabic’?
  - Arabic Script
  - Arabic Language
    - Modern Standard Arabic (MSA)
    - Arabic Dialects

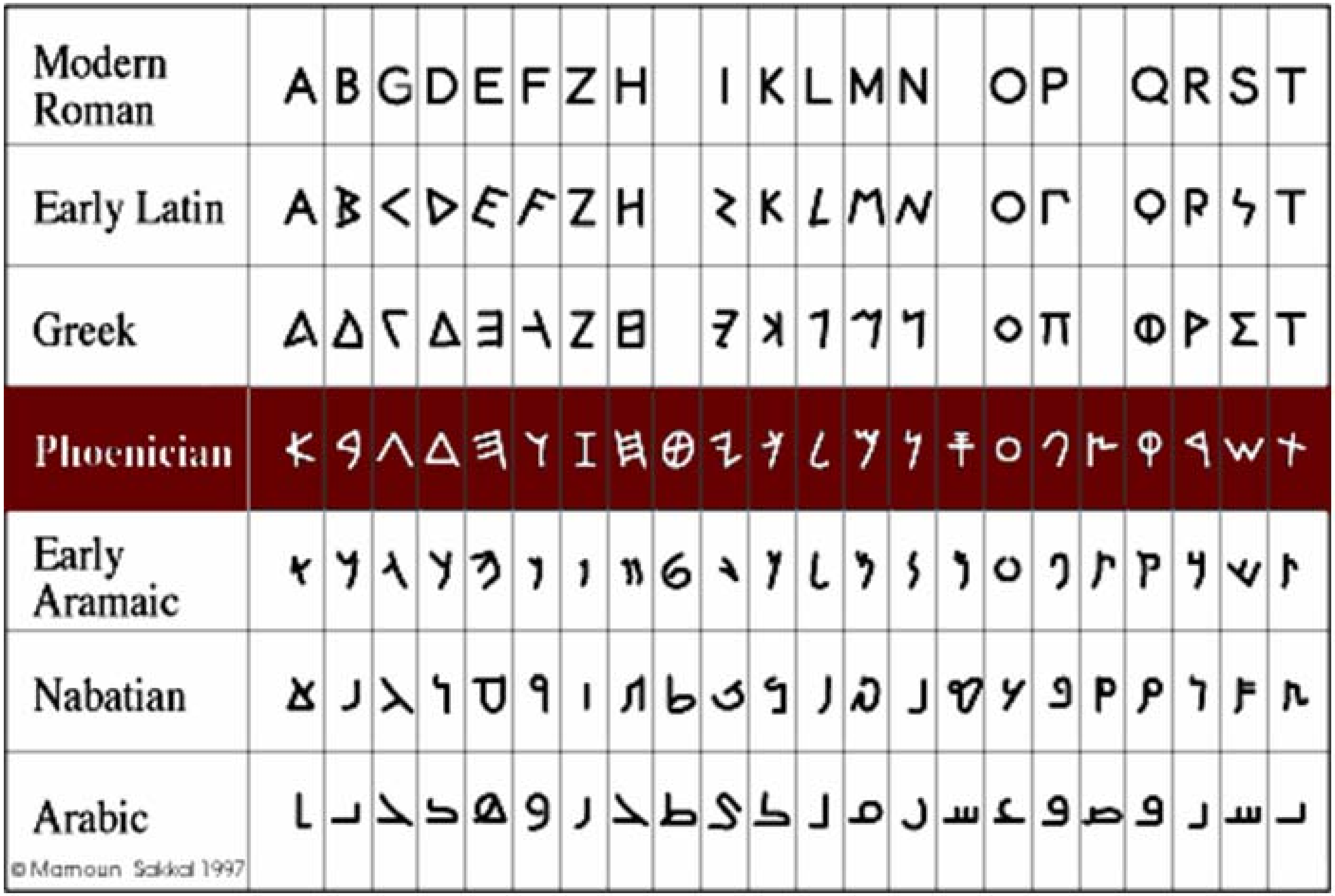
Road Map

- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
- Dialects

Road Map

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

Arabic Script


Modern Roman
A B G D E F Z H I K L M N O P Q R S T
Early Latin
Greek
Phoenician
Early Aramaic
Nabatian
Arabic
©Mamoun Sakkal 1997

Arabic Script

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

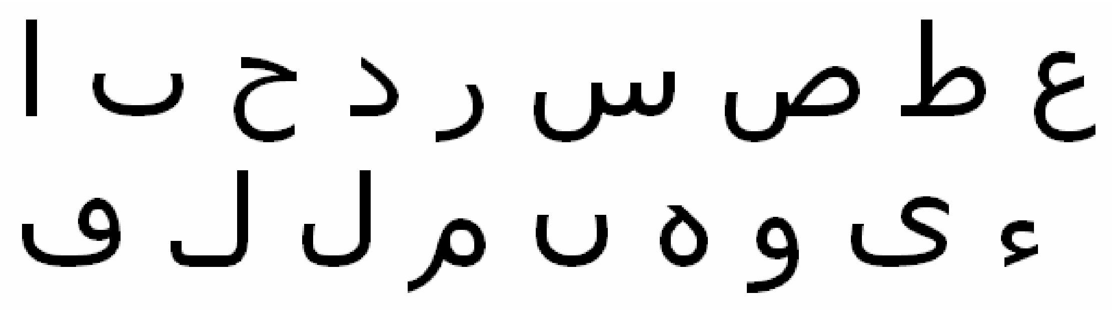
الخَطُ العَرَبِي

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

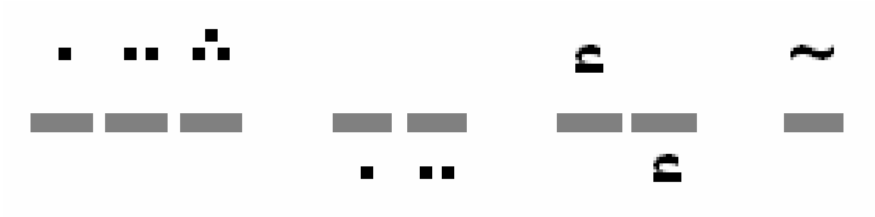
Arabic Script

Alphabet

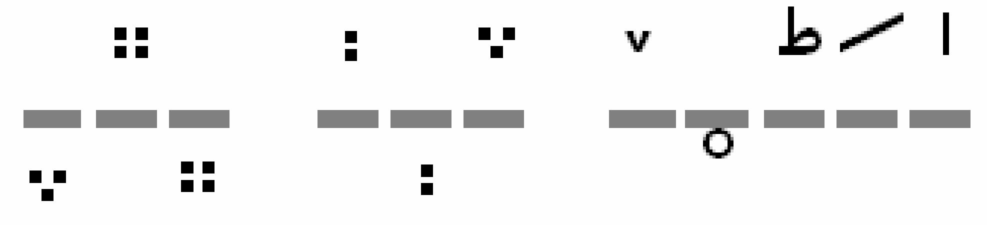
- letter forms



- letter marks
  - Arabic only



  - Other languages
    - Persian, Kurdish, Urdu, 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

ز	د	ا	ن	ب	ك	م	ش	غ	**Stand alone**
			نـ	بـ	كـ	مـ	شـ	غـ	**initial**
ـز	ـد	ـا	ـنـ	ـبـ	ـكـ	ـمـ	ـشـ	ـغـ	**medial**
			ـن	ـب	ـك	ـم	ـش	ـغ	**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/ عَ رَ ب ← عَرَب = عرب

West /ʁarb/ غَ رْ ب ← غَرْب = غرب

Ligatures

Peace /salām/ س ل ا م ← سلام ✗ سلام

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

- Different styles
- High fluidity
- Optional ligatures
- Vertical arrangements

Arabic	Muhammad	algebra
عربـ	محمد	الجبر
عربي	محمد	الجبر
عربي	محمد	الجبر
عربي	محمد	الجبر

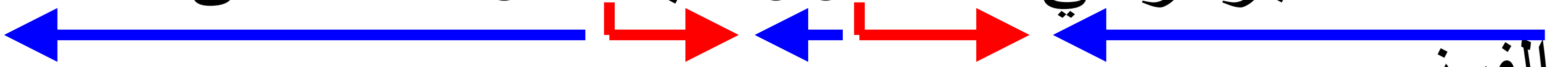
/ʕarabi/ /muħammad / /aldʒabr/

Arabic Script

“Arabic” Numerals

- Decimal system
- Numbers written left-to-right in right-to-left text

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

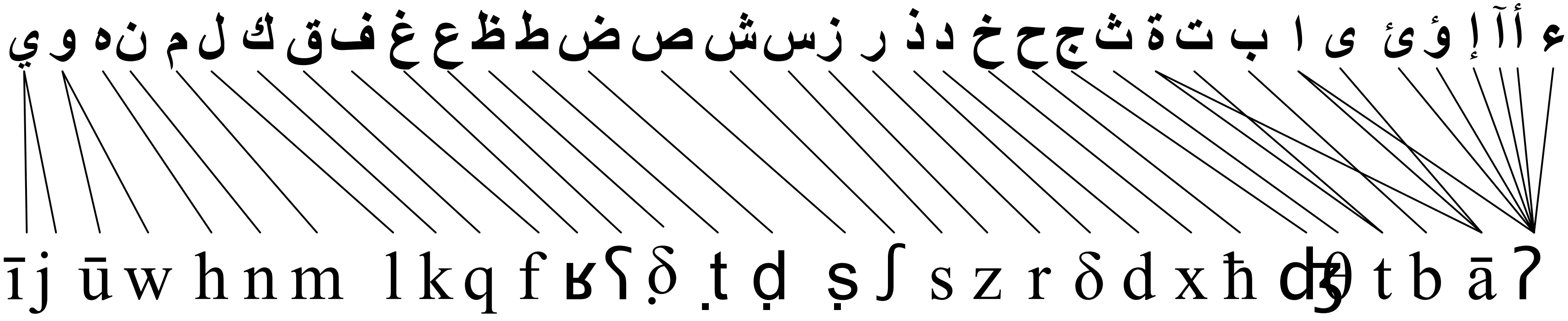
Western Arabic *Tunisia, Morocco, etc.*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Indo-Arabic *Middle East*	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Eastern Indo-Arabic *Iran, Pakistan, etc.*	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹

Road Map

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

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
  - Present diacritics mark syntactic/semantic distinctions
    - كتب /katab/ to write كُتب /kutib/ to be written
    - حُب /ħubb/ love حَب /ħabb/ seed
- 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
  - optional diacritics and dual use of letter
- 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

Road Map

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

Arabic Script
Other languages

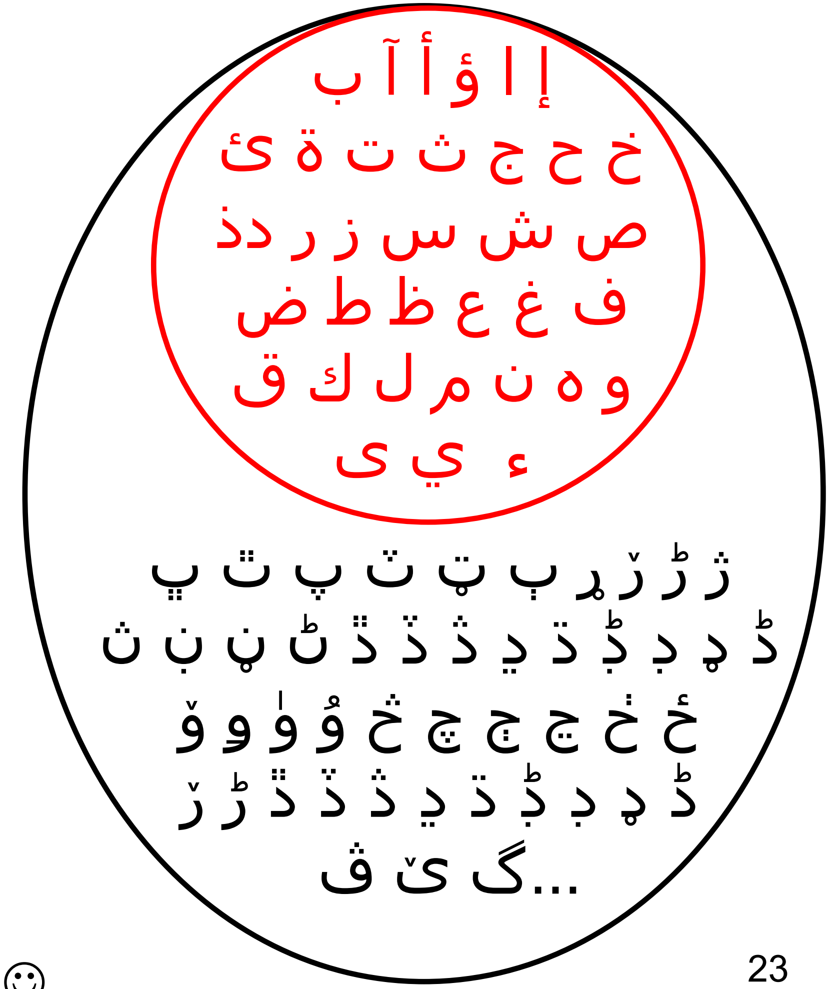
Arabic

- 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/, ڨ /v/, ڨ گ چ /g/, چ /tʃ/
    - regionally variable

Not Arabic

- Extra marks: haft (v), ring (o), taa (ط), four dots (::), vertical dots (:)
- Some Numerals (۴,۵,۶)

Once you learn the alphabet, it is easier ☺



إ ا ؤ أ آ ب
خ ح ج ث ت ة ئ
ص ش س ز ر د ذ
ف غ ع ظ ط ض
و ه ن م ل ك ق
ء ي ى

ژ ڑ ڒ ړ ب ټ ٽ پ ٿ ڀ
ڈ ډ ڊ ڋ ڌ ڍ ڎ ڏ ڐ ڻ ڼ ڹ ڽ
ځ ڂ ڃ ڄ ڇ څ ۇ ۈ ۅ ۆ
ڈ ډ ڊ ڋ ڌ ڍ ڎ ڏ ڐ ڑ ڒ
ڨ ێ گ...

بۆنه سووتێ جگه رو بۆچی نه بێ دڵ به که باب
بۆچی نه ڕوا له ته نم رۆحی ره وان میسلی شه هاب (١)
بۆله سه ر چاو ه یی چاو هه ڵنه قوڵێ ڕه شحه یی خوێن (٢)
بۆچ له فه ووار ه یی موژگان نه تکێ قه تر ه یی ئاب
بۆله به ر ناڵه نه بێ حه لقه ی حه لقم به سروود
بۆله به ر گریه نه بێ چه شمه ی چه شمم به سه راب
موونسی رۆژو شه ووم باعیسی ئارامی دڵم (٤)
رۆیی وو من له غه می که وتمه نێو به حری عه زاب
به وقووعی سه فه ری قادر ی ئوستاد خدری (٥)
به جه فا عه یشمی تاڵ کرد فه له کی خانه خه راب
چه نکَ ونه ی لێ مه د ه موتریب که له به ر فیرقه تی ئه و (٦)
رنه کی رۆحه له گوێم نه غمه ی ئاوازو روباب (٧)
ساغیر ی مه ی مه د ه ساقی که له به ر دووریی ئه و (٨)
تاڵه وه ک زه هری هه لایل له مه زاقم مه ی ئاب (٩)

□ Arabic
□ Not Arabic

□ **Arabic**

□ **Not Arabic**

سجل... انا عربي...
ورقم بطاقتي خمسون الف
واطفالي ثمانية
وتاسعهم سيأتي بعد صيف
فهل تغضب
سجل... انا عربي...
واعمل مع رفاق الكدح في محجر
واطفالي ثمانية
اسلّ لهم رغيف الخبز والاثواب والدفتر
من الصخر
ولا اتوسل الصدقات من بابك
ولا اصغر امام بلاط اعتابك
فهل تغضب

سجل... انا عربي للشاعر: محمود درويش

□ Arabic
□ Not Arabic

شیلی بیٹی کے نام

تجھے جب بھی کوئی دکھ دے
اس دکھ کا نام بیٹی رکھنا
جب میرے سفید بال
تیرے گالوں پر آن ہنسیں، رو لینا
میرے خواب کے دکھ پہ سو لینا
جن کھیتوں کو ابھی اگنا ہے
ان کھیتوں میں

Road Map

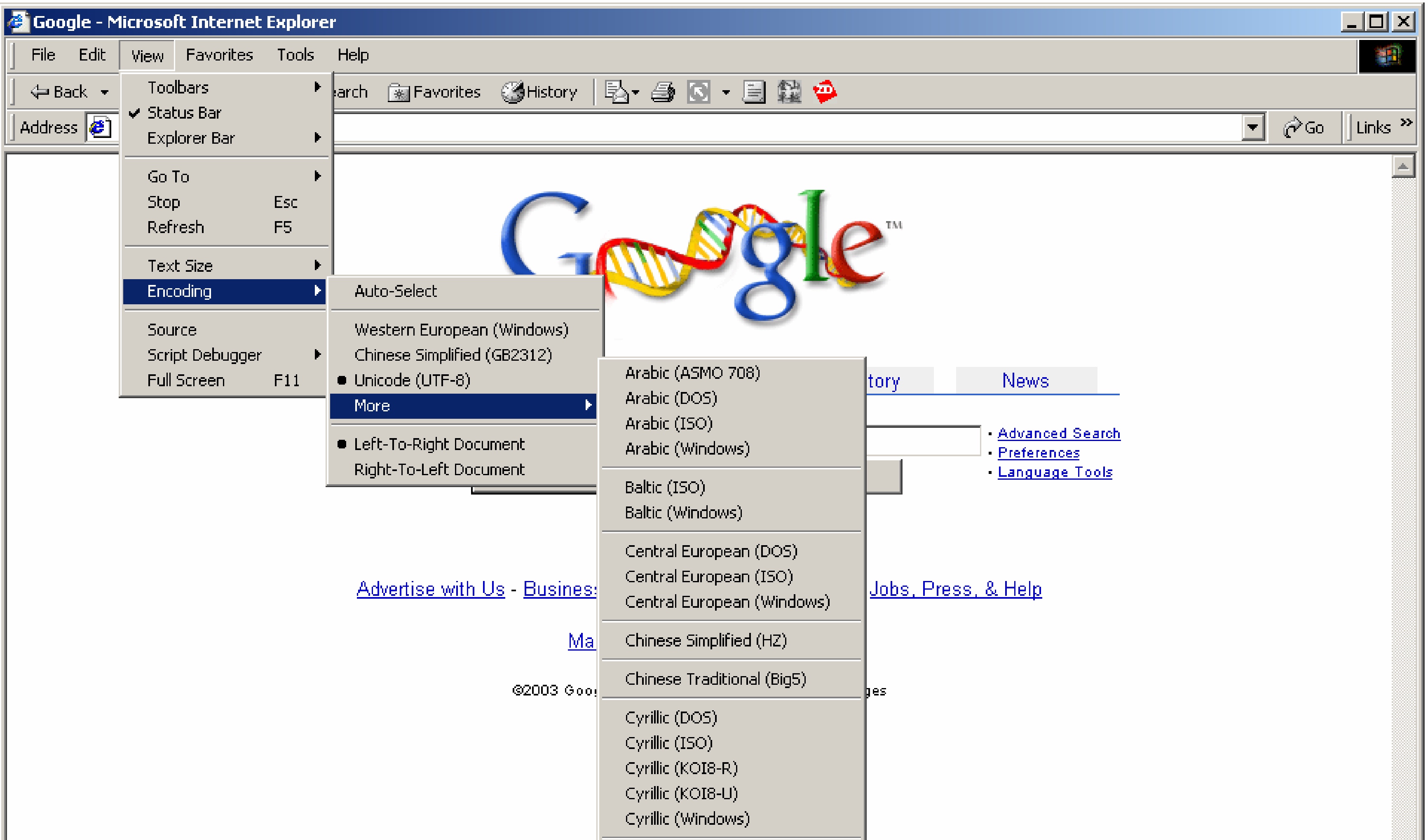
- Introduction
- Orthography
  - Arabic Script
  - MSA Phonology and Spelling
  - Recognizing Arabic vs. Persian/Urdu/Pashto/Kurdish/Sindhi/…
  - Encoding Issues
- Morphology
- Syntax
- Machine Translation Issues
- Dialects

Encoding Issues

- Encoding Arabic
  - Data entry, storage, and display
  - Ease of use for *Arabic-illiterate* users
  - 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

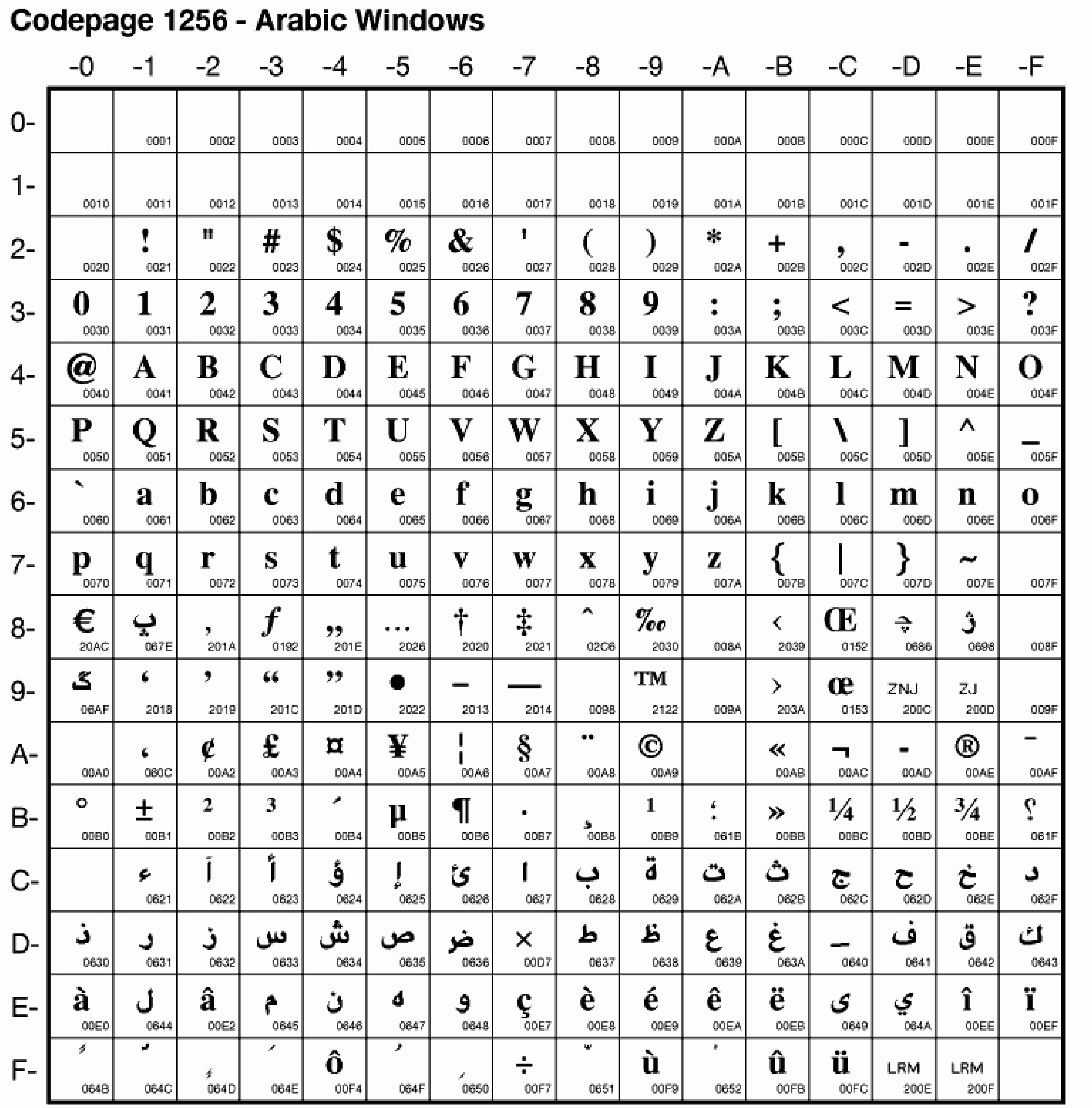
Encoding Issues

- 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)

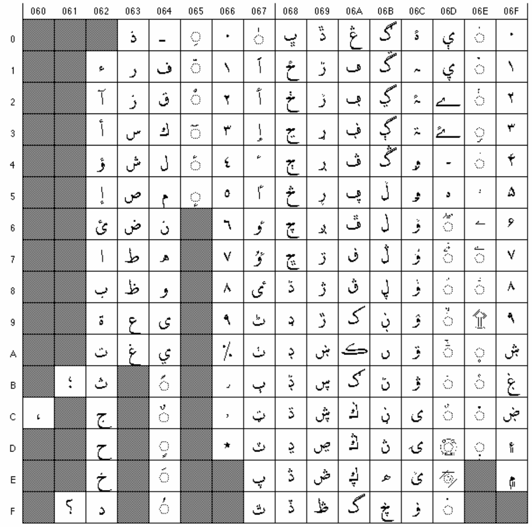


Codepage 1256 - Arabic Windows

	-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

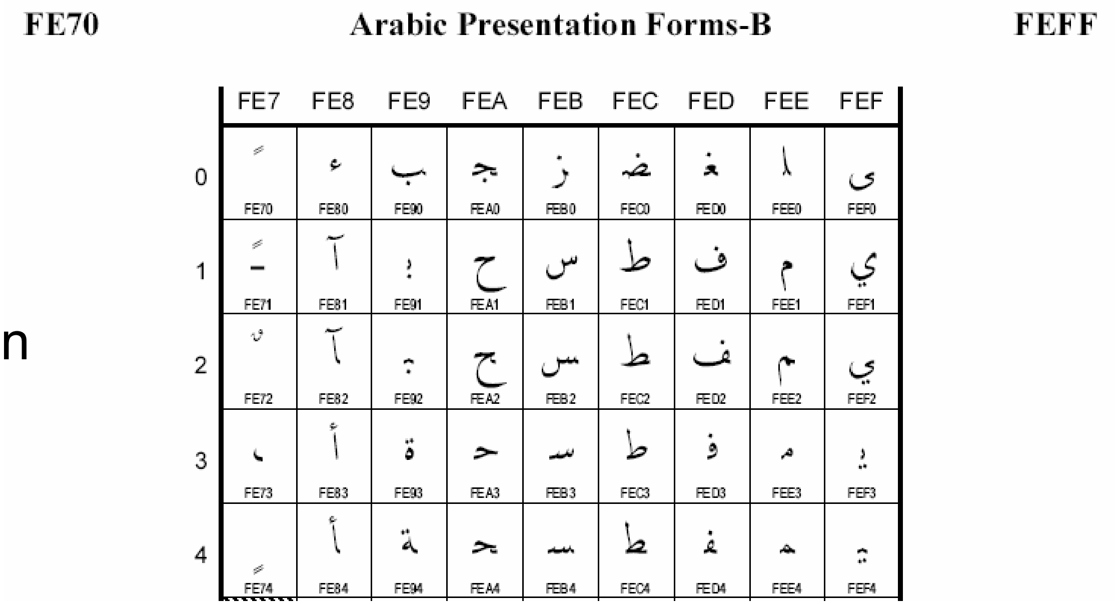
- Unicode
  - Becoming the standard more and more
  - 2-byte characters
  - Widely supported input/display
  - Supports extended Arabic characters
  - Multi-script representation



	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	◌۟		

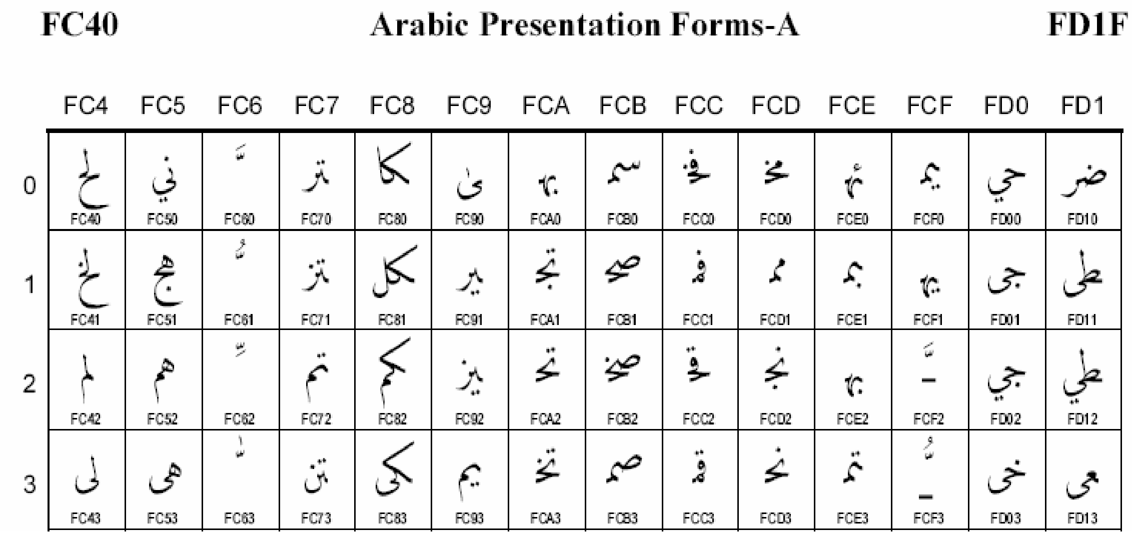
Encodings

- Unicode
  - Supports presentation forms (shapes and ligatures)



FE70 **Arabic Presentation Forms-B** FEFF

	FE7	FE8	FE9	FEA	FEB	FEC	FED	FEE	FEF
0	ﹰ FE70	ﺀ FE80	ﺐ FE90	ﺠ FEA0	ﺰ FEB0	ﻀ FEC0	ﻐ FED0	ﻠ FEE0	ﻰ FEF0
1	ﹱ FE71	ﺁ FE81	ﺑ FE91	ﺡ FEA1	ﺱ FEB1	ﻁ FEC1	ﻑ FED1	ﻡ FEE1	ﻱ FEF1
2	ﹲ FE72	ﺂ FE82	ﺒ FE92	ﺢ FEA2	ﺲ FEB2	ﻂ FEC2	ﻒ FED2	ﻢ FEE2	ﻲ FEF2
3	ﹳ FE73	ﺃ FE83	ﺓ FE93	ﺣ FEA3	ﺳ FEB3	ﻃ FEC3	ﻓ FED3	ﻣ FEE3	ﻳ FEF3
4	ﹴ FE74	ﺄ FE84	ﺔ FE94	ﺤ FEA4	ﺴ FEB4	ﻄ FEC4	ﻔ FED4	ﻤ FEE4	ﻴ FEF4



FC40 **Arabic Presentation Forms-A** FD1F

	FC4	FC5	FC6	FC7	FC8	FC9	FCA	FCB	FCC	FCD	FCE	FCF	FD0	FD1
0	ﱀ FC40	ﱐ FC50	ﱠ FC60	ﱰ FC70	ﲀ FC80	ﲐ FC90	ﲠ FCA0	ﲰ FCB0	ﳀ FCC0	ﳐ FCD0	ﳠ FCE0	ﳰ FCF0	ﴀ FD00	ﴐ FD10
1	ﱁ FC41	ﱑ FC51	ﱡ FC61	ﱱ FC71	ﲁ FC81	ﲑ FC91	ﲡ FCA1	ﲱ FCB1	ﳁ FCC1	ﳑ FCD1	ﳡ FCE1	ﳱ FCF1	ﴁ FD01	ﴑ FD11
2	ﱂ FC42	ﱒ FC52	ﱢ FC62	ﱲ FC72	ﲂ FC82	ﲒ FC92	ﲢ FCA2	ﲲ FCB2	ﳂ FCC2	ﳒ FCD2	ﳢ FCE2	ﳲ FCF2	ﴂ FD02	ﴒ FD12
3	ﱃ FC43	ﱓ FC53	ﱣ FC63	ﱳ FC73	ﲃ FC83	ﲓ FC93	ﲣ FCA3	ﲳ FCB3	ﳃ FCC3	ﳓ FCD3	ﳣ FCE3	ﳳ FCF3	ﴃ FD03	ﴓ FD13

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( داييبملوا يف )enitselaP( نيطسلف تكراش
```

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.

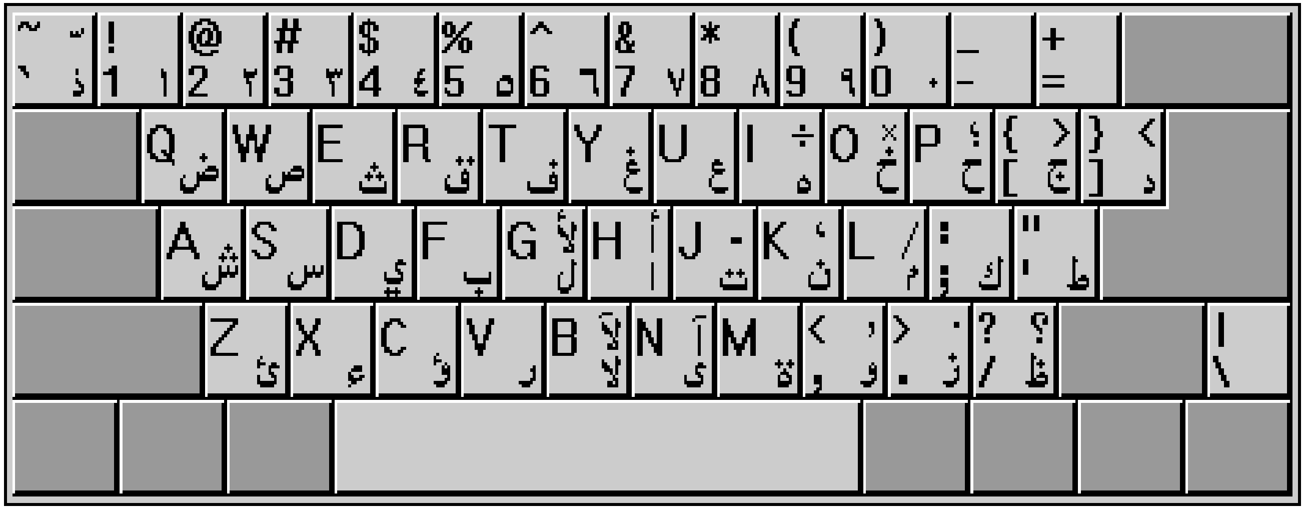
Display Problems

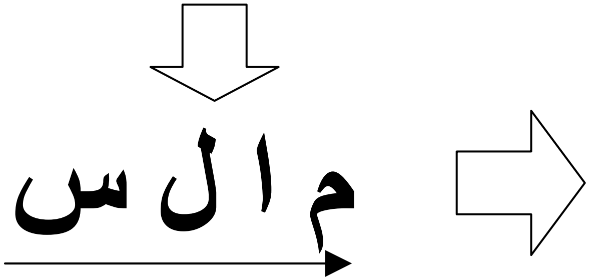
		Display Encoding			
		CP-1256	ISO-8859	Unicode	Western
Actual Encoding	CP-1256	تدشين منطقة حرة في دبي للتجارة الالكترونية	□ٍ ة حرة□تدشٍل كلظ ترنلٍة□دبٍ ففتجارة افاف	Y□□□ɠ□ɠ□ ψ□□Ŏɠāā□Ѧ□	ÊÏÔíä ãäØÞÉ ÍÑÉ Ýí ÏÈí ááÊÌÇÑÉ ÇáÇßÊÑæäíÉ
	ISO-8859	ة حرة â×و هوêتدش ننتجارة êدب êل ةêوèانانمتر	تدشين منطقة حرة في دبي للتجارة الالكترونية	Y□□□ ɠ□ɠ□ψ□□ŎɠǤǤ □□ɠ	ÊÏÔêæ åæ×âÉ ÍÑÉ áê ÏÈê ääÊÌÇÑÉ ÇäÇäãÊÑèæêÉ
	Unicode	ï» †ظٹظ´ط¯طھط؟ ©ط±طط ©ط‚ظ·ط†ظ…ظ ظٹظ¨ط¯ط ظپظٹ ©ط±ط§ط¬طھط„ظ„ظ ظ„ظ§ط„ظ§طƒ طھظ†ظˆظ±طھط©	ظ؟ ؛ ُ□ظ□ظ□ع□ع□ ع□ع□ظ□ ع□ظ□ ظ–ظ□ظ□ ع□ع□ ظ□ظ□ع□ ع□ع□ظ□ظ ، ظ□ظ□ظ □ ظ□ع□ظ□ع□ع□ظ□ظ □ع□ع□ع□ظ□	تدشين منطقة حرة في دبي للتجارة الالكترونية	ï»¿ØªØ¯Ø´ÙŠÙ† Ù…Ù†Ø·Ù‚Ø© ØØ±Ø© Ù□ÙŠ Ø¯Ø¨ÙŠ Ù„Ù„ØªØ¬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





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
- Commonly used in NLP
  - 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
- Machine Translation Issues
- Dialects

Morphology

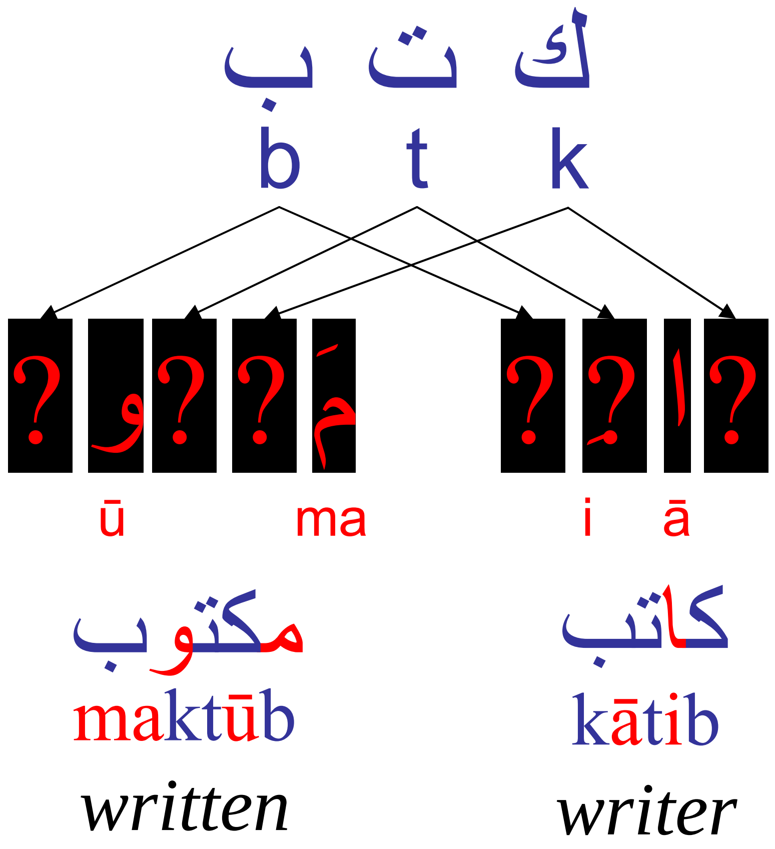
- Type
  - Concatenative: prefix, suffix, circumfix
  - Templatic: root+pattern
- Function
  - Derivational
    - Creating new words
    - *Mostly templatic*
  - Inflectional
    - Modifying features of words
      - Tense, number, person, mood, aspect
    - Mostly concatenative

Road Map

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

Derivational Morphology

- Templatic Morphology
  - Root
  - Pattern
  - Lexeme

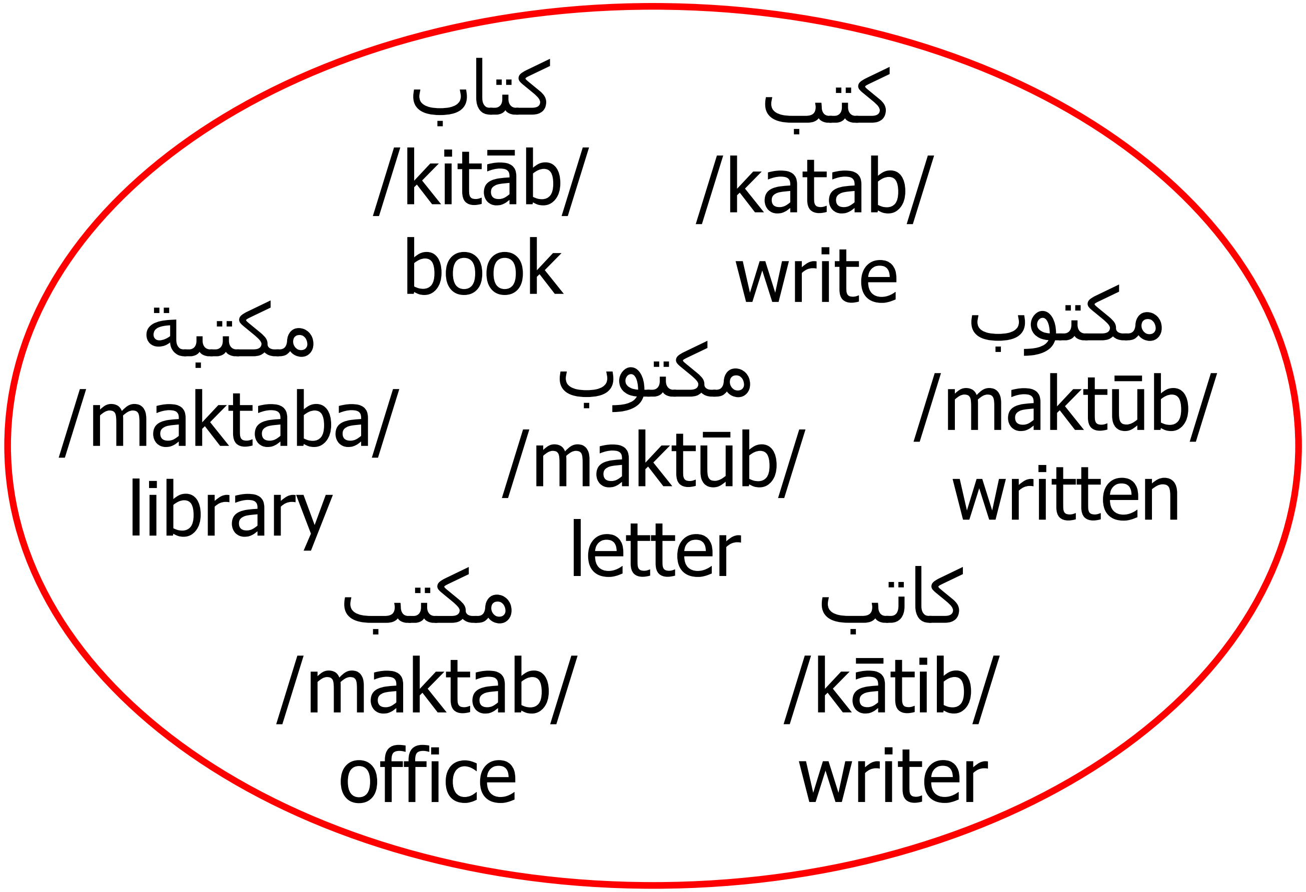


*Lexeme.Meaning = (Root.Meaning+Pattern.Meaning)*Idiosyncrasy.Random*

Derivational Morphology

Root Meaning

- ب ت ك KTB = notion of “*writing*”



Arabic	Transliteration	Meaning
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

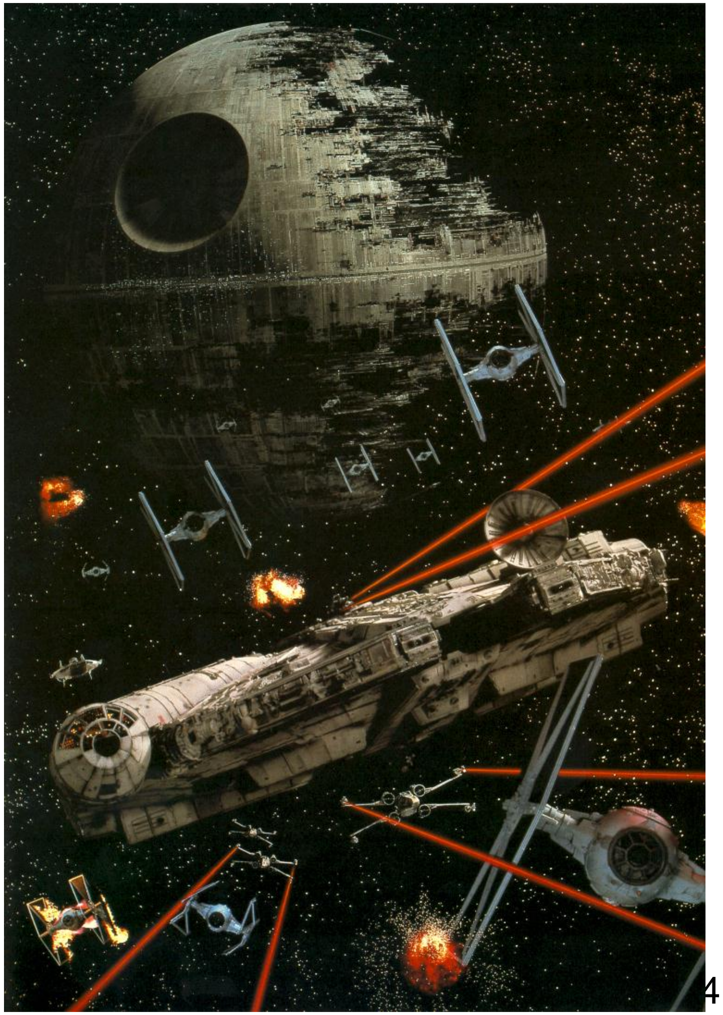
- LHM-1
- Notion of “meat”
  - لحم /laħm/
    - Meat
  - لحام /laħħām/
    - Butcher

لحم

laHm

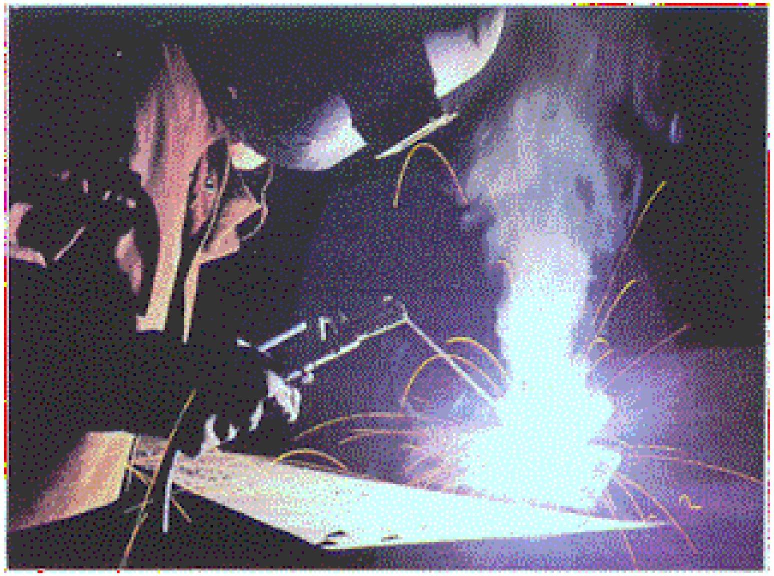
Derivational Morphology
Root Meaning

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



Derivational Morphology
Root Meaning

- LHM-3
- Notion of “soldering”
  - لحم /laħam/
    - Weld, solder, stick, cling
  - التحم /iltaħam/
    - Be welded/soldered/fused
  - ملتحم /multaħim/
    - Welded, soldered, fused



Derivational Morphology *Pattern Meaning*

- Verb Pattern Meaning is hard to define

Pattern	Pattern Meaning	Example	Gloss
I `1a2a3`	Basic sense of root	ktb → katab	write
II `1a22a3`	Intensification, causation	ktb → kattab	dictate
III `1aA2a3`	Interaction with others	ktb → kaAtab	correspond with
IV `Aa12a3`	Causation	jls → Ajlas	seat
V `ta1a22a3`	Reflexive of Pattern II	Elm → taEal~am	learn
VI `ta1aA2a3`	Reflexive of Pattern III	ktb → takaAtab	correspond
VII `Ain1a2a3`	Passive of Pattern I	ktb → Ainkatab	subscribe/enroll
VIII `Ai1ta2a3`	Acquiescence, exaggeration	ktb → Aiktatab	register
IX `Ai12a33`	Transformation	Hmr → AiHmarr	Turn red/blush
X `Aista12a3`	Requirement	ktb → Aistaktab	ask/make_write

Road Map

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

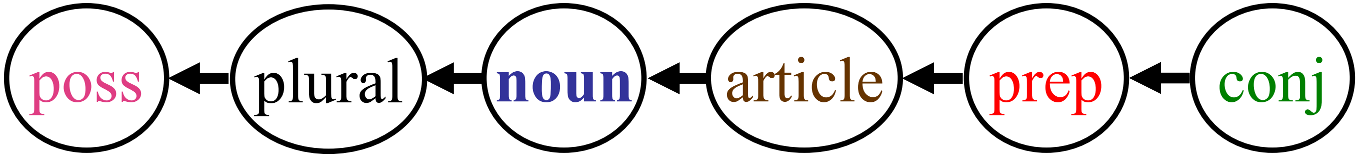
Inflectional Morphology

- 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


poss ← plural ← **noun** ← article ← prep ← conj

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

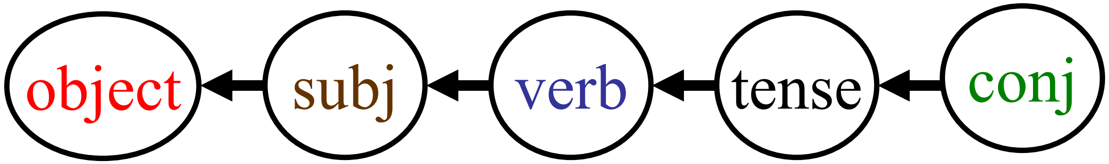
wa+li+al+maktaba+āt

and+for+the+library+plural

And for the libraries

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

Inflectional Morphology
Verbs



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
- Machine Translation Issues
- Dialects

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

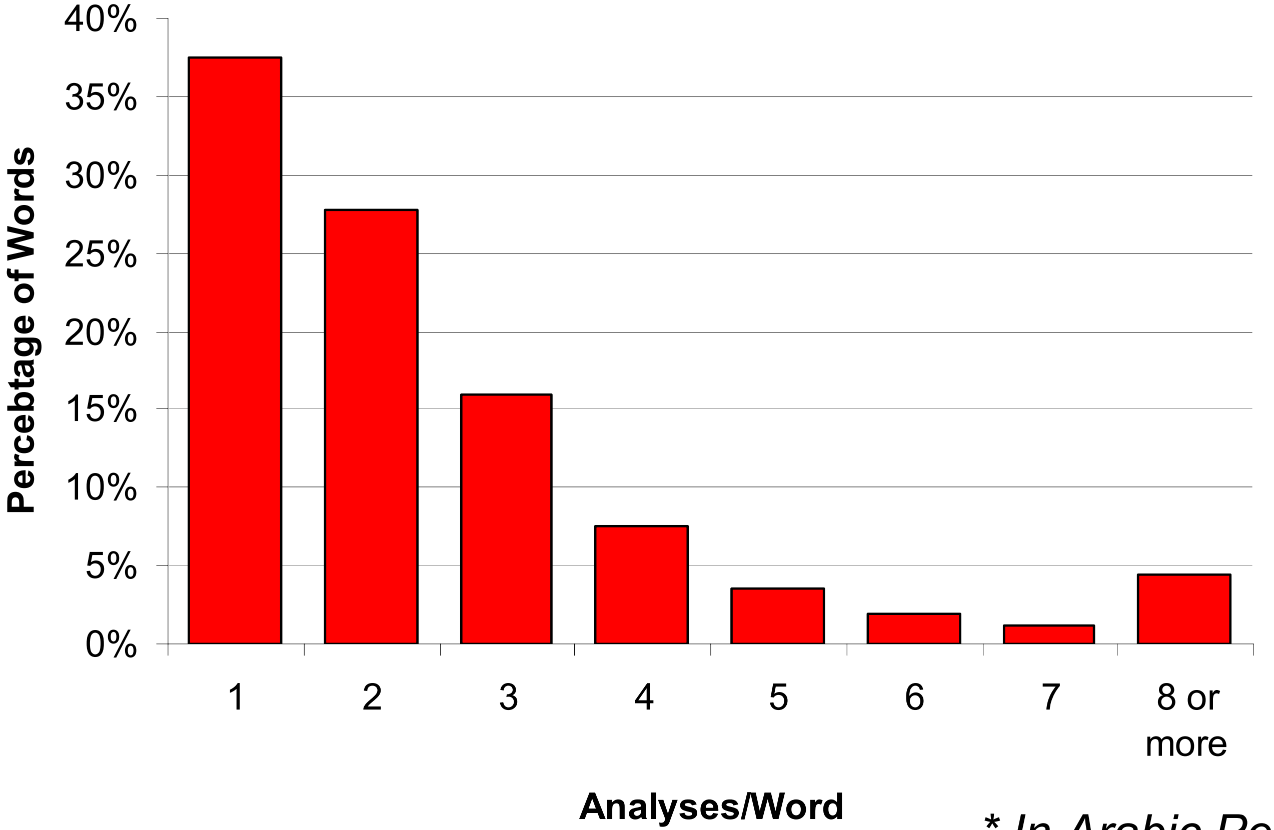
– /bayyana/	Verb	*he declared/demonstrated*
– /bayyanna/	Verb	*they [feminine] declared/demonstrated*
– /bayyin/	Adj	*clear/evident/explicit*
– /bayna/	Prep	*between/among*
– /biyin/	Proper Noun	*in Yen*
– /biyn/	Proper Noun	*Ben*

- 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)*

Morphological Ambiguity

- Average overall ambiguity* is 2.5 analyses/word
  - Compare to English ENGTWOL ambiguity (1.7-2.2 analyses/word)



** In Arabic Penn Treebank 1*

Road Map

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

Arabic Computational Morphology

- Representation units
  - Natural token وللـمـكـتـبــــات
    - 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) (Habash & Rambow 2005a) (Smith et al, 2005)
- Packages
  - AMIRA: Arabic SVM Toolkit (Diab et al, 2004)
  - MADA: Morphological Analysis and Disambiguation for Arabic (Habash and Rambow 2005a)
- 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
- Morphology
- Syntax
  - Morphology and Syntax
  - Sentence Structure
  - Phrase Structure
  - Computational Resources
- Machine Translation Issues
- Dialects

Morphology and Syntax

- Rich morphology crosses into syntax
  - Pro-drop / Subject conjugation
  - Verb subcategorization 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
    - **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

ل+المكتبات	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
- Machine Translation Issues
- Dialects

Sentence Structure

Two types of Arabic Sentences

- Verbal sentences
  - [Verb Subject Object] (VSO)
  - كتب الاولاد الاشعار
    Wrote the-boys the-poems
    The boys wrote the poems
- Copular sentences
  - [Topic Complement]
  - الاولاد شعراء
    the-boys poets
    The boys are poets

Sentence Structure

- Verbal sentences
  - Verb agreement with gender only
    - كتب الولد\الاولاد wrote$_{\text{3MascSing}}$ the-boy/the-boys
    - كتبت البنت\البنات wrote$_{\text{3FemSing}}$ the-girl/the-girls
  - Pronominal subjects are conjugated
    - كتبتُ wrote-you$_{\text{MascSing}}$
    - كتبتم wrote-you$_{\text{MascPlur}}$
    - كتبوا wrote-they$_{\text{MascPlur}}$
  - Passive verbs
    - Same structure: Verb$_{\text{passive}}$ Subject$_{\text{underlyingObject}}$
    - Agreement with surface subject

Sentence Structure

- Verbal sentences
  - Common structural ambiguity
    - *Third masculine/feminine singular are structurally ambiguous*
      - Verb$_{\text{3MascSingular}}$ Noun$_{\text{Masc}}$
        Verb subject=he object=Noun
        Verb subject=Noun
    - Passive and active forms are often similar in standard orthography
      - كتب /kataba/ he wrote
      - كُتب /kutiba/ it was written

Sentence Structure

- Copular sentences
  - [Topic Complement]
    Definite Topic, Indefinite Complement
    - الولد شاعر
      the-boy poet
      The boy is a poet
  - [Auxiliary Topic Complement]
    Auxiliaries (*kāna and her sisters*)
    - Tense, Negation, Transformation, Persistence
    - كان الولد شاعرا was the-boy poet *The boy was a poet*
    - ليس الولد شاعرا is-not the-boy poet *The boy is not a poet*
  - Inverted order is expected in certain cases
    - Indefinite topic
      عندي كتاب /ʕandi kitābun/ at-me a-book *I have a book*

Sentence Structure

- Copular sentences
  - Types of complements
    - Noun/Adjective/Adverb
      - الولد ذكي the-boy smart *The boy is smart*
    - Prepositional Phrase
      - الولد في المكتبة the-boy in the-library *The boy is in the library*
    - Copular-Sentence
      - الولد كتابه كبير [the-boy [book-his big]] *The boy, his book is big*
    - Verb-Sentence
      - الاولاد كتبوا الاشعار

        [the-boys [wrote-they poems]] The boys wrote the poems
      - Full agreement in this order (SVO)
      - الاشعار كتبها الاولاد

        [the-poems [wrote-it the boys]] The poems, the boys wrote

Road Map

- Introduction
- Orthography
- Morphology
- Syntax
  - Morphology and Syntax
  - Sentence Structure
  - Phrase Structure
  - Computational Resources
- Machine Translation Issues
- Dialects

Phrase Structure

- Noun Phrase
  - Determiner Noun Adjective PostModifier
    - هذا الكاتب الطموح القادم من اليابان

      this the-writer the-ambitious the-arriving from Japan

      This ambitious writer from Japan
  - Noun-Adjective agreement
    - number, gender, definiteness
      - الكاتبة الطموحة the-writer$_{fem}$ the-ambitious$_{fem}$
      - الكاتبات الطموحات the-writer$_{femPlur}$ the-ambitious$_{femPlur}$

Phrase Structure

- Noun Phrase
  - Idafa construction (اضافة)
    - **Noun1** ***of*** **Noun2** encoded structurally
    - Noun1-indefinite Noun2-definite
    - ملك الاردن

      king Jordan

      the king of Jordan / Jordan's king
  - Noun1 becomes definite
    - Agrees with definite adjectives
  - Idafa chains
    - $N^{1}_{indef}\,N^{2}_{indef}\ \ldots\ N^{n-1}_{indef}\,N^{n}_{def}$
    - ابن عم جار رئيس مجلس ادارة الشركة

      son uncle neighbor chief committee management the-company

      The cousin of the CEO's neighbor

Phrase Structure

- Morphological *definiteness* interacts with syntactic structure

		Word 1 كاتب *writer*	
		definite	Indefinite
Word 2 فنان *artist*	definite	***Noun Phrase*** الكاتب الفنان *The artist(ic) writer*	***Noun Compound*** كاتب الفنان The writer of the artist
	indefinite	***Copular Sentence*** الكاتب فنان The writer is an artist	***Noun Phrase*** كاتب فنان An artist(ic) writer

Road Map

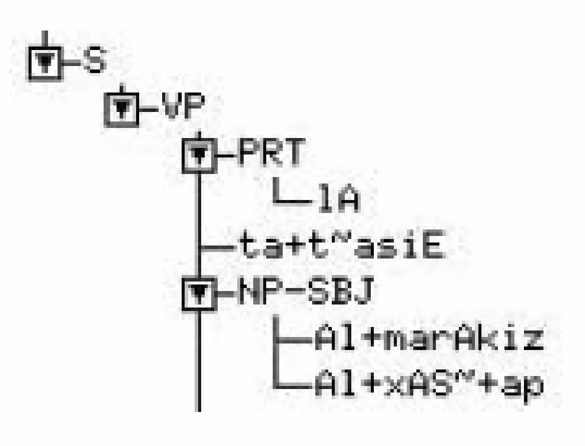
- Introduction
- Orthography
- Morphology
- Syntax
  - Morphology and Syntax
  - Sentence Structure
  - Phrase Structure
  - Computational Resources
- Machine Translation Issues
- Dialects

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)

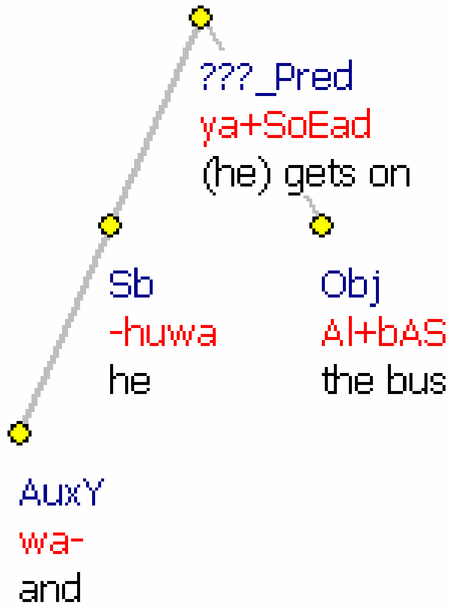
Computational Resources

- 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)



Computational Resources

- Prague Dependency Treebank
- Currently 100k words
- Partial overlap with PATB and Arabic Gigaword
  - Agence France Presse, AlHayat and Xinhua
- Morphological analysis
  - Similar to PATB
- Dependency representation



Graphic courtesy of Otakar Smrž: http://ckl.mff.cuni.cz/padt/PADT_1.0/docs/slides/2003-eacl-trees.ppt

Computational Resources

- Applications using Penn Arabic Treebank
  - Statsitical parsing
    - Bikel’s parser (Bikel 2003)
      - Same engine used with English, Chinese and Arabic
  - POS tagging and morphological disambiguation
    - (Diab et al, 2004) and (Habash and Rambow, 2005a)
- Arabic pos tagging (Khoja, 2001)
- Formalism conversion
  - Constituency to dependency (Žabokrtský and Smrž 2003)
  - Tree-adjoining grammar extraction (Habash and Rambow 2004)
- Automatic diacritization

Road Map

- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
  - Morphology and Translation
  - Translation Divergences
  - Computational Resources
- Dialects

Morphology and Translation

which level to go down to?

- Natural token ولـلـمـكـتـبـــات
- Word وللمكتبات
- Segmented Word و ل المكتبات
- Prefix + Stem + Suffix ولل+مكتب+ات
- Lexeme + Features مكتبة [+Plural +Def +ل +و]
- Root + Pattern + Features

ك ت ب + ةa3a21aم + [+Plural +Def +ل +و]

Morphology and Translation

What approach?

• Natural token	Not Appropriate
• Word	Statistical MT
• Segmented Word	Statistical MT
• Prefix + Stem + Suffix	Statistical/Symbolic
• Lexeme + Features	Symbolic MT
• Root + Pattern + Features	Too Abstract?

Morphology and Translation

What resources?

- Available resources may span different levels of representation!
- Most dictionaries are lexeme-based
- Buckwalter stem dictionary contains English glosses
- Statistical translation lexicons depend on the type of tokenization used before alignment
  - Word (no disambiguation necessary)
  - Segmented word (minimal disambiguation necessary)
  - Stem/Lexeme (machine/human disambiguation necessary)
- *Consistency is important*

(Lee, 2004), (Habash, 2006), (Habash and Sadat, 2006), (Habash et al. 2006)

Road Map

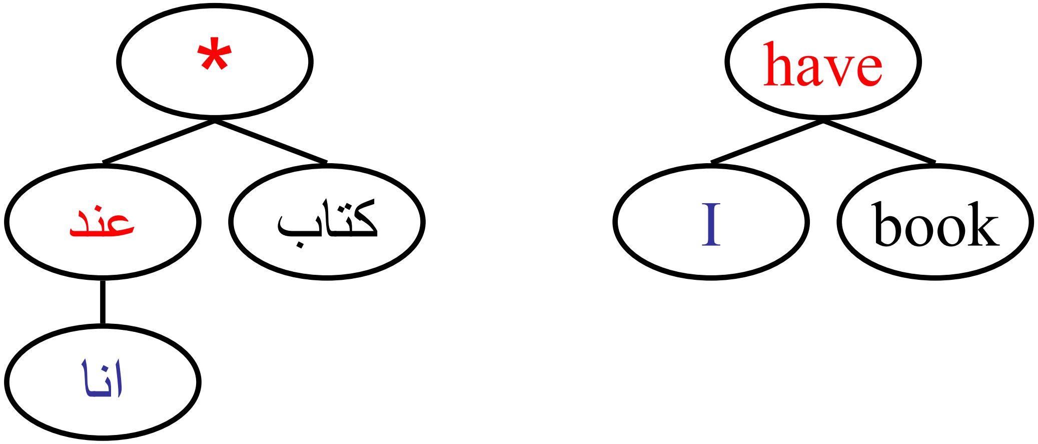
- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
  - Morphology and Translation
  - Translation Divergences
  - Computational Resources
- Dialects

Translation Divergences

- Beyond word-order variation
  - Arabic VSO - English SVO
  - Arabic N Adj - English Adj N
- Meaning of two translationally equivalent constituents is distributed differently in two languages
- Divergence dimensions
  - Categorial Variation *(develop → development)*
  - Conflation *(become frozen → freeze)*
  - Inflation *(freeze → become frozen)*
  - Structural *(enter the room → enter into the room)*
  - Head Swap *(swim across the river → cross the river swimming)*
  - Thematic *(John likes Mary → Mary pleases John)*

Translation Divergences

conflation

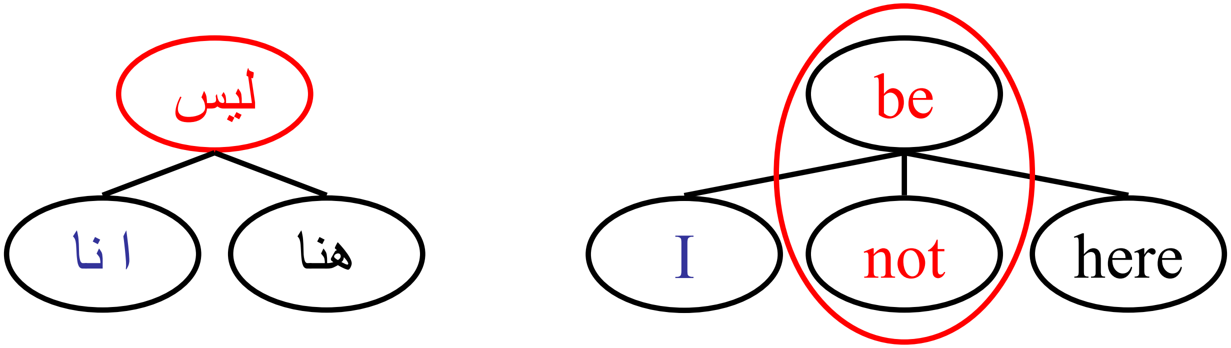


عندي كتاب

at-me book

I have a book

Translation Divergences
conflation

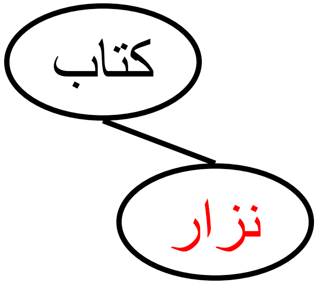


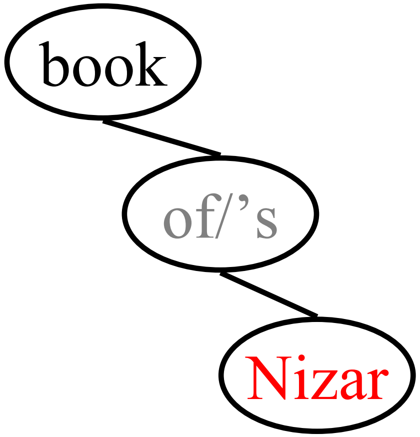
لست هنا

I-am-not here

I am not here

Translation Divergences *structural*

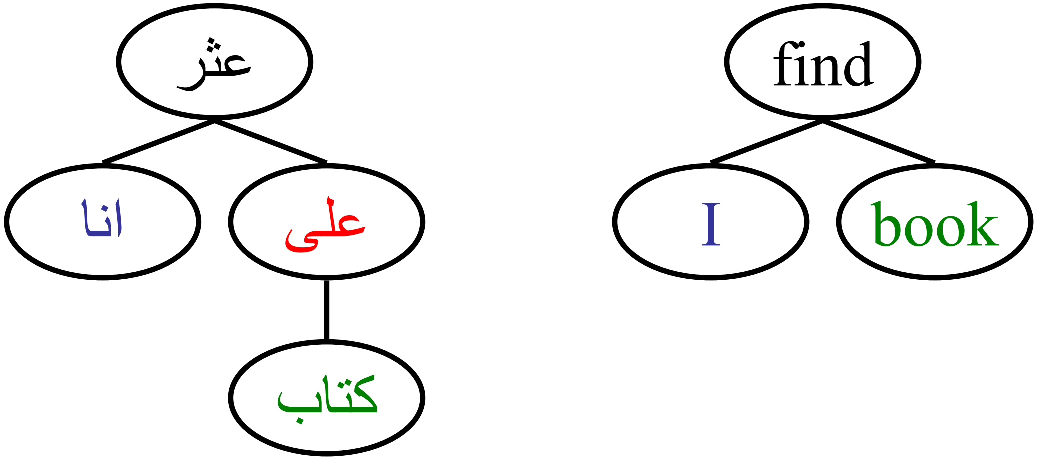




نزار كتاب
book Nizar

Nizar's book
Book of Nizar

Translation Divergences
structural



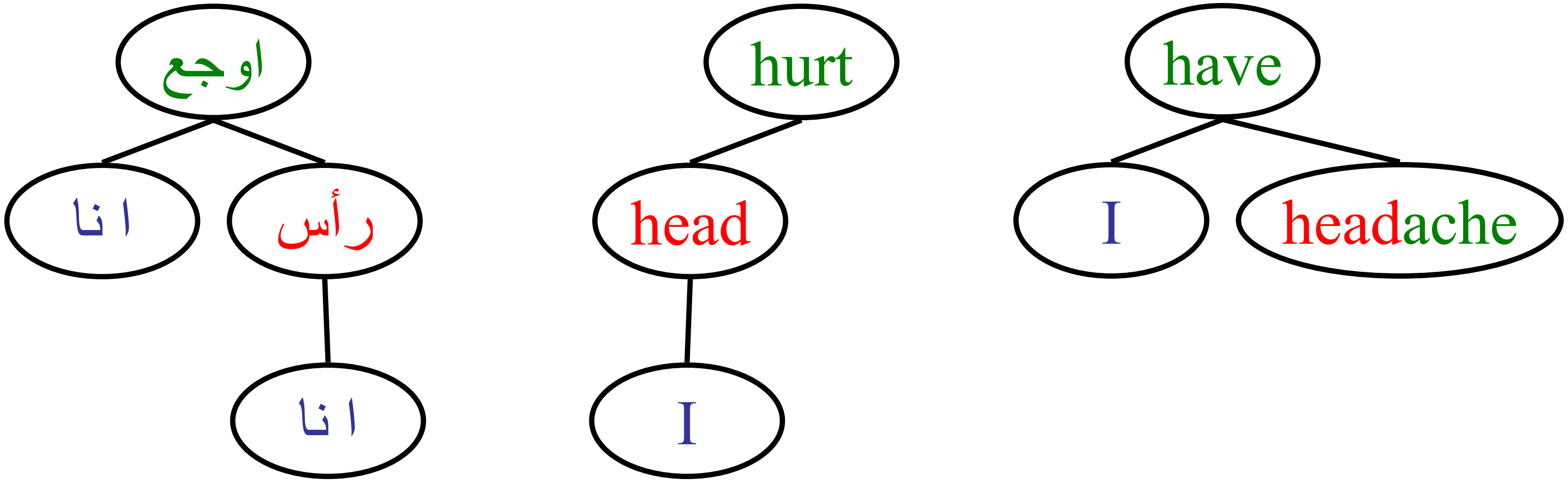
عثرت على الكتاب

found-I *upon* the-book

I found the book

Translation Divergences

thematic & conflational

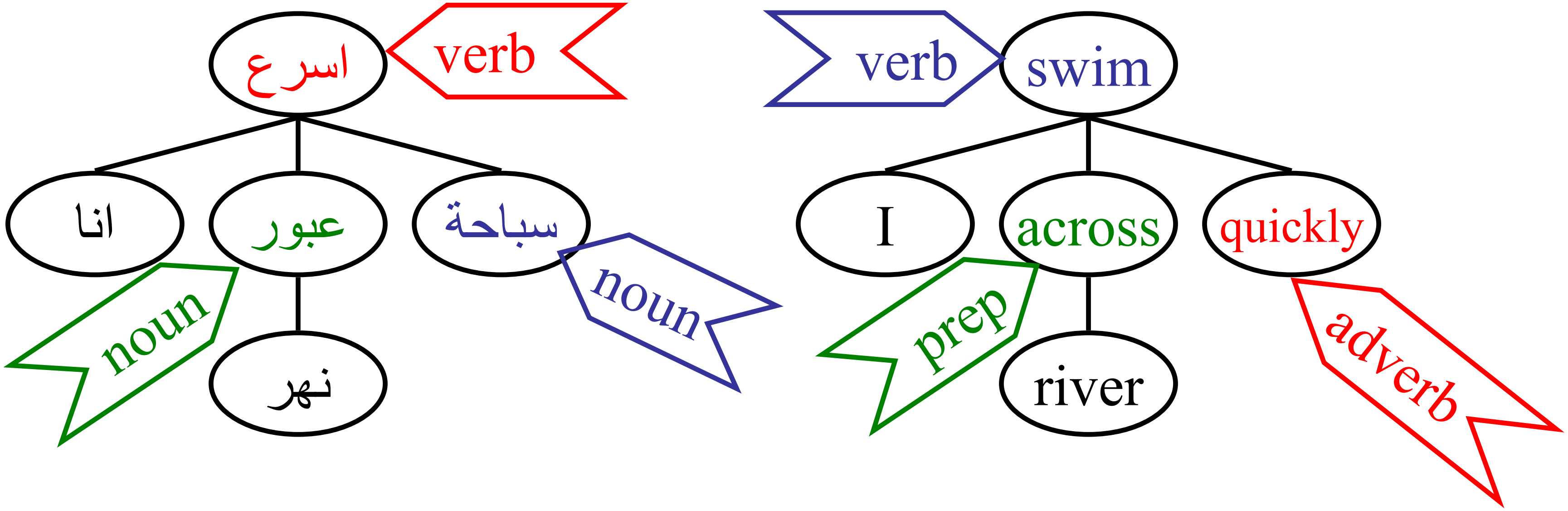


رأسي يوجعني

head-my hurts-me   my head hurts   I have a headache

Translation Divergences

head swap and categorial



اسرعت عبور النهر سباحة

I-sped crossing the-river swimming

I swam across the river quickly

Road Map

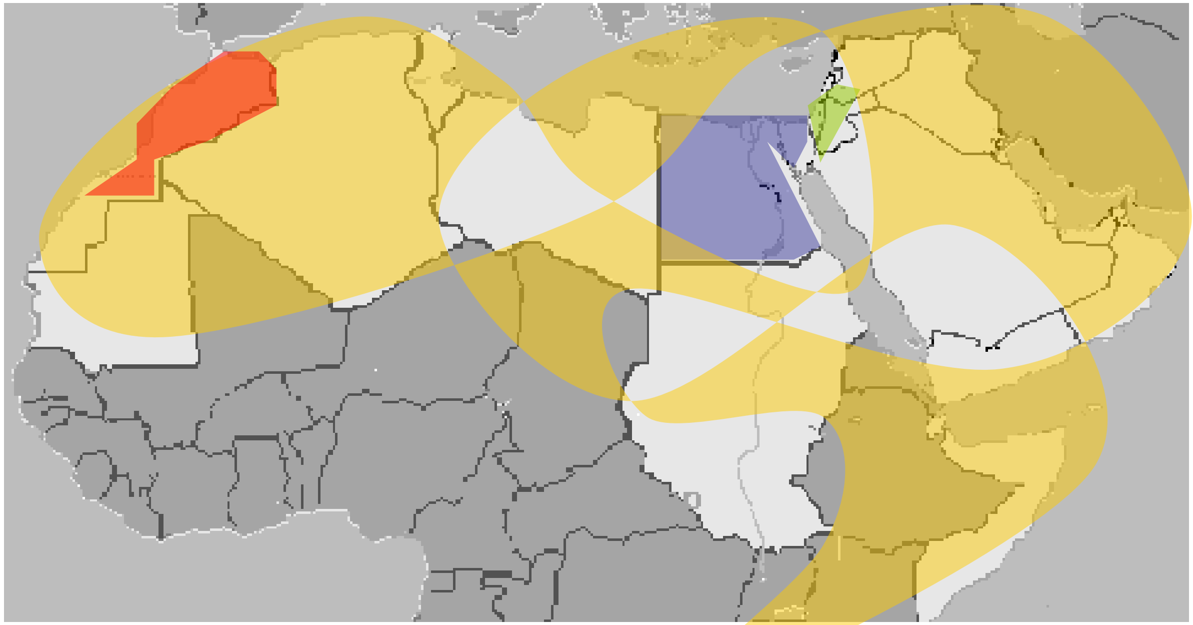
- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
  - Morphology and Translation
  - Translation Divergences
  - Computational Resources
- Dialects

Computational Resources

- Dictionaries
  - Buckwalter stem dictionary (LDC)
  - Salmone dictionary (Tufts university)
  - Online dictionaries – Ajeeb.com (Sakhr), Almisbar.com, Ectaco.com
- Parallel corpora (LDC)
  - United Nations Corpus (parallel with other UN languages)
  - Ummah Corpus (parallel with English)
  - Arabic News Translation Corpus
  - Arabic Treebank English Translation
  - *More on LDC webpage…*
- MT evaluation
  - Arabic-English Multi-translation Corpus (LDC)
  - NIST’s MT-EVAL

Road Map

- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
- Dialects
  - General Definitions
  - Phonological & Lexical Variation
  - Morphological Variation
  - Syntactic Variation
  - Code Switching
  - Computational Resources



lam jaʃtari nizār ṭawilatan ζadīdatan لم يشتر نزار طاولة جديدة

didn’t buy Nizar table new

nizār maʃtarāʃ ṭarabēza gidīda ● نزار ماشتراش طربيزة جديدة

nizār maʃtarāʃ ṭawile ζdīde ● نزار ماشتراش طاولة جديدة

nizar maʃrāʃ mida ζdīda ● نزار ماشراش ميدة جديدة

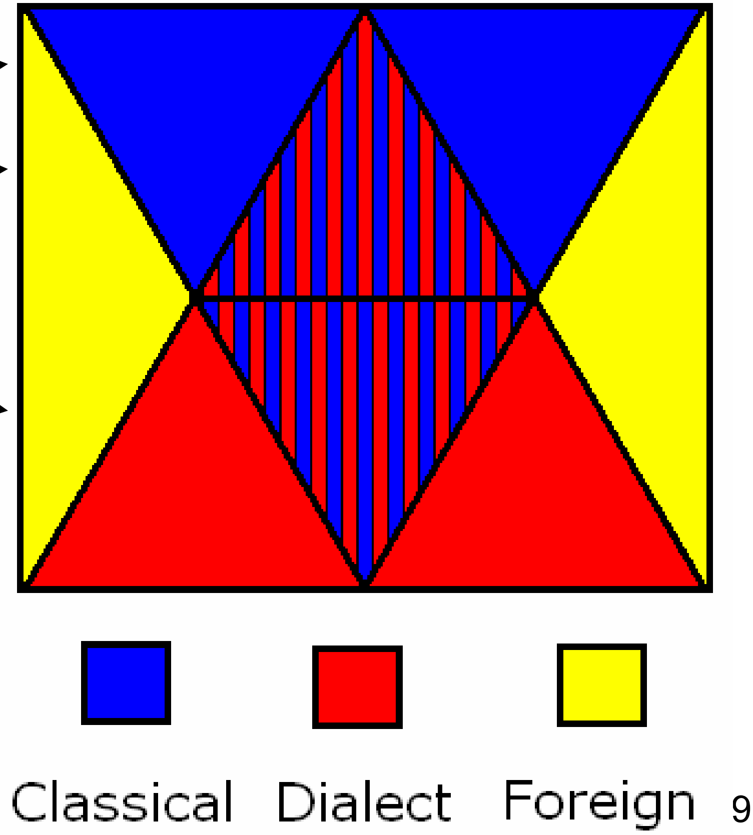
Nizar not-bought-not table new

General Definitions

- What is a ‘dialect’?
  - Political and Religious factors
- Modern Standard Arabic
- Regional Dialects
  - Egyptian Arabic (EGY)
  - Levantine Arabic (LEV)
  - Gulf Arabic (GULF)
  - North African Arabic (NOR)
  - Iraqi, Yemenite, Sudanese, Maltese?
- Social dialects
  - City
  - Peasant
  - Bedouin

General Definitions

- Diglossia
- Badawi’s levels
  - Traditional Arabic
  - Modern Arabic
  - Educated Colloquial
  - Literate Colloquial
  - Illiterate Colloquial
- Polyglossia

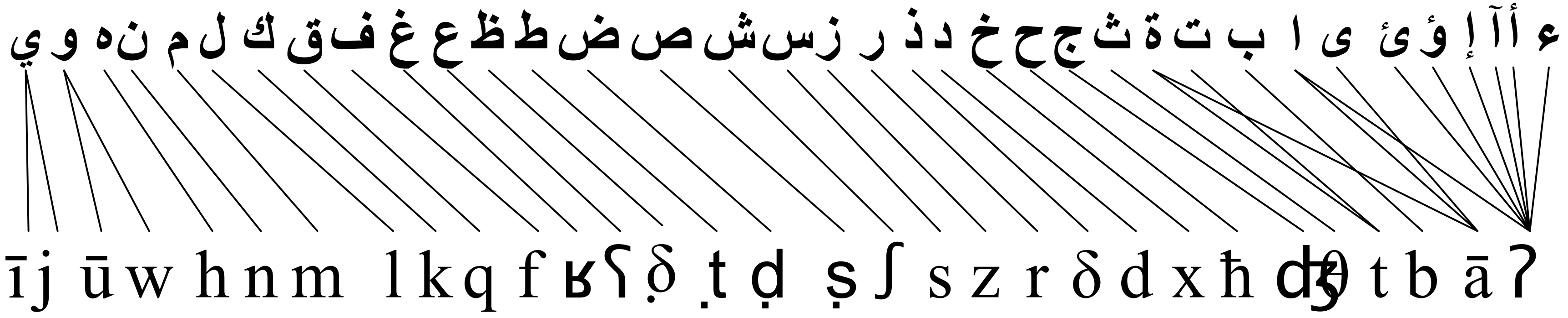


Road Map

- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
- Dialects
  - General Definitions
  - Phonological & Lexical Variation
  - Morphological Variation
  - Syntactic Variation
  - Code Switching
  - Computational Resources

Phonological Variation

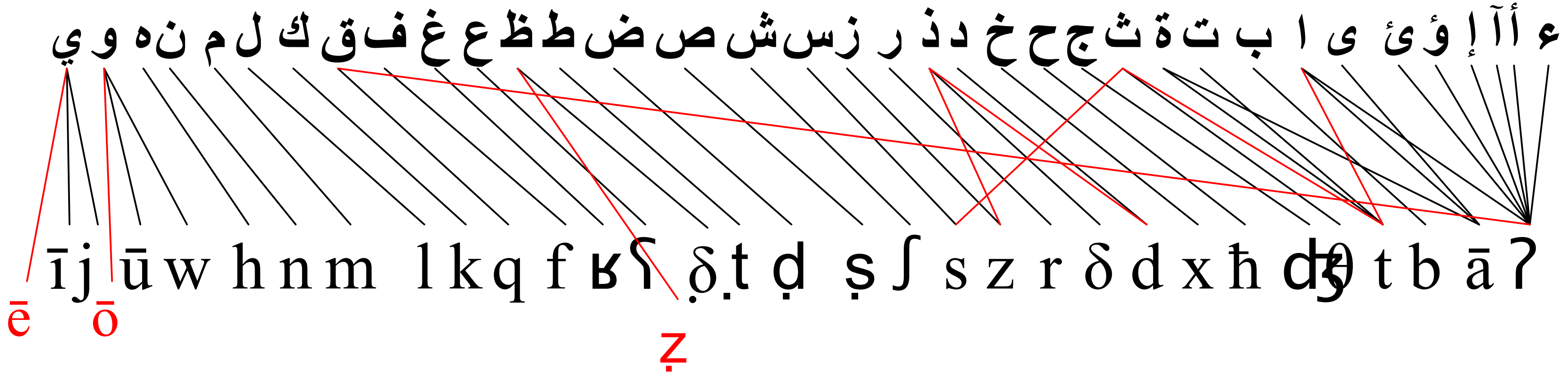
MSA



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

LEV



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

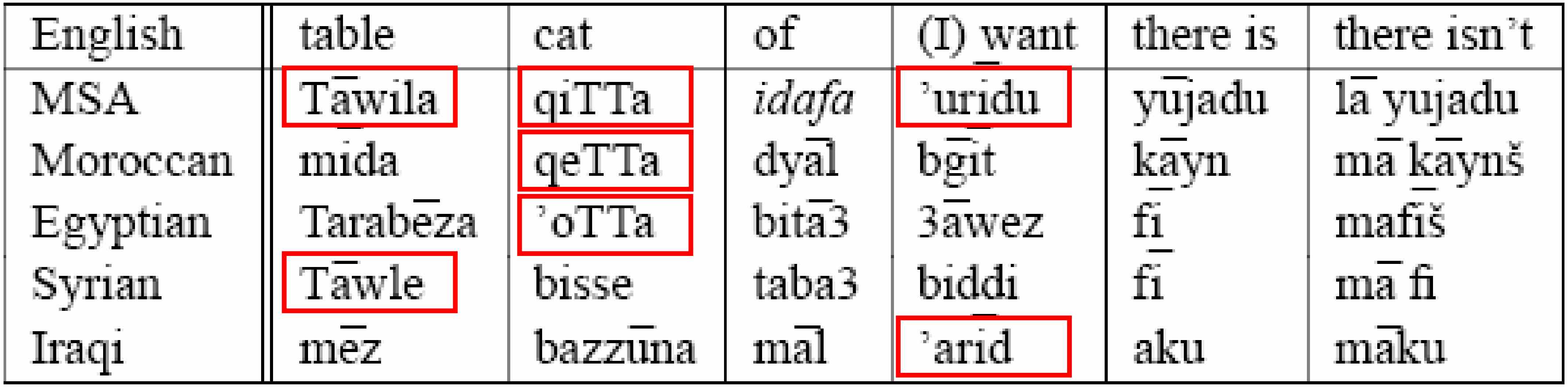
ī j ū w h n m l k q f ʁ ʕ δ̣ ṭ ḍ ṣ ʃ s z r δ d x ħ ʤ θ t b ā ʔ

ē ō ẓ

- No dialect-specific standard orthography

Lexical Variation

- Arabic Dialects vary widely lexically



English	table	cat	of	(I) want	there is	there isn’t
MSA	Tāwila	qiTTa	*idafa*	’urīdu	yūjadu	lā yujadu
Moroccan	mida	qeTTa	dyāl	bgīt	kāyn	mā kāynš
Egyptian	Tarabēza	’oTTa	bitā3	3āwez	fī	mafīš
Syrian	Tāwle	bisse	taba3	biddi	fī	mā fī
Iraqi	mēz	bazzūna	māl	’arīd	aku	māku

- Arabic orthography allows consolidating some variations

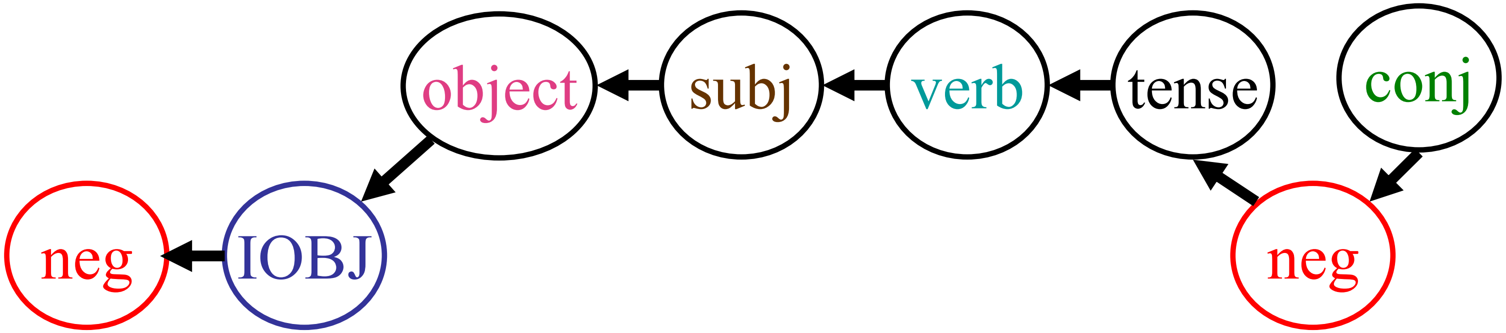
Road Map

- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
- Dialects
  - General Definitions
  - Phonological & Lexical Variation
  - Morphological Variation
  - Syntactic Variation
  - Code Switching
  - Computational Resources

Morphological Variation

- Nouns
  - No case marking
    - Word order implications
  - Paradigm reduction
    - Consolidating masculine & feminine plural
- Verbs
  - Paradigm reduction
    - Loss of dual forms
    - Consolidating masculine & feminine plural (2nd,3rd person)
    - Loss of morphological moods
      - Subjunctive/jussive form dominates in some dialects
      - Indicative form dominates in others
  - Other aspects increase in complexity

Morphological Variation
Verb Morphology



MSA
ولم تكتبوها له
walam taktubūhā lahu
wa+lam taktubū+hā la+hu
and+not_past write_you+it for+him

EGY
وماكتبتوهالوش
wimakatabtuhalūʃ
wi+ma+katab+tu+ha+lū+ʃ
and+not+wrote+you+it+for_him+not

And you didn’t write it for him

Morphological Variation

Verb conjugation

- Perfect verb derivation (*suffixes only*)

	1st Person Singular	2nd Person Singular ♂	2nd Person Singular ♀
MSA	كتبتُ katabtu	كتبتَ katabta	كتبتِ katabti
LEV	كتبت katabt		كتبتي katabti

- Imperfect verb derivation (*prefix+suffix*)

	1st Person Singular	2nd Person Singular ♂	2nd Person Singular ♀
MSA	اكتبُ aktubu	تكتبُ taktubu	تكتبينَ taktubīna تكتبي taktubī
LEV	اكتب aktob	تكتب toktob	تكتبي toktobi

Morphological Variation

Tense expression

	Perfect	***Imperfect***			
M **S** **A**	كتب *kataba* Past	يكتب *jaktubu* Present			سيكتب *sajaktubu* Future
L **E** **V**	كتب *katab* Past	يكتب *jiktob* 0-Tense	بيكتب *bjoktob* Present *habitual*	عم بيكتب ʕam *bjoktob* Present *progressive*	حيكتب *ħajiktob* Future

Road Map

- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
- Dialects
  - General Definitions
  - Phonological & Lexical Variation
  - Morphological Variation
  - Syntactic Variation
  - Code Switching
  - Computational Resources

Syntactic Variation

- Verbal sentences
  - The boys wrote the poems
  - MSA
    - Verb Subject Object (Partial agreement)

      كتب الاولاد الاشعار

      wrote$_{\text{masc}}$ the-boys the-poems
    - Subject Verb Object (Full agreement)

      الاولاد كتبوا الاشعار

      the-boys wrote$_{\text{mascPlural}}$ the-poems
  - LEV, EGY
    - Subject Verb Object

      الاولاد كتبو الاشعار

      The-boys wrote$_{\text{mascPlural}}$ the-poems
    - Less present: Verb Subject Object

      كتبو الاولاد الاشعار

      wrote$_{\text{mascPlural}}$ the-boys the-poems
    - Full agreement in both order

	V-S ***explicit subject***	**V(S)** ***pro dropped subject***	**S-V** ***explicit subject***
MSA	35%	30%	35%
LEV	10%	60%	30%

Verb-Subject distributions in the Levantine Arabic Treebank

(Maamouri et al, 2006)

Syntactic Variation

- Noun Phrase
  - Idafa construction
    - **Noun1** ***of*** **Noun2** encoded structurally
    - ملك الاردن
      king Jordan
      the king of Jordan / Jordan's king
  - Dialects have an additional common construct
    - **Noun1 <*particle*> Noun2**
    - LEV: الملك تبع الاردن the-king *belonging-to* Jordan
    - <particle> differs widely among dialects
  - Pre/post-modifying demonstrative article
    - MSA: هذا الرجل this the-man *this man*
    - EGY: الراجل ده the-man this *this man*

Road Map

- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
- Dialects
  - General Definitions
  - Phonological & Lexical Variation
  - Morphological Variation
  - Syntactic Variation
  - Code Switching
  - Computational Resources

Code Switching

MSA

LEV

MSA and Dialect mixing in speech

- phonology, morphology and syntax

لا أنا ما بعتقد لأنه عملية اللي عم بيعارضوا اليوم تمديد للرئيس لحود هم اللي طالبوا بالتمديد للرئيس الهراوي
وبالتالي موضوع منه موضوع مبدئي على الأرض أنا بحترم أنه يكون في نظرة ديمقراطية للأمور وأنه
يكون في احترام للعبة الديمقراطية وأن يكون في ممارسة ديمقراطية وبعتقد إنه الكل في لبنان أو أكثرية
ساحقة في لبنان تريد هذا الموضوع، بس بدي يرجع لحظة على موضوع إنجازات العهد يعني نعم نحكي
عن إنجازات العهد لكن هل النظام في لبنان نظام رئاسي النظام في لبنان من بعد الطائف ليس نظام رئاسي
وبالتالي السلطة هي عمليا بيد الحكومة مجتمعة والرئيس لحود أثبت خلال ممارسته الأخيرة بأنه لما بيكون
في شخص مسؤول في منصب معين وأنا عشت هذا الموضوع شخصيا بممارستي في موضوع الاتصالات
لما بياخد مواقف صالحة ضمن خطاب ومبادئ خطاب القسم هو إلى جانبه إنما مش مطلوب من رئيس
جمهورية هو يكون رئيس السلطة التنفيذية لأنه منه بقى في لبنان ما بعد إتفاق الطائف رئيس السلطة التنفيذية
عليه التوجيه عليه إبداء الملاحظات عليه القول ما هو خطأ وما هو صح عليه تثمير جهود الوطنية الشاملة
كي يظل في مصالحة وطنية كي يظل في توافق ما بين المسلم والمسيحي في لبنان يحتضن أبناء هذا البلد ما
يترك المسار يروح باتجاه الخطأ نعم إنما خطاب القسم كان موضوع مبادئ طرحت هو ملتزم فيها اللي
مشيوا معه وآمنوا فيها التزموا فيها أنا أثبت خلال الأربع سنوات بالممارسة الحكومية أني التزمت فيها ولما
التزمنا بهذا الموضوع كان الرئيس لحود إلى جنبنا في هذا الموضوع، أما الموضوع الديمقراطي أنا بتفهم
تماما هذا هالوجهة النظر بس ما ممكن نقول إنه الدستور أو تعديله هو أو إمكانية فتح إعادة انتخاب
ديمقراطي ضمن المجلس والتصويت إلى ما هنالك لرئيس جمهورية بولاية ثانية هو مسح هيئة في جوهر
الديمقراطية هذا بالأقل يعني قناعتي في هذا الموضوع.

Aljazeera Transcript http://www.aljazeera.net/programs/op_direction/articles/2004/7/7-23-1.htm

Road Map

- Introduction
- Orthography
- Morphology
- Syntax
- Machine Translation Issues
- Dialects
  - General Definitions
  - Phonological & Lexical Variation
  - Morphological Variation
  - Syntactic Variation
  - Code Switching
  - Computational Resources

Computational Resources

- Most work on Arabic dialects focuses on Automatic Speech Recognition
- Speech/transcript corpora
  - Egyptian and Levantine Arabic (LDC)
  - Moroccan and Tunisian Arabic (ELDA)
  - Gulf Arabic (Appen)
  - Many other…
- Few lexicons/morphology resources
  - CallHome Egyptian Arabic monolingual lexicon (LDC)
  - CallHome Egyptian Verb transducer (LDC)
- Work on multi-dialectic resources
  - Linguistic Data Consortium
  - Columbia University Arabic Dialect Modeling (CADIM) Group
    - Pan-Arab lexicon and Pan-Arab Morphology
- Novel Approaches to Arabic Speech Recognition (JHU summer workshop 2002) (Kirchhoff et al, 2002)
- Parsing Arabic Dialects (JHU summer workshop 2005)

(Rambow et al, 2005) , (Chiang et al., 2006)

Resources

Distributors

- Linguistic Data Consortium
- NEMLAR (Network for Euro-Mediterranean LAnguage Resources)
- ELSNET is the European Network of Excellence in Human Language Technologies
- ELDA Evaluation and Language resources Distribution Agency

Resources

Reports

- Mohamed Maamouri and Christopher Cieri. 2002. Resources for Natural Language Processing at the Linguistic Data Consortium. In Proceedings of the International Symposium on Processing of Arabic, pages 125--146, Manouba, Tunisia, April 2002.
- Mahtab Nikkhou and Khalid Choukri. Survey on Arabic Language Resources and Tools in the Mediterranean Countries.
- Arabic Information Retrieval and Computational Linguistics Resources (thanks to Doug Oard)

Resources

Monolingual Corpora

- Arabic Gigaword
- Arabic Newswire

Parallel Corpora

- United Nations Parallel Corpus
- Ummah Parallel Corpus
- Arabic News Translation
- Multiple-Translation Arabic

Treebanks

- Arabic Penn Treebank Webpage
  - Part 1 v 2.0, Part 2 v 2.0, Part 3 v 1.0, 10K-word English Translation
- Prague Arabic Dependency Treebank

Resources

Morphology

- **Buckwalter Arabic Morphological Analyzer**
  - **Version 1.0**, **Version 2.0**
- **Xerox Arabic Morphology (online)**

Dialect Resources

- CALLHOME Egyptian Arabic Transcripts
- CALLHOME Egyptian Arabic Speech
- Egyptian Colloquial Arabic Lexicon
- Levantine Arabic Resources
- http://www.orientel.org/
- http://www.appen.com.au/
- CADIM: http://www.ccls.columbia.edu/cadim

Resources

Dictionaries

- Buckwalter Stem Dictionary
- H. Anthony Salmone. An Advanced Learner's Arabic-English Dictionary encoded by the Perseus Project, Tufts University (contact: David Smith dasmith@perseus.tufts.edu)
- Ajeeb Arabic-English Dictionary (online)
- Al-Misbar Dictionary (online)
- Ectaco Bilingual Dictionary (online)

Online MT systems

- Ajeeb's Arabic-English Machine Translation (online)
- Al-Misbar English-Arabic Machine Translation (online)

Conferences and Workshops

with some focus on Arabic

- Parsing Arabic Dialects (JHU summer workshop 2005)
- ACL 2005 Workshop on Computational Approaches to Semitic Languages
- Arabic Language Resources and Tools Conference 2004 Cairo, Egypt
- WORKSHOP Computational Approaches to Arabic Script-based Languages (COLING 2004)
- Traitement Automatique du Langage Naturel (TALN ' 04)
- NIST MT EVAL (http://www.nist.gov/speech/tests/mt/)
- MT Summit IX Workshop on Machine Translation for Semitic Languages in 2003
- Novel Approaches to Arabic Speech Recognition (JHU summer workshop 2002)
- LREC 2002 Arabic Language Resources and Evaluation Workshop
- ACL 2002 Workshop on Computational Approaches to Semitic Languages
- International Symposium on Processing of Arabic 2002, Tunisia
- Workshop on ARABIC Language Processing: Status and Prospects (ACL/EACL 2001)
- Arabic Translation and Localisation Symposium (ATLAS 1999)
- Computational Approaches to Semitic Languages (COLING/ACL 1998)

References

Aljlayl, M. and O. Frieder. 2002. On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach. ACM Conference on Information and Knowledge Management.

Al-Sughaiyer, I. and I. Al-Kharashi. 2004. Arabic Morphological Analysis Techniques: A Comprehensive Survey. Journal of the American Society for Information Science and Technology. Volume 55 , Issue 3.

Beesley, K. 2001. Finite-State Morphological Analysis and Generation of Arabic at Xerox Research: Status and Plans in 2001. EACL workshop on Arabic Language Processing: Status and Prospects.

Bikel, D. 2002. Design of a Multi-lingual, Parallel-processing Statistical Parsing Engine. HLT.

Buckwalter, T. 2002. Buckwalter Arabic Morphological Analyzer Version 1.0. LDC catalog number LDC2002L49.

Cavalli-Sforza, V., A. Soudi, and T. Mitamura. 2000. Arabic Morphology Generation Using a Concatenative Strategy. ANLP.

Chiang, D., M. Diab, N. Habash, O. Rambow, and S. Shareef. 2006. Arabic Dialect Parsing. EACL.

Darwish, K. 2002. Building a Shallow Morphological Analyzer in One Day. ACL workshop on Computational Approaches to Semitic Languages.

Diab, M., K. Hacioglu and D. Jurafsky. 2004. Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. HLT-NAACL.

Fischer, W. 2001. A Grammar of Classical Arabic. Yale Language Series. Yale University Press. Translated by Jonathan Rodgers.

Habash, N. and O. Rambow. 2004. Extracting a Tree Adjoining Grammar from the Penn Arabic Treebank. TALN.

Habash, N. and O. Rambow. 2005a. Arabic Tokenization, Part-of-Speech Tagging in and Morphological Disambiguation One Fell Swoop. ACL.

References

Habash, N., O. Rambow and G. Kiraz. 2005b. Morphological Analysis and Generation for Arabic Dialects. *ACL workshop on Computational Approaches to Semitic Languages.*

Habash, N. 2004. Large Scale Lexeme Based Arabic Morphological Generation. TALN.

Habash, N. and F. Sadat. 2006. Arabic Preprocessing Schemes for Statistical Machine Translation. NAACL.

Habash, N. 2006. "Arabic Morphological Representations for Machine Translation." Book Chapter. In Arabic Computational Morphology: Knowledge-based and Empirical Methods. Editors A. van den Bosch and A. Soudi.

Habash, N., C. Mah, S. Imran, R. Calistri-Yeh, and P. Sheridan. 2006. The Design and Validation of an Arabic WordNet for Information Retrieval. LREC.

Khoja, S. 2001. APT: Arabic Part-of-Speech Tagger. NAACL Student Research Workshop.

Kiraz, G. 2001. Computational Nonlinear Morphology with Emphasis on Semitic Languages. Studies in Natural Language Processing. Cambridge University Press.

Kirchhoff, K., J. Bilmes, S. Das, N. Duta, M. Egan, G. Ji, F. He, J. Henderson, D. Liu, M. Noamany, P. Schone, R. Schwartz and D. Vergyri. 2003. Novel Approaches to Arabic Speech Recognition: Report from the 2002 Johns-Hopkins Summer Workshop. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing.

Lee, Y., K. Papineni, S. Roukos, O. Emam and H. Hassan. 2003. Language Model Based Arabic Word Segmentation. ACL.

Lee, Y. 2004. Morphological Analysis for Statistical Machine Translation. NAACL.

Maamouri, M., A. Bies, T. Buckwalter, M. Diab, N. Habash, O. Rambow, D. Tabessi. 2006. Developing and Using a Pilot Dialectal Arabic Treebank. LREC.

References

Rambow, O., D. Chiang, M. Diab, N. Habash, R. Hwa, K. Sima'an, V. Lacey, R. Levy, C. Nichols, and S. Shareef. 2005. Parsing Arabic Dialects. Final Report, JHU Summer Workshop.

Rogati, M., S. McCarley, and Y. Yang. 2003. Unsupervised Learning of Arabic Stemming Using a Parallel Corpus. ACL.

Smrž, O. and P. Zemánek. 2002. Sherds from an Arabic Treebanking Mosaic. Prague Bulletin of Mathematical Linguistics, (78).

Smith N., D. Smith, and R. Tromble. 2005. Context-Based Morphological Disambiguation with Random Fields. HLT-EMNLP.

Soudi, A., V. Cavalli-Sforza, and A. Jamari. 2001. A Computational Lexeme-Based Treatment of Arabic Morphology. ACL workshop on Arabic Natural Language Processing.

Xu J. 2002. UN Parallel Text (Arabic-English), LDC Catalog No.: LDC2002E15.

Žabokrtský, Z. and O. Smrž. 2003. Arabic Syntactic Trees: from Constituency to Dependency. EACL.

Zitouni, I., J. Olive, D. Iskra, K. Choukri, O. Emam, O. Gedge, M. Maragoudakis, H. Tropf, A. Moreno, A. Rodriguez, B. Heuft and R. Siemund. 2002. OrienTel: Speech-Based Interactive Communication Applications for the Mediterranean and the Middle East. ICSLP.

Conference/Institution Name Abbreviations

ANLP = Applied Natural Language Processing

ACL = Association for Computational Linguistics

ACM= Association for Computing Machinery

EMNLP = Empirical Methods to Natural Language Processing

EACL= European ACL

HLT = Human Language Technology Conference

ICSLP = International Conference on Spoken Language Processing

JHU = Johns Hopkins University

LREC = Language Resources and Evaluation *Conference*

LDC = Linguistic Data Consortium, University of Pennsylvania

NAACL = North American ACL

TALN =Traitement Automatique du Langage Naturel