MEADR 2009
Cairo, Egypt
April 21, 2009

Introduction to Arabic Natural Language Processing

Nizar Habash

Columbia University

Center for Computational Learning Systems

habash@ccls.columbia.edu

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

CADIM

CCLS

- 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

Road Map

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

Arabic Script

Arabic Script

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

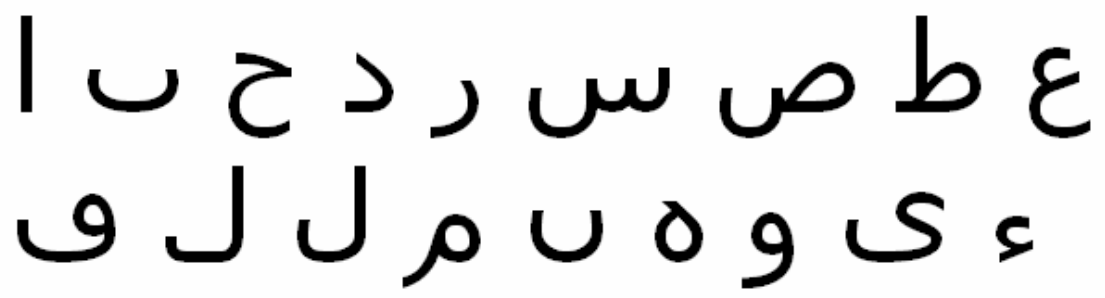
الخَطُّ العَرَبِي

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

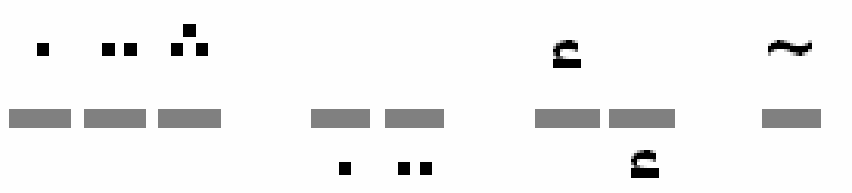
Arabic Script

Alphabet

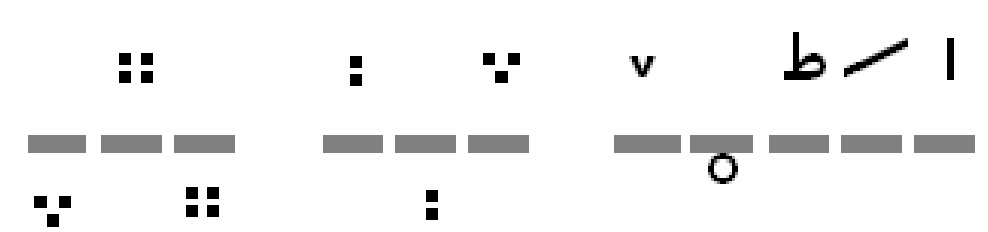
- letter forms



- letter marks
  - Arabic only



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



- *OCR output ambiguity; common spelling errors*

Arabic Script

Alphabet (MSA)

- letters (form+mark)
  - Distinctive

ب ت ث س ش

/b/ /t/ /θ/ /s/ /ʃ/

  - Non-distinctive

ا أ إ آ ى 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 (ا ؤ و ذ د ز ر ى ة) cause word-internal visual spacing

ز	د	ا	ن	ب	ك	م	ش	غ	**Stand alone**
ز	د	ا	ز	ب	ك	م	ش	غ	**initial**
ز	د	ا	ـن	ـب	ـك	ـم	ـش	ـغ	**medial**
ـز	ـد	ـا	ـن	ـب	ـك	ـم	ـش	ـغ	**final**

Arabic Script

Letter shaping

كتب = كتب ← ك ت ب

/katab/ k t b

to write

كتاب = كتاب ← ك ت ا ب

/kitāb/ k t ā b

book

Arabic Script

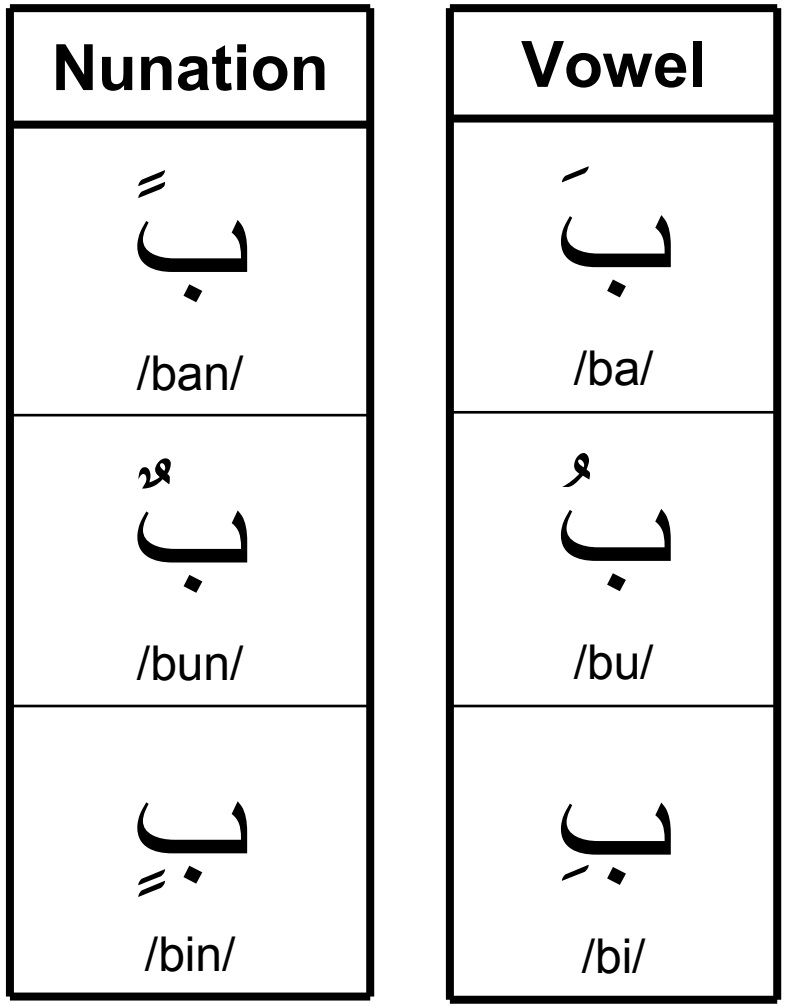
Diacritics

- Zero-width characters
- Used for short vowels

  كَتَبَ /katab/ *to write*

- Nunation is used for nominal indefinite marker in MSA

  كِتَابٌ /kitābun/ *a book*



Arabic Script

Diacritics

- No-vowel marker (*sukun*)

  مَكْتَب /maktab/ *office*

- Double consonant marker (*shadda*)

  كَتَّب /kattab/ *to dictate*

- Combinable بُّ بِّ بَّ

  /bbu/ /bbin/ /bban/

No Vowel
بْ
/b/

Double Consonant
بّ
/bb/

Arabic Script

Putting it together

Simple combination

Arab /ʕarab/ ع ر َ ب ← عَرَب = عرب

West /ʁarb/ غ ر ْ 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

حقـــوق الانســـان

حقـــــــوق الانســـــــان

حقـــــــــــوق الانســـــــــــان

human rights /ħuqūq alʔinsān/

Arabic Script

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

Arabic	Muhammad	algebra
عربي	محمد	الجبر
عربي	محمد	الجبر
عربي	محمد	الجبر
عربي	محمد	الجبر
/ʕarabi/	/muħammad /	/alʤabr/

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

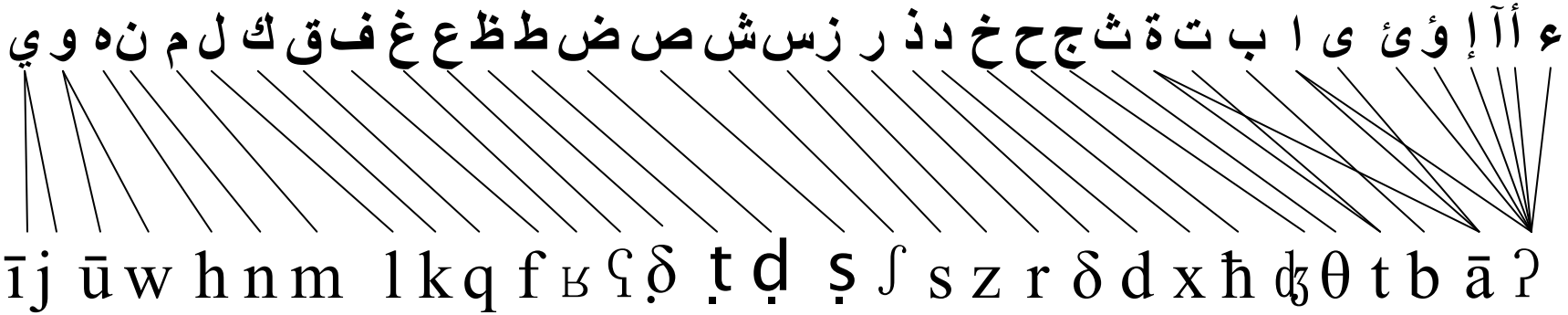
Western Arabic *Tunisia, Morocco, etc.*	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Indo-Arabic *Middle East*	٠	١	٢	٣	٤	٥	٦	٧	٨	٩
Eastern Indo-Arabic *Iran, Pakistan, etc.*	٠	١	٢	٣	۴	۵	۶	٧	٨	٩

Road Map

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



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 (/ʔ/,/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) بهاء 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

Proper Name Spelling

- The Qadafi-Schwarzenegger problem
  - Foreign Proper name spelling is often ad hoc
  - Multiplicity of spellings causes increased sparsity

قذافي	→	Gadafi Gaddafi Gaddfi Gadhafi Ghaddafi Kadaffy Qaddafi Qadhafi …
شوارزنيغر شوارزنغر شوارزنيجر شوارتزنجر	←	Schwarzenegger

Road Map

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

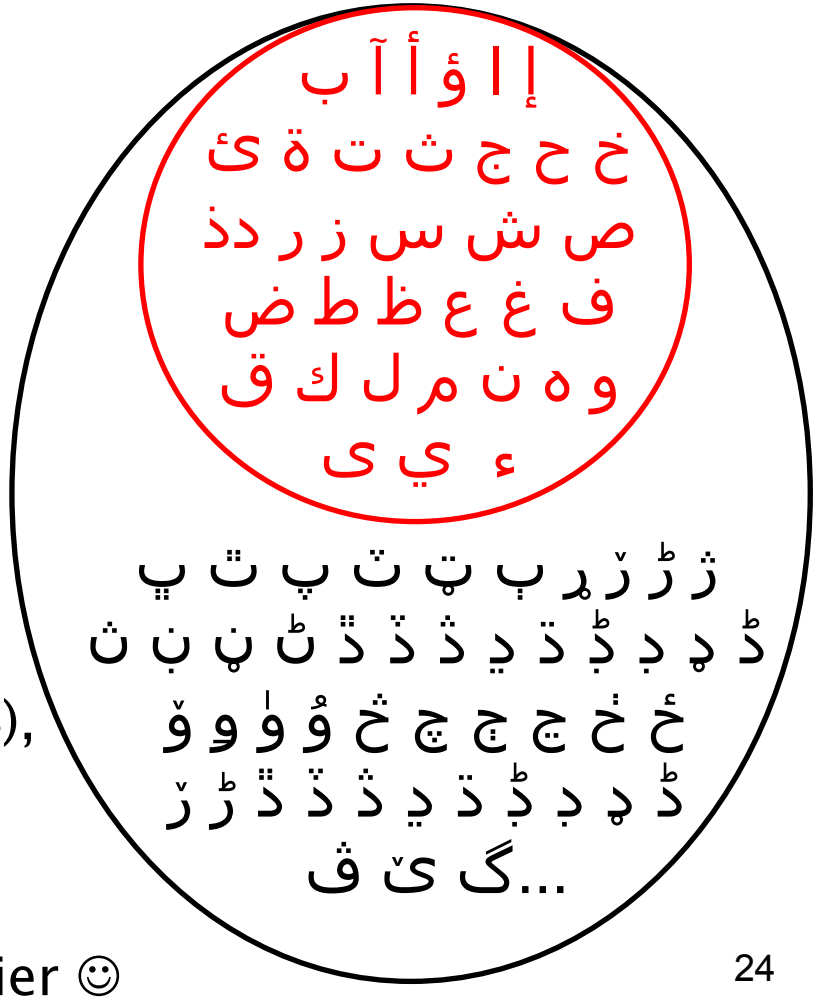
Arabic Script
Other languages

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 /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 ☺

☐ Not Arabic

بۆنه سووتێ جگه رو بۆچی نه بێ دڵ به که باب
بۆچی نه ڕوا له ته نم ڕۆحی ره وان میسلی شه هاب (١)
بۆله سه ر چاوه یی چاو هه ڵنه قوڵێ ڕه شحه یی خوێن (٢)
بۆچ له فه ووا ره یی موژگان نه تکێ قه تره 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 ڕۆژو شه ووم باعیسی ئارامی دڵم (٤)
ڕوێی وو من له غه می که وتمه نێو به حری عه زاب
به وقووعی سه فه ری قادری ئوستاد خدری (٥)
به جه فا عه یشمی تاڵ کرد فه له کی خانه خه راب
چه نکَ ونه ی لێ مه ده موتریب که له به ر فیرقه تی ئه و (٦)
رنه کی ڕۆحه له گوێم نه غمه ی ئاوازو روباب (٧)
ساغیر ی مه ی مه ده ساقی که له به ر دووریی نه و (٨)
تاڵه وه ک زه هری هه لایل له مه زاقم مه ی ناب (٩)

☐ 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

☐ 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

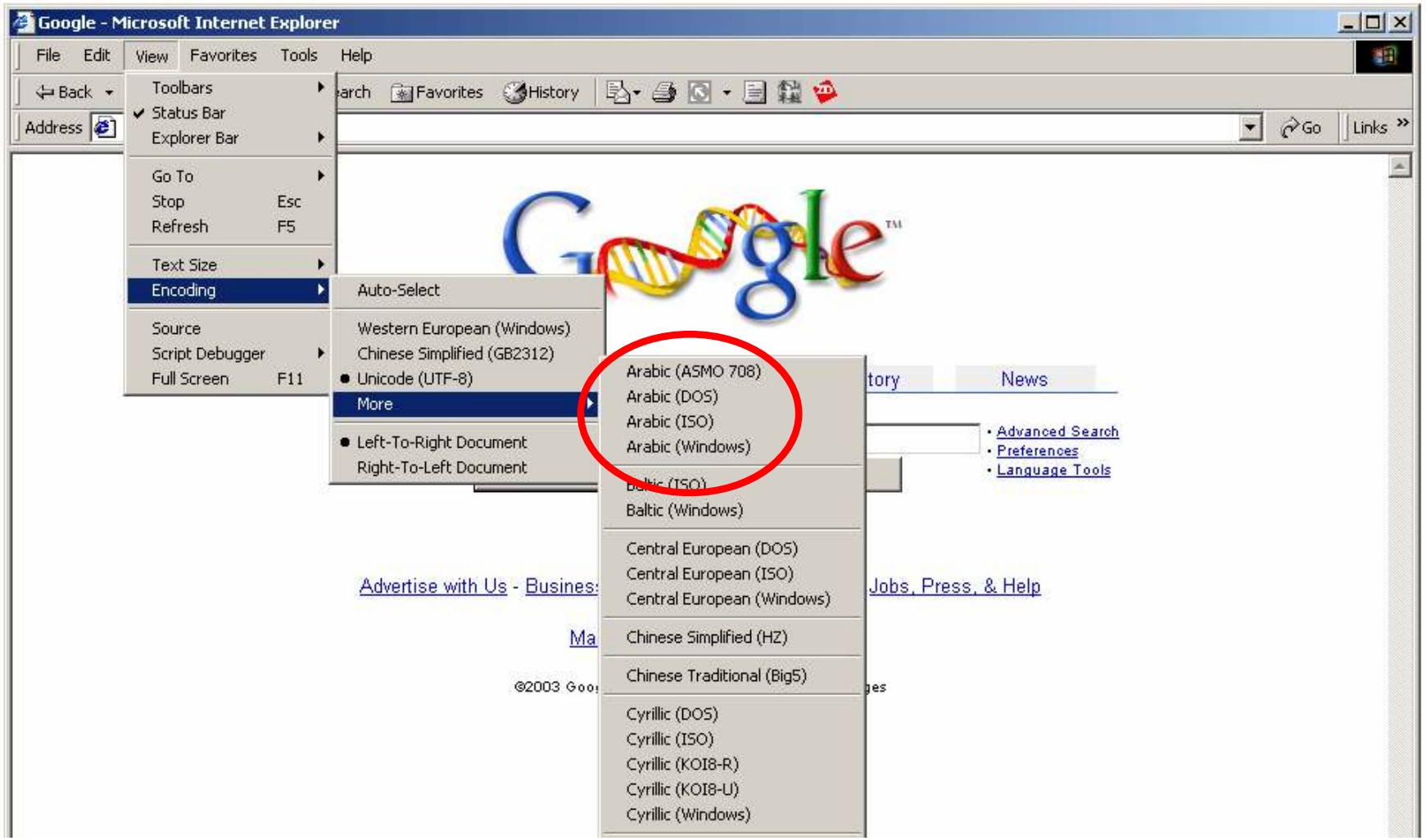
- Introduction
- Orthography
  - Arabic Script
  - MSA Phonology and Spelling
  - Recognizing Arabic vs. Persian/Urdu/Pashto/Kurdish/Sindhi/…
  - Encoding Issues
- Morphology
- Syntax

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

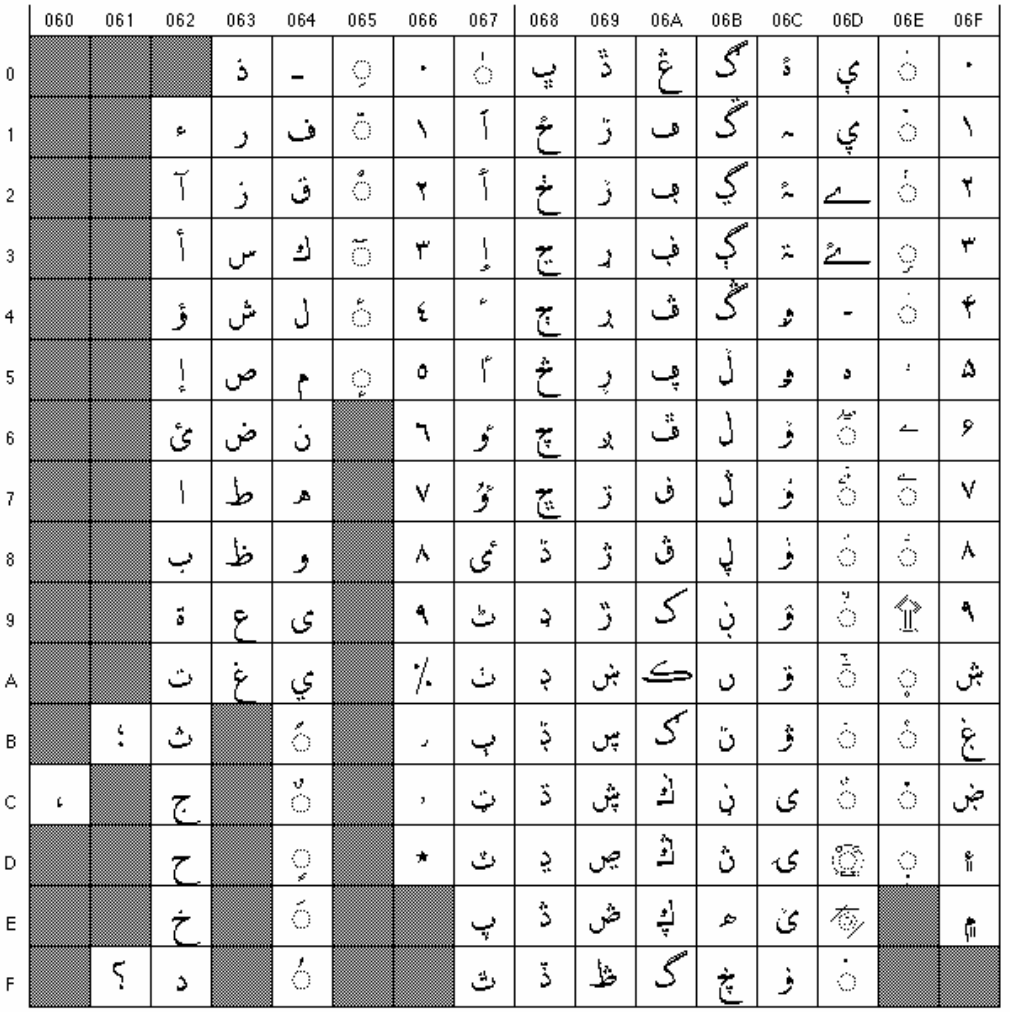
- CP-1256
  - Commonly used
  - 1-byte characters
  - Widely supported input/display
  - Minimal support for extended Arabic characters
  - bi-script support (Roman/Arabic)
  - Tri-lingual support: Arabic, French, English (ala ANSI)

Codepage 1256 - Arabic Windows

	-0	-1	-2	-3	-4	-5	-6	-7	-8	-9	-A	-B	-C	-D	-E	-F
0-	0000	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	ۜ 0686	ژ 0698	008F
9-	ک 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	00AA	« 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-	00C0	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



Encodings

- Unicode
  - Supports presentation forms (shapes and ligatures)

FE70 **Arabic Presentation Forms-B** **FEFF**

	FE7	FE8	FE9	FEA	FEB	FEC	FED	FEE	FEF
0	ً FE70	ء FE80	ب FE90	ج FEA0	ز FEB0	ض FEC0	غ FED0	ا FEE0	ى FEF0
1	ـً FE71	آ FE81	ب FE91	ح FEA1	س FEB1	ط FEC1	ف FED1	م FEE1	ي FEF1
2	ٌ FE72	آ FE82	ت FE92	ح FEA2	س FEB2	ط FEC2	ف FED2	م FEE2	ي FEF2
3	ـٌ FE73	أ FE83	ة FE93	ح FEA3	س FEB3	ط FEC3	ف FED3	م FEE3	ي FEF3
4	ٍ 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	ـَّ FC60	تر FC70	كا FC80	ى FC90	به FCA0	سم FCB0	فخ FCC0	مخ FCD0	ئه FCE0	يم FCF0	حي FD00	ضر FD10
1	لخ FC41	جح FC51	ـُّ FC61	تز FC71	كل FC81	ير FC91	تج FCA1	صح FCB1	فه FCC1	مم FCD1	به FCE1	يه FCF1	جى FD01	طى FD11
2	لم FC42	هم FC52	ـِّ 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	ـّ 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 دايبمل وا (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(في اولمبياد)enitselaP(شاركت فلسطين

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.

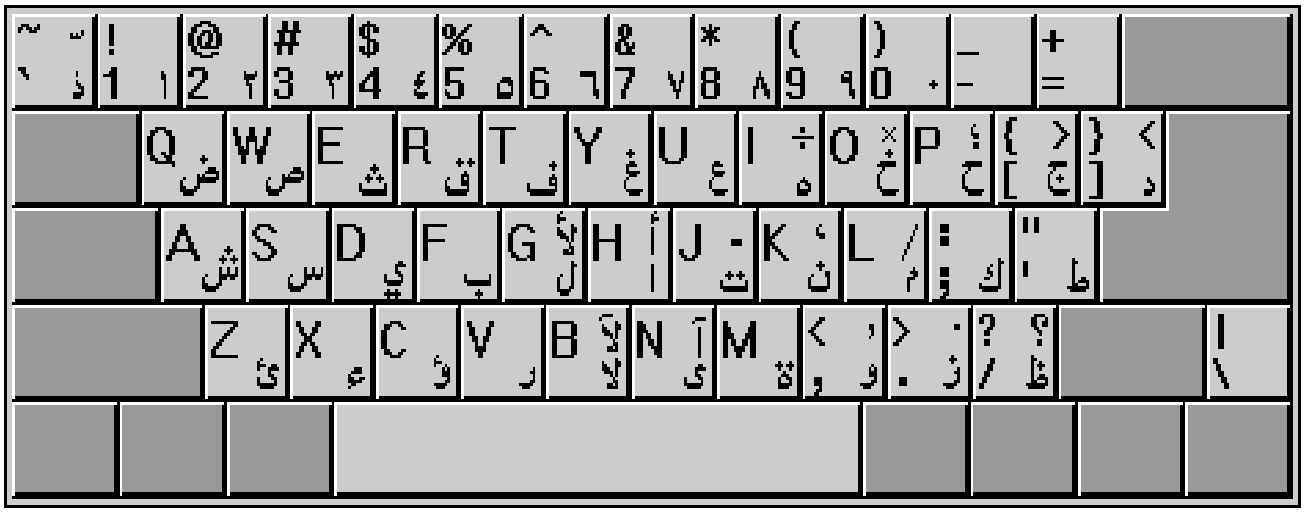
Display Problems

		Display Encoding			
		CP-1256	ISO-8859	Unicode	Western
Actual Encoding	CP-1256	تدشين منطقة حرة في دبي للتجارة الالكترونية	كلظ تدشِل□ة حرة ة □ٍ ترنلِة□دبٍ ففتجارة افاف	Y□ □□ɠ□ ɠ□ ψ□□Ŏɠāā□Ѧ□	ÊÏÔíä ãäØÞÉ ÍÑÉ Ýí ÏÈí áäÊÌÇÑÉ ÇáÇáßÊÑæäíÉ
	ISO-8859	ة حرة â×و هوêتدش ننتجارة êدب êل ةêوèانانمتر	تدشين منطقة حرة في دبي للتجارة الالكترونية	Y□楽既 ɠ□ ɠ□ψ□ □ŎɠGG ㈱親ɠ	ÊÏÔêæ åæ×âÉ ÍÑÉ áê ÏÈê ääÊÌÇÑÉ ÇäÇäãÊÑèæêÉ
	Unicode	ï» †ظٹظ´ط¯ط¨ط؟ ©ط±طط ©ط‚ط·ط†ظ…ظ ظٹط¨ط¯ط ظپظٹ ©ط±ط§ط¬ط¨ط„ظ„ظ ظ„ظ§ط„ظ§طƒ ط©ظ†ظˆظ±ط¨ط©	ظ؟؛□ظ□ظ□ع□ع□ عع□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»¿Ø ªØ¯Ø´ÙŠÙ† Ù…Ù†Ø·Ù'Ø© ØØ±Ø© Ù□ÙŠ 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



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	ً F
ة p	ظ Z	ٌ N
ت t	ع E	ٍ K
ث v	غ g	َ a
ج j	ـ _	ُ u
ح H	ف f	ِ i
خ x	ق q	ّ ~
د d	ك k	ْ o

Road Map

- Introduction
- Orthography
- Morphology
- Syntax

Morphology

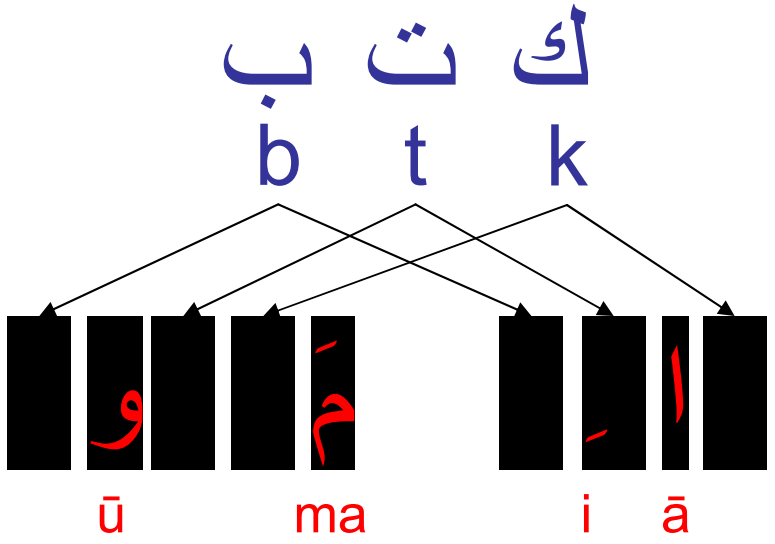
- Type
  - Concatenative: prefix, suffix, circumfix
  - Templatic: root+pattern
- Function
  - Derivational
    - Creating new words
    - *Mostly templatic*
  - Inflectional
    - Modifying features of words
      - Tense, number, person, mood, aspect
    - *Mostly concatenative*

Road Map

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

Derivational Morphology

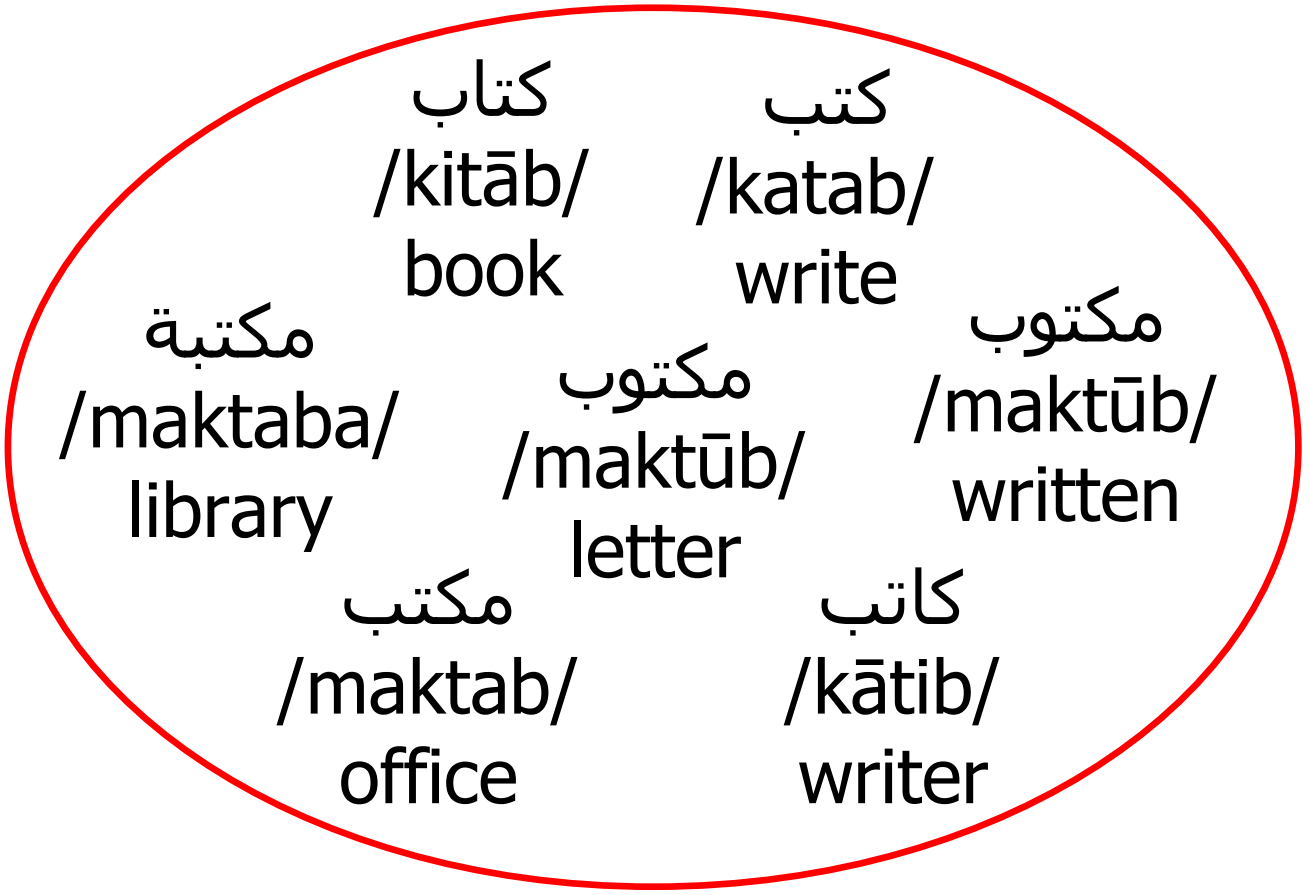
- Templatic Morphology
  - Root
  - Pattern
  - Lexeme



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

Derivational Morphology
Root Meaning

- ك ت ب KTB = notion of *"writing"*



Derivational Morphology
Root Meaning

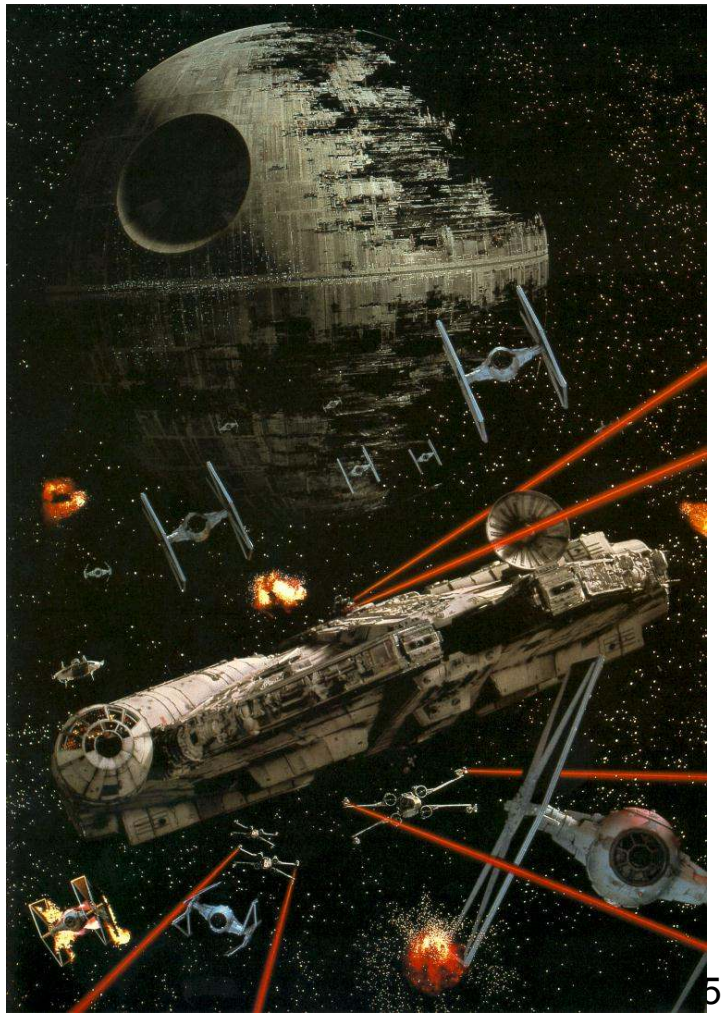
- LHM-1
- Notion of “meat”
  - لحم /laħm/
    - Meat
  - لحام /laħħām/
    - Butcher

لحم

laHm

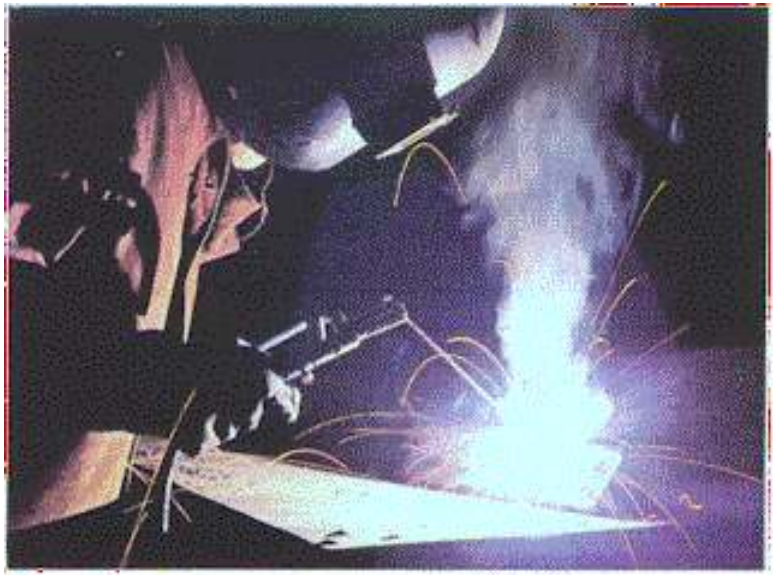
Derivational Morphology
Root Meaning

- 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

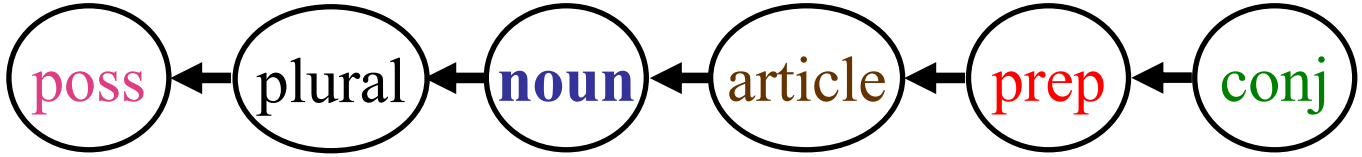
Inflectional Morphology

- Derivational Morphology
  - Lexeme ≈ Root + Pattern
- Inflectional Morphology
  - Word = Lexeme + Features
- Part-of-speech
  - *Traditional*: Noun, Verb, Particle
  - *Computational*: N, PN, V, Adj, Adv, P, Pron, Num, Conj, Det, Aux, Pun, IJ, and others
  - Many tag sets exist ranging from 3 to over 22K tags
- Noun-specific Features
- Verb-specific Features
- Other Features

Inflectional Morphology

- **Noun-specific Features**
  - Number: singular, dual, plural, collective
  - Gender: masculine, feminine
  - Definiteness: definite, indefinite
  - Case: nominative, accusative, genitive
  - Possessive clitic
- **Verb-specific Features**
  - Aspect: perfective, imperfective, imperative
  - Voice: active, passive
  - Tense: past, present, future
  - Mood: indicative, subjunctive, jussive
  - Subject (Person, Number, Gender)
  - Object clitic
- **Other Features**
  - Single-letter conjunctions
  - Single-letter prepositions

Inflectional Morphology
Nouns

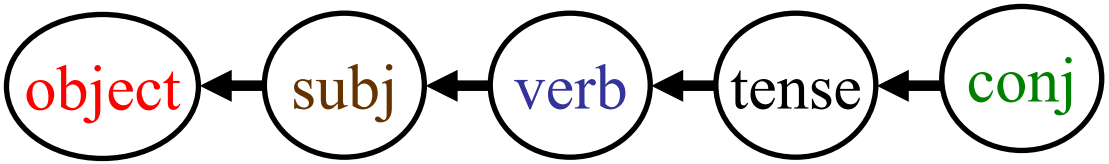


وكبيوتنا
/wakabiyūtinā/
و+ ك + بيوت + نا
wa+ka+biyūt+nā
and+like+houses+our
And like our houses

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



وسنقولها	فقلناها
/wasanaqūluhā/	/faqulnāhā/
و+ س+ ن+ قول + ها	ف+ قال+ نا+ ها
wa+sa+na+qūl+u+hā	fa+qul+na+hā
and+will+we+say+it	so+said+we+it
And we will say it	*So we said it*

- Morphotactics
- Subject conjugation (suffix or circumfix)

Inflectional Morphology

katab 'to write'

- **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 /taktub/: you write, she writes
  - Segmentation ambiguity
    - وجد: he found; و+جد: and+grandfather
    - للغة: لغة+ل: for a language; اللغة+ل: for the language

Morphological Ambiguity

- 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
- Multiple sources of ambiguity

بين

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

Morphological Disambiguation *in English*

- Select a morphological tag that fully describes the morphology of a word
- Complete English morphological tag set (Penn Treebank): 48 tags

  Verb:

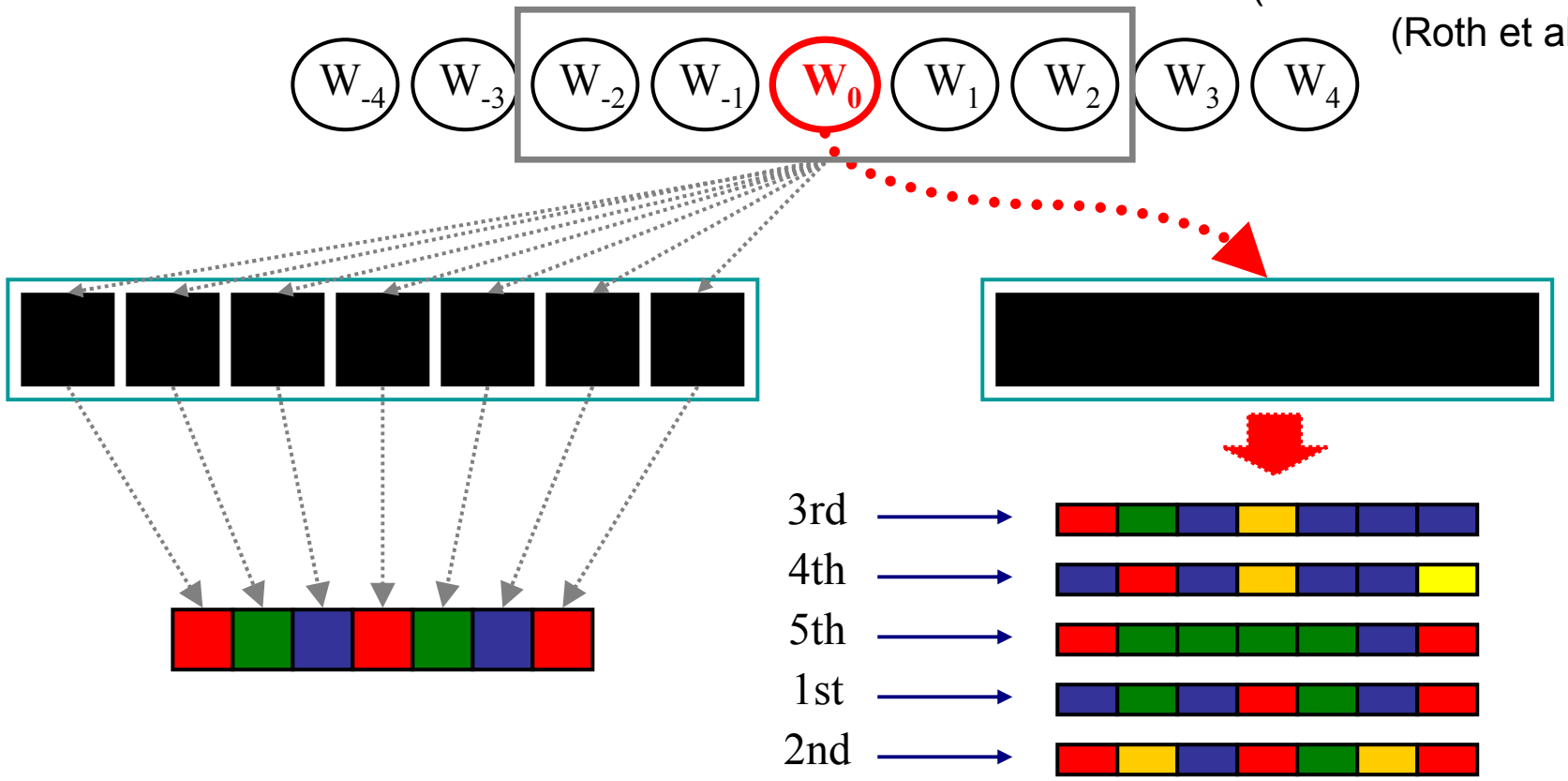
VB	VBD	VBG	VBN	VBP	VBZ
go	went	going	gone	go	goes

- Same as “POS Tagging” in English

Morphological Disambiguation *in Arabic*

- Morphological tag has 14 subtags corresponding to different linguistic categories
  - Example:Verb
    Gender(2), Number(3), Person(3), Aspect(3), Mood(3), Voice(2), Pronominal clitic(12), Conjunction clitic(3)
- 22,400 possible tags
  - Different possible subsets
- 2,200 appear in Penn Arabic Tree Bank Part 1 (140K words)
- Example solution: MADA (Habash&Rambow 2005)

MADA (Habash&Rambow 2005)
(Habash&Rambow 2007)
(Roth et al. 2008)



MORPHOLOGICAL CLASSIFIERS	RANKER	MORPHOLOGICAL ANALYZER
• Multiple independent classifiers • Corpus-trained	• Heuristic or corpus-trained	• Rule-based • Human-created

(Habash&Rambow 2005)
(Habash&Rambow 2007)
(Roth et al. 2008)

MADA

Morphological Analysis and Disambiguation for Arabic

```
;;; SENTENCE AsbAnyA tnfy tjmyd AlmsAEdp AlmmnwHp llmgrb
;;WORD AsbAnyA
;;MADA: AsbAnyA art-NO aspect-NA case-NOCASE clitic-NO conj-NO def-DEF mood-NA n
um-SG part-NO per-3 pos-PN voice-NA
*0.78571 <isobAniyA=[<isobAniyA_1 POS:PN BW:+<isobAniyA/NOUN_PROP+]=Spain
^0.71429 >asobAniyA=[<isobAniyA_1 POS:PN BW:+>asobAniyA/NOUN_PROP+]=Spain
_0.50000 <isobAniy~A=[<isobAniy~_2 POS:AJ +MASC +DU +NOM +POSS BW:+<isobAniy~/ADJ+A/NSUFF_MASC_DU_
_0.50000 <isobAniy~AF=[<isobAniy~_2 POS:AJ +ACC +INDEF BW:+<isobAniy~/ADJ+AF/CASE_INDEF_ACC]=Spanish/S
_0.57143 <isobAniy~A=[<isobAniy~_1 POS:N +MASC +DU +NOM +POSS BW:+<isobAniy~/NOUN+A/NSUFF_MASC_DU
_0.57143 <isobAniy~AF=[<isobAniy~_1 POS:N +ACC +INDEF BW:+<isobAniy~/NOUN+AF/CASE_INDEF_ACC]=Spanish/
--------------
;;WORD tnfy
;;MADA: tnfY art-NA aspect-IV case-NA clitic-NO conj-NO def-NA mood-I num-SG par
t-NO per-3 pos-V voice-ACT
*1.00000 tanofiy=[nafaY_1 POS:V +IV MOOD:I +S:3FS BW:ta/IV3FS+nofiy/IV+(null)/IVSUFF_MOOD:I]=disavow/deny/reje
_0.76923 tunofayo=[nafA-u_1 POS:V +IV +PASS MOOD:SJ +S:2FS
      BW:tu/IV2FS+nof/IV_PASS+ayo/IVSUFF_SUBJ:2FS_MOOD:SJ]=be_rejected/be_refuted/be_denied
_0.84615 tanif~iy=[naf~-i_1 POS:V +IV MOOD:SJ +S:2FS BW:ta/IV2FS+nif~/IV+iy/IVSUFF_SUBJ:2FS_MOOD:SJ]=blow_t
_0.84615 tanofiy=[nafA-u_1 POS:V +IV MOOD:SJ +S:2FS BW:ta/IV2FS+nof/IV+iy/IVSUFF_SUBJ:2FS_MOOD:SJ]=refute/
_0.84615 tanofiy=[nafaY_1 POS:V +IV MOOD:SJ +S:2FS BW:ta/IV2FS+nof/IV+iy/IVSUFF_SUBJ:2FS_MOOD:SJ]=disavow
_0.84615 tanofiya=[nafaY_1 POS:V +IV MOOD:S +S:2MS BW:ta/IV2MS+nofiy/IV+a/IVSUFF_MOOD:S]=disavow/deny/reje
_0.92308 tanofiy=[nafaY_1 POS:V +IV MOOD:I +S:2MS BW:ta/IV2MS+nofiy/IV+(null)/IVSUFF_MOOD:I]=disavow/deny/rej
_0.92308 tanofiya=[nafaY_1 POS:V +IV MOOD:S +S:3FS BW:ta/IV3FS+nofiy/IV+a/IVSUFF_MOOD:S]=disavow/deny/reje
--------------
```

Road Map

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

Arabic Computational Morphology

- Representation units
  - Natural token ولـلـمـكـتـبـــات
    - 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
    - وللـ+مكتب+ات
    - Can create more ambiguity
  - Lexeme + Features
    - مكتبة[+Plural +Def +و+ ل]
  - Root + Pattern + Features
    - كتب + مa3a21aة + [+Plural +Def +ل +و]
    - Very abstract
  - Root + Pattern + Vocalism + Features
    - كتب + م321aة + a.a.a + [+Plural +Def +ل +و]
    - Very very abstract

(Habash&Sadat 2006)

TOKAN

- A generalized tokenizer
- Assumes disambiguated morphological analysis
  - a la MADA
- Declarative specification of tokenization scheme

```
wsyktbhA=[katab_1 POS:V +IV w+ s+ +S:3MS +O:3FS]
```

Example	Scheme	Specification
w+ syktbhA	D1	w+ f+ REST
w+ s+ yktbhA	D2	w+ f+ b+ k+ l+ s+ REST
w+ s+ yktb +hA	D3	w+ f+ b+ k+ l+ s+ Al+ REST +P: +O:
w+ syktb +hA	TB	w+ f+ b+ k+ l+ REST +P: +O:
w+ s+ ktb/VBZ S:3MS +hA	EN	w+ f+ b+ k+ l+ s+ Al+ LEXEME + BIESPOS +S:

- Uses generator (Habash 2004)

Arabic Computational Morphology

- Approaches
  - Finite state machines (Beesely,2001) (Kiraz,2001) (Habash&Rambow 2006)
  - Concatenative analysis/generation (Smrz, 2007) (Buckwlater,2002) (Cavalli-Sforza et al, 2000)
  - Lexeme+Feature analysis/generation (Habash, 2004) (Habash&Rambow 2006)
  - 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)
  - Survey article: (Al-Sughaiyer&Al-Kharashi, 2004)
- 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

Morphology and Syntax

- Rich morphology crosses into syntax
  - Pro-drop / Subject conjugation
  - Verb sub-categorization and object clitics
    - $Verb_{transitive}$+subject+object
    - $Verb_{intransitive}$+subject *but not* $Verb_{intransitive}$+subject+object
    - $Verb_{passive}$+subject *but not* $Verb_{passive}$+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

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

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 *(aka nominal sentences)*
  - [Topic Complement]
  - الاولاد شعراء
    the-boys poets
    The boys are poets

Sentence Structure

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

Sentence Structure

- Verbal sentences
  - Common structural ambiguity
    - *Third masculine/feminine singular is structurally ambiguous*
      - $\text{Verb}_{\text{3MascSingular}}$ $\text{Noun}_{\text{Masc}}$
        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 /ʕindi 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$_{3rdMascPlur}$ poems]] The boys wrote the poems
      - Full agreement in this order (SVO)
      - الاشعار كتبها الاولاد

        [the-poems [wrote$_{3rdMascSing}$-them the boys]] The poems, the boys wrote

Road Map

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

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$_{FemSing}$ the-ambitious$_{FemSing}$
      - الكاتبات الطموحات the-writer$_{FemPlur}$ the-ambitious$_{FemPlur}$
    - Exception: Plural non-persons
      - definiteness agreement; feminine singular default
      - المكتب الجديد the-office$_{MascSing}$ the-new$_{MascSing}$
      - المكتبة الجديدة the-library$_{FemSing}$ the-new$_{FemSing}$
      - المكاتب الجديدة the-offices$_{MascBPlur}$ the-new$_{FemSing}$
      - المكتبات الجديدة the-libraries$_{FemPlur}$ the-new$_{FemSing}$

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

Agreement in Arabic

- Verb-Subject agreement
  - Verb agrees with subject in full (gender,number)
    - Exception: partial agreement (number=singular) in VSO order
    - Exception: partial agreement (number=singular; gender=feminine) for non-person plural subjects regardless of order
- Noun-Adjective
  - Adjective agrees with noun in full (gender, number, definiteness and case)
    - Exception: partial agreement (number=singular; gender=feminine) for non-person plural nouns
- Noun-Number
  - Number is the syntactic-case head
  - for numbers [3..10]: Noun is plural+genitive (idafa); number gender is inverted gender of noun!
  - for numbers [11..99]: Noun is singular+accusative (tamyiyz/specification); number gender is even more complicated ☺
  - for numbers [100,1K,1M]: Noun is singular+genitive (idafa)

bnyt *'was built'*	**vlAv *'three'***	**jAmEAt *'universities'***	**jdydp *'new'***
Fem+Sg	**Masc+Sg+Nom**	**Fem+PL+Gen**	**Fem+Sg+Gen**
Verbs in VSO order are always Sg and agree in gender only	**Numbers agrees by gender inversion**		**Adjectives of plural non-person nouns are Fem+Sg**

Road Map

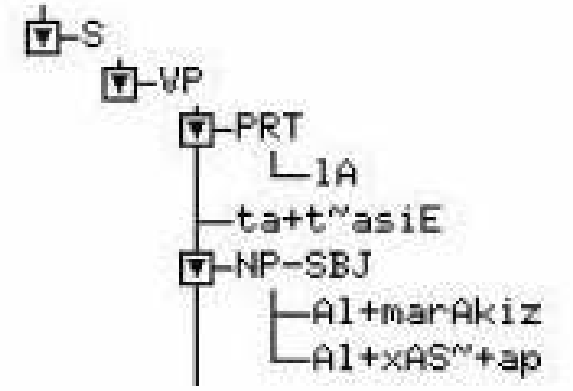
- Introduction
- Orthography
- Morphology
- Syntax
  - Morphology and Syntax
  - Sentence Structure
  - Phrase Structure
  - Computational Resources

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)

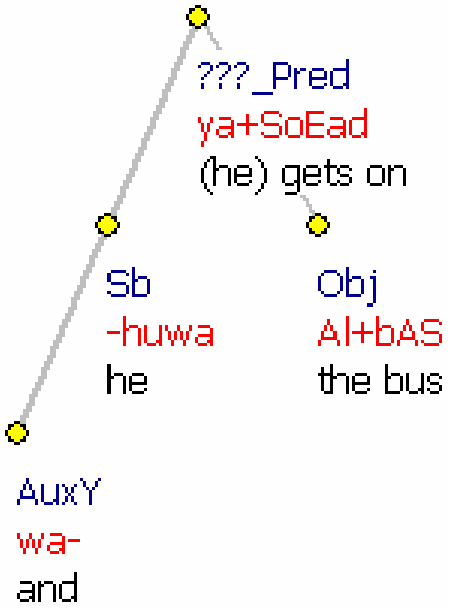
Computational Resources

- 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)



Computational Resources

- Prague Dependency Treebank
- Partial overlap with PATB and Arabic Gigaword
  - Agence France Presse, AlHayat and Xinhua
- Morphological analysis
  - Extends on PATB
- Dependency representation

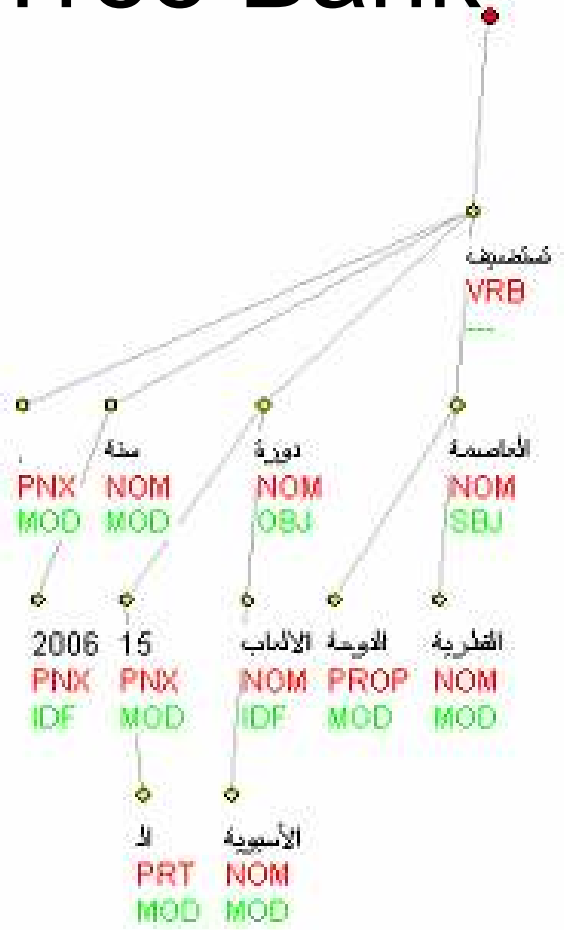


Graphic courtesy of Otakar Smrž: http://ckl.mff.cuni.cz/padt/PADT_1.0/docs/slides/2003-eacl-trees.ppt

(Habash, 2009)

CATiB:Columbia Arabic Tree Bank

- ***CATiB Representation***
  - ***Lite Dependency Syntax***
  - ***Tokenization***
    - CONJ PART BASE PRON
  - ***Part of Speech Tag set***
    - Six tags: VRB, VRB-pass, NOM, PROP, PRT, PNX
  - ***Syntactic Annotation***
    - Eight dependency relations: SBJ, OBJ, TPC, PRD, IDF, TMZ, MOD, FLAT



- ***Practical Considerations***
  - Less information to annotate
  - Dependencies easier to annotate than phrase structure
  - Terminology and representation close to traditional Arabic grammar

Computational Resources

- Applications using Arabic treebanks
  - Statistical parsing
    - Bikel's parser (Bikel 2003)
      - Same engine used with English, Chinese and Arabic
    - Nivre's MALT parser (Nivre et al. 2006)
  - Base-phrase Chunking
    - (Diab et al, 2004; Diab et al. 2007)
  - POS tagging and morphological disambiguation
    - (Diab et al, 2004; Diab et al. 2007; Habash and Rambow, 2005a; Smith et al., 2005; Roth et al. 2008)
    - Other non-treebank-based POS tagging efforts: (Khoja, 2001)
- Formalism conversion
  - Constituency to dependency (Žabokrtský and Smrž 2003; Habash et al. 2007; Tounsi et al., 2009)
  - Tree-adjoining grammar extraction (Habash and Rambow 2004)
- Automatic diacritization
  - Zitouni et al. (2006); Habash&Rambow (2007); Shaalan et al (2008) among others
  - Diacritization for MT (Diab et al. 2007)

Other Tutorial Slides

- **Columbia's Arabic Dialect Modeling Group (CADIM)**
  - http://www1.ccls.columbia.edu/~cadim/
    - Presentations

MEADR 2009
Cairo, Egypt
April 21, 2009

Introduction to Arabic Natural Language Processing

Nizar Habash

Columbia University

Center for Computational Learning Systems

habash@ccls.columbia.edu

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

CADIM

CLS