AMTA Association for Machine Translation in the Americas

AMTA – 2006 CONFERENCE

TUTORIAL ON

Arabic Dialect Processing

Presenters:
Mona Diab and Nizar Habash
Columbia University

BOSTON MARRIOTT CAMBRIDGE
CAMBRIDGE, MA

8 – 12 AUGUST 2006

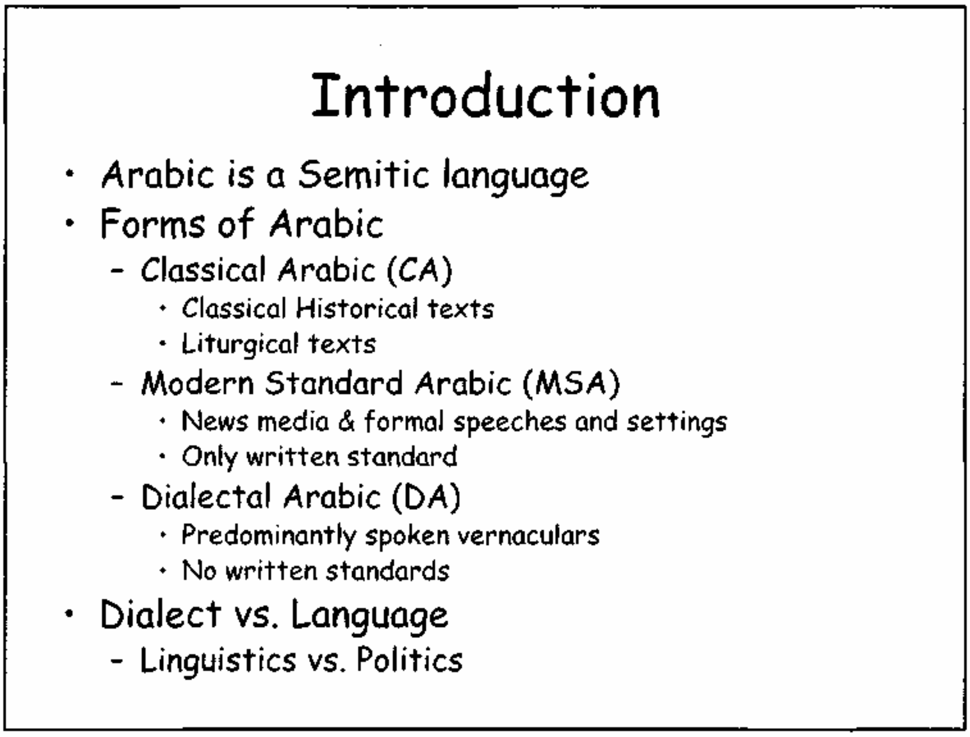
AMTA 2006, Boston, USA

Arabic Dialect Processing

Mona Diab   Nizar Habash
Center for Computational Learning Systems
Columbia University

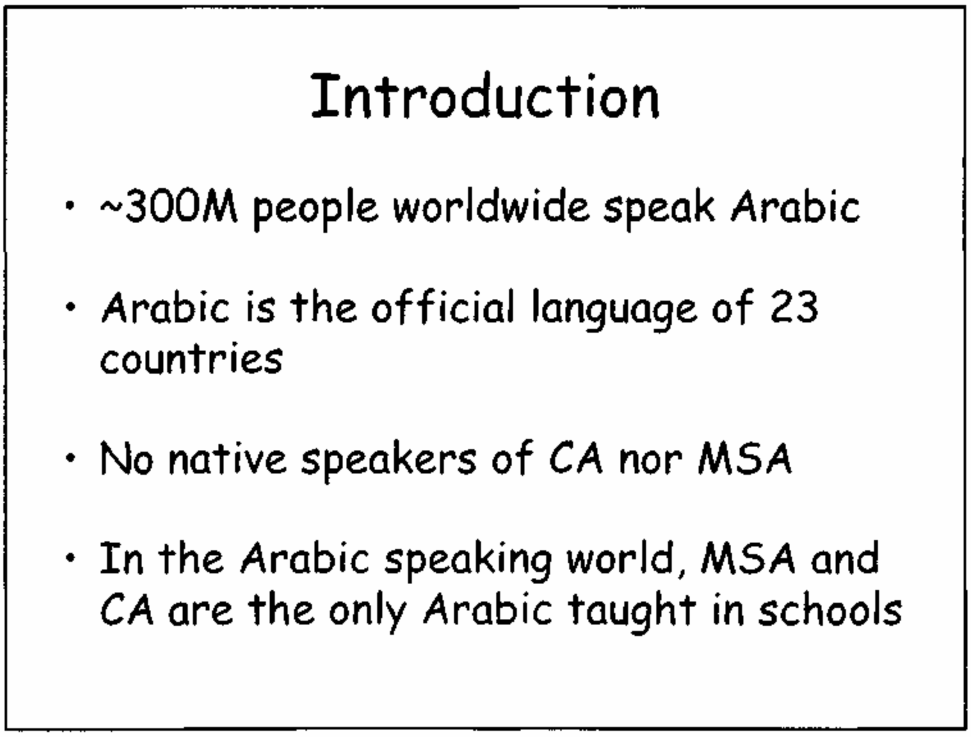
Tutorial Contents

- Introduction
- Description of MSA Phenomena
- Description of Dialectal Phenomena
- Sample Applications
- Resources and References



Introduction

- Arabic is a Semitic language
- Forms of Arabic
  - Classical Arabic (CA)
    - Classical Historical texts
    - Liturgical texts
  - Modern Standard Arabic (MSA)
    - News media & formal speeches and settings
    - Only written standard
  - Dialectal Arabic (DA)
    - Predominantly spoken vernaculars
    - No written standards
- Dialect vs. Language
  - Linguistics vs. Politics



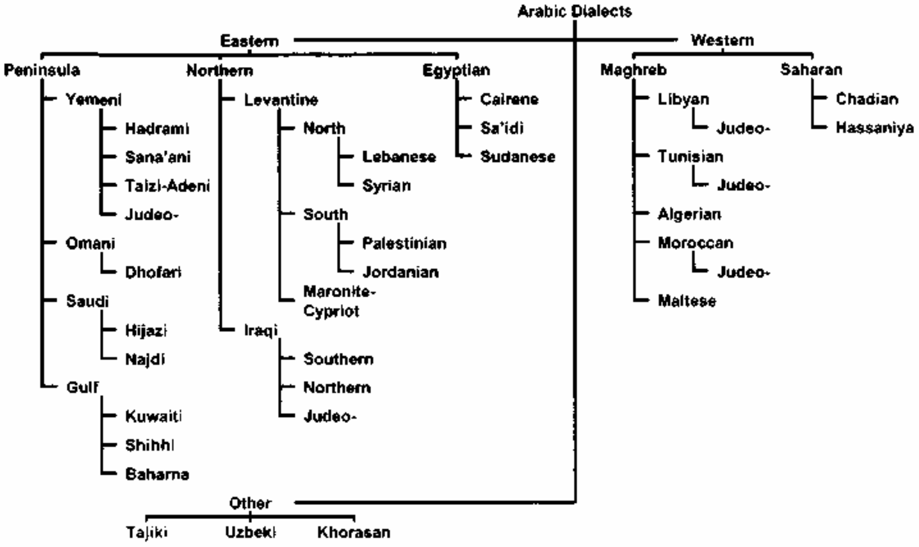
Introduction

- ~300M people worldwide speak Arabic
- Arabic is the official language of 23 countries
- No native speakers of CA nor MSA
- In the Arabic speaking world, MSA and CA are the only Arabic taught in schools

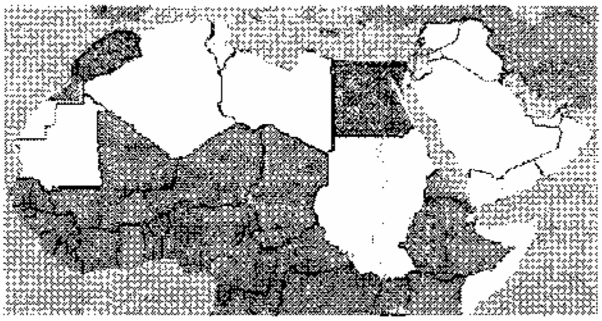
Introduction

- Arabic Diglossia
  - Diglossia is where two forms of the language exist side by side
  - MSA is the formal public language
    - Perceived as "language of the mind"
  - Dialectal Arabic is the informal private language
    - Perceived as "language of the heart"
- General Arab perception: dialects are a deteriorated form of Classical Arabic
- Continuum of dialects

Geographical Continuum



- Arabic Dialects
  - Eastern
    - Peninsula
      - Yemeni
        - Hadrami
        - Sana'ani
        - Taizi-Adeni
        - Judeo-
      - Omani
        - Dhofari
      - Saudi
        - Hijazi
        - Najdi
      - Gulf
        - Kuwaiti
        - Shihhi
        - Baharna
    - Northern
      - Levantine
        - North
          - Lebanese
          - Syrian
        - South
          - Palestinian
          - Jordanian
        - Maronite-Cypriot
      - Iraqi
        - Southern
        - Northern
        - Judeo-
    - Egyptian
      - Cairene
      - Sa'idi
      - Sudanese
  - Western
    - Maghreb
      - Libyan
        - Judeo-
      - Tunisian
        - Judeo-
      - Algerian
      - Moroccan
        - Judeo-
      - Maltese
    - Saharan
      - Chadian
      - Hassaniya
  - Other
    - Tajiki
    - Uzbeki
    - Khorasan



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

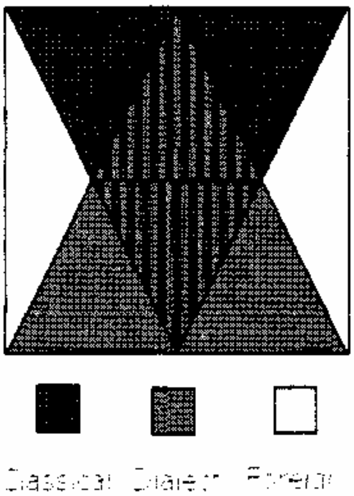
Nizar not-bought-not table new

Social Continuum

- Factors affecting dialect
  - Lifestyle
    - Bedouin, urban, rural
  - Education & Social Class
  - Religion
    - Muslim, Christian, Jewish, Druze, etc.
  - Gender

Social Continuum

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



Why Study Arabic Dialects?

- **Almost no** native speakers of Arabic sustain continuous spontaneous production of MSA
- Ubiquity of Dialect
  - Dialects are the primary form of Arabic used in all unscripted spoken genres: conversational, talk shows, interviews, etc.
  - Dialects are increasingly in use in new written media (newsgroups, weblogs, etc.)
  - Dialects have a direct impact on MSA phonology, syntax, semantics and pragmatics
  - Dialects lexically permeate MSA speech and text
- Substantial Dialect-MSA differences impede direct application of MSA NLP tools

Why Study Arabic Dialects?

- Degrees of linguistic distance

	Syntax	Morphology	Lexicon	Phonology
MSA-Dialect	++	+++	++++	++++
Inter-Dialect	+	+++	++++	++++
Intra-Dialect	0	0	+	+

- Lack of standards for the dialects
- Lack of written resources

A Note on Romanization

- Phonological Transcription
  - IPA
- Transliteration
  - Strict (one-to-one)
    - Buckwalter Encoding →
  - Loose
    - Many spelling variants
      - Qadafi, kadaphi, kaddafy, etc.
- This tutorial's examples are in
  - Arabic script سلام
  - Transcription (IPA) /salām/
  - Transliteration (Buckwalter) slAm

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

Tutorial Contents

- Introduction
- Description of MSA Phenomena
  - Orthography
  - Morphology
  - Syntax
- Description of Dialectal Phenomena
- Sample Applications
- Resources and References

Arabic Script

الخَطُ العَرَبِي

- An alphabet
- Written right-to-left
- Letter have allographic variants
- Optional diacritics
- Common ligatures
- Used to write many languages in addition to Arabic: Persian, Kurdish, Urdu, Pashto, etc.

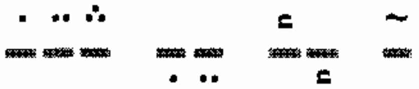
Arabic Script

Alphabet

- letter forms

ع ط ص س ر د ح ٮ ا
ء ى و ه ں م ل ٯ ڡ

- letter marks



Arabic Script

Alphabet

- letters (form+mark)
  - Distinctive

ب ت ث   س ش

/ʃ/ /s/ /θ/ /t/ /b/

  - Non-distinctive

ا أ إ آ ئ ؤ ء

/ʔ/

glottal stop aka hamza

Arabic Script

Diacritics

- Zero-width characters
- Used for short vowels

  كَتَب /katab/ *to write*

- Nunation is used for nominal indefinite marker in MSA

  كِتَابٌ /kitābun/ *a book*

Nunation
بً /ban/
بٌ /bun/
بٍ /bin/

Vowel
بَ /ba/
بُ /bu/
بِ /bi/

Arabic Script

Diacritics

- No-vowel marker (*sukun*)

  مَكْتَب /maktab/ *office*

- Double consonant marker (*shadda*)

  كَتَّب /kattab/ *to dictate*

- Combinable

بُّ	بٍّ	بًّ
/bbu/	/bbin/	/bban/

No Vowel
بْ /b/

Double Consonant
بّ /bb/

Arabic Script

Putting it together

Simple combination

Arab /ʕarab/ عرب = عَرَب ← عَ رَ ب

West /ʁarb/ غرب = غَرْب ← غَ رْ ب

Ligatures

Peace /salām/ سلام ~~سلام~~ ← س ل ا م

Arabic Script

"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.

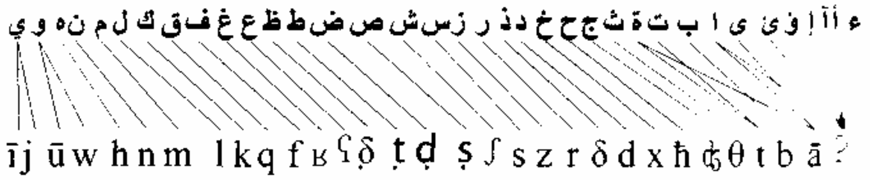
Algeria achieved its independence in 1962 after 132 years of French occupation.

- 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 IndoArabic *Iran, Pakistan, etc.*	۰	۱	۲	۳	۴	۵	۶	۷	۸	۹

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

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
    - كَتَب /katab/ to write كُتِب /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/)

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)

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

Tutorial Contents

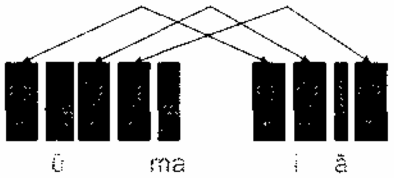
- Introduction
- Description of MSA Phenomena
  - Orthography
  - *Morphology*
  - Syntax
- Description of Dialectal Phenomena
- Sample Applications
- Resources and References

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

Derivational Morphology

- Templatic Morphology
  - Root ب ت ك
    b t k
  - Pattern

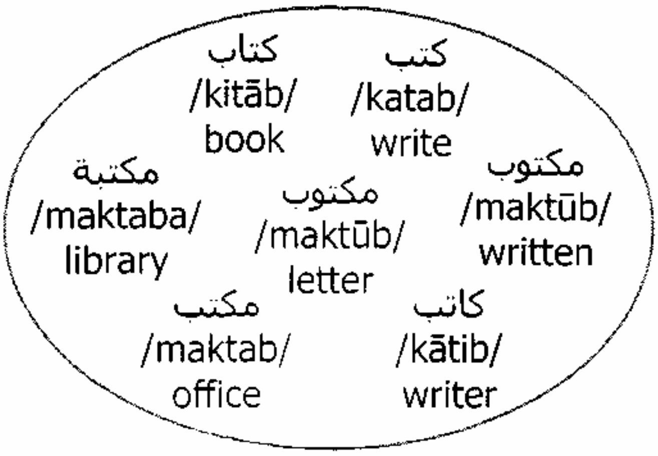


  - Lexeme مكتوب maktūb *written* كاتب kātib *writer*

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

Derivational Morphology
Root Meaning

- ب ت ك KTB = notion of *"writing"*



كتاب
/kitāb/
book

كتب
/katab/
write

مكتبة
/maktaba/
library

مكتوب
/maktūb/
letter

مكتوب
/maktūb/
written

مكتب
/maktab/
office

كاتب
/kātib/
writer

Root Polysemy

LHM-1 لحم

"meat"

لحم /laħm/

Meat

لحام /laħħām/

Butcher

LHM-2 لحم

"battle"

ملحمة /malħama/

Fierce battle

Massacre

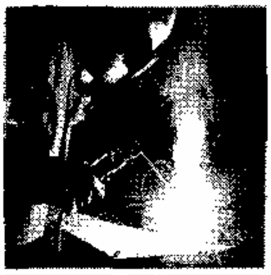
Epic

LHM-3 لحم

"soldering"

لحم /laħam/

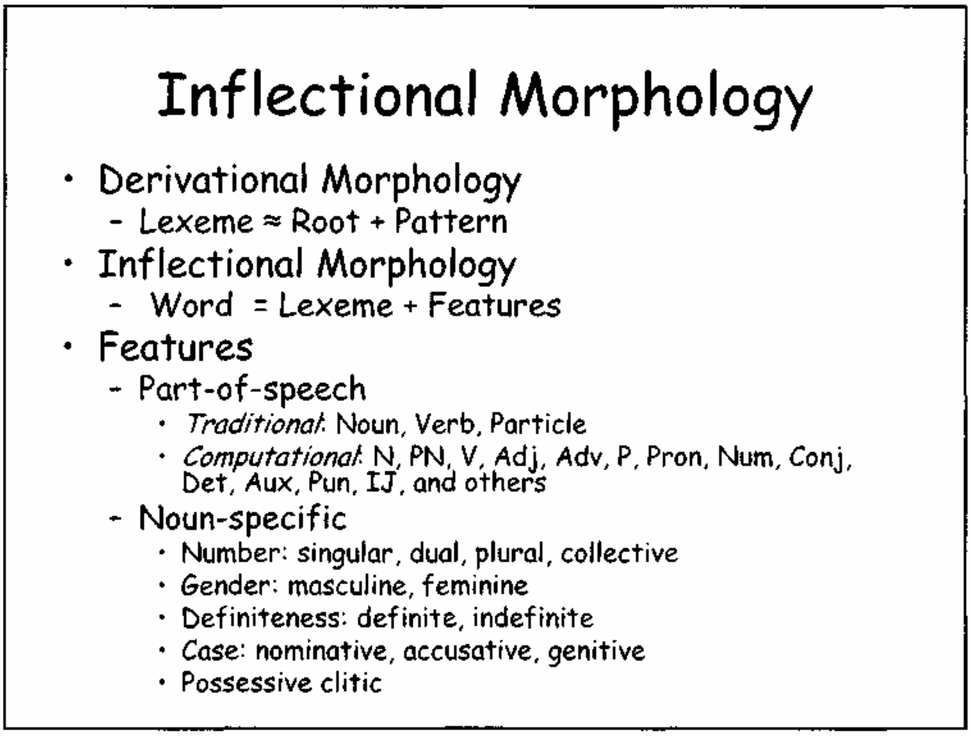
Weld, solder, stick, cling



Derivational Morphology
Pattern Meaning

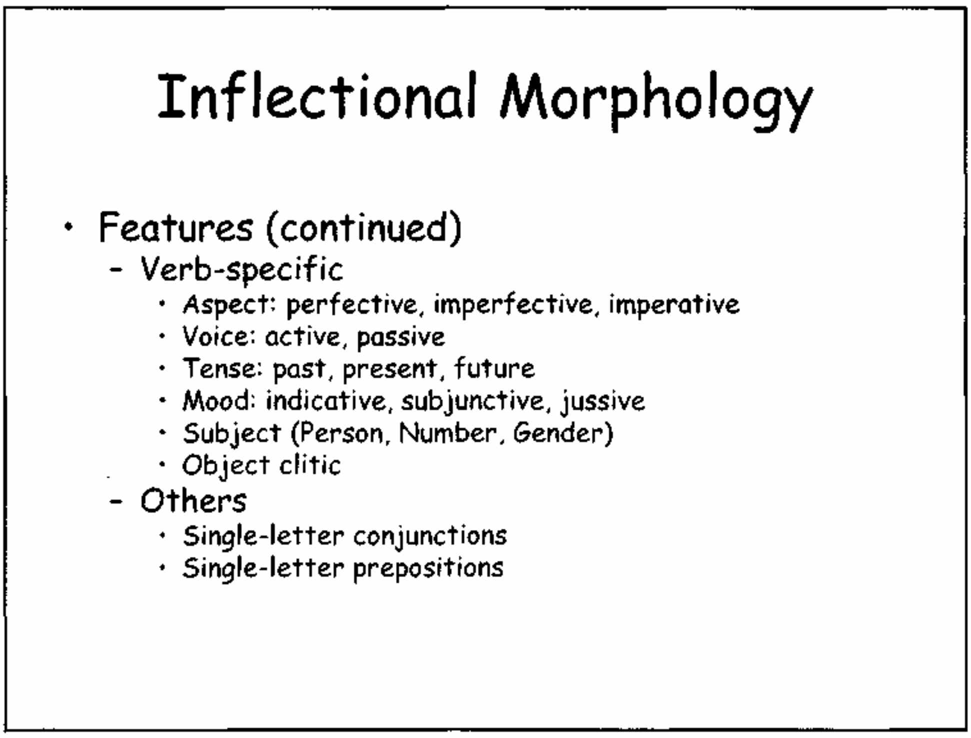
- Verb Pattern Meaning is hard to define systematically

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



Inflectional Morphology

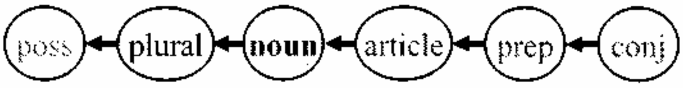
- Derivational Morphology
  - Lexeme ≈ Root + Pattern
- Inflectional Morphology
  - Word = Lexeme + Features
- Features
  - Part-of-speech
    - *Traditional:* Noun, Verb, Particle
    - *Computational:* N, PN, V, Adj, Adv, P, Pron, Num, Conj, Det, Aux, Pun, IJ, and others
  - Noun-specific
    - Number: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



Inflectional Morphology

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

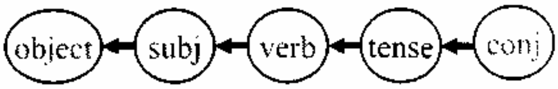
Inflectional Morphology
Nouns



وكبيوتنا	وللمكتبات
/wakabiyūtinā/	/walilmaktabāt/
و+ ك + بيوت + نا	و+ل+ال+مكتبة+ات
wa+ka+biyūt+nā	wa+li+al+maktaba+āt
and+like+houses+our	and+for+the+library+plural
And like our houses	*And for the libraries*

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

Inflectional Morphology
Verbs



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

Feminine form and other verb moods not shown

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
- 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*

Levels of Representation

Which level for NLP applications ?

• 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 + ŏa3a21aم + [+Plural +Def +ل +و]

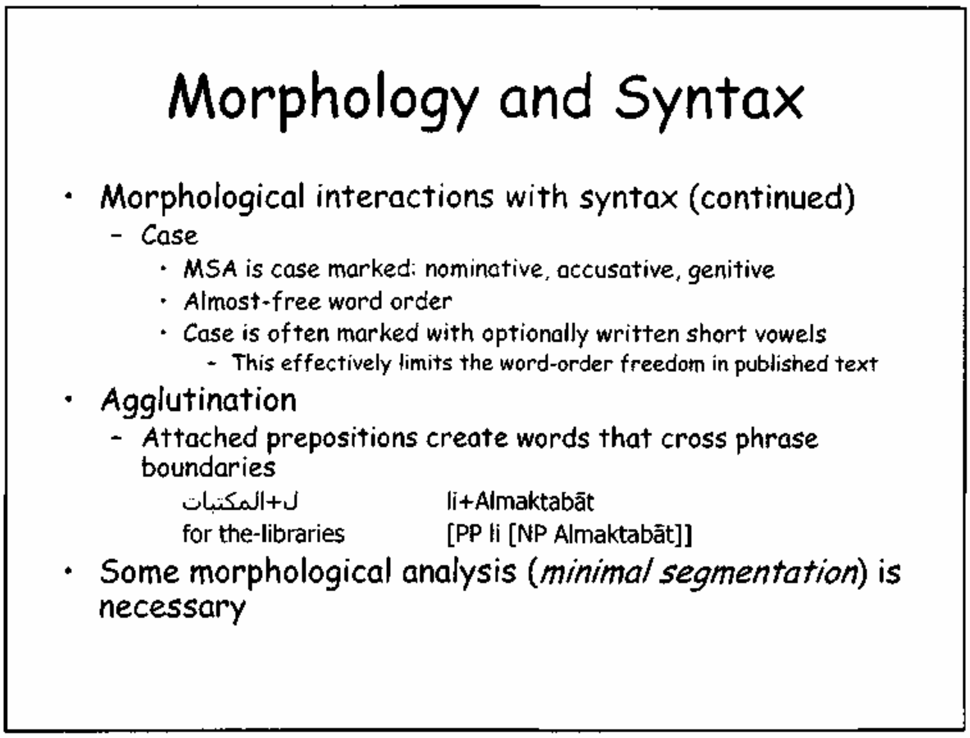
• etc.

Tutorial Contents

- Introduction
- Description of MSA Phenomena
  - Orthography
  - Morphology
  - Syntax
- Description of Dialectal Phenomena
- Sample Applications
- Resources and References

Morphology and Syntax

- Rich morphology crosses into syntax
  - Pro-drop / Subject conjugation
  - Verb subcategorization and object clitics
    - Verb$_{transitive}$+subject+object
    - Verb$_{intransitive}$+subject
    - Verb$_{passive}$+subject
- 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.

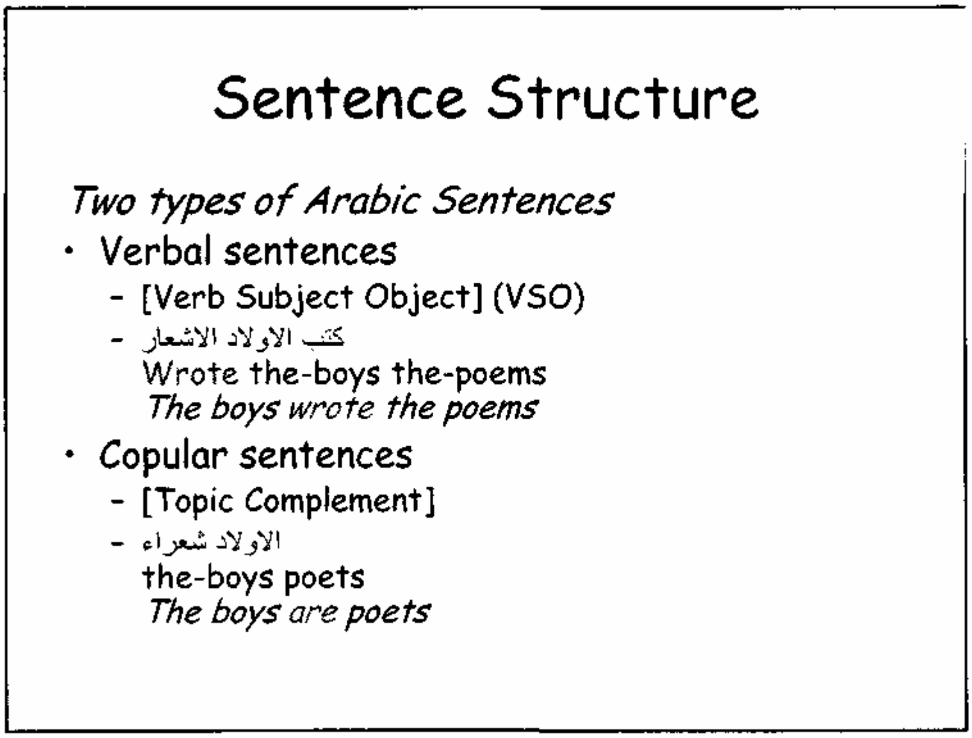


Morphology and Syntax

- Morphological interactions with syntax (continued)
  - Case
    - MSA is case marked: 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 (*minimal segmentation*) is necessary



Sentence Structure

Two types of Arabic Sentences

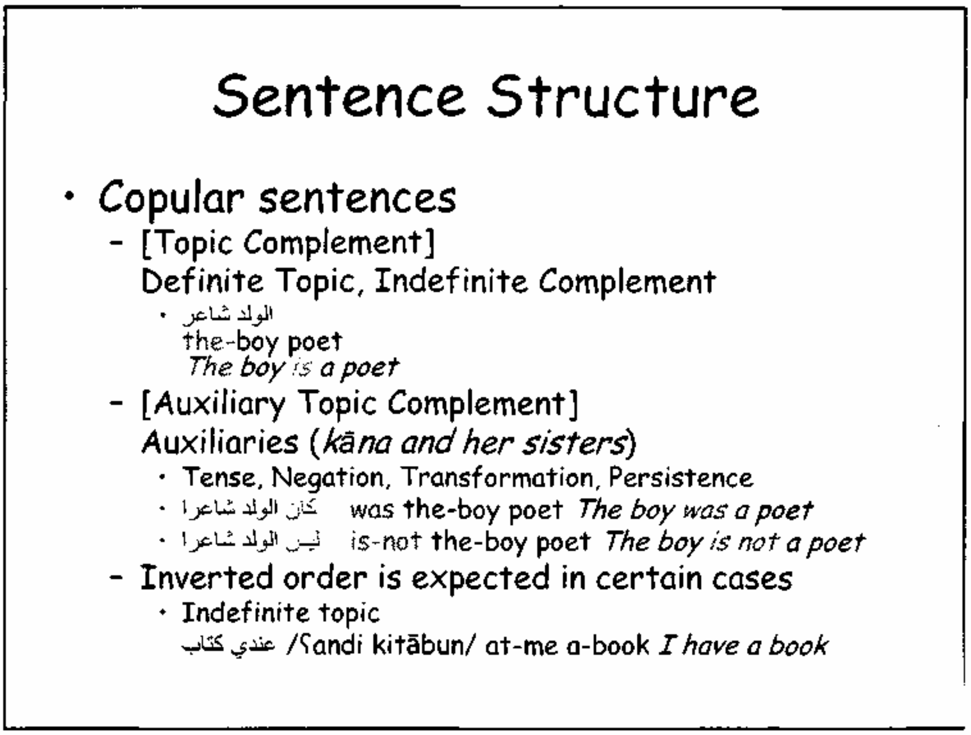
- Verbal sentences
  - [Verb Subject Object] (VSO)
  - كتب الاولاد الاشعار
    Wrote the-boys the-poems
    The boys wrote the poems
- Copular sentences
  - [Topic Complement]
  - الاولاد شعراء
    the-boys poets
    The boys are poets

Sentence Structure

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

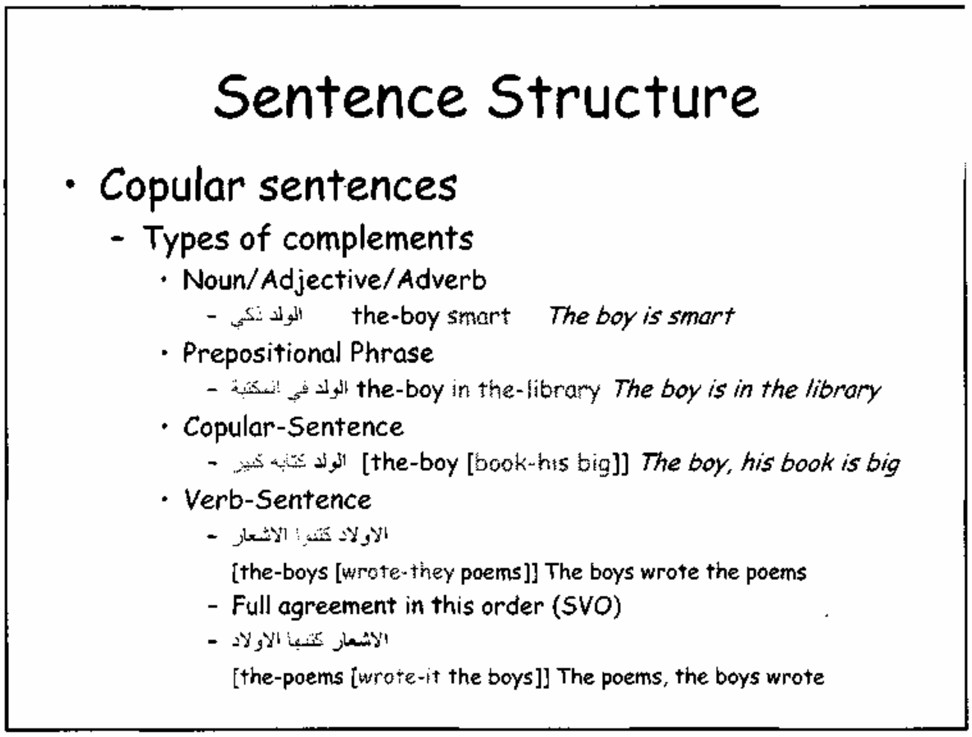
Sentence Structure

- Verbal sentences
  - Common structural ambiguity
    - *Third masculine/feminine singular are structurally ambiguous*
      - Verb$_{3MascSingular}$ Noun$_{Masc}$
        Verb subject=he object=Noun
        Verb subject=Noun
    - Passive and active forms are often similar in standard orthography
      - كَتَب /kataba/ he wrote
      - كُتِب /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

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

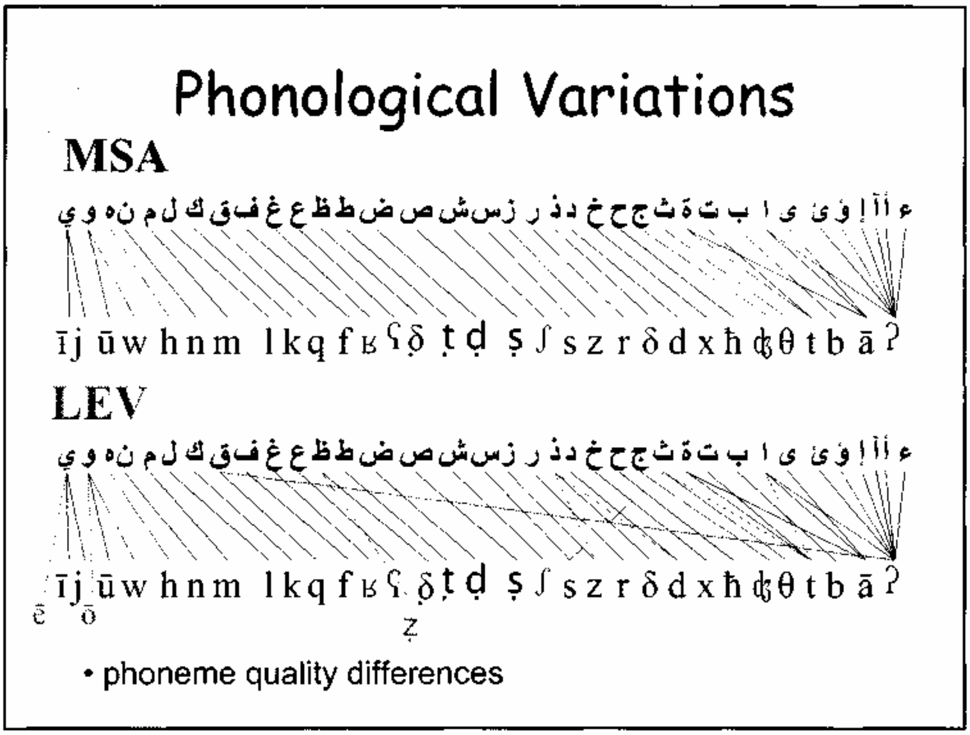
Phrase Structure

- 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

Tutorial Contents

- Introduction
- Description of MSA Phenomena
- Description of Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Resources and References



Phonological Variations

MSA

ء أ آ إ ئ ؤ 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

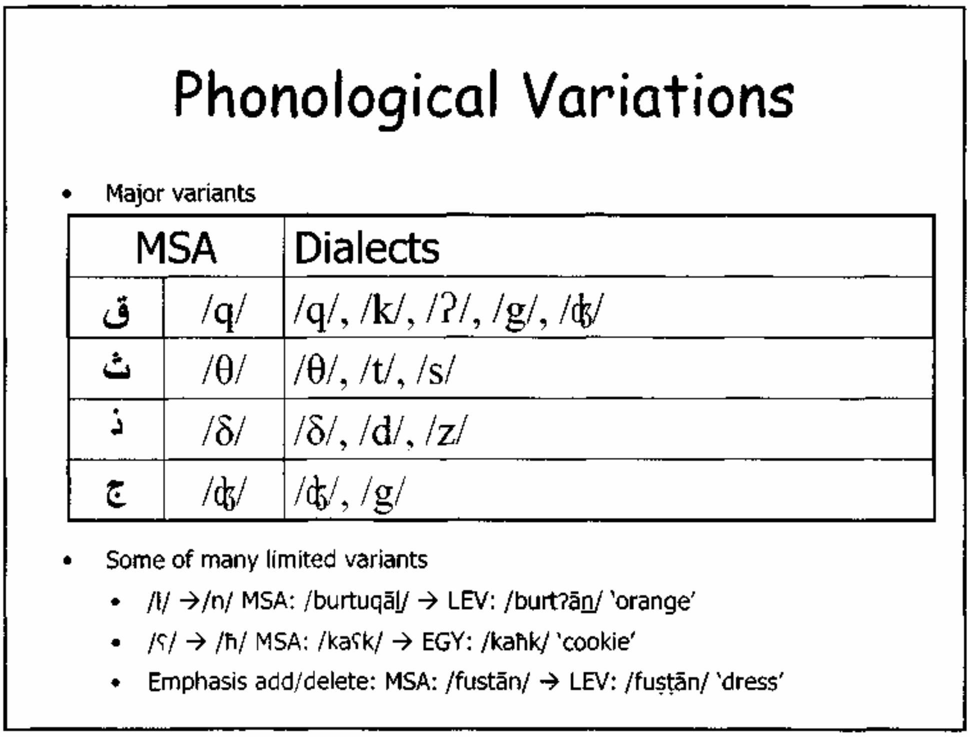
LEV

ء أ آ إ ئ ؤ 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

- phoneme quality differences



Phonological Variations

- Major variants

MSA		Dialects
ق	/q/	/q/, /k/, /ʔ/, /g/, /ʤ/
ث	/θ/	/θ/, /t/, /s/
ذ	/δ/	/δ/, /d/, /z/
ج	/ʤ/	/ʤ/, /g/

- Some of many limited variants
  - /l/ →/n/ MSA: /burtuqāl̲/ → LEV: /burtʔān̲/ 'orange'
  - /ʕ/ → /ħ/ MSA: /kaʕk/ → EGY: /kaħk/ 'cookie'
  - Emphasis add/delete: MSA: /fustān/ → LEV: /fuṣṭān/ 'dress'

Script Choices

- Arabic script:
  - + continuity with MSA
  - + masks the vocalic and some consonantal difference across dialects
  - - ambiguity
- Latin script
  - + precision
  - - lose connections among dialects (within dialects)
  - - politically loaded
- Other scripts
  - - Hebrew and Syriac
  - - Different religious/ethnic preferences

Arabic Script Orthographic Variants

	IRQ	LEV	EGY	TUN	MOR
/ʤ/	ج	ج	چ	ج	ج
/g/	گ	چ	ج	ڨ	ڭ
/tʃ/	چ	تش	تش	تش	تش
/p/	پ	پ	پ	پ	پ
/v/	ڤ	ڤ	ڤ	ڥ	ڥ

- Historical variants: MSA (ف, ق) = MOR (ڢ, ڧ)
- Modern proposals: LEV /ʔ/ قٔ , /ē/ يٖ, /ō/ وٚ (Habash 1999)

Syrian Arabic in Arabic Script

رح إحكي عنا نحن السوريين ..المعروفين بمأكولاتنا الشهية
واللذيذة والمميزة...مو بس هيك كل الخير فيها..دسمة وتقيلة
وعين الله ما بينقصها شي من المكسرات و..و...و..واللي لا
يمكن ترحمنا إذا ما رحمنا حالنا ..فبتلاقينا منهجم عالأكل يا
قاتل يا مقتول حتى التلت اللي لازم نتركه للنفس بديق بعيننا
و منعبيه أكل

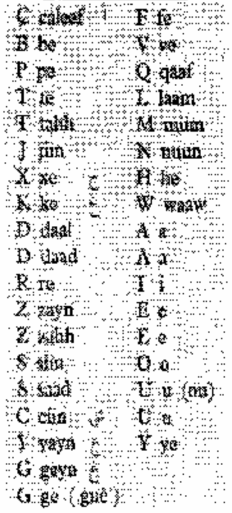
http://www.soriagate.net/showthread.php?t=32678

Latin Script

- Several proposals to the Arabic Language Academy in the 1940s
- Said Akl Experiment (1961)
- Web Arabic (Arabish, Franco-arabe)
  - No standard, but common conventions

عربي	IPA	Latin	عربي	IPA	Latin
الأعوزى	/ʔ/	` 2 0	ث	/θ/	th
ة	/a/,/t/	a t	ط	/ṭ/	t T 6
ح	ħ	H h 7	ع	/ʕ/	` 3 O
خ	/x/	kh 7' x 8	غ	/ʁ/	g gh 3'
ذ	/ð/	th	ق	/q/	q
ش	/ʃ/	sh ch	ي	/y/ /ay/ /I/ /ē/	y,i,e, ai,ei,...

Akl 1961



Ç caleef	ا	F fe
B be		V ve
P pe		Q qaaf
T te		L laam
Ŧ tahh		M mim
J jim		N nun
X xe	ح	H he
K ke	خ	W waaw
D daal		A a
Đ dhad		À a
R re		I i
Z zayn		E e
Ƶ zahh		È e
S sin		O o
Ş saad		U u (ou)
C cin	ج	Ü u
Y yayn	ع	Y ye
G gayn	غ	
G ge (guê)		

Egyptian Arabic in Latin Script

nadeity bsho2 nadeit
olteely ta3ala geit
laha3atbek 3alli fat
wala 7atta haloom 3aleiky
adeeni rge3telek
adeeni bein edeiky
kefaya dmoo3 ba2a
mush 3aref ashoof 3eneiky

http://www.jencomics.com/artist_a/amr_diab_lyrics/adeeni_rege3tilik_lyrics.htm

The Case of Maltese

- An Arabic dialect that is considered a separate language
- Standardized Latin-based orthography

Kulħadd hu intitolat għal dawn il-jeddijiet u l-libertajiet imxandra f'din l-Istqarrija, bla ebda għażla, bħal ta' razza, lewn, sess, ilsien, reliġjon, opinjoni politika jew kull opinjoni oħra, oriġini nazzjonali jew soċjali, proprjetà, twelid jew kull qagħda oħra. Mhux biss, iżda l-ebda għażla m'għandha ssir fuq bażi tal-qagħda politika, ġuridika jew internazzjonali tal-pajjiż jew territorju li minnu tiġi l-persuna kemm jekk ikun indipendenti, kemm jekk ikun fdat lil xi pajjiż ieħor, m'għandux gvern tiegħu jew għandu xi limiti oħra fis-sovranitŕ tiegħu.

http://www.language-museum.com/m/maltese.htm

Hebrew Script

- Example from Tunisian Judeo-Arabic

"The Ballad of Hannah and her Seven Sons "

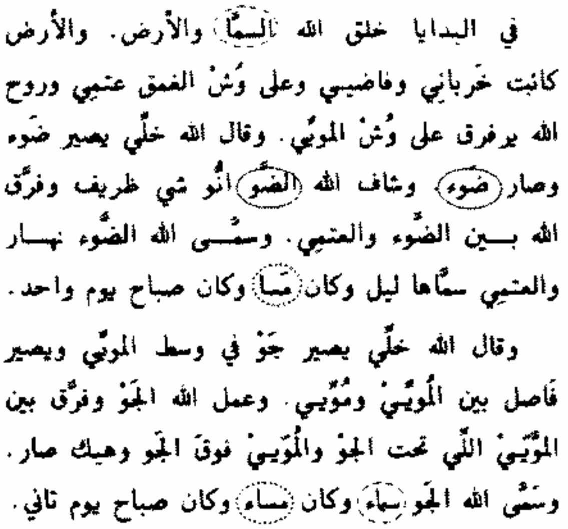
קצת חנה וזכריה
א   אסמעה קולי אנא חנה ואנע'רד מא ג'רא לי
לי סבע בנין באל כרם ועז ובאל דלאלי. וכאן
ביהום ולד זג'יר וגהו יע'וי כאל הלאלי ווקעו פי יד כאפר
מא יכאף מן רב אל עאלי. יעלח לנא לנבכי טול אל
איאם ולייאלי

http://www.uwm.edu/~corre/jatexts/

Lack of Orthographic Standards

- Orthographic inconsistency
- Egyptian /mabinʔulhalakʃ/
  - mA binquwlhA lak$ ما بنقولها لكش
  - mAbin&ulhalak$ مابنؤلهالكش
  - mA bin}ulhAlak$ ما بنئلهالكش
  - mA binqulhA lak$ ما بنقلها لكش
  - ...

Spelling Inconsistency I



في البدايا خلق الله السمَا والأرض. والأرض
كانت خَربانِي وفاضيي وعلى وُشْ الغمق عتمِي وروح
الله يرفرق على وُشْ المويّ. وقال الله خلّي يصير ضَوء
وصار ضوء. وشاف الله الضَّو أنُّو شي ظريف وفرّق
الله بين الضّوء والعتمِي. وسمّى الله الضّوء نهار
والعتمِي سمّاها ليل وكان مَسا وكان صباح يوم واحد.

وقال الله خلّي يصير جَوْ في وسط المويّ ويصير
فَاصل بين المُويّيْ ومْوّيي. وعمل الله الجَوْ وفرّق بين
المْوّيْ اللّي تحت الجوْ والمُوّيي فوقَ الجَو وهيك صار.
وسَمّى الله الجَو سَماء وكان مساء وكان صباح يوم تاني.

http://www.language-museum.com/a/arabic-north-levantine-spoken.ph

Spelling Inconsistency II

- ya alain lesh el 2aza
  ti7keh 3anneh kaza w kaza
  iza bidallak ti7keh hek
  2areeban ra7 troo7 3al 3aza

  chi3rik 3emilleh na2zeh
  li2anneh manneh mi2zeh
  bass law baddik yeha 7arb
  fikeh il layleh ra7 3azzeh

http://www.onelebanon.com/forum/archive/index.php/t-8236.html

Tutorial Contents

- Introduction
- Description of MSA Phenomena
- Description of Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Resources and References

Lexical Variation

- Arabic Dialects vary widely lexically

English	Table	Cat	Of	I_want	There_is	There_isn't
MSA	Tāwila طاولة	qiTTa قطة	*idafa* Ø	‘uridu اريد	yūjadu يوجد	lā yujadu لا يوجد
Moroccan	mida ميدة	qeTTa قطة	dyāl ديال	byīt بغيت	kāyn كاين	mā kāynš ما كاينش
Egyptian	Tarabēza طربيزة	‘oTTa قطة	bitāç بتاع	çāwez عاوز	fī في	mafīš مفيش
Syrian	Tāwle طاولة	bisse بسة	tabaç تبع	biddi بدي	fī في	mā fī ما في
Iraqi	mēz ميز	bazzūna بزونة	māl مال	‘arīd اريد	aku اكو	māku ما

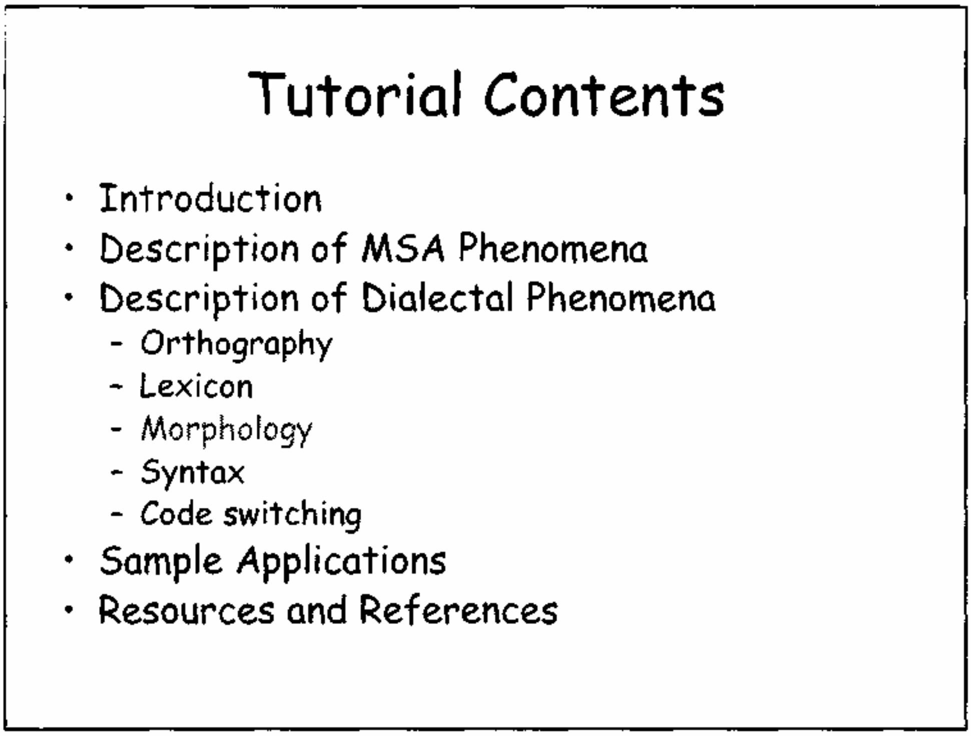
- Arabic orthography allows consolidating some variations

Lexical Variation

خلف	EGY:reproduce - GLF: give condolences
مكوى	EGY:press iron - GLF:buttocks
براد	EGY:kettle - LEV:fridge
مرا	EGY:prostitute - LEV:woman
ماشي	EGY/LEV:okay - MOR:not
بسط	EGY/LEV:make happy - IRQ:beat up
العافية	EGY/LEV:health - MOR:hell fire
بلش	LEV:start - SUD:end

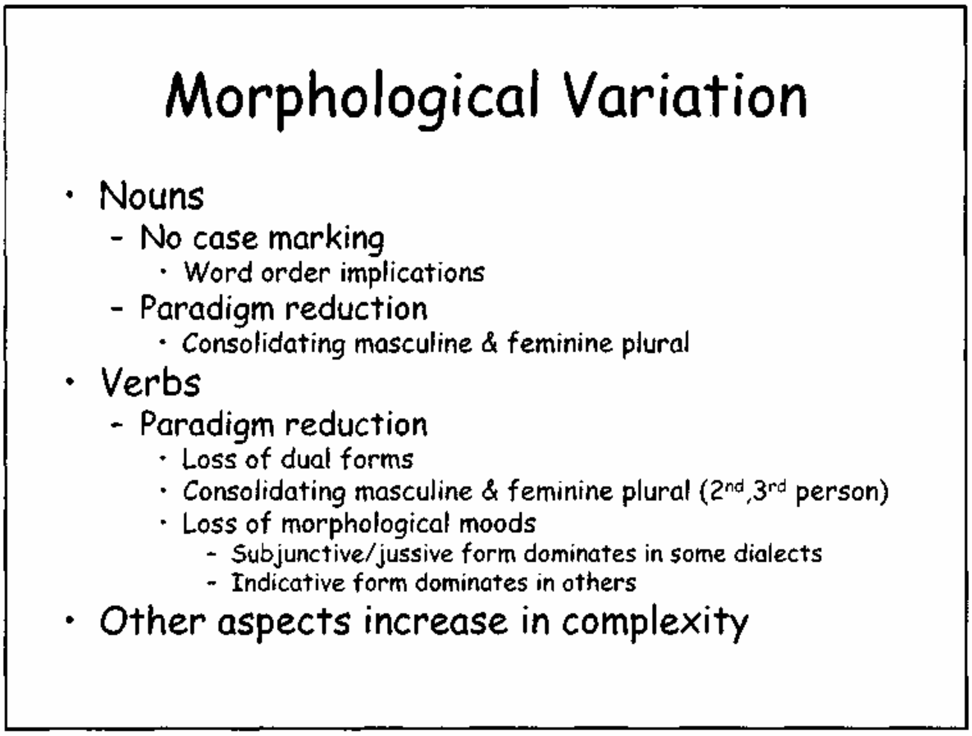
Foreign Borrowings

أوكي	>wky	okay
مرسي	mrsy	merci
بندورة	bndwrp	pomodoro (italian)
بيرا	byrA	birra (italian)
فرمت	frmt	format
تلفون	tlfwn	telephone
تلفن	talfan	to phone



Tutorial Contents

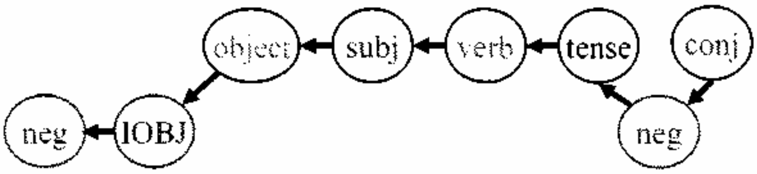
- Introduction
- Description of MSA Phenomena
- Description of Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Resources and References



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

	Perfect	***Imperfect***			
	Past	Subjunctive	Present habitual	Present progressive	Future
MSA	كتب /kataba/	يكتب /jaktuba/	يكتب /jaktubu/		سيكتب /sajaktubu/
LEV	كتب /katab/	يكتب /jiktob/	بيكتب /bjoktob/	عم بيكتب /ʕam bjoktob/	حيكتب /ħajiktob/
EGY	كتب /katab/	يكتب /jiktib/	بيكتب /bjiktib/		حيكتب /ħajiktib/
IRQ	كتب /kitab/	يكتب /jiktib/	ديكتب /dajiktib/		رح يكتب /raħ jiktib/
MOR	كتب /kteb/	يكتب /jekteb/	كيكتب /bjiktib/		غيكتب /ɣajekteb/

Morphological Variation

Verb conjugation

	Perfect			Imperfect		
	1S	2S♂	2S♀	1S	1P	2S♀
MSA	كتبت /katabtu/	كتبت /katabta/	كتبت /katabti/	كتب /aktubu/	نكتب /nəktubu/	تكتبين /taktubīna/ تكتبي /taktubī/
LEV	كتبت /katabt/		كتبتي /katabti/	كتب /aktob/	نكتب /noktob/	تكتبي /toktobi/
IRQ	كتبت /kitabit/		كتبتي /kitabti/	اكتب /aktib/	نكتب /niktib/	تكتبين /tikitbīn/
MOR	كتبت /ktebt/	كتبتي /ktebti/		نكتب /nekteb/	نكتبو /nektebu/	تكتبي /tektebi/

Tutorial Contents

- Introduction
- Description of MSA Phenomena
- Description of Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Resources and References

Idafa Construction

- *Genitive/Possessive Construction*
- Both *MSA* and dialects
  - **Noun1   Noun2**
  - ملك الاردن

    king Jordan

    the king of Jordan / Jordan's king
- Ta-marbuta allomorphs

	Idafa	No Idafa	Waqf
MSA	+at		+a
EGY	+it	+a	

- Dialects have *an additional* common construct
  - **Noun1 *<exponent>* Noun2**
  - LEV: الملك تبع الاردن the-king *belonging-to* Jordan
  - <expontent> differs widely among dialects

Demonstrative Articles

- Forms

	Proclitic	Word	
		Proximal	Distal
MSA	-	هذا,هذه,هؤلاء	ذلك,تلك,اولئك
Egyptian	-	ده, دي, دول	
Levantine	+هـ	هدا, هادي, هدول	هداك, هديك, هدوك

- Word Order (Example: *this man*)

	Pre-nominal	Post-nominal
MSA	هذا الرجل	X
EGY	X	الراجل ده
LEV	هدا الرجال	الرجال هدا

Negation of Declarative Verbal Sentences

	Pre	Circum	Post
MSA	لا, لم, لن, ما mA, lm, ln, lA	X	X
Egyptian	مش m$	ما ... ش mA ... $	X
Levantine	ما, مش mA, m$	ما ... ش mA ... $	ش $

Sentence Word Order

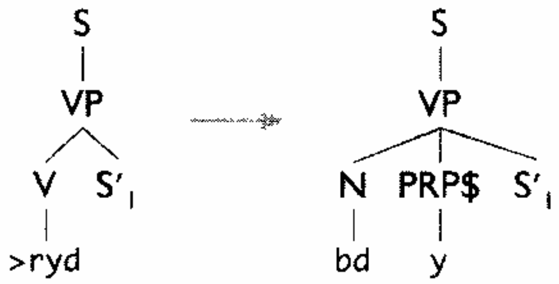
- Verbal sentences
  - The boys wrote the poems
  - MSA
    - Verb Subject Object (Partial agreement)
      كتب الاولاد الاشعار
      wrote$_{masc}$ the-boys the-poems
    - Subject Verb Object (Full agreement)
      الاولاد كتبوا الاشعار
      the-boys wrote$_{mascPlural}$ the-poems
  - LEV, EGY
    - Subject Verb Object
      الاولاد كتبوا الاشعار
      The-boys wrote$_{mascPlural}$ the-poems
    - Less present: Verb Subject Object
      كتبوا الاولاد الاشعار
      wrote$_{mascPlural}$ the-boys the-poems
    - Full agreement in both orders

	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)

Lexico-syntactic Variation

- 'want' (Levantine)



Tutorial Contents

- Introduction
- Description of MSA Phenomena
- Description of Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Resources and References

Code Switching

MSA

LEV

MSA and Dialect mixing in speech

- phonology, morphology and syntax

لا لأن ما بتعتقد بأنه عملية اللي عم بيعترضوا اليوم تمديد للرئيس لحود هم اللي طالبوا بالتمديد للرئيس الهراوي
وبالتالي موضوع ... موضوع مبدئي على الأرض أنا بحترم انه يكون في نظرة ديمقراطية للأمور وانه
يكون في حرية للشعب الديمقراطية وأن يكون في ممارسة ديمقراطية وبعتقد انه الكل في لبنان أو أكثرية
ساحقة في لبنان تريد هذا الموضوع، بس بدي يرجع لحظة على موضوع انجازات العهد يعني بعد نحكي
عن انجازات 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 بيأخذ مواقف صالحة ضمن خطاب ومبادئ خطاب القسم هو اللي جانبه انما 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 ابداء 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 الى جنبنا في هذا الموضوع، أما الموضوع الديمقراطي لا بتقيد
تماما هذا هالتوجيهة النظر بس ما ممكن نقول انه الدستور أو تعديله هو أو امكانية فتح إعادة انتخاب
ديمقراطي ضمن المجلس والتصويت الى 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

Code Switching with English

- Iraqi Arabic Example
  - ya ret 3inde hech sichena tit7arrak wa77ad-ha , 7atta ma at3ab min asawwe zala6a yomiyya :D
  - 3ainee Zainab, tara hathee technology jideeda, they just started selling it !! Lets ask if anybody knows where do they sell them ! :

http://www.aliraqi.org/forums/archive/index.php/t-16137.html

Dialectal Impact on MSA

- Loss of case endings and nunation in read MSA

  /fī bajt ʤadīd/

  instead of /fī bajtin ʤadīdin/

  'in a new house'
- A shift toward SVO rather than VSO in written MSA

Dialectal Impact on MSA

- Structure borrowing
- Example: monies and properties of the company
  - اموال الشركة وممتلكاتها
    - /ʔamwālu ʃʃarikati wamumtalakātuhā/
    - *monies the-company and-properties-its*

  instead of
  - اموال وممتلكات الشركة
    - /ʔamwālu wamumtalakātu ʃʃarikati/
    - *monies and-properties the-company*

Dialectal Impact on MSA

- Code switching in written MSA
- Dialectal lexical and structural uses
  - Example Newswire Alnahar newspaper (ATB3 v.2)

فأخذ على خاطر الأخوان ومن حقهم ان يزعلوا

f>x ElY xATr* AlAxwAn wmn hqhm An *yzElw*

then-was-taken upon self the-brothers and-from right-their to be-angry

'they were upset, and they had the right to be angry'

Tutorial Contents

- Introduction
- Description of MSA Phenomena
- Description of Dialectal Phenomena
- Sample Applications
  - Automatic speech recognition
  - Dictionary creation
  - Morphological analysis
  - Part-of-speech tagging
  - Syntactic parsing
  - Machine translation
- Resources and References

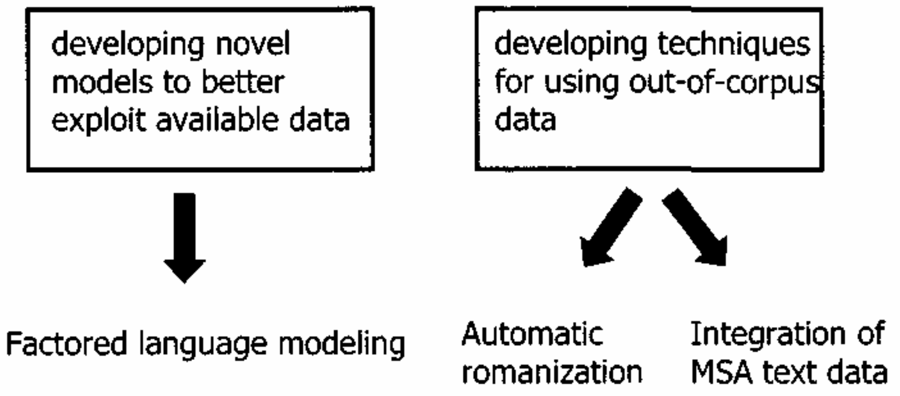
Arabic ASR:
State of the Art

- BBN TIDESOnTap: 15.3% WER
- BBN CallHome system: 55.8% WER
- JHU WS 2002: 53.8% WER
- WER on conversational speech noticeably higher than for other languages

  (eg. 30% WER for English CallHome)

JHU WS02 Approach

improvements to Arabic ASR through



developing novel models to better exploit available data

developing techniques for using out-of-corpus data

Factored language modeling

Automatic romanization

Integration of MSA text data

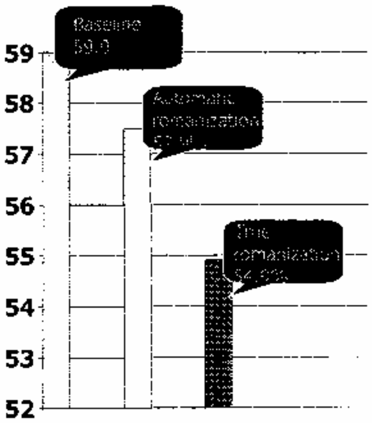
Slide courtesy of (Kirchhoff et al.2002)

Approach Details

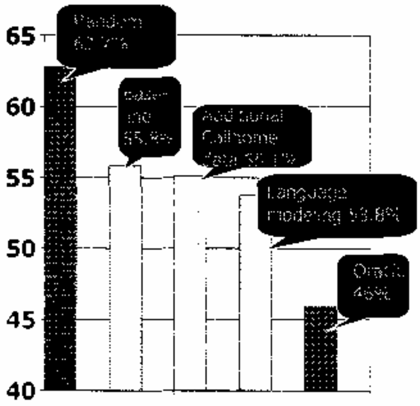
- Factored Language Models
  - complex morphological structure leads to large number of possible word forms
  - break up word into separate components
  - build statistical n-gram models over individual morphological components rather than complete word forms
- Automatic Vowelization
  - try to predict vowelization automatically from data and use result for recognizer training
- Integrate data from MSA written sources

JHU WS02 Results (WER)

Grapheme-based recognizer



Phone-based recognizer



S ide courtesy of (Kirchhoff et al.2002)

Dialect-MSA Dictionary

- Problem: total lack of Dialect-MSA resources
  - No Dialect-MSA parallel text
  - No paper dictionaries for Dialect-MSA
- Dialect-MSA dictionary is required for many NLP applications exploiting MSA resources
  - e.g., to translate dialect sentences to MSA before parsing them with an MSA parser

Levantine-MSA Dictionary

- The Automatic-Bridge dictionary (AB)
  - English as a bridge language between MSA and LA
- The Egyptian-Cognate dictionary (EC)
  - Levantine-Egyptian cognate words in Columbia University Egyptian-MSA lexicon (2,500 lexeme pairs)
- The Human-Checked dictionary (HC)
  - Human cleanup of the union of AB and EC
  - Using lexemes speeded up the process of dictionary cleaning
    - reducing the number of entries to check
    - minimizing word ambiguity decisions
  - Morphological analysis and generation are required to map from inflected LA to inflected MSA
- The Simple-Modification dictionary (SM)
  - Minimal modification to LA inflected forms to look more MSA-like
  - Form modification: (أغنيا >gnyA 'rich pl.') is mapped to (أغنياء >gnyA')
  - Morphology modification: (بشرب b$rb 'I drink') is mapped to (أشرب >$rb)
  - Full translation: (كمان kmAn 'also') is mapped to (ايضا AyDAF)

(Maamouri el al. 2006)

Dialectal Morphological Analysis

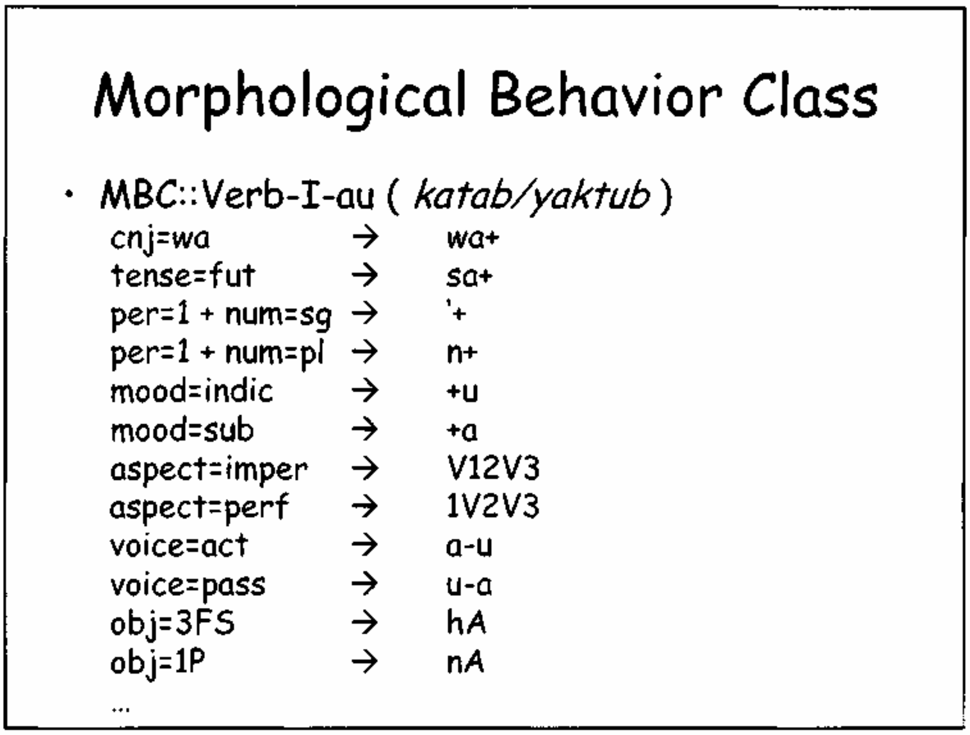
- **MAGEAD** **(Habash and Rambow 2006)**
- **Levels of Morphological Representation**
  - Lexeme Level

    Aizdahar$_1$ PER:3 GEN:f NUM:sg ASPECT:perf
  - Morpheme Level

    [zhr,1tV2V3,iaa] +at
  - Surface Level
    - Phonology: /izdaharat/
    - Orthography: Aizdaharat (إزْدَهَرت)

The Lexeme

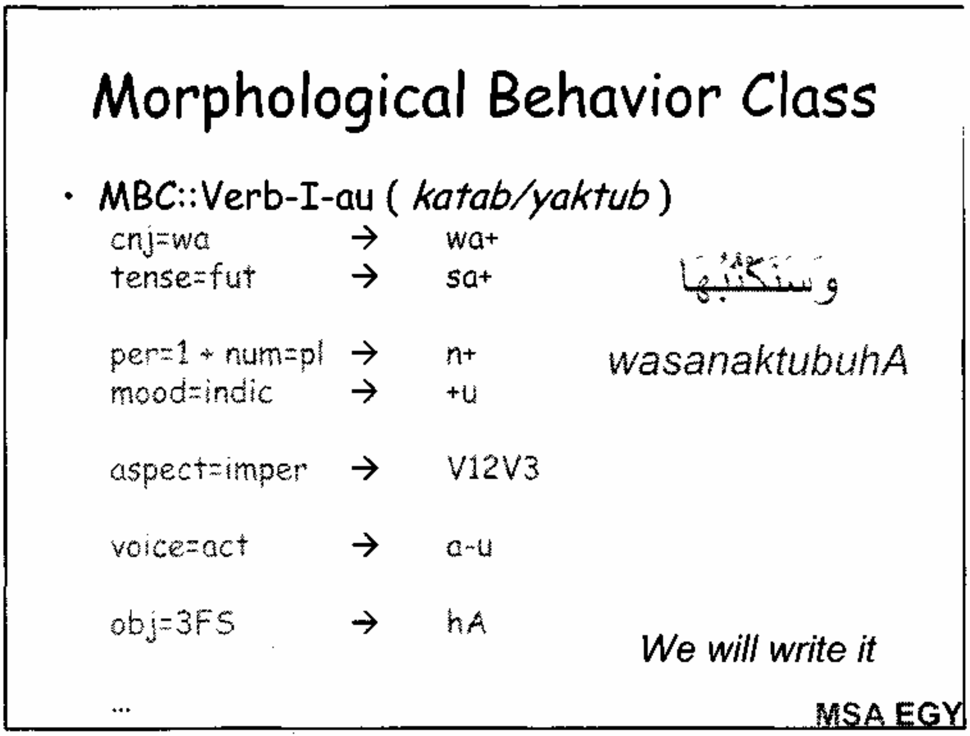
- Lexeme is an abstraction of all inflectional variants of a word
  - |كِتاب1| comprises كتابان الكتابين كتبهم للكتب كُتُب كِتاب ...
- For us, lexeme is formally a triple
  - Root or NTWS
  - Morphological behavior class (MBC)
    - {بيت ابيات}'verse' vs. {بيت بيوت} 'house'
  - Meaning index
    - |قاعدة1| : {قاعدة قواعد} 'rule'
    - |قاعدة2| : {قاعدة قواعد} 'military base'



Morphological Behavior Class

- MBC::Verb-I-au (*katab/yaktub*)

cnj=wa	→	wa+
tense=fut	→	sa+
per=1 + num=sg	→	'+
per=1 + num=pl	→	n+
mood=indic	→	+u
mood=sub	→	+a
aspect=imper	→	V12V3
aspect=perf	→	1V2V3
voice=act	→	a-u
voice=pass	→	u-a
obj=3FS	→	hA
obj=1P	→	nA
...		



Morphological Behavior Class

- MBC::Verb-I-au (*katab/yaktub*)

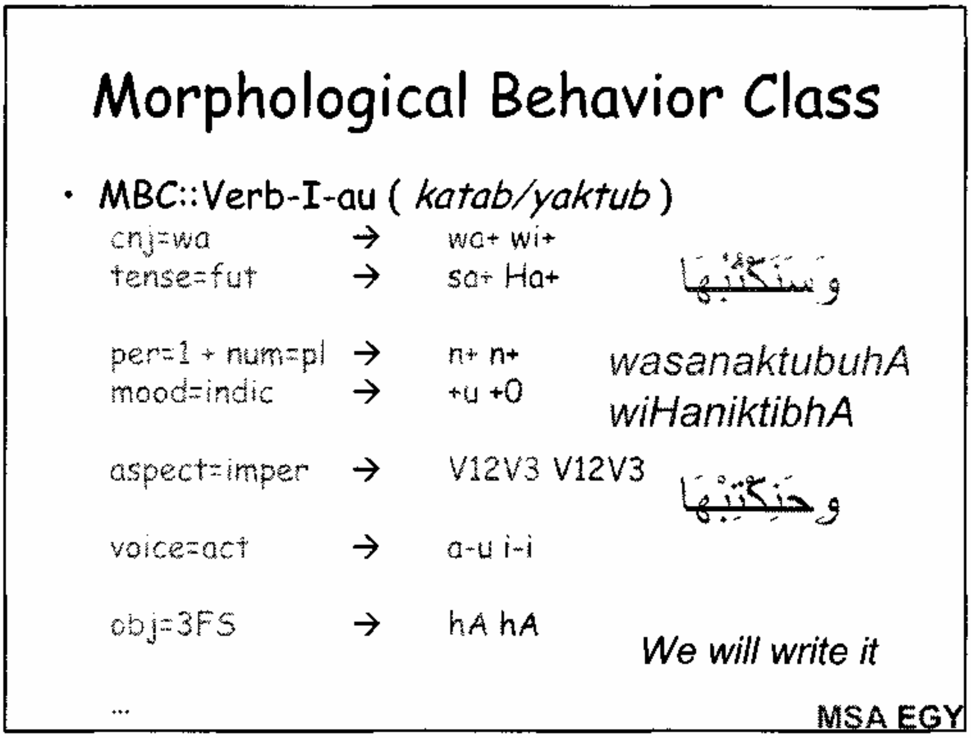
cnj=wa	→	wa+
tense=fut	→	sa+
per=1 + num=pl	→	n+
mood=indic	→	+u
aspect=imper	→	V12V3
voice=act	→	a-u
obj=3FS	→	hA
...		

وَسَنَكْتُبُها

wasanaktubuhA

We will write it

MSA EGY



Morphological Behavior Class

- MBC::Verb-I-au (*katab/yaktub*)

cnj=wa	→	wa+ wi+
tense=fut	→	sa+ Ha+
per=1 + num=pl	→	n+ n+
mood=indic	→	+u +0
aspect=imper	→	V12V3 V12V3
voice=act	→	a-u i-i
obj=3FS	→	hA hA

...

وَسَنَكْتُبُها

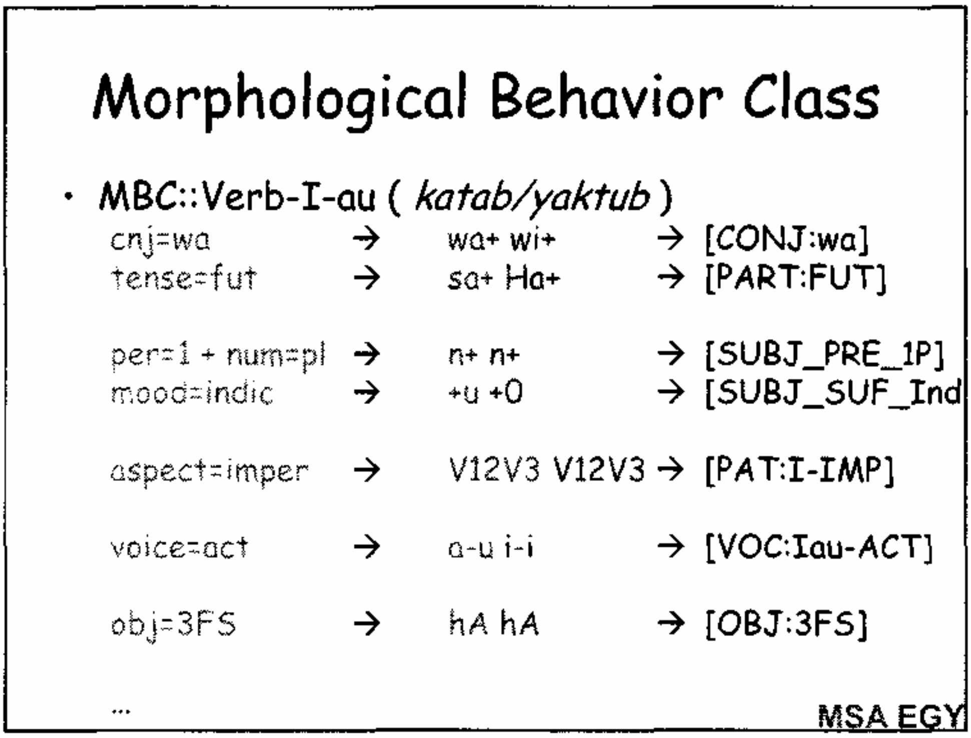
wasanaktubuhA

wiHaniktibhA

وحَنِكْتِبْها

We will write it

MSA EGY



Morphological Behavior Class

- MBC::Verb-I-au (*katab/yaktub*)

cnj=wa	→	wa+ wi+	→	[CONJ:wa]
tense=fut	→	sa+ Ha+	→	[PART:FUT]
per=1 + num=pl	→	n+ n+	→	[SUBJ_PRE_1P]
mood=indic	→	+u +0	→	[SUBJ_SUF_Ind
aspect=imper	→	V12V3 V12V3	→	[PAT:I-IMP]
voice=act	→	a-u i-i	→	[VOC:Iau-ACT]
obj=3FS	→	hA hA	→	[OBJ:3FS]

...

MSA EGY

Morphological Behavior Class

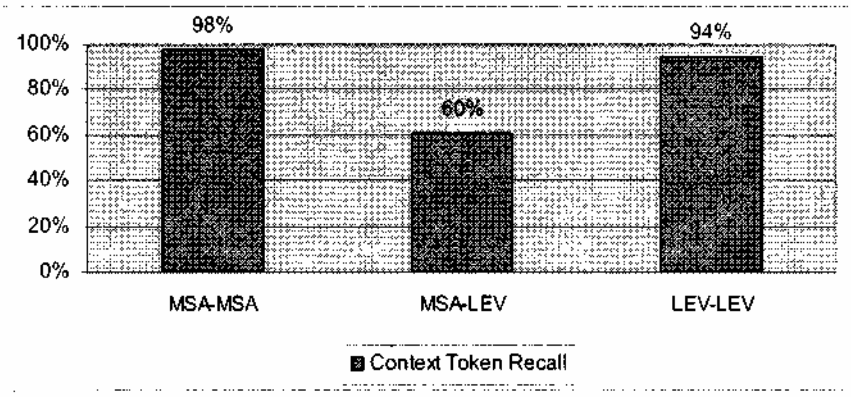
- MBC::Verb-I-au (*katab/yaktub*)

cnj=wa	→	[CONJ:wa]
tense=fut	→	[PART:FUT]
per=1 + num=pl	→	[SUBJ_PRE_1P]
mood=indic	→	[SUBJ_SUF_Ind]
aspect=imper	→	[PAT:I-IMP]
voice=act	→	[VOC:Iau-ACT]
obj=3FS	→	[OBJ:3FS]

...

Levantine Evaluation

- Results on Levantine Treebank



Arabic Dialect POS Tagging

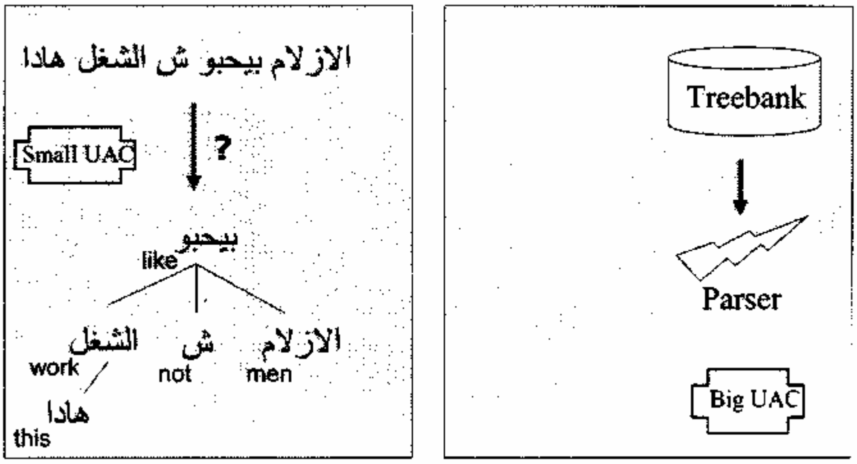
- Duh and Kirchhoff 2005;Duh and Kirchhoff 2006
  - Egyptian Arabic and Levantine Arabic
  - Minimal supervision
    - dialectal text
    - and MSA morphological analyzer
  - Cross-dialect sharing techniques
- Rambow et al. 2005
  - Levantine Arabic
  - LEV-MSA transduction using LEV-MSA lexicon
  - MSA POS Tagging
  - Projection of MSA tags unto LEV

Arabic Dialect Parsing

- Possible Approaches
  - Annotate corpora ("Brill Approach")
    - Too expensive
  - Leverage existing MSA resources
    - Difference MSA/dialect not enormous
    - Linguistic studies of dialects exist
    - Too many dialects: even with dialects annotated, still need leveraging for other dialects

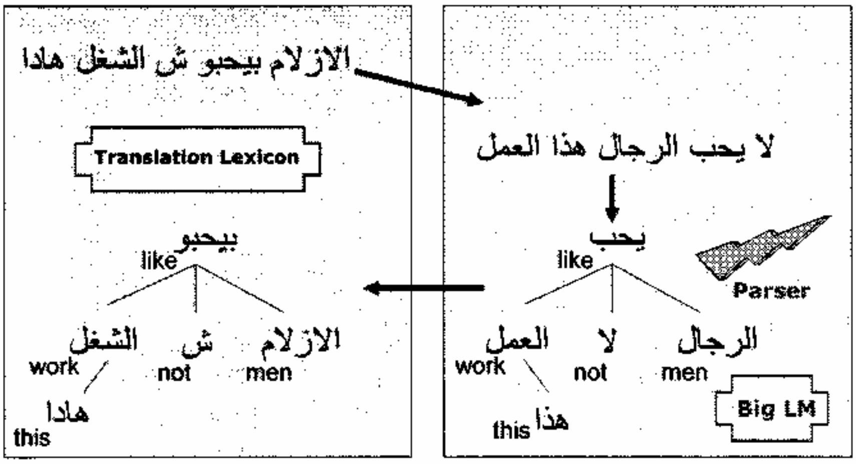
Parsing Arabic Dialects:
The Problem

- Dialect - - MSA -



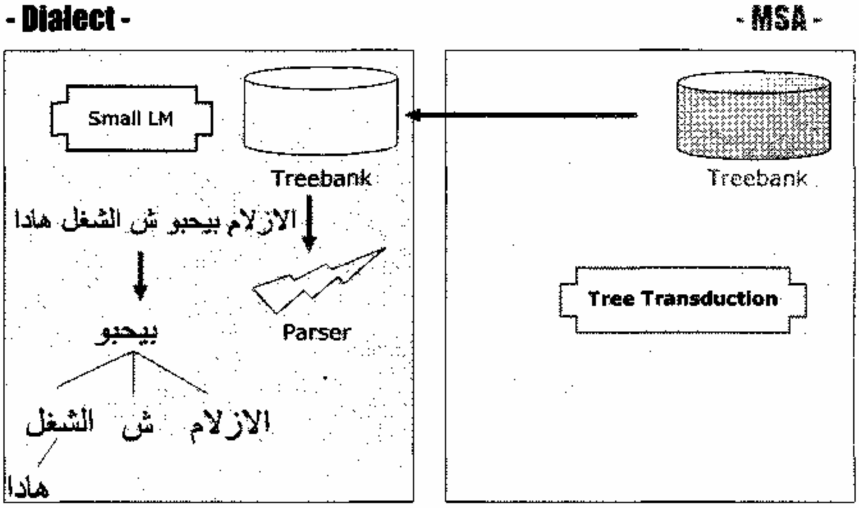
Sentence Transduction Approach

- Dialect - - MSA -



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

MSA Treebank Transduction



- Dialect -

- MSA -

Small LM

Treebank

الازلام بيحبو ش الشغل هادا

بيحبو

الازلام ش الشغل

هادا

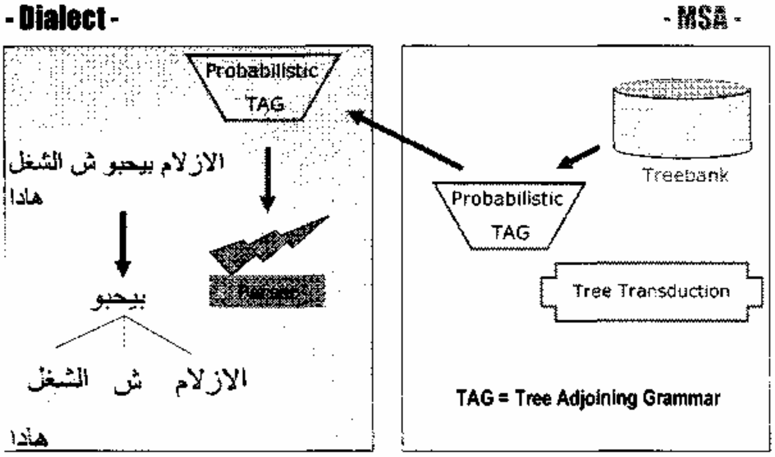
Parser

Treebank

Tree Transduction

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

Grammar Transduction



- Dialect -

- MSA -

Probabilistic TAG

الازلام بيحبو ش الشغل هادا

بيحبو

الازلام ش الشغل

هادا

Parser

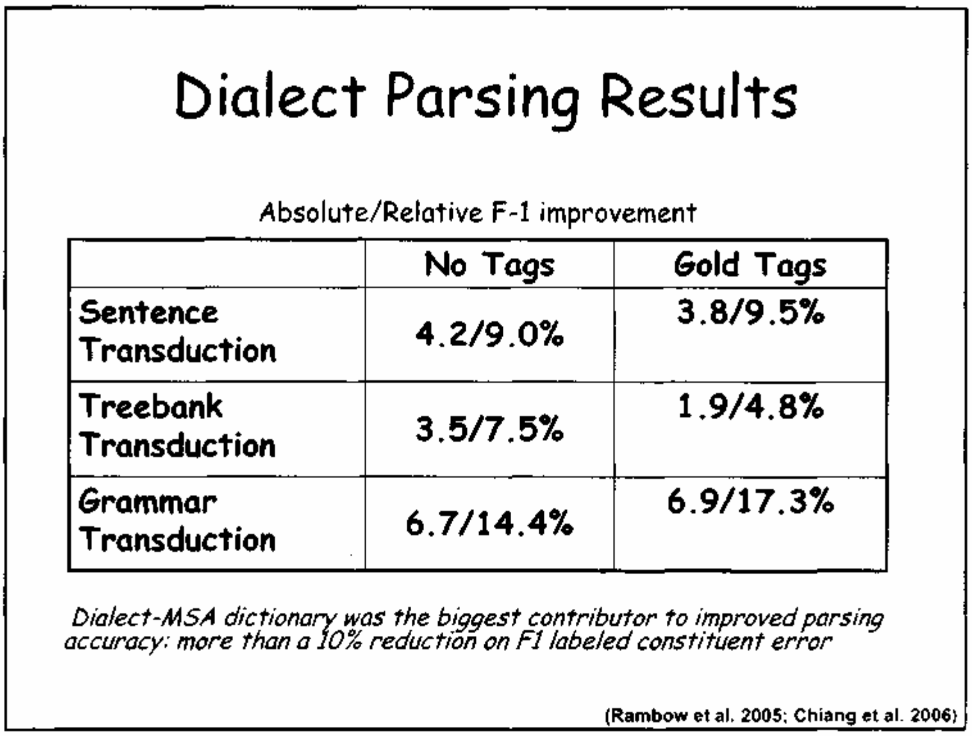
Treebank

Probabilistic TAG

Tree Transduction

TAG = Tree Adjoining Grammar

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



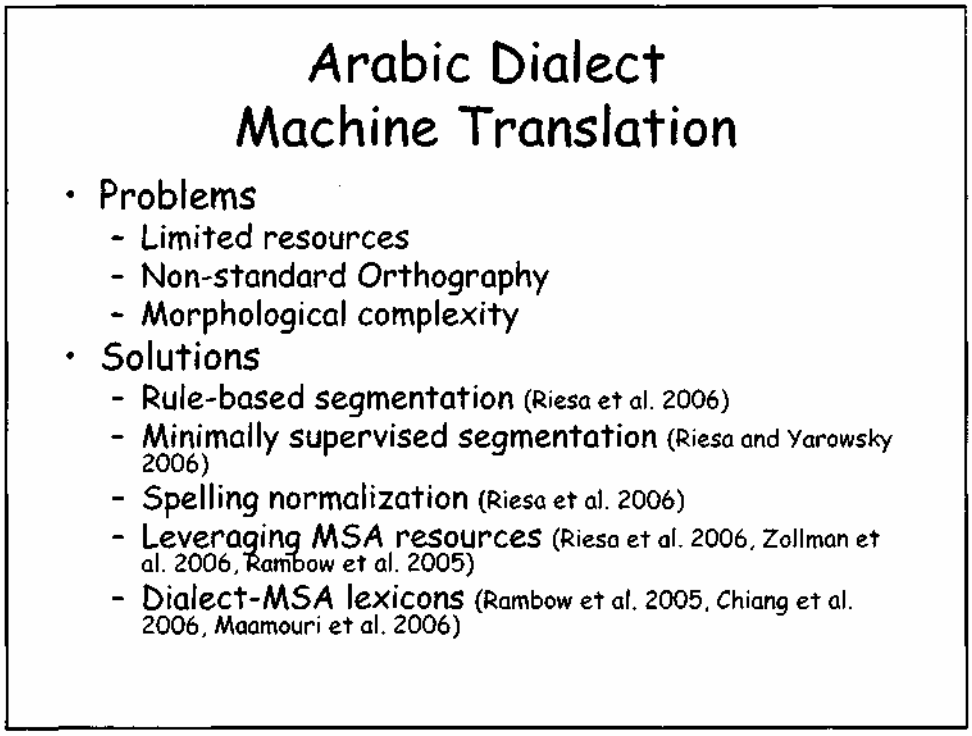
Dialect Parsing Results

Absolute/Relative F-1 improvement

	No Tags	Gold Tags
Sentence Transduction	**4.2/9.0%**	**3.8/9.5%**
Treebank Transduction	**3.5/7.5%**	**1.9/4.8%**
Grammar Transduction	**6.7/14.4%**	**6.9/17.3%**

Dialect-MSA dictionary was the biggest contributor to improved parsing accuracy: more than a 10% reduction on F1 labeled constituent error

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



Arabic Dialect Machine Translation

- Problems
  - Limited resources
  - Non-standard Orthography
  - Morphological complexity
- Solutions
  - Rule-based segmentation (Riesa et al. 2006)
  - Minimally supervised segmentation (Riesa and Yarowsky 2006)
  - Spelling normalization (Riesa et al. 2006)
  - Leveraging MSA resources (Riesa et al. 2006, Zollman et al. 2006, Rambow et al. 2005)
  - Dialect-MSA lexicons (Rambow et al. 2005, Chiang et al. 2006, Maamouri et al. 2006)

Arabic Dialect Machine Translation

- **TransTac: DARPA Program on Translation System for Tactical Use**
  - Iraqi <> English speech-to-speech MT
  - Phraselator: http://www.phraselator.com/
- MT as a component
  - JHU Workshop on Parsing Arabic dialect (Rambow et al. 2005, Chiang et al. 2006)

Dialect 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 Speech **and** Transcripts
- Egyptian Colloquial Arabic Lexicon
- Levantine Arabic Resources
- http://www.orientel.org/
- http://www.appen.com.au/
- LDC Arabic EARS
- **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

Books

Bateson, Mary Catherine. Arabic Language Handbook. Georgetown University Press. 2003.

Brustad, Kristen E. The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects. Georgetown University Press. 2000.

Holes, Clive. Modern Arabic: Structures, Functions, and Varieties. Georgetown University Press. 2004.

Conference Papers and Journal Articles

Alexandrescu, A. and K. Kirchhoff. 2006. Factored Neural Language Models. HLT.

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.

References

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

Diab, M., K. Hacioglu and D. Jurafsky. "Automated Methods for Processing Arabic Text: From Tokenization to Base Phrase Chunking." Book Chapter. In Arabic Computational Morphology: Knowledge-based and Empirical Methods. Editors A. van den Bosch and A. Soudi.

Duh, K. and K. Kirchhoff. 2005. POS Tagging of Dialectal Arabic: A Minimally Supervised Approach. ACL Workshop on Semitic Languages.

Duh, K. and K. Kirchhoff. 2006. Lexicon Acquisition for Dialectal Arabic Using Transductive Learning. EMNLP.

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

Habash, N. Issues in Palestinian Arabic Spelling Standardization. NACAL 27, 1999. Baltimore, MD.

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

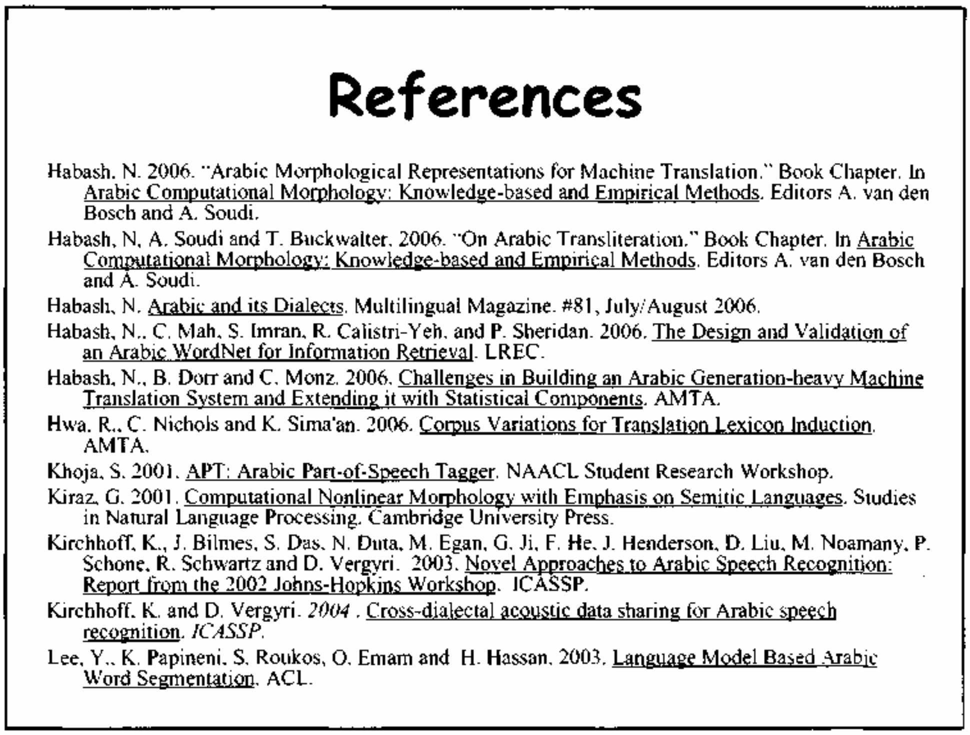
Habash, N. and O. Rambow. 2005a. Arabic Tokenization, Part-of-Speech Tagging in and Morphological Disambiguation One Fell Swoop. ACL.

Habash, N. and O. Rambow. 2006. MAGEAD: A Morphological Analyzer and Generator for the Arabic Dialects. ACL.

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.



References

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, A. Soudi and T. Buckwalter. 2006. "On Arabic Transliteration." Book Chapter. In Arabic Computational Morphology: Knowledge-based and Empirical Methods. Editors A. van den Bosch and A. Soudi.

Habash, N. Arabic and its Dialects. Multilingual Magazine. #81, July/August 2006.

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.

Habash, N., B. Dorr and C. Monz. 2006. Challenges in Building an Arabic Generation-heavy Machine Translation System and Extending it with Statistical Components. AMTA.

Hwa, R., C. Nichols and K. Sima'an. 2006. Corpus Variations for Translation Lexicon Induction. AMTA.

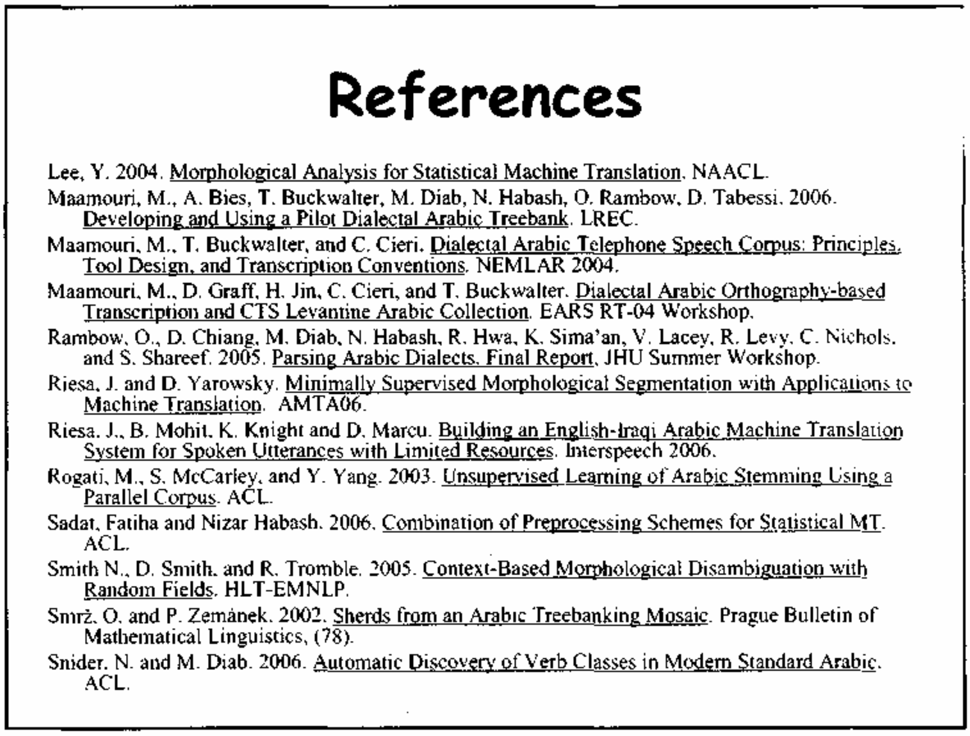
Khoja, S. 2001. APT: Arabic Part-of-Speech Tagger. NAACL Student Research Workshop.

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 Workshop. ICASSP.

Kirchhoff, K. and D. Vergyri. *2004* . Cross-dialectal acoustic data sharing for Arabic speech recognition. *ICASSP*.

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



References

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.

Maamouri, M., T. Buckwalter, and C. Cieri. Dialectal Arabic Telephone Speech Corpus: Principles, Tool Design, and Transcription Conventions. NEMLAR 2004.

Maamouri, M., D. Graff, H. Jin, C. Cieri, and T. Buckwalter. Dialectal Arabic Orthography-based Transcription and CTS Levantine Arabic Collection. EARS RT-04 Workshop.

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.

Riesa, J. and D. Yarowsky. Minimally Supervised Morphological Segmentation with Applications to Machine Translation. AMTA06.

Riesa, J., B. Mohit, K. Knight and D. Marcu. Building an English-Iraqi Arabic Machine Translation System for Spoken Utterances with Limited Resources. Interspeech 2006.

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

Sadat, Fatiha and Nizar Habash. 2006. Combination of Preprocessing Schemes for Statistical MT. ACL.

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

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

Snider, N. and M. Diab. 2006. Automatic Discovery of Verb Classes in Modern Standard Arabic. ACL.

References

Snider, N. and M. Diab. 2006. Unsupervised Induction of Arabic Verb Classes. NAACL.

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

Vergyri, D., K. Kirchhoff, K. Duh and A. Stolcke. 2004. Morphology-based language modeling for Arabic speech recognition. ICSLP.

Vergyri. D. and K. Kirchhoff. 2004. Automatic Diacritization of Arabic for Acoustic Modeling in Speech Recognition. *COLING Workshop on Arabic-script Based Languages.*

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

Yang, M. and K. Kirchhoff. Phrase-Based Backoff Models for Machine Translation of Highly Inflected Languages. EACL.

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.

Zollmann. A., A. Venugopal and S. Vogel. 2006. Bridging the Inflection Morphology Gap for Arabic Statistical Machine Translation. NAACL.

References

Conference/Institution/Program 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

ICASSP=International Conference on Acoustics, Speech and Signal 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

NACAL = North America Conference on Afro-asiatic Languages

EARS = DARPA Program (Efficient, Affordable, Reusable Speech-to-Text)