MEDAR 2009

Cairo, Egypt
April 21, 2009

Arabic Dialect Processing

Mona Diab   Nizar Habash
Center for Computational Learning Systems
Columbia University
{mdiab,habash}@ccls.columbia.edu

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

CADIM

CCLS

Tutorial Contents

- Introduction
- Dialectal Phenomena
- Sample Applications
- Dialect Resources

Introduction

- 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

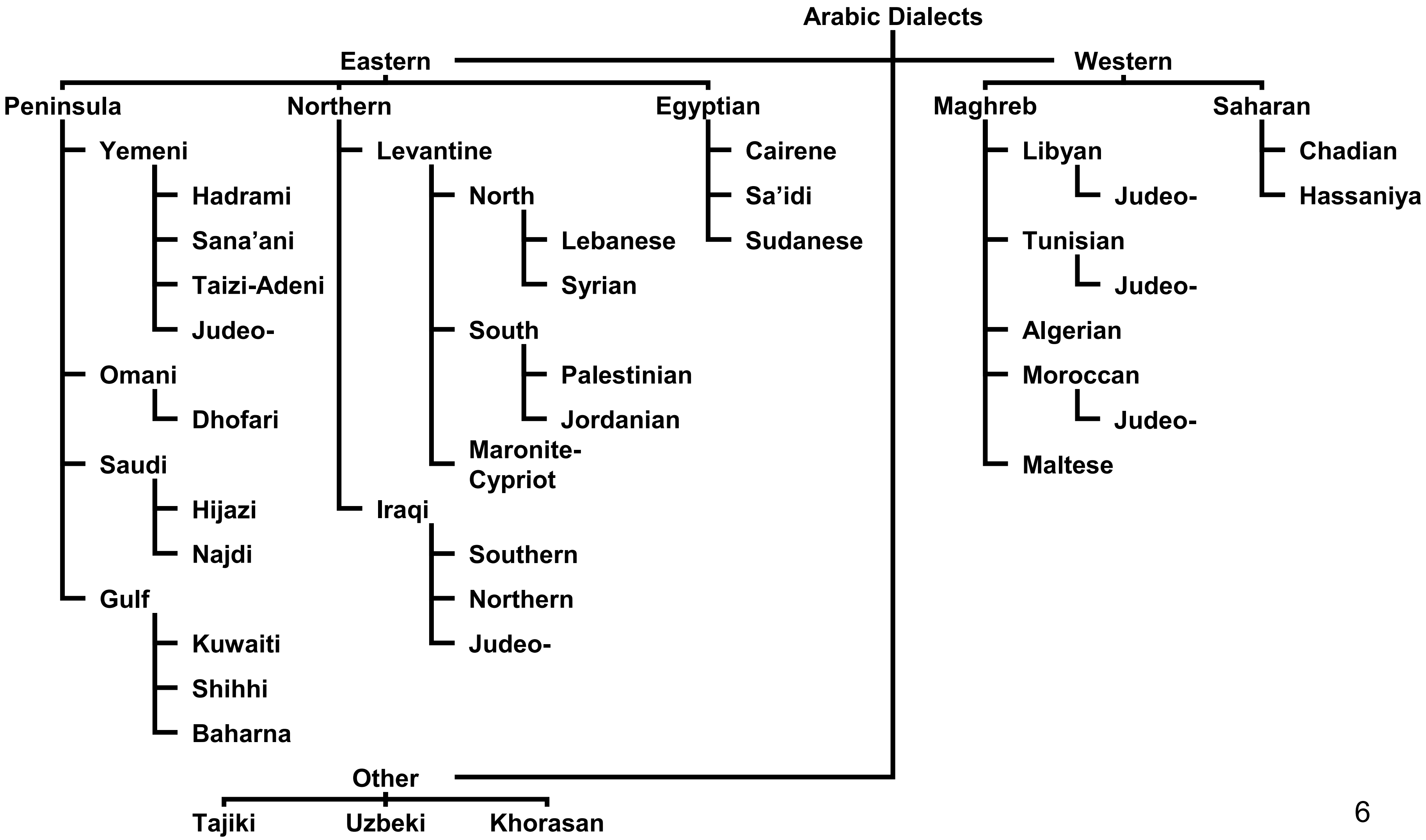
Introduction

- ~300M people worldwide speak Arabic
- Arabic is the/an 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

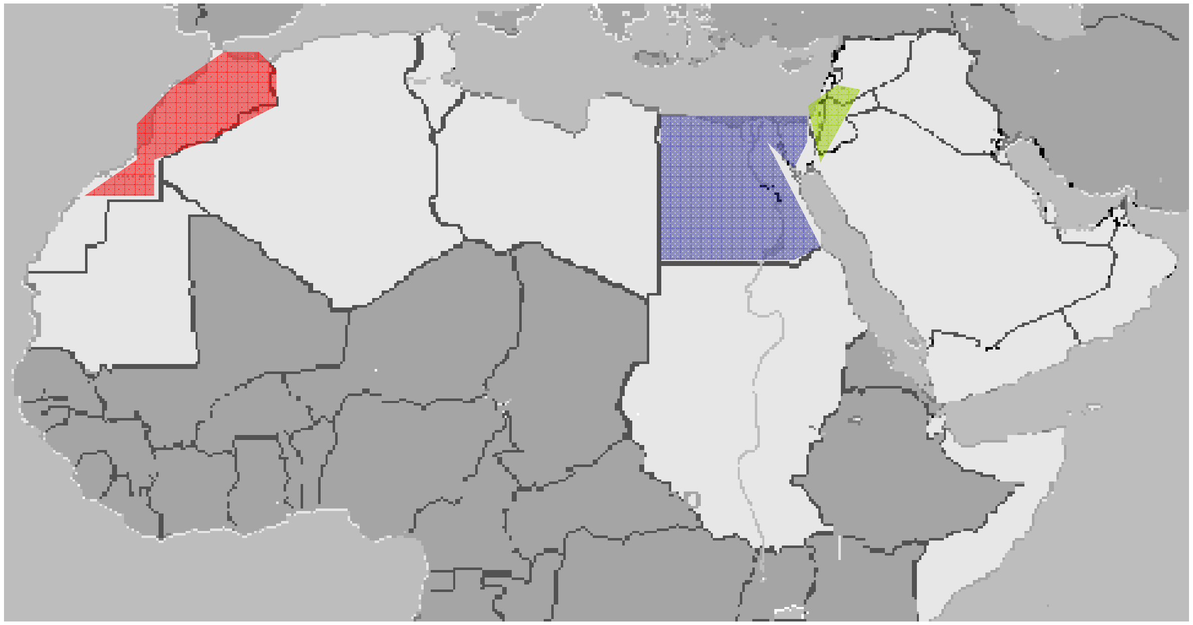
Introduction

- 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 ʒadīdatan لم يشتر نزار طاولة جديدة

didn’t buy Nizar table new

nizār maʃtarāʃ ṭarabēza gidīda 🔵 نزار ماشتراش طربيزة جديدة

nizār maʃtarāʃ ṭawile ʒdīde 🟢 نزار ماشتراش طاولة جديدة

nizar maʃrāʃ mida ʒdīda 🔴 نزار ماشراش ميدة جديدة

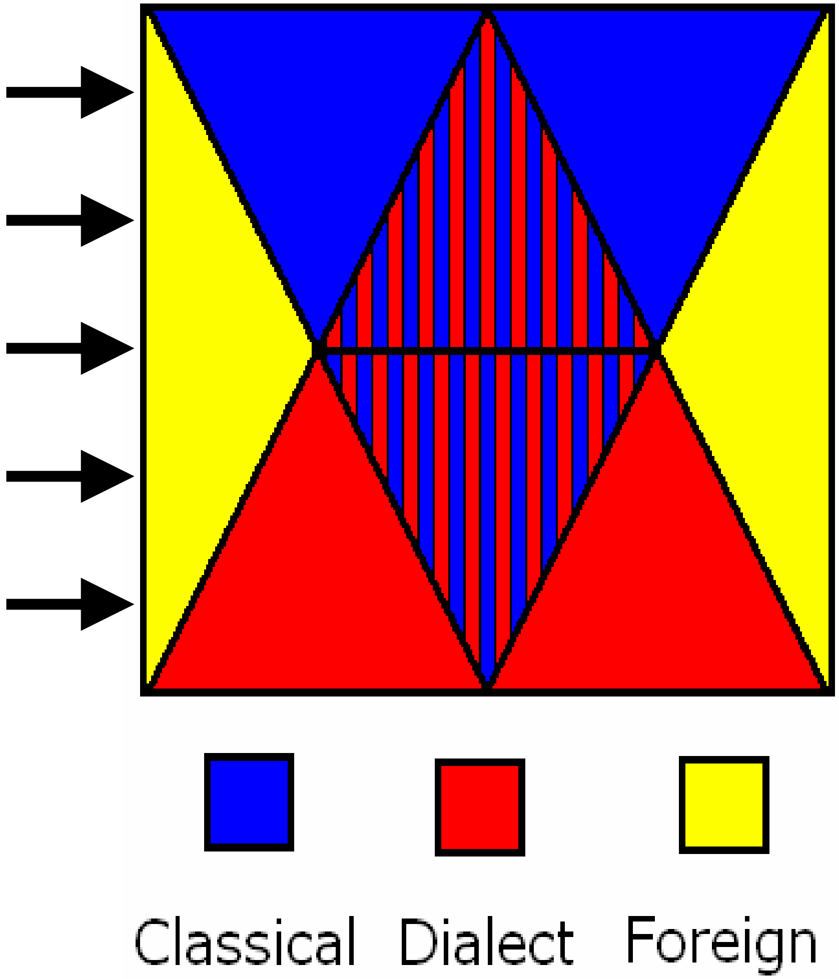
Nizar not-bought-not table new

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**

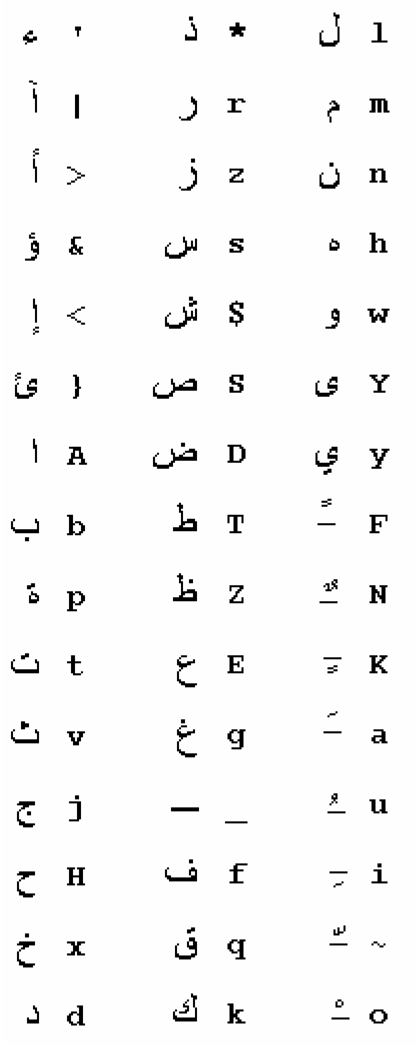
- **Lack of written resources**

Tutorial Contents

- Introduction
- Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Dialect Resources

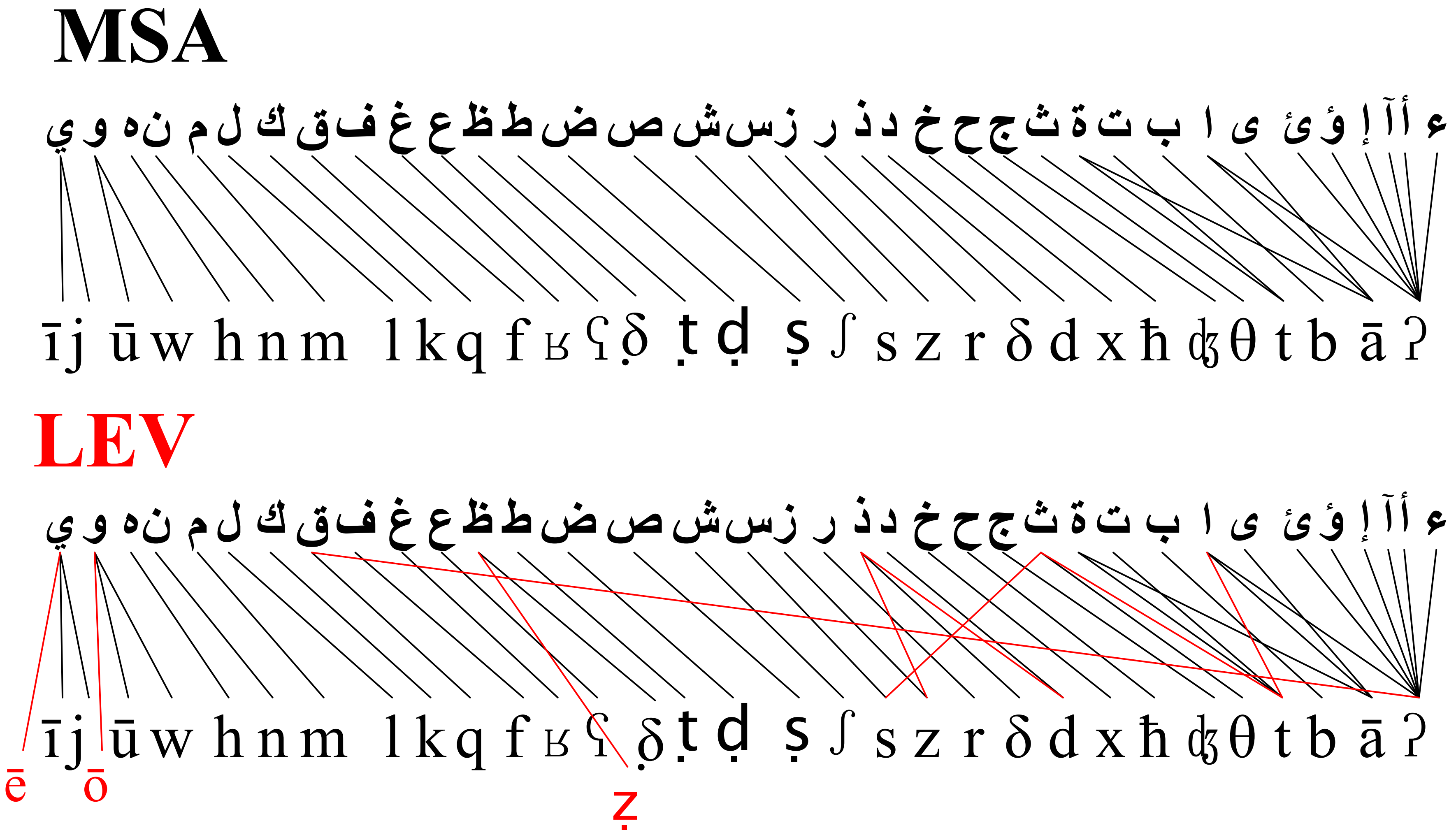
A Note on Romanization

- 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

Phonological Variations


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

- 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 (ـف, ق) = MOR (ـڢ, ڧ)
- Modern proposals: LEV /ʔ/ قٔ , /ē/ ى̭, /ō/ قٚ (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

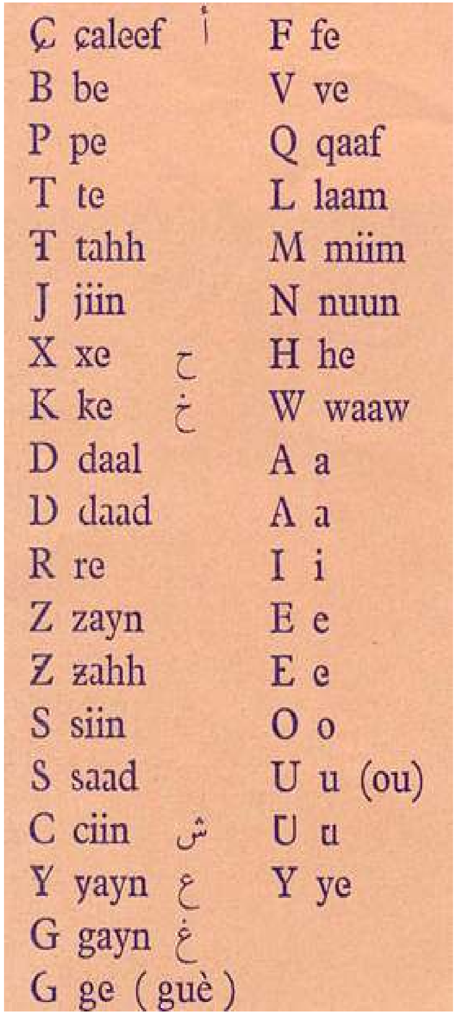
http://www.soriagate.net/showthread.php?t=32678

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
  - www.yamli.com

عربي	IPA	Latin	عربي	IPA	Latin
أإآءؤئ	/ʔ/	\` 2 Ø	ث	/θ/	th
ة	/a/,/t/	a t	ط	/tˤ/	t T 6
ح	ħ	H h 7	ع	/ʕ/	\` 3 Ø
خ	/x/	kh 7′ x 8	غ	/ʁ/	g gh 3′
ذ	/ð/	th	ق	/q/	q
ش	/ ʃ/	sh ch	ي	/y/ /ay/ /ī/ /ē/	y,i,e, ai,ei,…

Akl 1961



Ç	çaleef	أ	F	fe
B	be		V	ve
P	pe		Q	qaaf
T	te		L	laam
Ŧ	tahh		M	miim
J	jiin		N	nuun
X	xe	ح	H	he
K	ke	خ	W	waaw
D	daal		A	a
Đ	daad		A	a
R	re		I	i
Z	zayn		E	e
Ƶ	ƶahh		E	e
S	siin		O	o
S	saad		U	u (ou)
C	ciin	ش	U	u
Y	yayn	ع	Y	ye
G	gayn	غ		
G	ge (guè)			

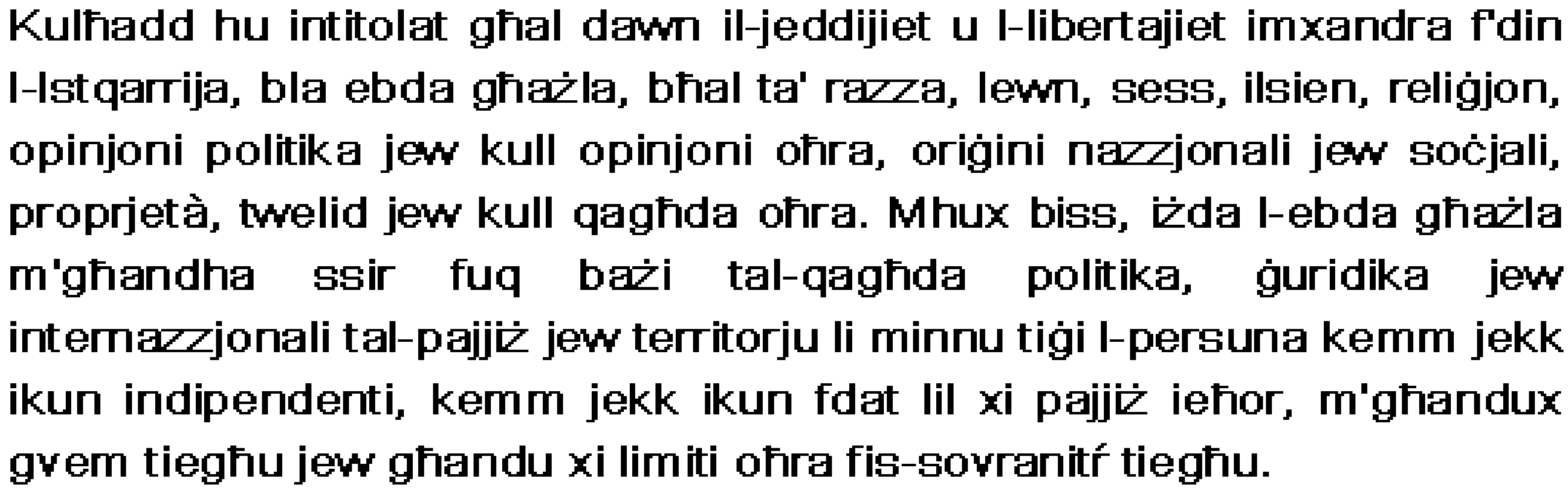
Egyptian Arabic in Latin Script

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.html

The Case of Maltese

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



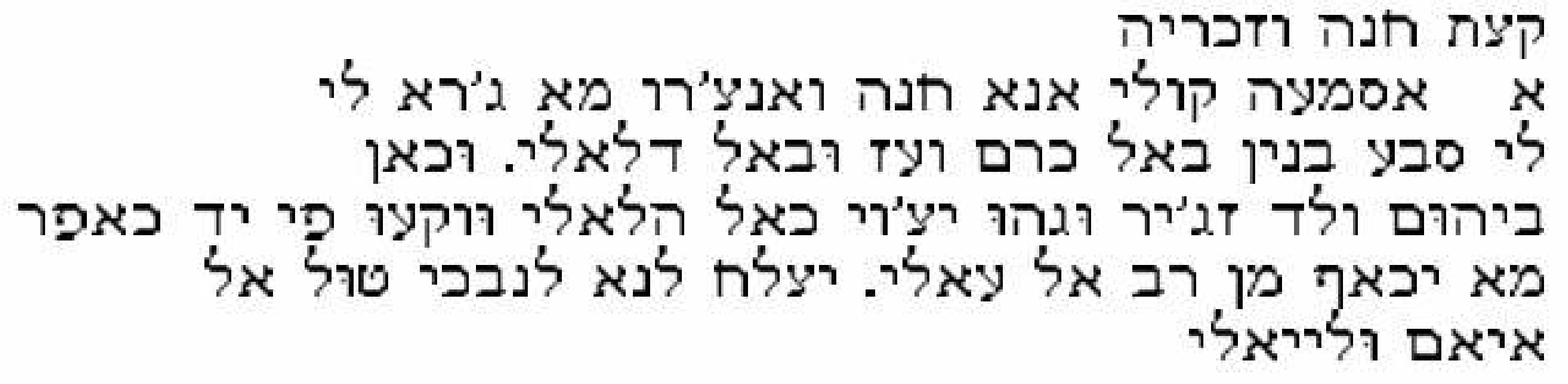
Kulħadd hu intitolat għal dawn il-jeddijiet u l-libertajiet imxandra f'din l-Istqarrija, bla ebda għażla, bħal ta' razza, lewn, sess, ilsien, reliġjon, opinjoni politika jew kull opinjoni oħra, oriġini nazzjonali jew soċjali, proprjetà, twelid jew kull qagħda oħra. Mhux biss, iżda l-ebda għażla m'għandha 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 "



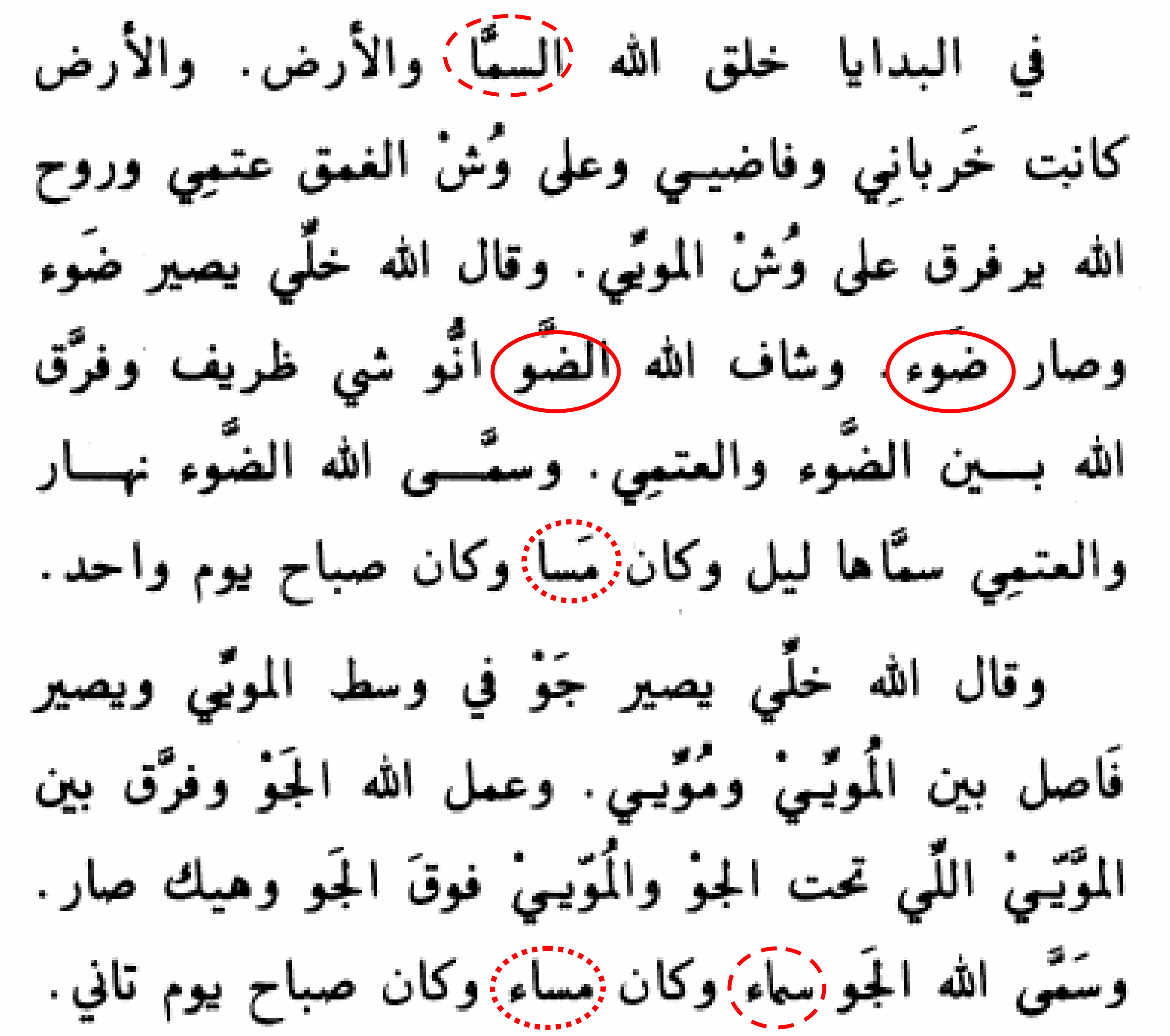
קצת חנה וזכריה
א   אסמעה קולי אנא חנה ואנצ’רו מא ג’רא לי
לי סבע בנין באל כרם ועז וּבאל דלאלי. וּכאן
ביהוּם ולד זג’יר וּגהוּ יצ’וי כאל הלאלי וּוקעוּ פי יד כאפר
מא יכאף מן רב אל עאלי. יצלח לנא לנבכי טוּל אל
איאם וּלייאלי

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 السمًّا والأرض. والأرض

كانت خَربانِي وفاضيي وعلى وُشْ الغمق عتمِي وروح

الله يرفرق على وُشْ الموّي. وقال الله خلّي يصير ضَوء

وصار ضوء. وشاف الله الضّو انّو شي ظريف وفرّق

الله بين الضّوء والعتمِي. وسمّى الله الضّوء نهار

والعتمِي سمّاها ليل وكان مَسا وكان صباح يوم واحد.

وقال الله خلّي يصير جَوّ في وسط الموّي ويصير

فَاصل بين المُويّيْ ومُوّيي. وعمل الله الجَوْ وفرّق بين

الموّييْ اللّي تحت الجوْ والمُوّييْ فوقَ الجَو وهيك صار.

وسَمّى الله الجَو سماء وكان مساء وكان صباح يوم تاني.

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

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
- Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Dialect Resources

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	bγīt بغيت	kāyn كاين	mā kāynš ما كاينش
Egyptian	Tarabēza طربيزة	‘oTTa قطة	bitāς بتاع	ςāwez عاوز	fī في	mafīš مفيش
Syrian	Tāwle طاولة	bisse بسة	tabaς تبع	biddi بدي	fī في	mā fi 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

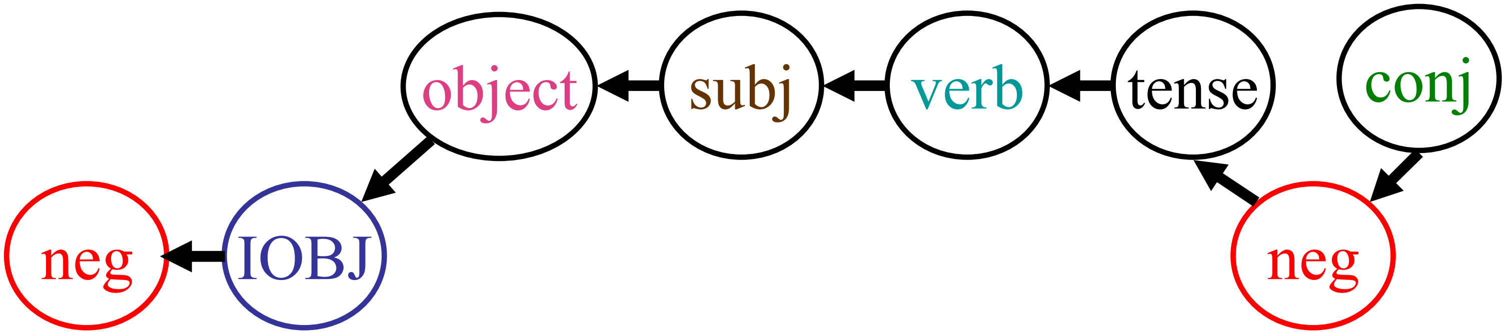
Tutorial Contents

- Introduction
- Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Dialect Resources

Morphological Variation

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

Morphological Variation
Verb Morphology



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

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

And you didn’t write it for him

Morphological Variation

	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/		هيكتب /hajiktib/
IRQ	كتب /kitab/	يكتب /jiktib/	ديكتب /dajiktib/		رح يكتب /raħ jiktib/
MOR	كتب /kteb/	يكتب /jekteb/	كيكتب /kjekteb/		غيكتب /ʁajekteb/

Morphological Variation

Verb conjugation

	Perfect			Imperfect		
	1S	2S♂	2S♀	1S	1P	2S♀
MSA	كتبتُ /katabtu/	كتبتَ /katabta/	كتبتِ /katabti/	اكتبُ /aktubu/	نكتبُ /naktubu/	تكتبينَ /taktubīna/ تكتبي /taktubī/
LEV	كتبت /katabt/		كتبتي /katabti/	اكتب /aktob/	نكتب /noktob/	تكتبي /toktobi/
IRQ	كتبت /kitabit/		كتبتي /kitabti/	اكتب /aktib/	نكتب /niktib/	تكتبين /tikitbīn/
MOR	كتبت /ktebt/	كتبتي /ktebti/		نكتب /nekteb/	نكتبوا /nektebu/	تكتبي /tektebi/

Tutorial Contents

- Introduction
- Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Dialect Resources

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
EGY	-	ده, دي, دول	
LEV	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 lA, lm, ln, mA	X	X
EGY	مش m$	ما ... ش mA ... $	X
LEV	ما, مش mA, m$	ما ... ش mA ... $	ش $

Sentence Word Order

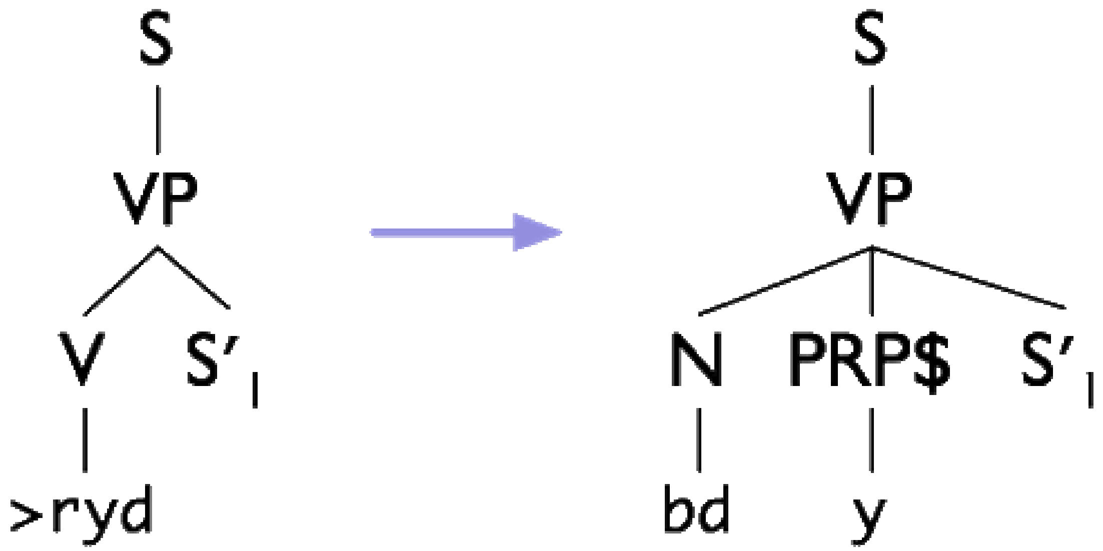
- Verbal sentences
  - The boys wrote the poems
  - MSA
    - Verb Subject Object (Partial agreement)
      كتب الاولاد الاشعار
      wrote$_{\text{masc}}$ the-boys the-poems
    - Subject Verb Object (Full agreement)
      الاولاد كتبوا الاشعار
      the-boys wrote$_{\text{mascPl}}$ the-poems
  - LEV, EGY
    - Subject Verb Object
      الاولاد كتبو الاشعار
      The-boys wrote$_{\text{mascPl}}$ the-poems
    - Less present: Verb Subject Object
      كتبو الاولاد الاشعار
      wrote$_{\text{mascPl}}$ 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
- Dialectal Phenomena
  - Orthography
  - Lexicon
  - Morphology
  - Syntax
  - Code switching
- Sample Applications
- Dialect Resources

Code Switching

MSA

LEV

MSA and Dialect mixing in speech

- phonology, morphology and syntax

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

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

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’

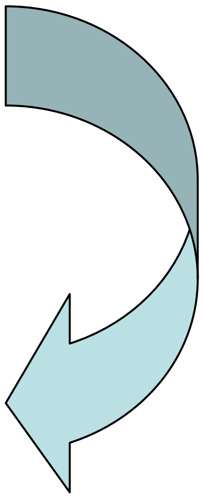
- A shift toward SVO rather than VSO in written MSA

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*

  - 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
- Dialectal Phenomena
- Sample Applications
  - Automatic speech recognition
  - Dictionary creation
  - Morphological analysis
  - Part-of-speech tagging
  - Syntactic parsing
  - Machine translation
- Dialect Resources

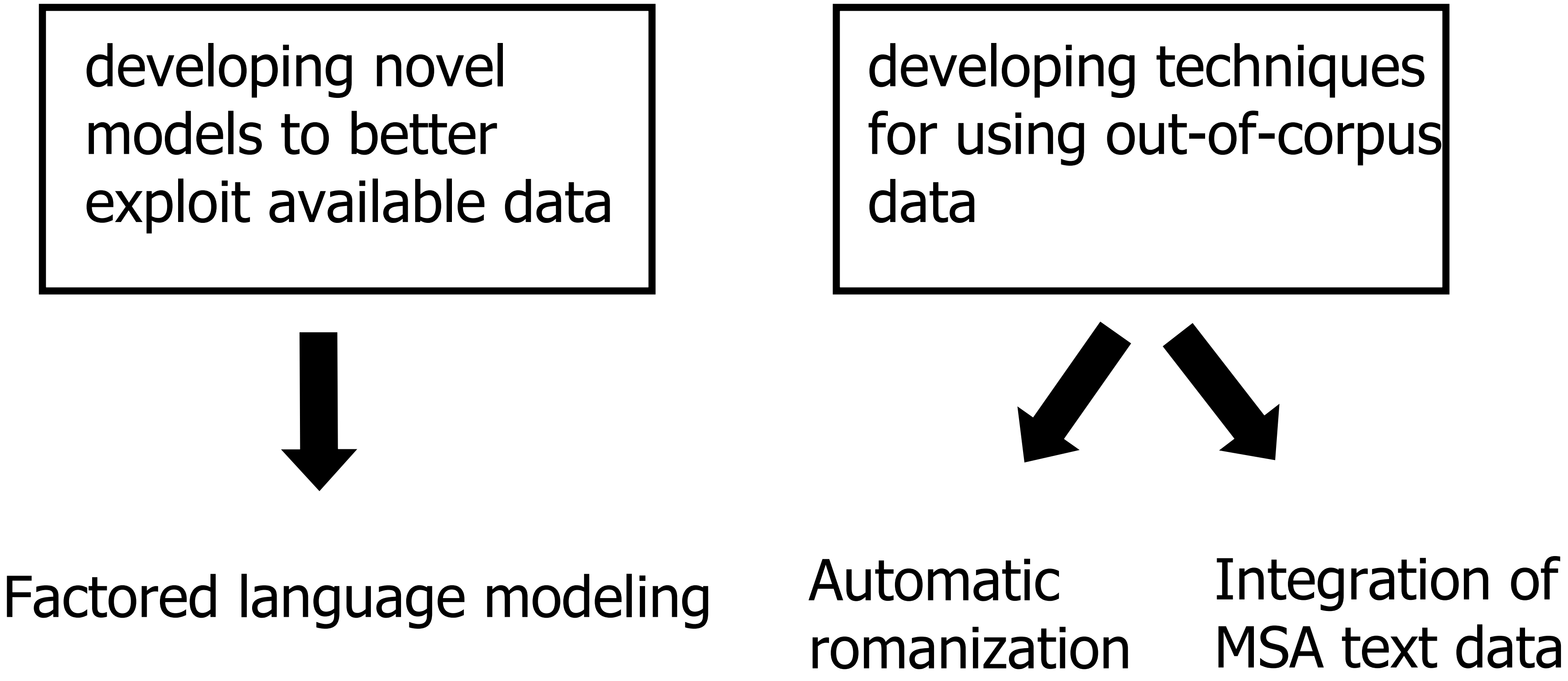
Arabic ASR: State of the Art

- 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

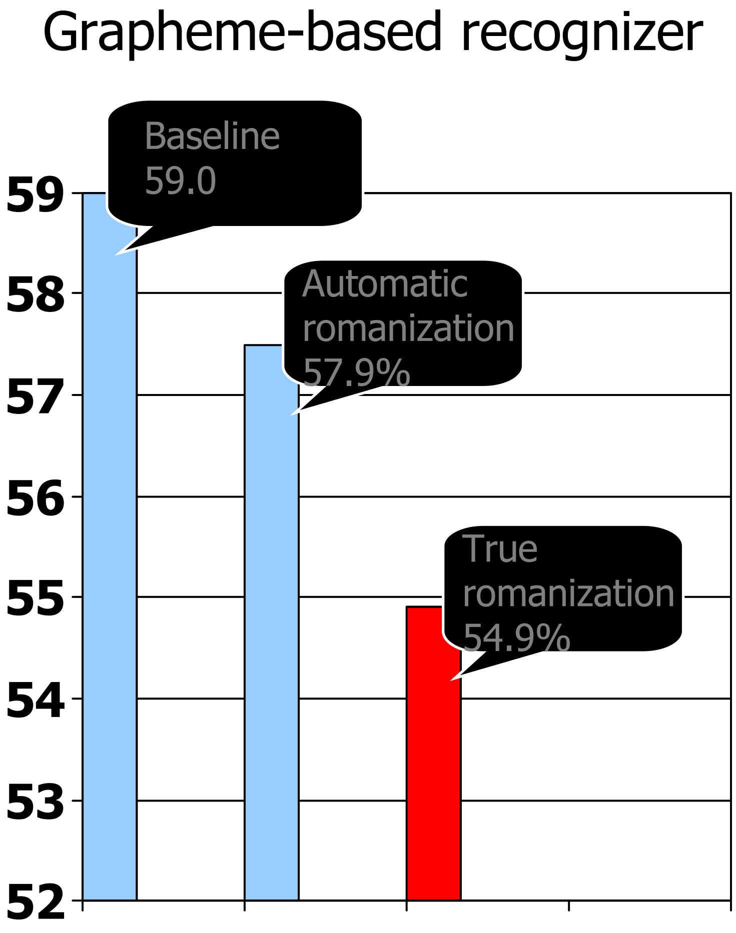
Integration of MSA text data

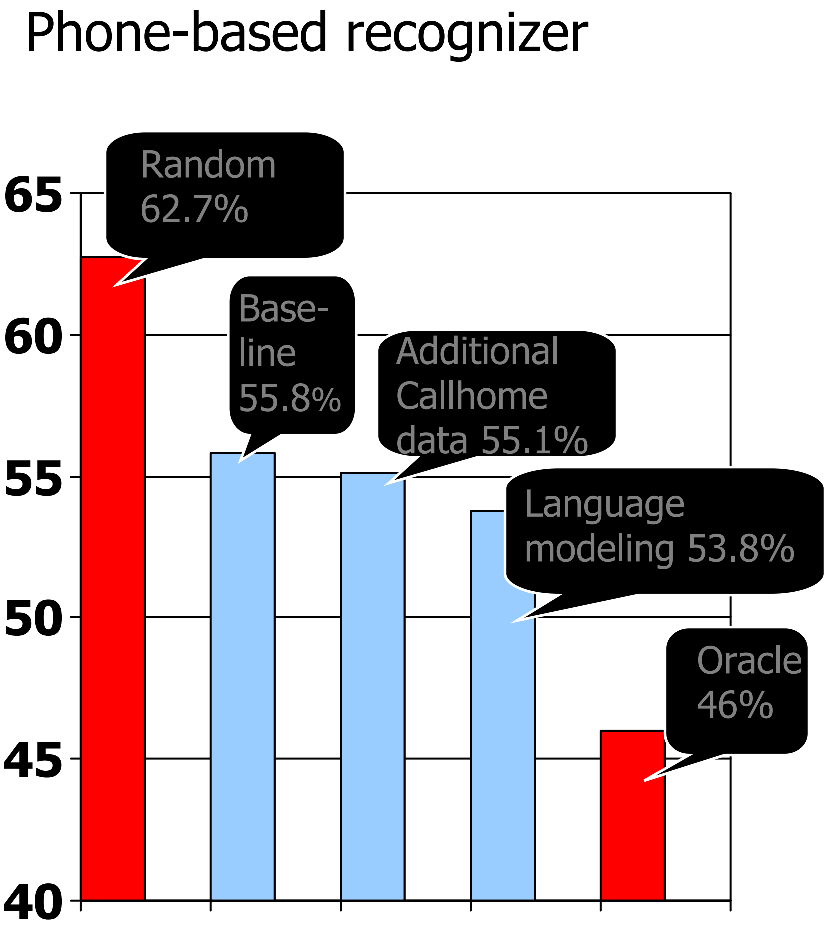
Slide courtesy of (Kirchhoff et al.2002)

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/Diacritization
  - try to predict vowelization automatically from data and use result for recognizer training
- Integrate data from MSA written sources

JHU WS02 Results (WER)





Slide courtesy of (Kirchhoff et al.2002)

Tutorial Contents

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

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 et al. 2006)

Tutorial Contents

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

Dialectal Morphological Analysis

- **MAGEAD** (Habash and Rambow 2006)
  - Morphological Analysis and GEneration for Arabic and its Dialects
- **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 (اِزْدَهَرَت)

The Lexeme

- Lexeme is an abstraction of all inflectional variants of a word
  - |كِتاب| 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=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

…

وَسَنَكْتُبُهَا

wasanaktubuhA

We will write it

Morphological Behavior Class

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

cnj=wa	→	wa+	wi+
tense=fut	→	sa+	Ha+
per=1 + num=sg	→	'+	
per=1 + num=pl	→	n+	n+
mood=indic	→	+u	+0
mood=sub	→	+a	
aspect=imper	→	V12V3	V12V3
aspect=perf	→	1V2V3	
voice=act	→	a-u	i-i
voice=pass	→	u-a	
obj=3FS	→	hA	hA
obj=1P	→	nA	

…

وَسَنَكْتُبُهَا

wasanaktubuhA

wiHaniktibhA

وِحَنِكْتِبْهَا

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=sg	→	'+			
per=1 + num=pl	→	n+	n+	→	[SUBJ_PRE_1P]
mood=indic	→	+u	+0	→	[SUBJ_SUF_Ind
mood=sub	→	+a			
aspect=imper	→	V12V3	V12V3	→	[PAT:I-IMP]
aspect=perf	→	1V2V3			
voice=act	→	a-u	i-i	→	[VOC:Iau-ACT]
voice=pass	→	u-a			
obj=3FS	→	hA	hA	→	[OBJ:3FS]
obj=1P	→	nA			

…

MSA EGY

Morphological Behavior Class

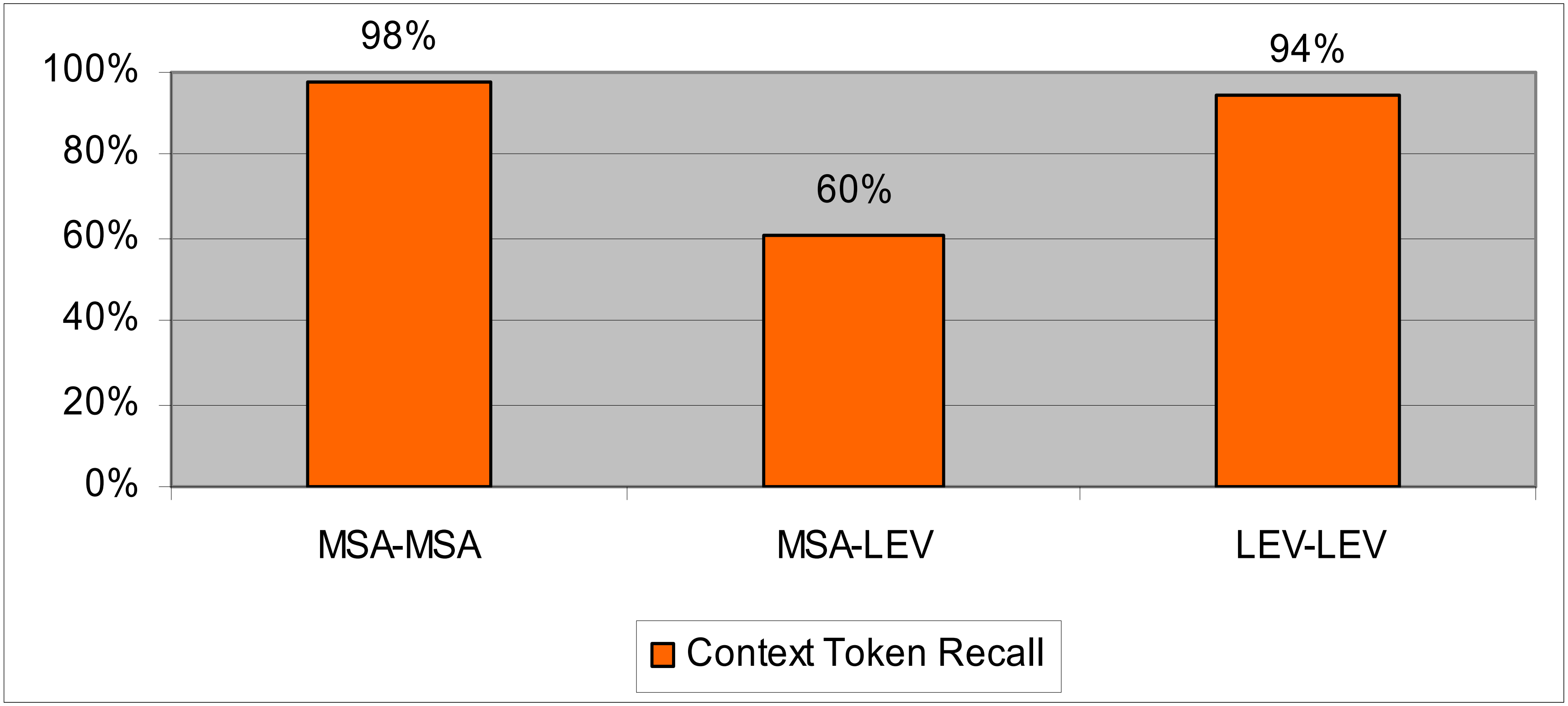
- MBC::Verb-I-au (*katab/yaktub*)

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



Tutorial Contents

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

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

Tutorial Contents

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

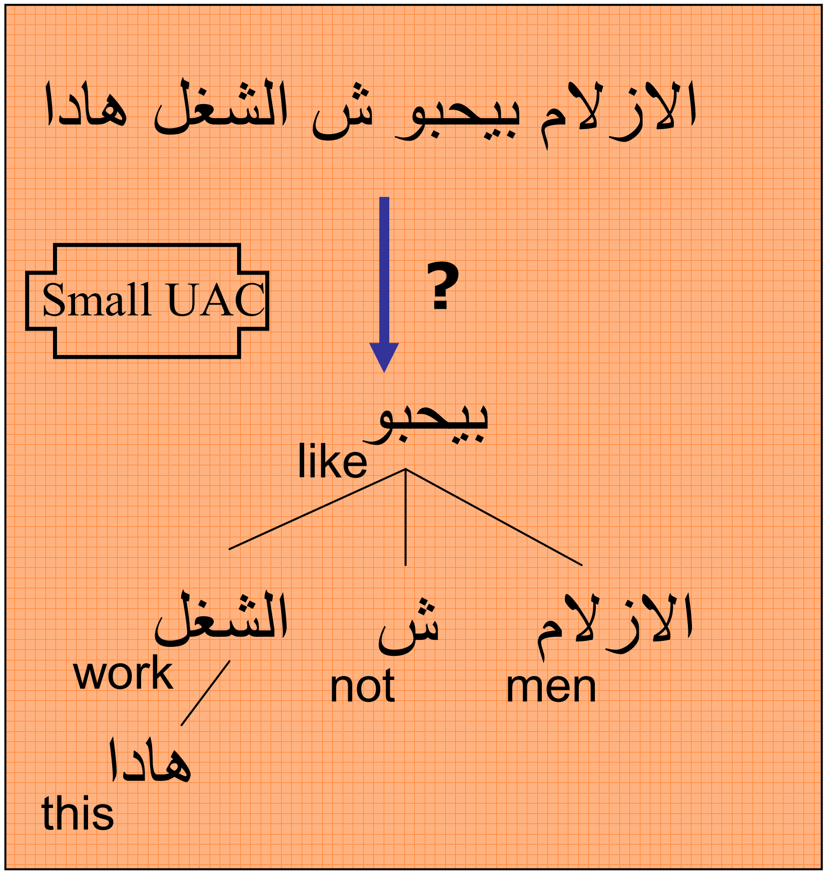
Arabic Dialect Parsing

- 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 -



الازلام بيحبو ش الشغل هادا

Small UAC

?

بيحبو
like

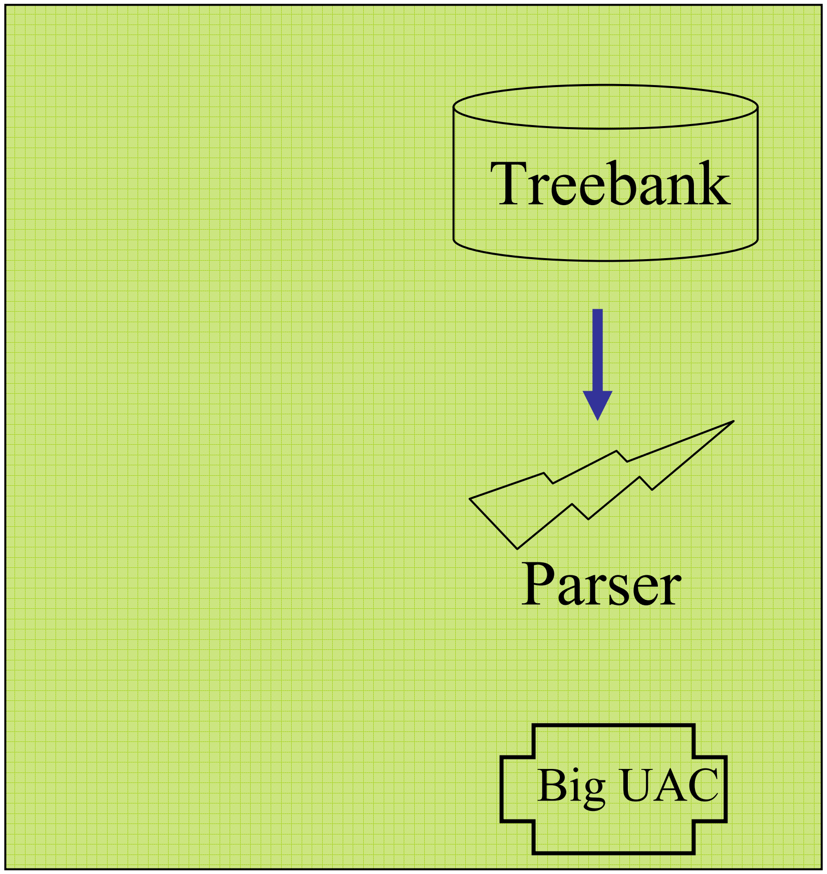
الشغل
work

ش
not

الازلام
men

هادا
this

- MSA -



Treebank

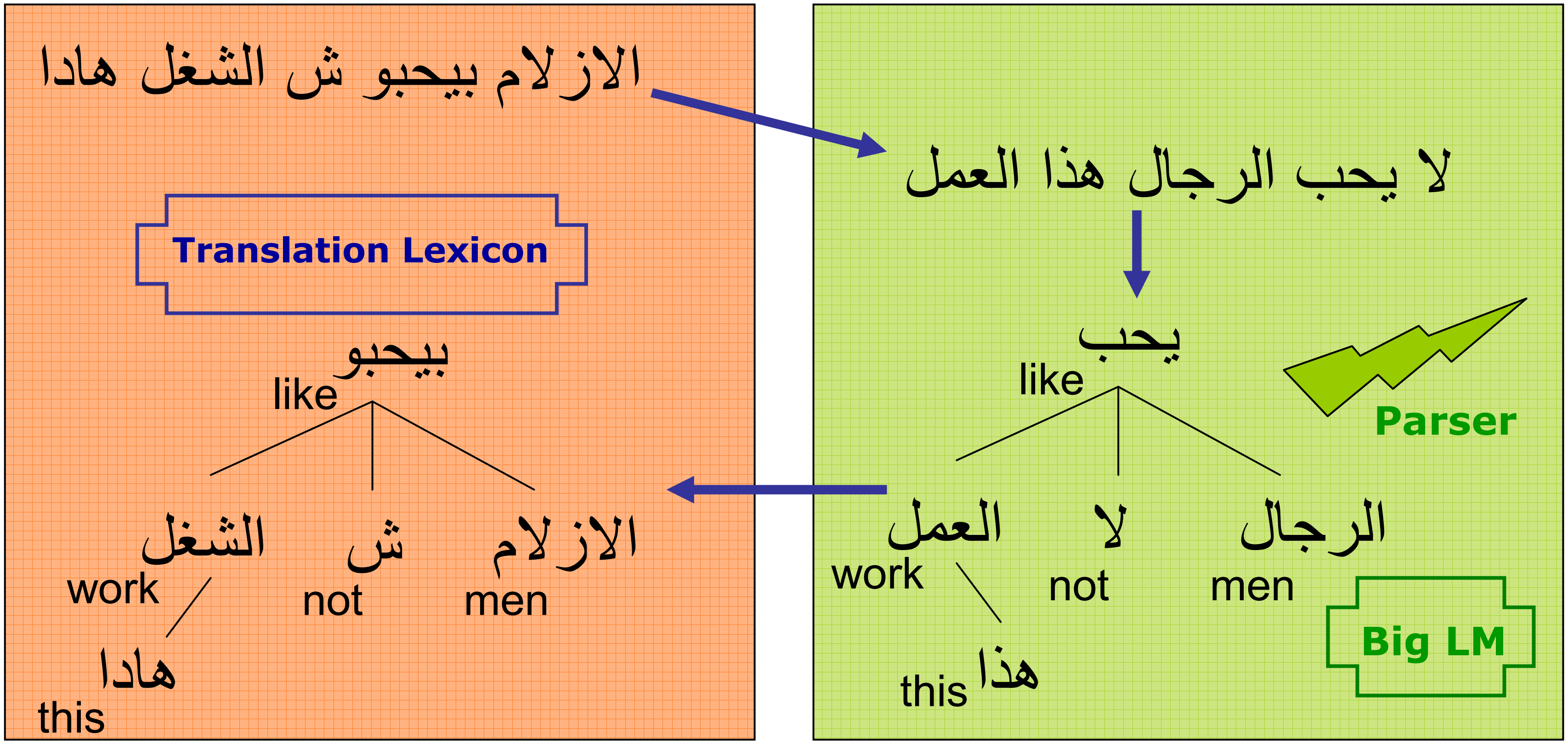
Parser

Big UAC

Sentence Transduction Approach

- Dialect -

- MSA -



الازلام بيحبو ش الشغل هادا

Translation Lexicon

بيحبو like

الشغل work

ش not

الازلام men

هادا this

لا يحب الرجال هذا العمل

يحب like

العمل work

لا not

الرجال men

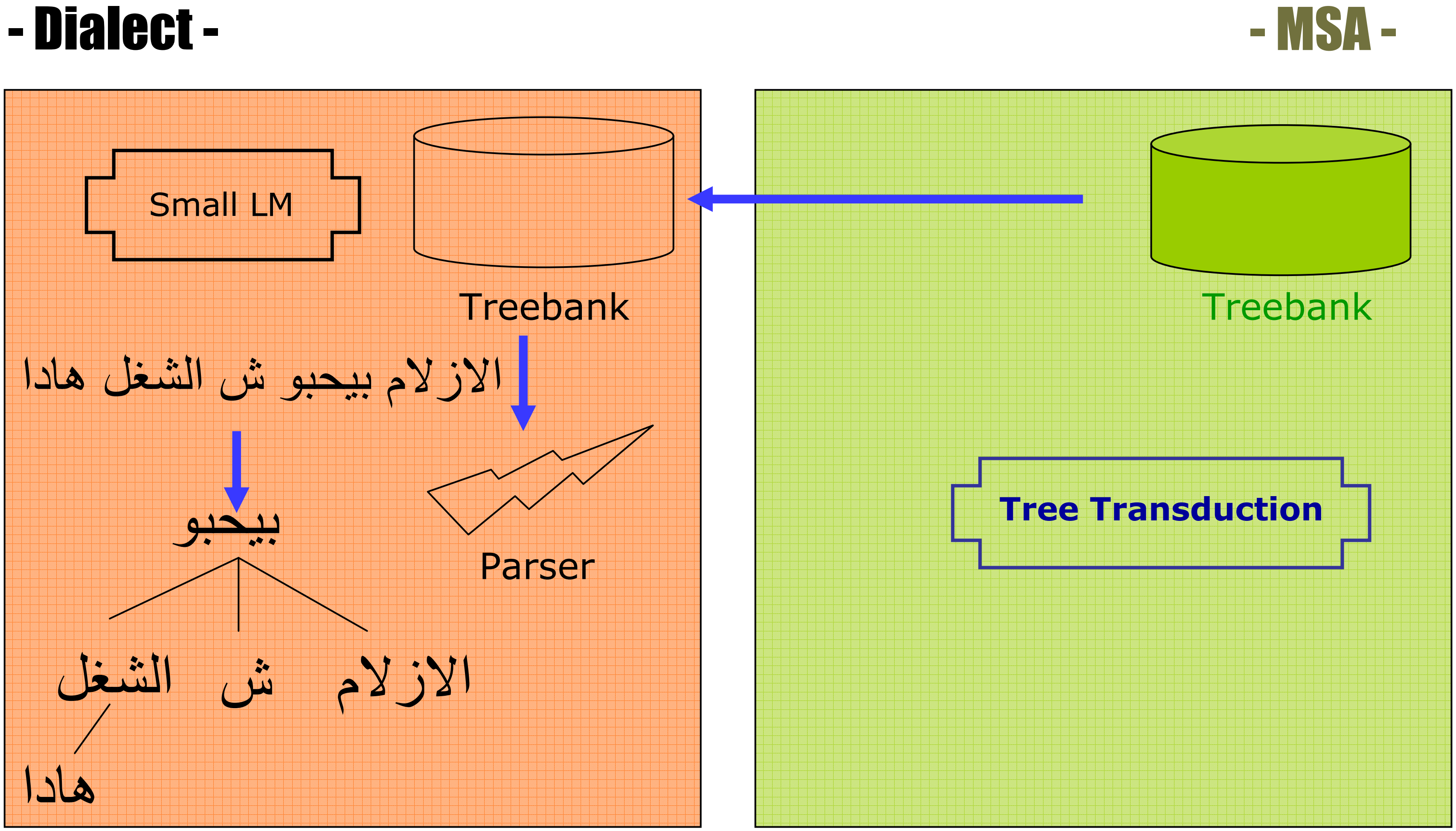
هذا this

Parser

Big LM

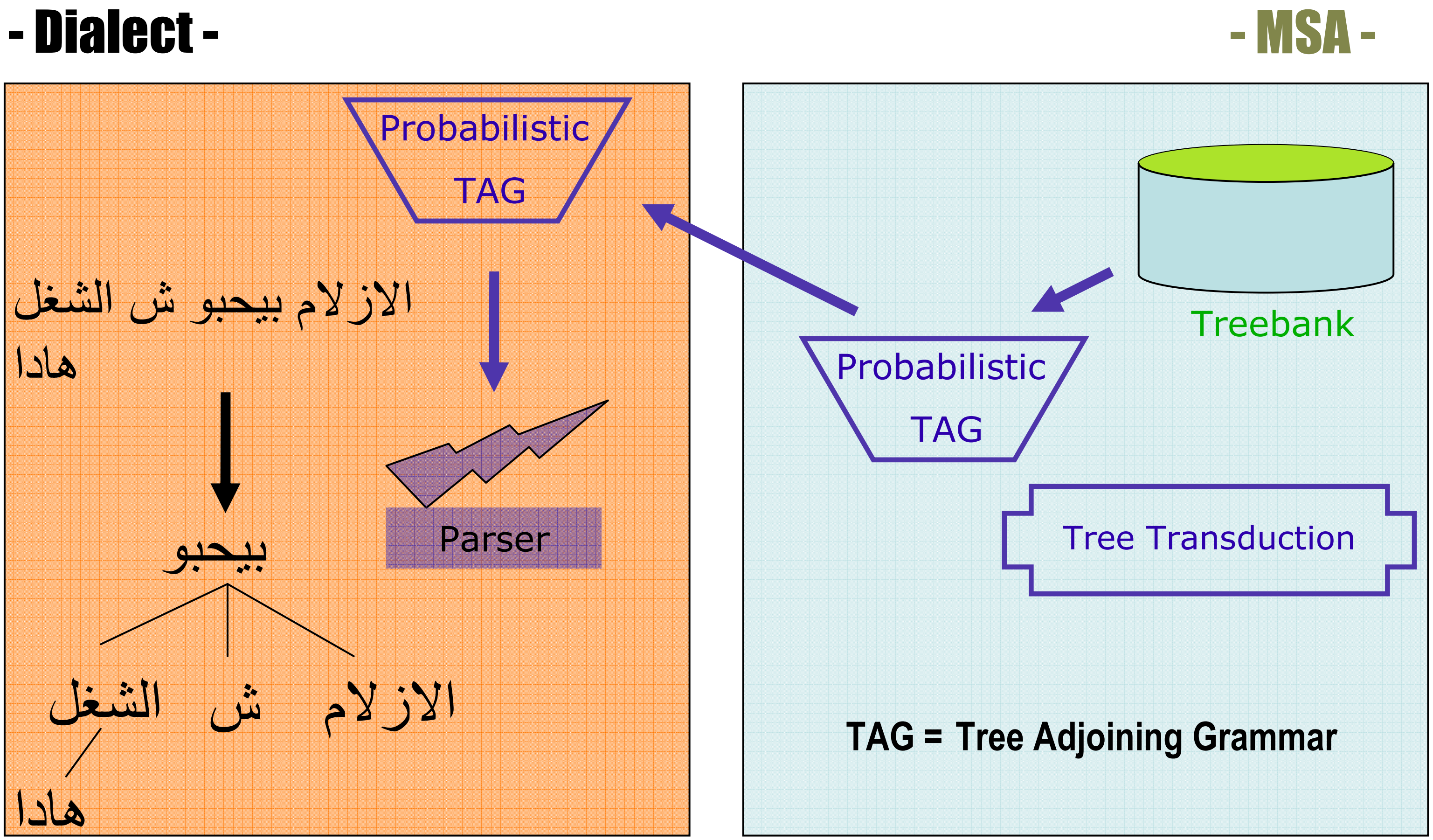
(Rambow et al. 2005; Chiang et al. 2006)

MSA Treebank Transduction



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

Grammar Transduction



(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)

Tutorial Contents

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

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)
- Dialect-MSA translation
  - (Rambow et al. 2005; Abo-Bakr et al., 2008)

Arabic Dialect Machine Translation

- **TransTac: DARPA Program on Trans**lation System for **Tac**tical 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)

Tutorial Contents

- Introduction
- Dialectal Phenomena
- Sample Applications
- Dialect Resources

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)

Other Tutorial Slides

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

MEDAR 2009

Cairo, Egypt
April 21, 2009

Arabic Dialect Processing

Mona Diab   Nizar Habash
Center for Computational Learning Systems
Columbia University
{mdiab,habash}@ccls.columbia.edu