EMNLP 2014, Doha, Qatar

Tutorial

Natural Language Processing of Arabic and its Dialects

Mona Diab

The George Washington University

mtdiab@gwu.edu

Nizar Habash

New York University Abu Dhabi

nizar.habash@nyu.edu

CADIM
Columbia Arabic Dialect Modeling

- Founded in 2005 at Columbia University
  - Center for Computational Learning Systems
- Arabic-focused Natural Language Processing
- Research Scientists
  - Mona Diab, Nizar Habash and Owen Rambow
  - Formal degrees in both Computer Science and Linguistics
  - Over 200 publications & numerous software releases
- **CADIM is now a multi-university consortium**
  - **Columbia U. (Rambow), George Washington U. (Diab) and New York U. Abu Dhabi (Habash)**

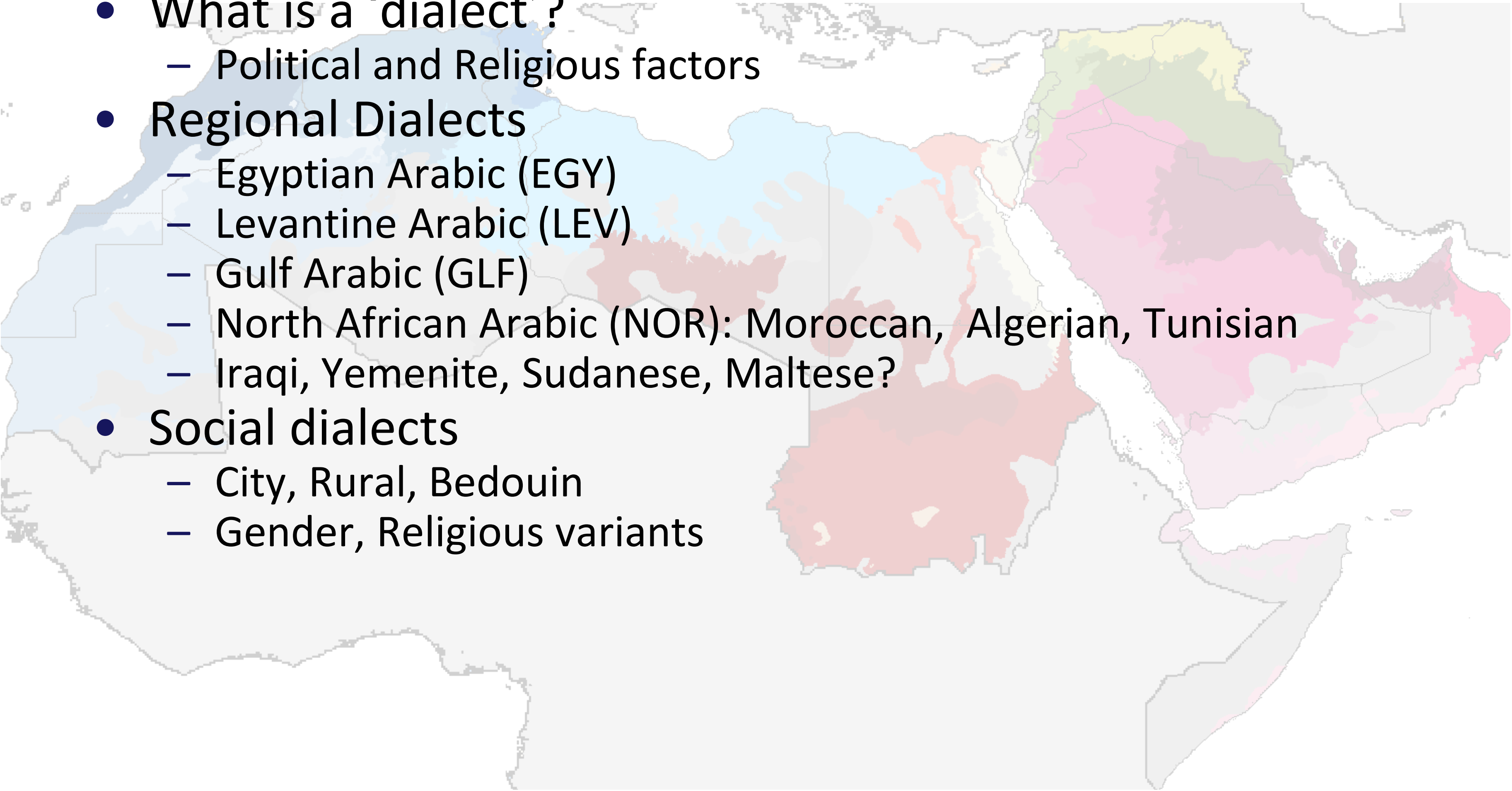
Tutorial Contents

- Introduction
  - The many forms of Arabic
- Orthography
  - Script, phonology and spelling, dialectal variations, spelling inconsistency, automatic spelling correction and conventionalization, automatic transliteration
- Morphology
  - Derivation and inflection, ambiguity, dialectal variations, automatic analysis and disambiguation, tokenization
- Syntax
  - Arabic syntax basics, dialectal variations, treebanks, parsing Arabic and its dialects
- Lexical Variation and Code Switching
  - Dialectal variation, lexical resources, code switching, automatic dialect identification
- Machine Translation
  - Tokenization, out-of-vocabulary reduction, translation from and into Arabic, dialect translation

Introduction

- Arabic is a Semitic language
- ~300M speakers
- 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

Arabic and its Dialects



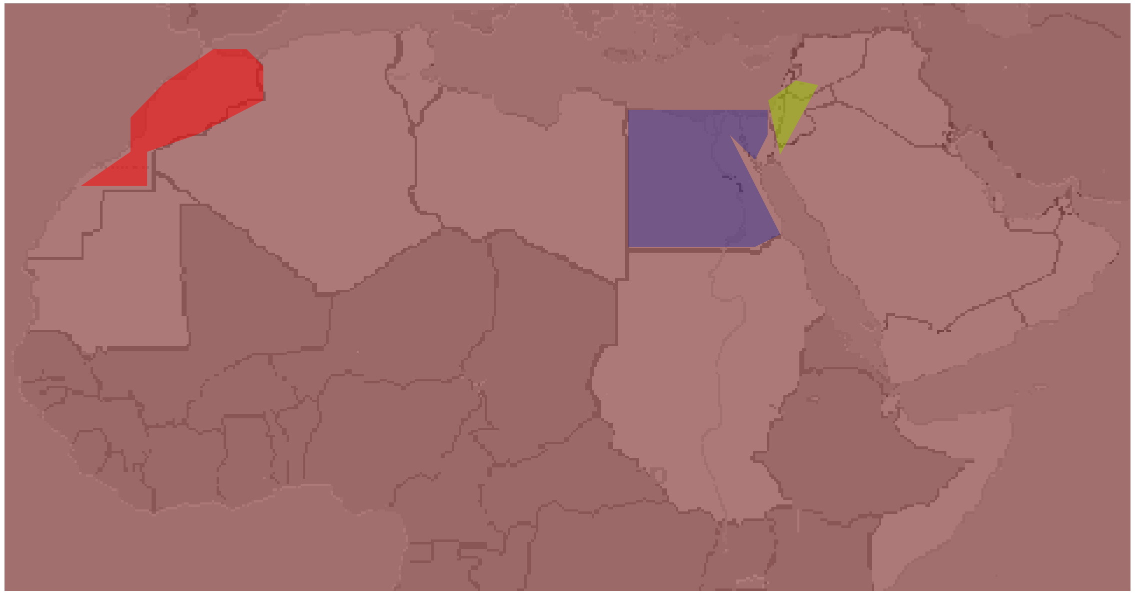
- Official language: Modern Standard Arabic (MSA)
  - ➢ No one's native language
- What is a 'dialect'?
  - – Political and Religious factors
- Regional Dialects
  - – Egyptian Arabic (EGY)
  - – Levantine Arabic (LEV)
  - – Gulf Arabic (GLF)
  - – North African Arabic (NOR): Moroccan, Algerian, Tunisian
  - – Iraqi, Yemenite, Sudanese, Maltese?
- Social dialects
  - – City, Rural, Bedouin
  - – Gender, Religious variants

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

Arabic Diglossia

	Formal	Informal
MSA	**Typical** MSA	*Telenovela Arabic* *MSA L2*
Dialect	Formal Spoken Arabic	**Typical** Dialect



lam jaʃtari kamāl ţawilatan ζadīdatan لم يشتر كمال طاولة جديدة

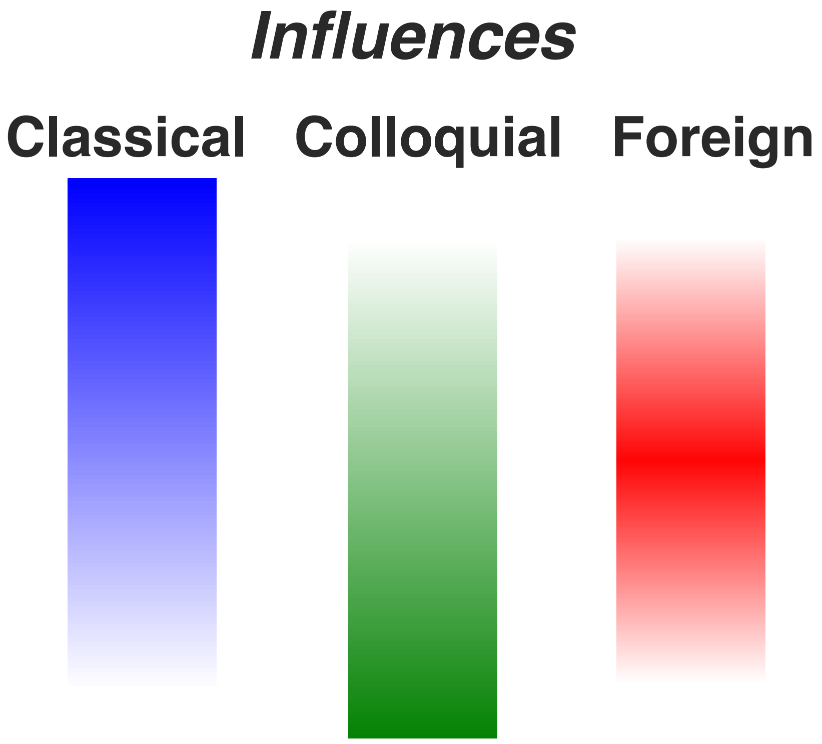
didn’t buy Kamel table new

kamāl maʃtarāʃ ţarabēza gidīda	●	كمال ماشتراش طربيزة جديدة
kamāl maʃtarāʃ ţawile ζdīde	●	كمال ماشتراش طاولة جديدة
kamāl maʃrāʃ mida ζdīda	●	كمال ماشراش ميدة جديدة

Kamel not-bought-not table new

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 is Arabic processing hard?

	Arabic	English
Orthographic ambiguity	More	Less
Orthographic inconsistency	More	Less
Morphological inflections	More	Less
Morpho-syntactic complexity	More	Less
Word order freedom	More	Less
Dialectal variation	More	Less

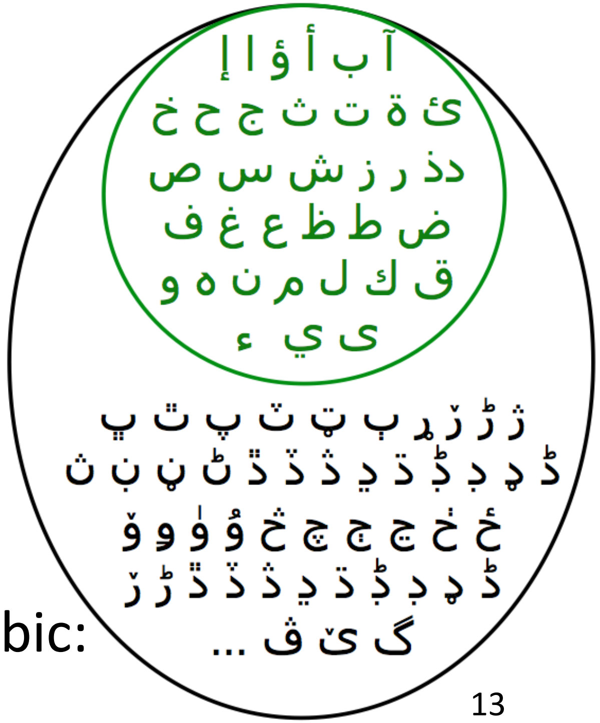
Tutorial Contents

- Introduction
  - The many forms of Arabic
- Orthography
  - Script, phonology and spelling, dialectal variations, spelling inconsistency, automatic spelling correction and conventionalization, automatic transliteration
- Morphology
  - Derivation and inflection, ambiguity, dialectal variations, automatic analysis and disambiguation, tokenization
- Syntax
  - Arabic syntax basics, dialectal variations, treebanks, parsing Arabic and its dialects
- Lexical Variation and Code Switching
  - Dialectal variation, lexical resources, code switching, automatic dialect identification
- Machine Translation
  - Tokenization, out-of-vocabulary reduction, translation from and into Arabic, dialect translation

Arabic Script

الخَطُ العَرَبِي

- An alphabet
- Written right-to-left
- Letters have allographic variants
- No concept of “capitalization”
- Optional diacritics
- Common ligatures
- Used to write many languages besides Arabic: Persian, Kurdish, Urdu, Pashto, etc.



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

ژ ڑ ڙ ړ ٻ ټ ٽ پ ٿ ڀ
ڈ ډ ڊ ڋ ڌ ې ڎ ڏ ڐ ٹ ڼ ٻ ٺ
ځ ڂ ڃ ڄ چ څ ۇ ۉ ۄ ۆ
ڈ ډ ڊ ڋ ڌ ې ڎ ڏ ڐ ڑ ڙ
گ ێ ڤ ...

Arabic Script

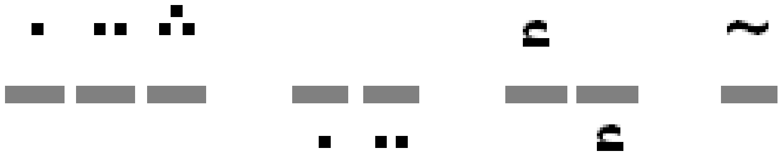
Alphabet

- letter forms

ا ٮ ح د ر س ص ط ع

ڡ ك ل م ں ه و ى ء

- 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

- Arabic script uses a set of optional diacritics
  - 6.8 diacritizations/word
  - Only 1.5% of words have at least one diacritic

Vowel				Nunation			Gemination
بَ	بُ	بِ	بْ	بً	بٌ	بٍ	بّ
/ba/	/bu/	/bi/	/b/	/ban/	/bun/	/bin/	/bb/

  - Combinable
    - /kattab/ *to dictate* كَتَّب

Arabic Script

Putting it together

Simple combination

Arab /ʕarab/ عَ رَ ب ← عَرَب = عرب

West /ʁarb/ غَ رْ ب ← غَرْب = غرب

Ligatures

Peace /salām/ س ل ا م ← ~~سلـام~~ سلام

اسبانيا تنفي تجميد المساعدة الممنوحة للمغرب
مدريد 1 - 11 (اف ب)- اكد رئيس الحكومة الاسبانية خوسيه ماريا
اثنار اليوم الخميس ان اسبانيا لم توقف المساعدة التي تقدمها للمغرب خلافا
لما اكده امس الاربعاء وزير الشؤون الخارجية والتعاون المغربي محمد بن
عيسى امام مجلس النواب المغربي . وقال رئيس الحكومة الاسبانية في
مؤتمر صحافي ان التعاون بين اسبانيا والمغرب لم يتوقف ابدا ولم يجمد.

اِسْبانِيا تَنْفِي تَجْمِيدَ المُساعَدَةَ المَمْنُوحَةَ لِلمَغْرِبِ
مَدْرِيد 1 - 11 (اِف ب)- اَكَّدَ رَئِيسُ الحُكُومَةِ الاِسْبانِيَّةُ خُوسِيه مارِيا
اثنار اليَوْمَ الخَمِيسَ اَنَّ اِسْبانِيا لِمَ تَوَقَّفَ المُساعَدَةُ الَّتِي تُقَدِّمُها لِلمَغْرِبِ خِلافاً
لِما اَكَّدَهُ اَمْسِ الاَرْبِعاءَ وَزِيرَ الشُؤُونِ الخارِجِيَّةِ وَالتَعاوُنِ المَغْرِبِيِّ مُحَمَّد بِن
عِيسَى اَمامَ مَجْلِسِ النُوّابِ المَغْرِبِيِّ . وَقالَ رَئِيسُ الحُكُومَةِ الاِسْبانِيَّةِ فِي
مُؤْتَمَرٍ صَحافِيٍّ اَنَّ التَعاوُنَ بَيْنَ اِسْبانِيا وَالمَغْرِبِ لِمَ يَتَوَقَّفْ اَبَداً وَلِمَ يُجَمِّدْ.

Arabic Script

Tatweel

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

حقوق الانسان

حقـوق الانسـان

حقـــوق الانســـان

حقـــــوق الانســـــان

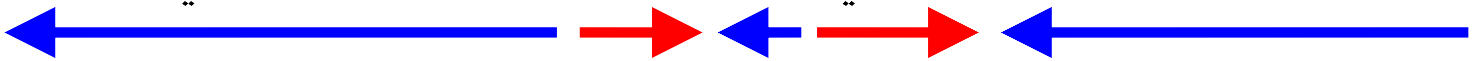
human rights /ħuqūq alʔinsān/

Arabic Script

“Arabic” Numerals

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

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



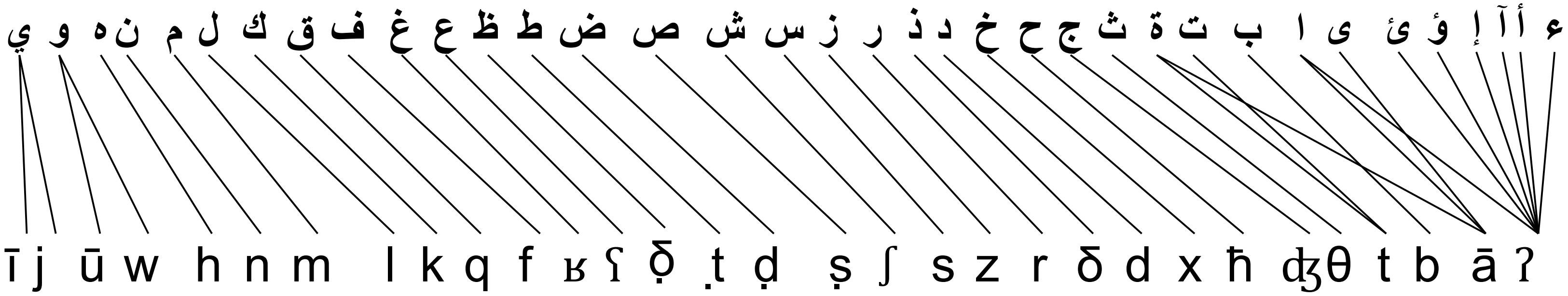
Algeria achieved its independence in 1962 after 132 years of French occupation.

- 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 ʤθ 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

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

دور	/dawr/ role,part
	/dūr/ houses
	/dawwar/ to rotate

Phonology and Spelling

- Arabic spelling is mostly phonemic …

Except for (continued)

- Morphophonemic characters
  - Ta-Marbuta feminine marker ة

/kabīr/ (big ♂)	كبير
/kabīra/ (big ♀)	كبيرة

  - Alif-Maqsura derivation marker

to disobey	عصى
a stick	عصا

- Hamza variants: 6 characters (ء أآإؤئ) for one phoneme (/’/)!

baha’ +3MascSing (his glory)	بهاءه بهاؤه بهائـه

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 deterministically inferable

  ths is wht an Arbc txt lks lik wth no vwls

Will revisit ambiguity in more detail again under morphology discussion

Proper Name Transliteration

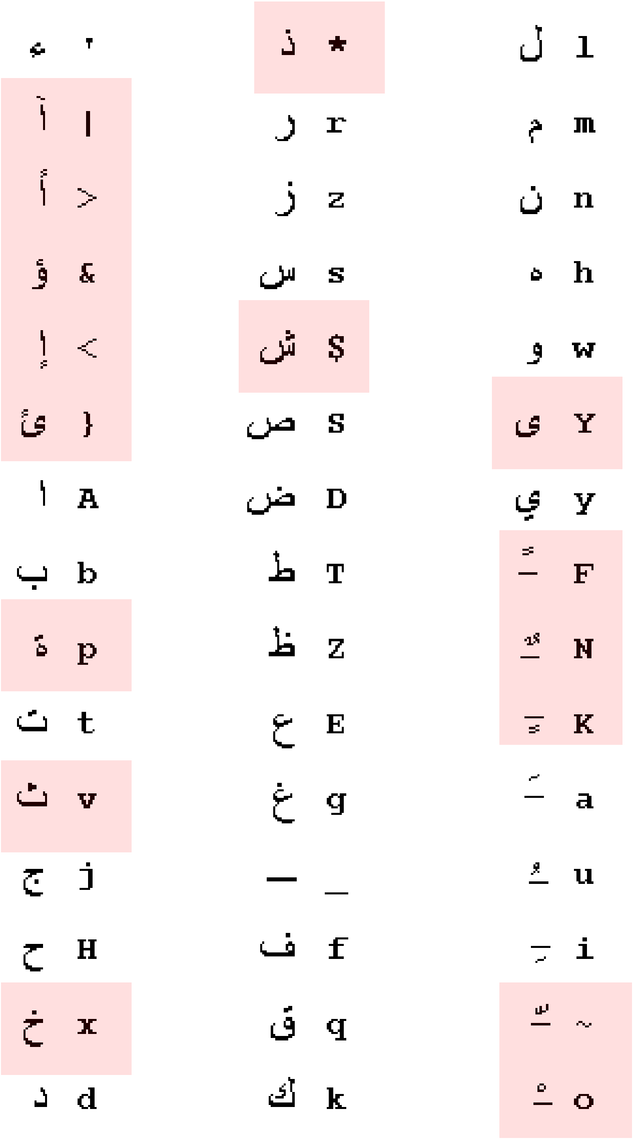
- The Qaddafi-Schwarzenegger problem
  - Foreign Proper name spelling is often ad hoc
  - Multiplicity of spellings causes increased sparsity

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

Transliteration

Buckwalter's Scheme

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



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

Dialectal 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: /fust̲ān/ → LEV: /fus̲tān/ ‘dress’

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

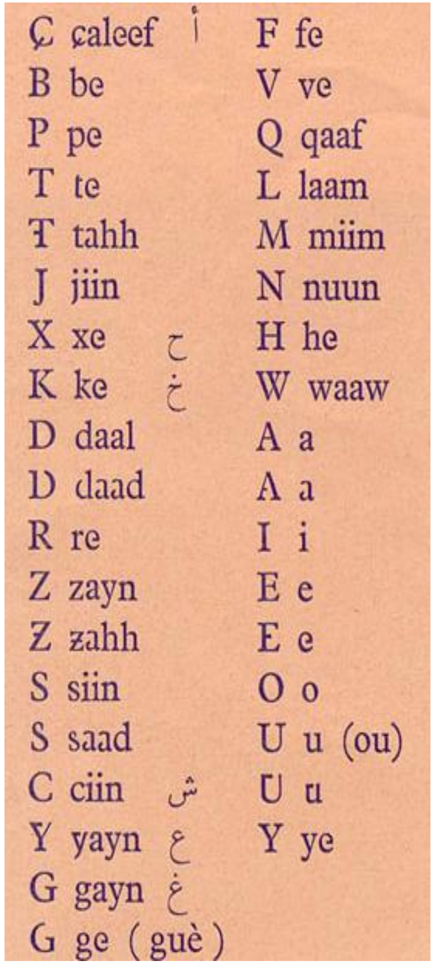
- Historical variants: MSA (ف, ق) = MOR (ڢ, ڧ)
- Modern proposals: LEV /ʔ/ ٯٔ , /ē/ ى̭ , /ō/ ۆ (Habash 1999)

Latin Script for Arabic?

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

عربي	IPA	Latin	عربي	IPA	Latin
أإآءؤئ	/ʔ/	‘ 2 ∅	ث	/θ/	th
ة	/a/,/t/	a t	ط	/ṭ/	t T 6
ح	ħ	H h 7	ع	/ʕ/	‘ 3 ∅
خ	/x/	kh 7’ x 8	غ	/ʁ/	g gh 3’
ذ	/δ/	th	ق	/q/	q
ش	/ ʃ/	sh ch	ي	/y/ /ay/ /ī/ /ē/	y,i,e, ai,ei,…

Akl 1961



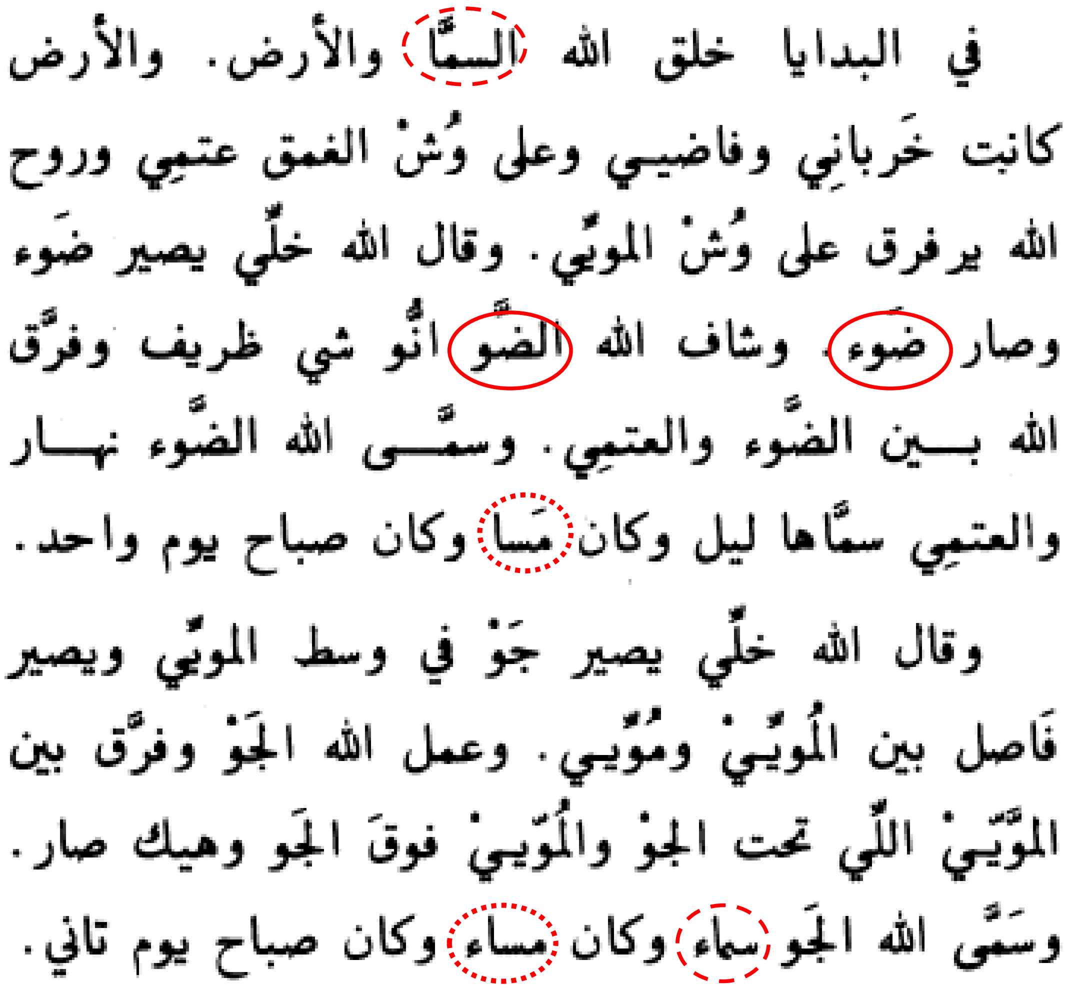
Ç çaleef	أ	F fe
B be		V ve
P pe		Q qaaf
T te		L laam
Ŧ tahh		M miim
J jiin		N nuun
X xe	ح	H he
K ke	خ	W waaw
D daal		A a
D daad		A a
R re		I i
Z zayn		E e
Ƶ ƶahh		E e
S siin		O o
S saad		U u (ou)
C ciin	ش	U u
Y yayn	ع	Y ye
G gayn	غ	
G ge (guè)		

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 السمّاً والأرض. والأرض
كانت خَربانِي وفاضيـي وعلى وُشْ الغمق عتمِي وروح
الله يرفرق على وُشْ المويّي. وقال الله خلّي يصير ضَوء
وصار ضوء. وشاف الله الضّو انُّو شي ظريف وفرّق
الله بــين الضّوء والعتمِي. وسمّـــى الله الضّوء نهـــار
والعتمِي سمّاها ليل وكان مَسا وكان صباح يوم واحد.

وقال الله خلّي يصير جَوّ في وسط المويّي ويصير
فَاصل بين الْمويّيْ ومُوّيي. وعمل الله الجَوْ وفرّق بين
المّويّيْ اللّي تحت الجوْ والْمُوّييْ 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

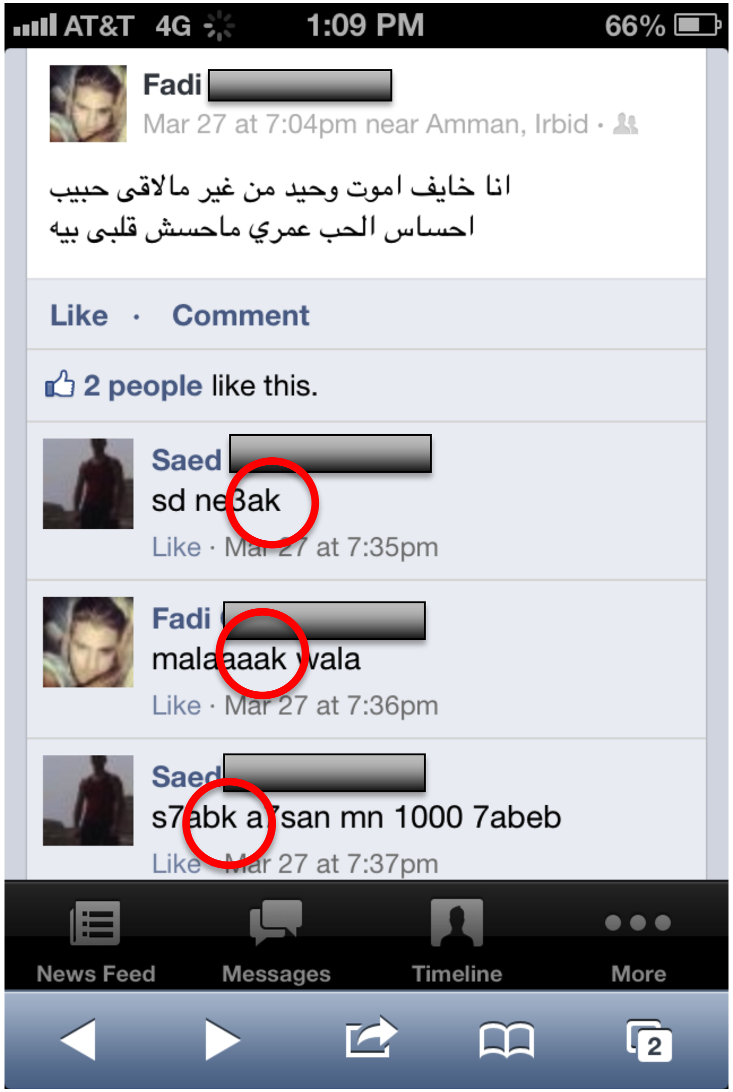
- ya alain lesh el 2aza
  ti7keh 3anneh kaza w kaza
  iza bidallak ti7keh hek
  2areeban ra7 troo7 3al 3aza

  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

Spelling Inconsistency III

- Social media spelling variations
  - +ak
  - +aaaaak
  - +k



Fadi [redacted]
Mar 27 at 7:04pm near Amman, Irbid ·

انا خايف اموت وحيد من غير مالاقى حبيب
احساس الحب عمري ماحسش قلبى بيه

2 people like this.

Saed [redacted]
sd ne3ak
Like · Mar 27 at 7:35pm

Fadi [redacted]
malaaaak wala
Like · Mar 27 at 7:36pm

Saed [redacted]
s7abk a7san mn 1000 7abeb
Like · Mar 27 at 7:37pm

CODA: A Conventional Orthography for Dialectal Arabic

- Developed by CADIM for computational processing
- Objectives
  - CODA covers all DAs, minimizing differences in choices
  - CODA is easy to learn and produce consistently
  - CODA is intuitive to readers unfamiliar with it
  - CODA uses Arabic script
- Inspired by previous efforts from the LDC and linguistic studies

CODA Examples

CODA	الامتحانات	قبل	اللي	الفترة	صحابي	ما شفتش
gloss	*the exams*	*before*	*which*	*the period*	*my friends*	*I did not see*
Spelling variants	الإمتحانات	أبل	اللـــى	الفتره	صحابـــى	ماشفتش
	الـمتحانات	ابل	إللي	الفـطـرة	صـوحابي	مـشفتش
	الامتــحــنات	abl	إللـــى	الفـطـره	صـوحابـــى	ما شـوفتش
	الإمتــحــنات	qbl	الـــي	ilftra	Su7abi	ماشـوفتش
	الـمتــحــنات	qabl	الى		sohaby	مـشـوفتش
	ilimti7anat		إلـــي			mashoftish
	limtihanaat		إلى			
			illi			

CODA Examples

Phenomenon	Original	CODA
Spelling Errors	الاجابه	الإجابة
Typos	شبب	سبب
Speech effects	كبييييييير	كبير
Merges	اليومبريستيج	اليوم بريستيج
Splits	المع روف	المعروف
MSA Root Cognate	آلب، كلب	قلب
Dialectal Clitic	عهلبيت	عهالبيت
Guidelines	مشفناش	ما شافناش
Unique Dialect Words	بردو، برضو	برضه

CODAFY

Raw Orthography to CODA Converter Egyptian Arabic

- What:

- Converts from raw DA orthography to CODA
- Corrects typos and various speech effects

- CODA Conventions:

– **Phonology**:
relate some DA words to their MSA cognates

– **Morphology**:
preserve DA morphology with consistent choices

– **Lexicon**:
select a spelling convention for DA-only words

- Example:

Input	مشفتش صحابى الفتره الى فاتت **m$ft$ SHAbY Alftrh AlY fAtt**
Output	ما شفتش صحابي الفترة اللي فاتت **mA $ft$ SHAby Alftrp Ally fAtt**

- Evaluation:

CODAfication	Accuracy (tokens)	A/Y Norm. Accuracy (tokens)
Baseline (doing nothing)	76.8%	90.5%
CODAFY v0.4	91.5%	95.2%

MT (no tokenization)	BLEU
Baseline	22.1
CODAFY v0.4	22.6

- Used In: MADA-ARZ
  - Accessed through the MADA-ARZ configuration file

3arrib

CADIM’s Arabizi-to-Arabic Conversion

- We developed a system for automatic mapping of Arabizi to Arabic script
  1. train finite state machines to map Arabizi to Arabic
     113K words of Arabizi-Arabic (Bies et al., 2014 – EMNLP Arabic NLP Workshop)
  2. restrict choices using the CALIMA-ARZ morphological analyzer
  3. rerank using a 5-gram Egyptian Arabic LM
  4. tag punctuation, emoticons, sounds, foreign words and names
- Evaluation
  - test 32K words
  - transliteration correct 83.6% of Arabic words and names.

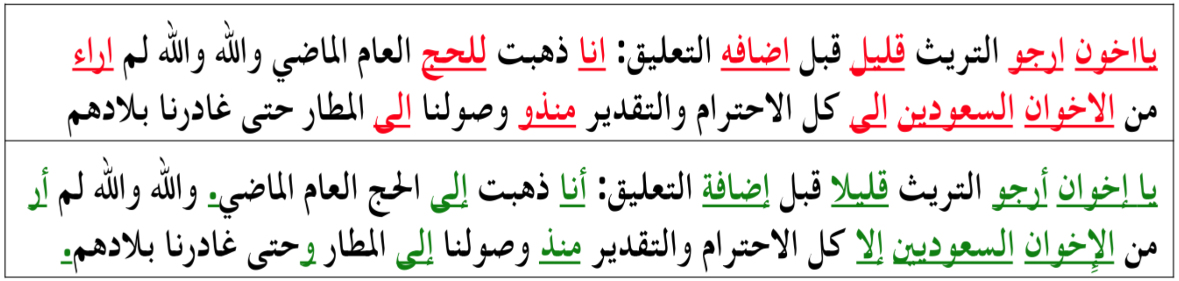
ana msh 3aref a2ra elly enta katbo AnA m$ EArf AqrA Ally Ant kAtbh انا مش عارف اقرا اللي انت كاتبه	w fel aa5er tele3 fshenk w mab2raash **arabic** w fl Axr TlE f$nk w mab2raash ArAbyk و+ فال+ اخر طلع فشنك و mab2raash ارابيك

(Al-Badrashiny et al., CONLL 2014; Eskander et al., EMNLP CodeSwitch Workshop 2014)

Qatar Arabic Language Bank

- Spelling errors in unedited Standard Arabic text

32% WER



ياخون ارجو التريث قليل قبل اضافه التعليق: انا ذهبت للحج العام الماضي والله والله لم اراء من الاخوان السعودين الي كل الاحترام والتقدير منذو وصولنا الي المطار حتى غادرنا بلادهم
يا إخوان أرجو التريث قليلا قبل إضافة التعليق: أنا ذهبت إلى الحج العام الماضي. والله والله لم أر من الإخوان السعوديين إلا كل الاحترام والتقدير منذ وصولنا إلى المطار وحتى غادرنا بلادهم.

- QALB – Qatar Arabic Language Bank
  - A collection of 2M words of unedited native and non-native text
  - The largest portion of the corpus is from Aljazeera comments
  - Manually corrected by a team of annotators
  - Data is public (from shared task site)
- Project site: http://nlp.qatar.cmu.edu/qalb/
- EMNLP 2014 Arabic NLP Shared Task
  - Nine teams participated
  - http://emnlp2014.org/workshops/anlp/shared_task.html

(Zaghouani et al., LREC 2014; Mohit et al., EMNLP Arabic NLP W., 2014)

Tutorial Contents

- Introduction
  - The many forms of Arabic
- Orthography
  - Script, phonology and spelling, dialectal variations, spelling inconsistency, automatic spelling correction and conventionalization, automatic transliteration
- Morphology
  - Derivation and inflection, ambiguity, dialectal variations, automatic analysis and disambiguation, tokenization
- Syntax
  - Arabic syntax basics, dialectal variations, treebanks, parsing Arabic and its dialects
- Lexical Variation and Code Switching
  - Dialectal variation, lexical resources, code switching, automatic dialect identification
- Machine Translation
  - Tokenization, out-of-vocabulary reduction, translation from and into Arabic, dialect translation

Morphology

- Form
  - 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

    ك ت ب

    k=1 t=2 b=3

  - Pattern

ma12ū3 *passive participle*	1ā2i3 *active participle*

  - Lexeme

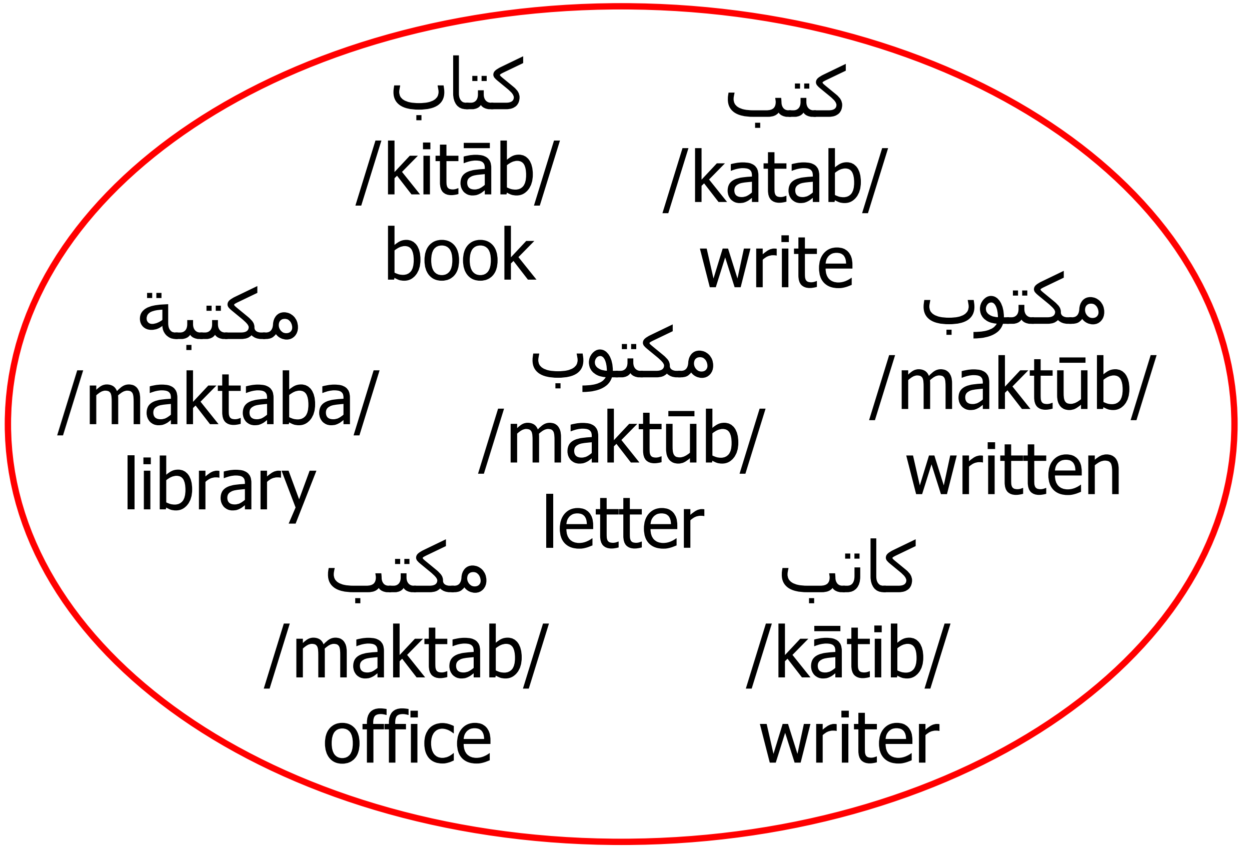
مكتوب maktūb *written*	كاتب kātib *writer*

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

Derivational Morphology

Root Meaning

ك ت ب KTB = notion of “*writing*



كتاب
/kitāb/
book

كتب
/katab/
write

مكتبة
/maktaba/
library

مكتوب
/maktūb/
letter

مكتوب
/maktūb/
written

مكتب
/maktab/
office

كاتب
/kātib/
writer

Root Polysemy

LHM-1 لحم

“meat”

لحم /laħm/

Meat

لحام /laħħām/

Butcher

LHM-2 لحم

“battle”

ملحمة /malħama/

Fierce battle

Massacre

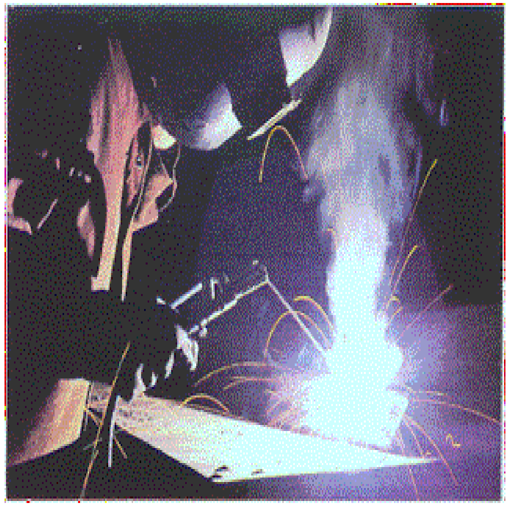
Epic

LHM-3 لحم

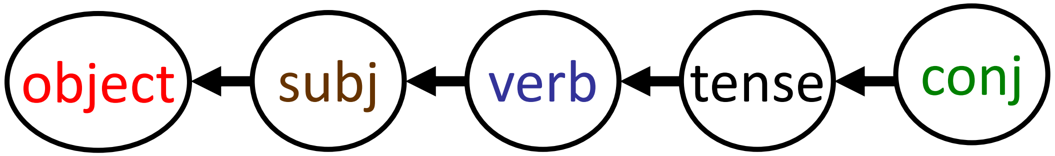
“soldering”

لحم /laħam/

Weld, solder, stick, cling



MSA Inflectional Morphology
Verbs



object ← subj ← verb ← tense ← conj

فقلناها

/faqulnāhā/

ف+ قال+ نا+ ها

fa+qul+na+hā

so+said+we+it

So we said it.

وسنقولها

/wasanaqūluhā/

و+ س+ ن+ قول + ها

wa+sa+na+qūl+u+hā

and+will+we+say+it

And we will say it

- Morphotactics
- Subject conjugation (suffix or circumfix)

Inflectional Morphology

***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

Inflectional Morphology

Terminology

Word	A space/punctuation delimited string	lilmaktabapi
Lexeme	The **set** of all inflectionally related words	maktabap, lilmaktabapi, Almaktabapu, walimaktabatihA, etc.
Lemma	An ad hoc word form used to represent the lexeme	maktabap
Features	The space of variation of words in a lexeme	Clitics: li_prep, Al_det, Gen:f, num:s, stt:d, cas:g
Root جذر	The root morpheme of the Lexeme	k-t-b
Stem جذع	The core root+pattern substring; it does not include any affixes	maktab
Segmentation	A shallow separation of affixes	li+l+maktab+ap+i
Tokenization	Segmentation + morpheme recovery	li+Al+maktab+ap+i

Inflectional Features

	Feature Name		(Some Important) Feature Values	
PER	Person	الشخص	1st, 2nd, 3rd, na	متكلم، مخاطب، غائب، غ/م
ASP	Aspect	الزمن	perfect, imperfect, command, na	ماضي، مضارع، أمر، غ/م
VOX	Voice	البناء	active, passive, na	للمعلوم، للمجهول، غ/م
MOD	Mood	الصيغة	indicative, subjunctive, jussive, na	مرفوع، منصوب، مجزوم، غ/م
GEN	Gender	الجنس	feminine, masculine, na	مؤنث، مذكر، غ/م
NUM	Number	العدد	singular, dual, plural, na	مفرد، مثنى، جمع، غ/م
STT	State	التعريف	indefinite, definite, construct, na	نكرة، معرفة، مضاف، غ/م
CAS	Case	الحالة	nominative, accusative, genitive, na	مرفوع، منصوب، مجرور، غ/م

Cliticization Features

Feature Name			(Some Important) Feature Values	
PRC3	Proclitic 3	سابقة 3	<a_ques, 0	أداة استفهام،0
PRC2	Proclitic 2	سابقة 2	fa_conj, wa_conj, 0	حروف عطف،0
PRC1	Proclitic 1	سابقة 1	bi_prep, li_prep, sa_fut, 0	حروف جر، سين الاستقبال، 0
PRC0	Proclitic 0	سابقة 0	Al_det, mA_neg, 0	ال التعريف، أداة نفي،0
ENC0	Enclitic	لاحقة 0	3ms_dobj, 3ms_poss, …, 0	ضمير مفعول به مباشر مفرد مذكر للغائب، ضمير ملكية مفرد مذكر للغائب، ... ، 0

Part-of-Speech

- *Traditional POS tagset*: Noun, Verb, Particle
- Many tag sets exist (from size 3 to over 22K tags)
  - Core Computational POS tags (~34 tags)
    - NOUN, ADJ, ADV, VERB, PREP, CONJ, etc.
    - Collapse or refine core POS
    - Extend tag with some or all morphology features
  - Buckwalter's Tagset (170 morphemes, 500 tokenized tags, 22K untokenized tags)
    - DET+ADJ+NSUFF_FEM_SG+CASE_DEF_NOM (الجميلة)
  - Bies' Reduced Tagset (24)
  - Kulick's Reduced Tageset (43)
  - Diab's Extended Reduced Tagset (72)
  - Habash's CATiB tagset (6)

Exampleويستمر

```
<morph_feature_set
   diac="وَيَسْتَمِرُّ" lemma="ٱِسْتَمَرّ_1"
   bw="wa/CONJ+ya/IV3MS+sotamir~/IV+u/IVSUFF_MOOD:I"
   gloss="continue;last_(time)"
   pos="verb"
   prc3="0" prc2="wa_conj" prc1="0" prc0="0”
   per="3" asp="i" vox="a" mod="i" gen="m"
   num="s” stt="na" cas="na" enc0="0" stem="سْتَمِرّ"/>
```

Example الغياب

```
<morph_feature_set
          diac="الغِيابُ" lemma="غِياب_1"
          bw=”Al/DET+giyAb/NOUN+u/CASE_DEF_NOM"
          gloss="absence;disappearance"
          pos="noun"
          prc3="0" prc2="0" prc1="0" prc0="Al_det" per="na"
          asp="na" vox="na" mod="na" gen="m" num="s”
          stt="d" cas="n" enc0="0" stem="غِياب"/>
```

Form / Function Discrepancy

Word	Gloss	Morphemes	Form-based Features	Functional Features
كِتاب	*book*	kitab+Ø	MS	MS
مَكْتَبَة	*library*	maktab+ap	FS	FS
كاتِبُون	*writers*	kAtib+uwn	MP	MP
عَين	*eye*	Eayn+Ø	MS	FS
خَلِيفَة	*caliph*	xaliyf+ap	FS	MS
رِجال	*men*	rijAl+Ø	MS	MP
سَحَرَة	*wizards*	saHar+ap	FS	MP
اِمْتِحانات	*exams*	AimtiHAn+At	FP	MP

M=Masculine F=Feminine S=Singular P=Plural

Morphological Ambiguity

- Morphological richness
  - Token Arabic/English = 80%
  - Type Arabic/English = 200%

- Morphological ambiguity
  - Each word: 12.3 analyses and 2.7 lemmas

- Derivational ambiguity
  - qAEdap: basis/principle/rule, military base, Qa'ida/Qaeda/Qaida

Morphological Ambiguity

- Inflectional ambiguity
  - *taktub*: you write, she writes
  - Segmentation ambiguity
    - wjd: *wajada* he found; *wa+jad~u*: and+grandfather

- Spelling ambiguity
  - Optional diacritics
    - kAtb: *kAtib* writer; *kAtab* to correspond
  - Suboptimal spelling
    - Hamza dropping: أ, إ → ا
    - Undotted ta-marbuta: ة → ه
    - Undotted final ya: ي → ى

Analysis vs. **Disambiguation**

Will Ben Affleck be a good Batman?

هل سينجح بين أفليك في دور باتمان؟

	PV+PVSUFF_SUBJ:3MS	bay~an+a	He demonstrated
	PV+PVSUFF_SUBJ:3FP	bay~an+~a	They demonstrated (f.p)
*	**NOUN_PROP**	**biyn**	**Ben**
	ADJ	bay~in	Clear
	PREP	bayn	Between, among

Morphological Analysis is out-of-context
Morphological Disambiguation is in-context

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:

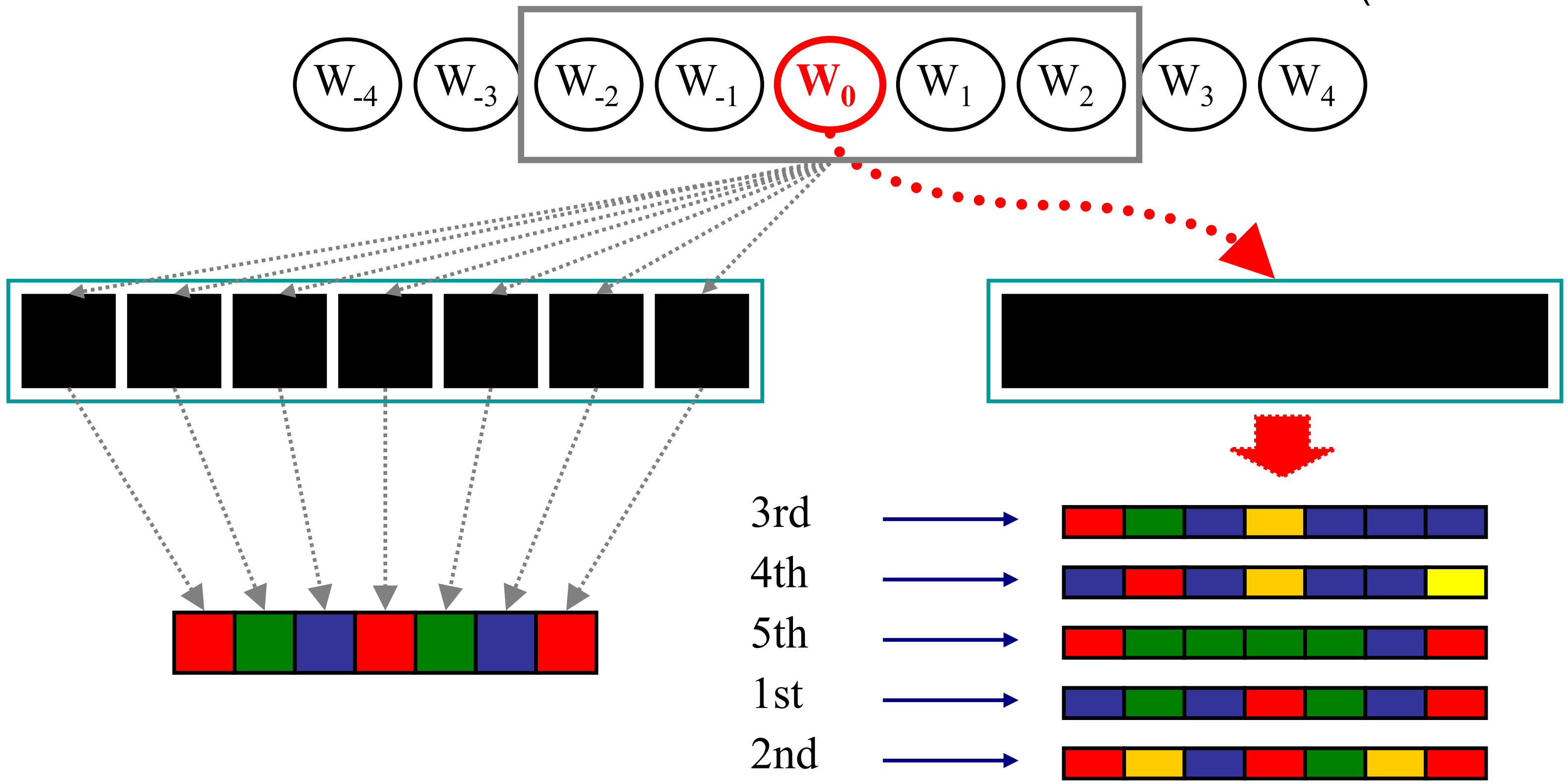
VB	VBD	VBG	VBN	VBP	VBZ
go	went	going	gone	go	goes

- 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;Roth et al. 2008)
MADAMIRA (Pasha et al., 2014)



MORPHOLOGICAL CLASSIFIERS

- Multiple independent classifiers
- Corpus-trained

RANKER

- Heuristic or corpus-trained

MORPHOLOGICAL ANALYZER

- Rule-based
- Human-created

MADA 3.2 (MSA) Evaluation

Accuracy	PATB 3 Blind Test		
	Baseline	MADA	Error ↓
All	74.8%	84.3%	38%
POS + Features	76.0%	85.4%	39%
All Diacritics	76.8%	86.4%	41%
Lemmas	90.4%	96.1%	60%
Partial Diacritics	90.6%	95.3%	50%
Base POS	91.1%	96.1%	56%
Segmentation	96.1%	99.1%	77%

Baseline: most common analysis per word in training

وكاتب wkAtb

and (the) writer of

wakAtibu

kAtib_1

pos:noun

prc3:0 prc2:wa_conj prc1:0 prc0:0 per:3 asp:na vox:na mod:na gen:m num:s stt:c cas:n enc0:0

w+ kAtb

Tokenization (TOKAN)

- Deterministic, generalized tokenizer
- **Input:** disambiguated morph. analysis + tokenization scheme
- **Output:** highly-customizable tokenized text

```
wsyktbhA = lex:katab-u_1 gloss:write pos:verb prc3:0
     prc2:wa_conj prc1:sa_fut prc0:0 enc0:3fs_dobj
```

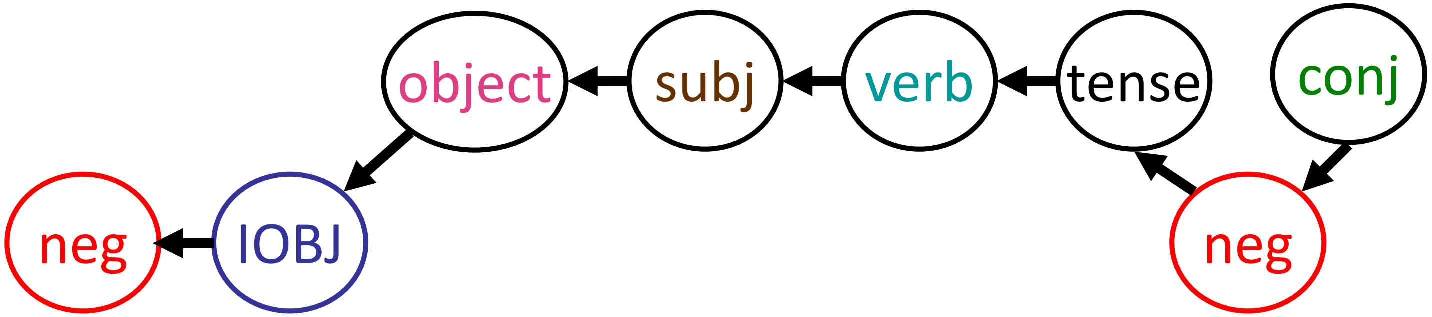
Example	**Scheme**	**Specification**
w+ syktbhA	D1	prc3 prc2 REST
w+ s+ yktbhA	D2	prc3 prc2 prc1 REST
w+ s+ yktb +hA	D3	prc3 prc2 prc1 prc0 REST enc0
w+ syktb +hA	ATB	prc3 prc2 prc1 prc0:lA prc0:mA REST enc0
w+•w+•wa+ syktbhA•syktbhA•katab	D1-3tier	prc3 prc2 REST ::FORM0 WORD ::FORM1 WORD NORM:AY ::FORM2 LEXEME

(Habash&Sadat 2006; Pasha et al., 2014)

Dialectal Arabic 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

DA 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

DA 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/

DA 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/

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 (اِ زْ دَ هَرَت)

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+	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	

…

وَسَنَكتُبُهَا

wasanaktubuhA

wiHaniktibhA

وِحَنِكتِبْهَا

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

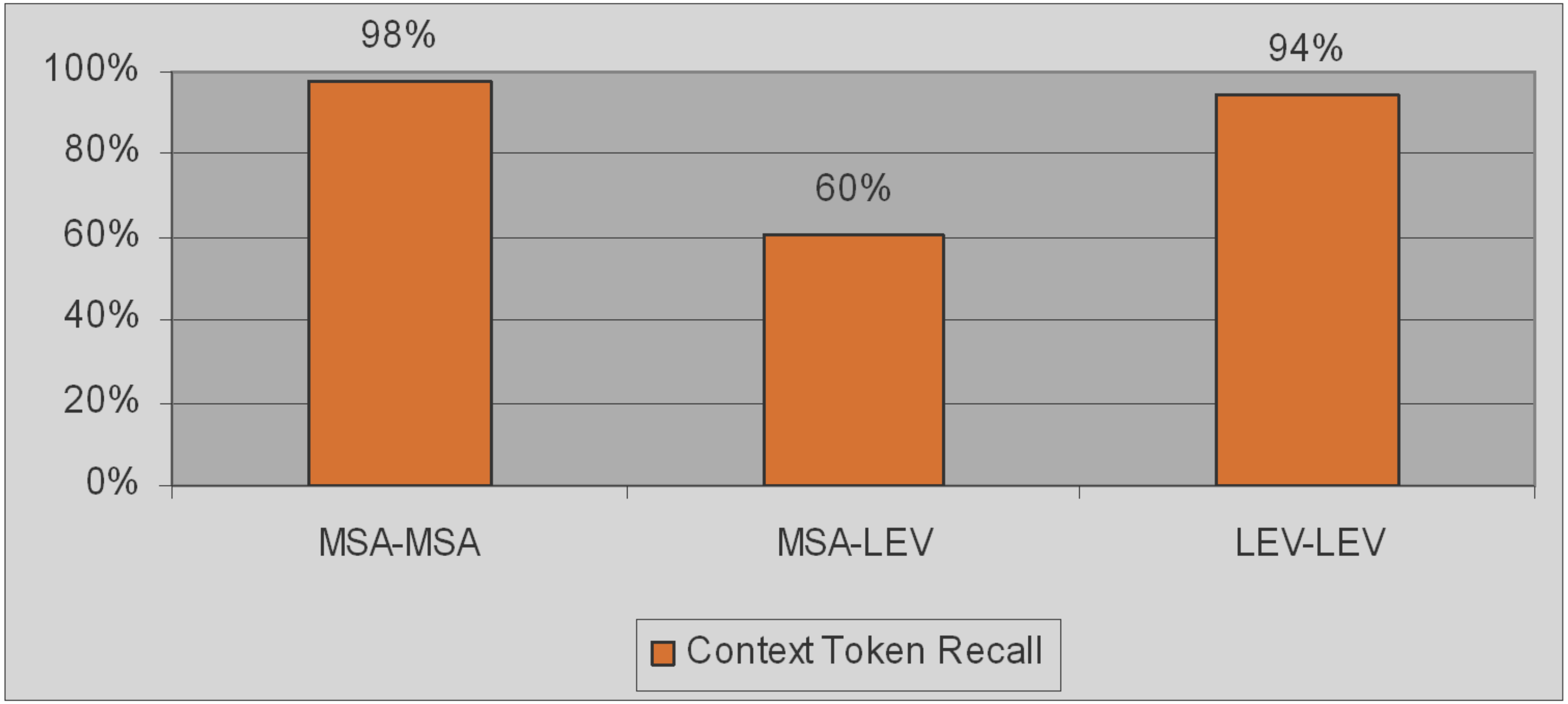
- MBC::Verb-I-au (*katab/yaktub*)

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



CALIMA-ARZ

- CALIMA is the Columbia Arabic Language Morphological Analyzer
- CALIMA-ARZ (ARZ = Egyptian Arabic)
  - Extends the Egyptian Colloquial Arabic Lexicon (ECAL) (Kilany et al., 2002) and Standard Arabic Morphological Analyzer (SAMA) (Graff et al., 2009).
  - Follows the part-of-speech (POS) guidelines used by the LDC for Egyptian Arabic (Maamouri et al., 2012b).
  - Accepts multiple orthographic variants and normalizes them to CODA (Habash et al., 2012).
  - Incorporates annotations by the LDC for Egyptian Arabic.

Building CALIMA-ARZ

- Starting with 66K inflected entries in ECAL
  - Example: (He doesn’t call him)
  - **Orthography** mbyklmw$ مبيكلموش
  - **Phonology** mabiykallimUš
  - **Morphology** kallim:verb+pres-3rd-masc-sg+DO-3rd-masc-sg+neg
- Convert entries to LDC guidelines fromat
  - **CODA** mA_biyikl~imhuw$ ما_بيكلمهوش
  - **Lemma** kal~im_1
  - **Morphemes** mA#bi+yi+kal~im+huw+$
  - **POS** NEG_PART#PROG_PART+IV3MS+IV+IVSUFF_DO:3MS+NEG_PART

Building CALIMA-ARZ

- Prefix/stem/suffix given class categories automatically
- Class categories are designed to
  - support extending paradigm coverage
    - Hab~**+ayt (Suff-PV-ay-SUBJ)→ +aynA, +ayty, +aytwA**
      +aynA+hA, +ayty+hA, +aytw+hA
      +aynA+hA+š, +ayty+hA+š, etc.
  - enforce morphotactic constraints

qalb**+ahA**	qalb**+ik**	**(Suff-NOM-stem-CC-POSS)**
kitAb**+hA**	kitAb**+ik**	**(Suff-NOM-stem-VC-POSS)**
hawA**+hA**	hawA**+kiy**	**(Suff-NOM-stem-V-POSS)**

Building CALIMA-ARZ

- Extending clitics and POS tags
  - Ea+ +ع (on), fi+ +ف (in), closed classes
- Non CODA support
  - The variant +w of the suffix +hu (his/him)
  - The variant ha+ of the prefix Ha+ (will)
  - Variants for specific frequent stems, e.g., the variants brDw and brdh of the stem brDh (also)

  Example: The word ***hyktbw*** هيكتبوreturns the analysis of the word ***Hyktbh*** حيكتبه (he will write it) among other analyses.
- With all the extensions, CALIMA-ARZ Egyptian core increases coverage from 66K to 48M words

CALIMA-ARZ Example

mktbtlhA$ مكتبتلهاش

Lemma	katab_1
CODA	mA_katab**t**_lahA$
POS	mA/NEG_PART+katab/PV+**t/PVSUFF_SUBJ:2MS**+ +li/PREP+hA/PRON_3FS+$/NEG_PART
Gloss	not + write + **you** + to/for + it/them/her + not

Lemma	katab_1
CODA	mA_katab**it**_lahA$
POS	mA/NEG_PART+katab/PV+**it/PVSUFF_SUBJ:3FS** +li/PREP+hA/PRON_3FS+$/NEG_PART
Gloss	not + write + **she/it/they** + to/for + it/them/her + not

CALIMA-ARZ v 0.5

- Incorporates LDC ARZ annotations (p1-p6)
  - 251K tokens, 52K types
  - Annotation clean up needed
    - Many rejected entries; ongoing clean up effort

System	Token Recall	Type Recall
SAMA-MSA v 3.1	67.7%	59.7%
CALIMA-ARZ v0.5 (Egyptian core)	88.7%	75.8%
CALIMA-ARZ v0.5 (++ SAMA dialect extensions)	92.6%	81.5%

MADA-ARZ

- Built on basic MADA framework with differences
- Uses CALIMA-ARZ as morphological analyzer
- Classifiers and language models trained using
  - LDC Egyptian Arabic annotated corpus (ARZ p1-p6)
  - LDC MSA PATB3 v3.1
- Non-Egyptian feature models dropped
  - case, mood, state, voice, question proclitic

MADA-ARZ Intrinsic Evaluation

System	MADA-MSA		MADA-ARZ	
Training Data	MSA	MSA	ARZ	MSA+ARZ
Test Set	MSA	Egyptian Arabic (ARZ)		
All	84.3%	27.0%	**75.4%**	64.7%
POS + Features	85.4%	35.7%	**84.5%**	75.5%
Full Diacriticization	86.4%	32.2%	**83.2%**	72.2%
Lemmatization	96.1%	67.1%	**86.3%**	82.8%
Base POS-tagging	96.1%	82.1%	91.1%	**91.4%**
ATB Segmentation	99.1%	90.5%	97.4%	**97.5%**

CALIMA-IRQ

Morphological Analysis for Iraqi Arabic

- What:
  - Morphological analyzer for Iraqi Arabic
  - Given a word, it returns all analyses/tokenizations out of context
  - Built by extending the LDC's Iraqi Arabic Morphological Lexicon (IAML) developed for Transtac
  - Currently has "approximate" stem-based lemmas

- Example : شدتقول $dtqwl

Lemma	qAl_1
Diac	$datquwl
POS	$/INTERROG_PART+ da/PROG_PART+t/IV2MS+quwl/IV
Gloss	what + [pres. tense] + you + say

- Evaluation

Analyzability (1.4M word Iraqi corpus)

System	Type	Token
SAMA-MSA-v3.1	78.0%	91.5%
CALIMA-IRQ v0.1	94.5%	99.5%

- Last Release: v 0.1

CALIMA-IRQ-TOK

Morphological Analysis and Tokenization for Iraqi Arabic

- What:
  - Tokenizer for Iraqi Arabic
  - Simple model of morpheme probabilities (no context)
  - Tokenization is deterministic given an analysis
  - Very fast tokenization required by the BOLT B/C performers
- Example

Input: بنفس المكان بالمستودع اللي هو مركز عملياتهم
bnfs AlmkAn bAlmstwdE Ally hw mrkz EmlyAthm
Output: ب# نفس ال# مكان ب# ال# مستودع اللي هو مركز عمليات +هم
b# nfs Al# mkAn b# Al# mstwdE Ally hw mrkz EmlyAt +hm

- Intrinsic Evaluation

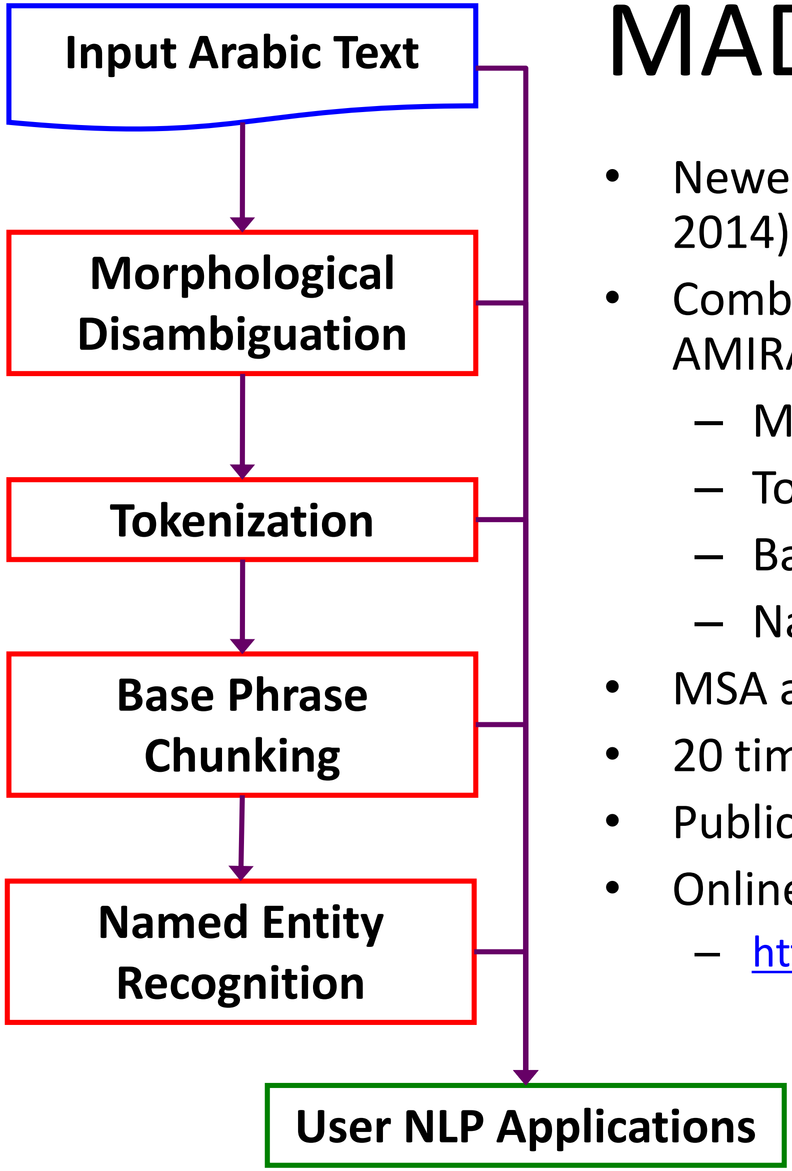
  On a 100 sentence (543 word) gold tokenized set
  - 98.7% have correct segmentation
  - 92.6% have correct tokenization

- Extrinsic Evaluation

  Transtac Data (Train 5M words)

Preprocessing	BLEU	METEOR	TER
None	27.4	30.7	53.4
CALIMA-TOK-IRQ	28.7	31.6	52.9

- Latest Release: v 0.1



MADAMIRA

- Newest tool from the CADIM group (Pasha et al., 2014)
- Combines MADA (Habash&Rambow, 2005) and AMIRA (Diab et al., 2004)
  - Morphological disambiguation
  - Tokenization
  - Base phrase chunking
  - Named entity recognition
- MSA and Egyptian Arabic modes
- 20 times faster than MADA, but same quality
- Publicly available (with some restrictions)
- Online demo
  - http://nlp.ldeo.columbia.edu/madamira/

Arabic Computational Morphology

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

Arabic Computational Morphology

- Representation units (continued)
  - Prefix + Stem + Suffix

    wll+mktb+At ولل+مكتب+ات

    – Can create more ambiguity
  - Lexeme + Features
    – [maktabap_1   +Plural +Def w+ l+]
  - Root + Pattern + Features
    – Very abstract
  - Root + Pattern + Vocalism + Features
    – Very very abstract

Arabic Computational Morphology

- Tools
  - Morphological Analyzers
    - Given a word ***out of context***, render all possible analyses
  - Morphological Segmenters (Tokenizers)
    - Given a word ***in context***, render best possible segmentation
  - Morphological Disambiguators (POS taggers)
    - Given a word ***in context***, render best possible analysis
- Considerations
  - Appropriateness of level of representation for an application
    - Tokenization Level, POS tag set for Machine Translation vs. Information Retrieval vs. Natural Language Generation
    - Arabic spelling vs. phonetic spelling
  - Coverage, extendibility, availability

Arabic Computational Morphology: Tools and Approaches

- Morphological Analyzers
  - MSA finite state machines [Beesely,2001]**,** [Kiraz,2001]
  - MSA Concatenative analysis/generation: **BAMA/SAMA** [Buckwalter 2000, Maamouri et al., 2009], **ALMOR** [Habash, 2004], **ELIXIRFM** [Smrz, 2007]
  - Dialectal Analyzers: **MAGEAD** [Habash&Rambow 2006], **ADAM** [Salloum & Habash, 2011], **CALIMA** [Habash et al., 2012]
- Tokenizers
  - Rule Based: Shallow stemming [Aljlayl and Frieder 2002], [Darwish,2002], [Larkey, 2003]
  - Machine learning (ML): [Lee et al,2003], [Rogati et al, 2003], **AMIRA** [Diab et al, 2004], **MADA+TOKAN** [Habash & Rambow 2005, Habash et al., 2009]
- Morphological Disambiguators/ POS Taggers
  - Supervised ML: **AMIRA** [Diab et al., 2004, 2007], **MADA** [Habash&Rambow, 2005], **MADAMIRA** [Pasha et al., 2014]
  - Semisupervised ML [Duh & Kirchhoff, 2005, 2006]
  - Unsupervised ML & Projections [Rambow et al., 2005]

Tutorial Contents

- Introduction
  - The many forms of Arabic
- Orthography
  - Script, phonology and spelling, dialectal variations, spelling inconsistency, automatic spelling correction and conventionalization, automatic transliteration
- Morphology
  - Derivation and inflection, ambiguity, dialectal variations, automatic analysis and disambiguation, tokenization
- Syntax
  - Arabic syntax basics, dialectal variations, treebanks, parsing Arabic and its dialects
- Lexical Variation and Code Switching
  - Dialectal variation, lexical resources, code switching, automatic dialect identification
- Machine Translation
  - Tokenization, out-of-vocabulary reduction, translation from and into Arabic, dialect translation

Morphology and Syntax

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

Morphology and Syntax

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

ل+المكتبات	li+Almaktabāt
for the-libraries	[PP li [NP Almaktabāt]]

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

MSA 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

MSA 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
    - wrote-you$_{MascSing}$ كتبـــتُ
    - wrote-you$_{MascPlur}$ كتبـــتم
    - wrote-they$_{MascPlur}$ كتبـــوا
  - Passive verbs
    - Same structure: Verb$_{passive}$ Subject$_{underlyingObject}$
    - Agreement with surface subject

MSA 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
    - o Passive and active forms are often similar in standard orthography
      - o كتب /kataba/ he wrote
      - o كُتب /kutiba/ it was written

MSA 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*

MSA 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

MSA 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
      - the-writer$_{FemSing}$ the-ambitious$_{FemSing}$
      - 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$_{FemSing}$

MSA 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

MSA 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'***	**>rbE** ***'four'***	**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**

Dialectal Arabic Variation
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$_{mascPl}$ the-poems
  - LEV, EGY
    - Subject Verb Object
      الاولاد كتبو الاشعار
      The-boys wrote$_{mascPl}$ the-poems
    - Less present: Verb Subject Object
      كتبو الاولاد الاشعار
      wrote$_{mascPl}$ 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]

Dialectal Arabic Variation
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

<table>
<tr><th></th><th>Idafa</th><th>No Idafa</th><th>Waqf</th></tr>
<tr><td><b>MSA</b></td><td colspan="2"><b>+at</b></td><td><b>+a</b></td></tr>
<tr><td><b>EGY</b></td><td><b>+it</b></td><td colspan="2"><b>+a</b></td></tr>
</table>

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

Dialectal Arabic Variation
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*)

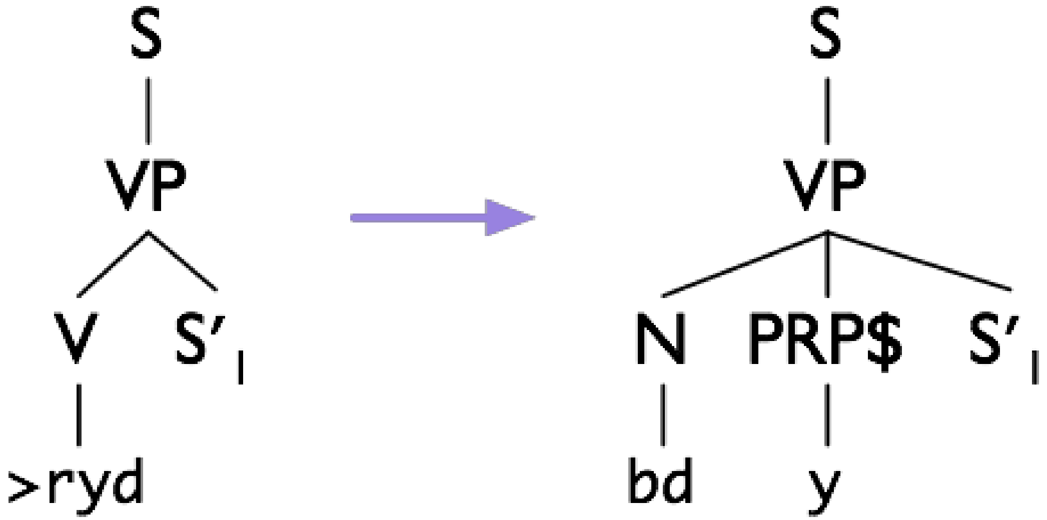
	Pre-nominal	Post-nominal
MSA	هذا الرجل	X
EGY	X	الراجل ده
LEV	هدا الرجال	الرجال هدا

Dialectal Arabic Variation
Negation Particles

	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 $

Dialectal Arabic
Lexico-syntactic Variation

- ‘want’ (Levantine)



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)
- Treebanks ...

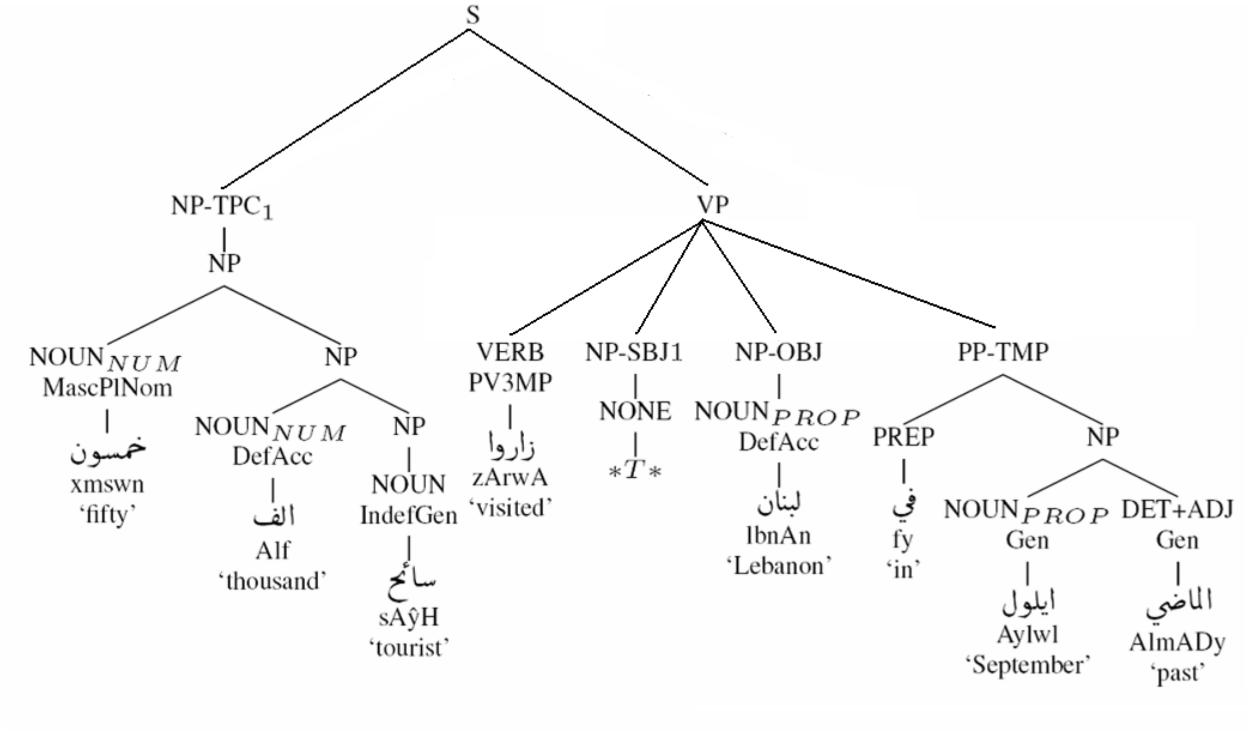
(Maamouri et al, 2004; Maamouri et al, 2006)

Penn Arabic Treebank

- Penn Arabic Treebank (PATB)
  - Started in 2001
  - Goal is 1 Million words
  - Currently 650K words (public)
    - 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)

(Maamouri et al, 2004; Maamouri et al, 2006)

Penn Arabic Treebank



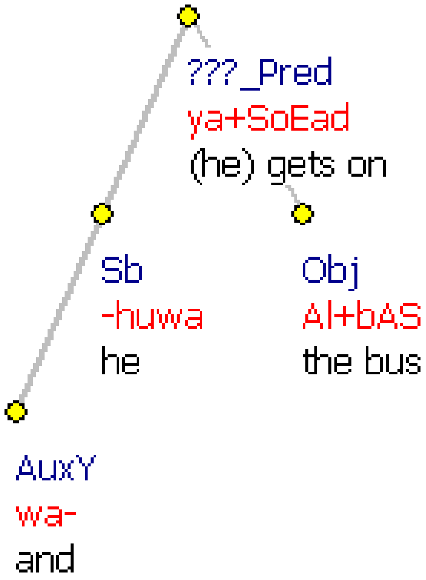
خمسون الف سائح زاروا لبنان في ايلول الماضي

Fifty thousand tourists visisted Lebanon in last September

(Smrž&Zemánek., 2002; Hajič et al., 2004; Smrž 2007)

Prague Arabic Dependency Treebank

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

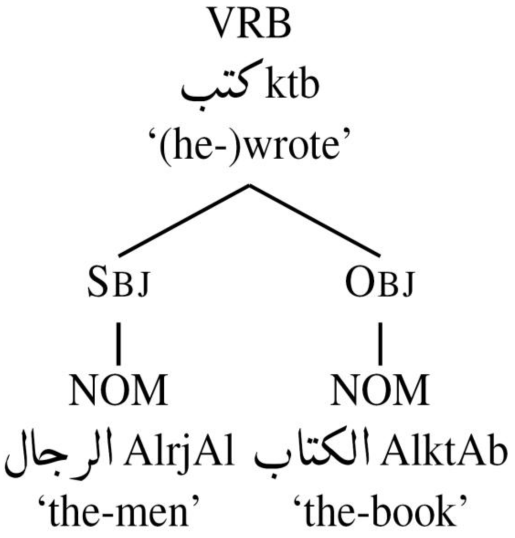


Graphic courtesy of Otakar Smrž: http://ckl.mff.cuni.cz/padt/PADT_1.0/docs/slides/2003-eacl-trees.ppt

Resource: Columbia Arabic Treebank

- Syntactic dependency
  - Six POS tags, eight relations
  - Inspired by traditional Arabic grammar
- Emphasis on annotation speed
  - Challenge: 200K words in 6 months
  - 540-700 w/h end-to-end
    - Penn Arabic Treebank (250-300) w/h
- Automatic enrichment of tags
  - Form 6 tags to full tagset (95.3% accuracy)
- CATiB in parsing shared task (2013)
  - Workshop for Parsing of Morphologically Rich Languages

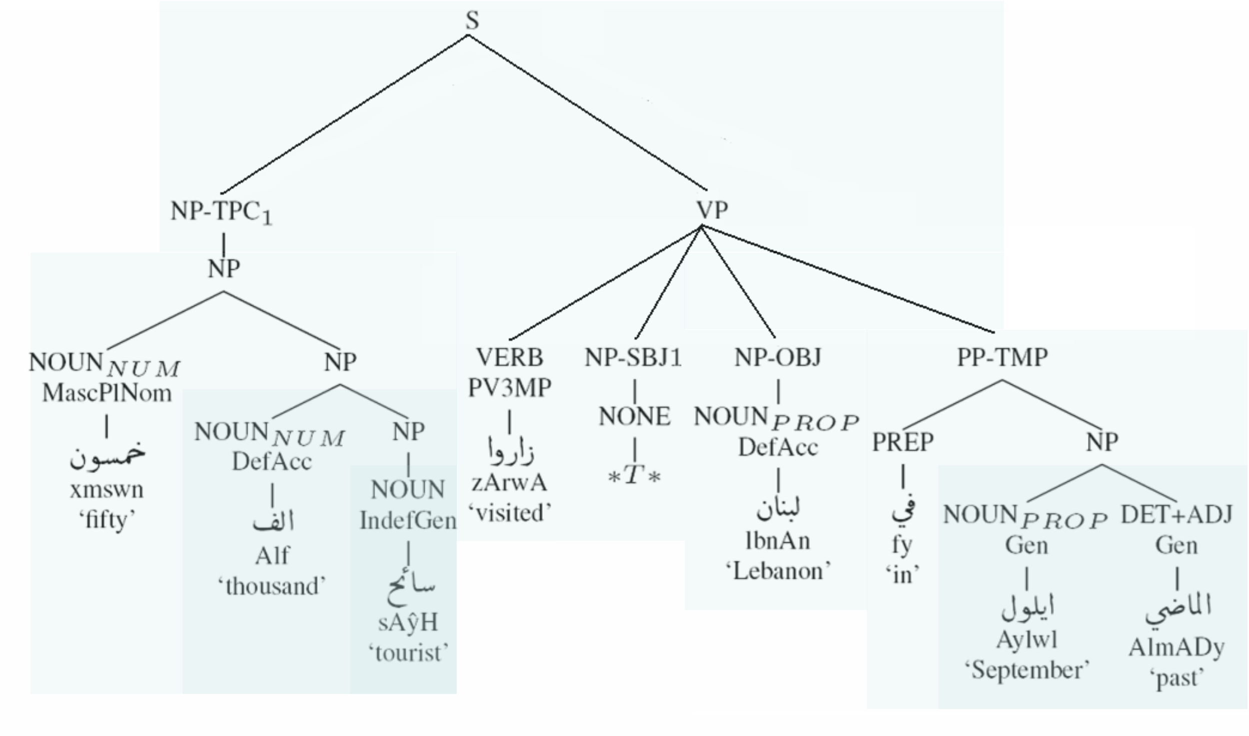
CATiB

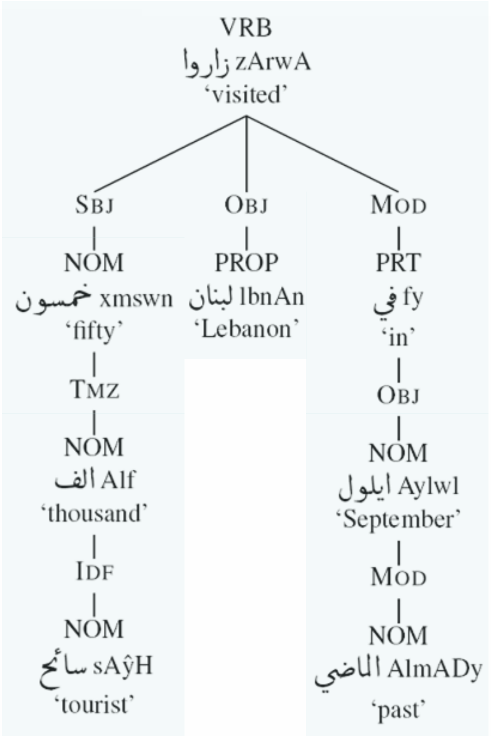


(Habash & Roth, 2009; Alkuhlani & Habash, 2013)

Constituency vs. Dependency

PATB vs. CATiB





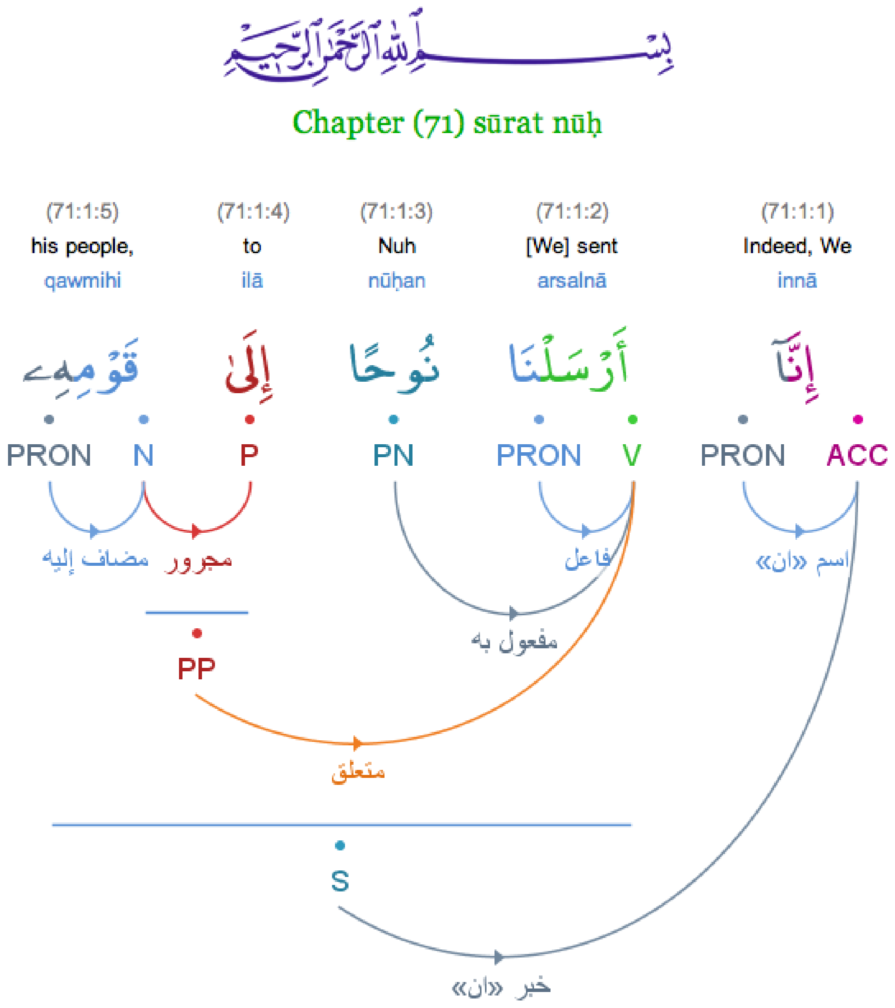
خمسون الف سائح زاروا لبنان في ايلول الماضي

Fifty thousand tourists visisted Lebanon in last September

(Dukes&Habash, 2010; Dukes& Buckwalter, 2010; Dukes et al., 2010)

The Quranic Arabic Corpus

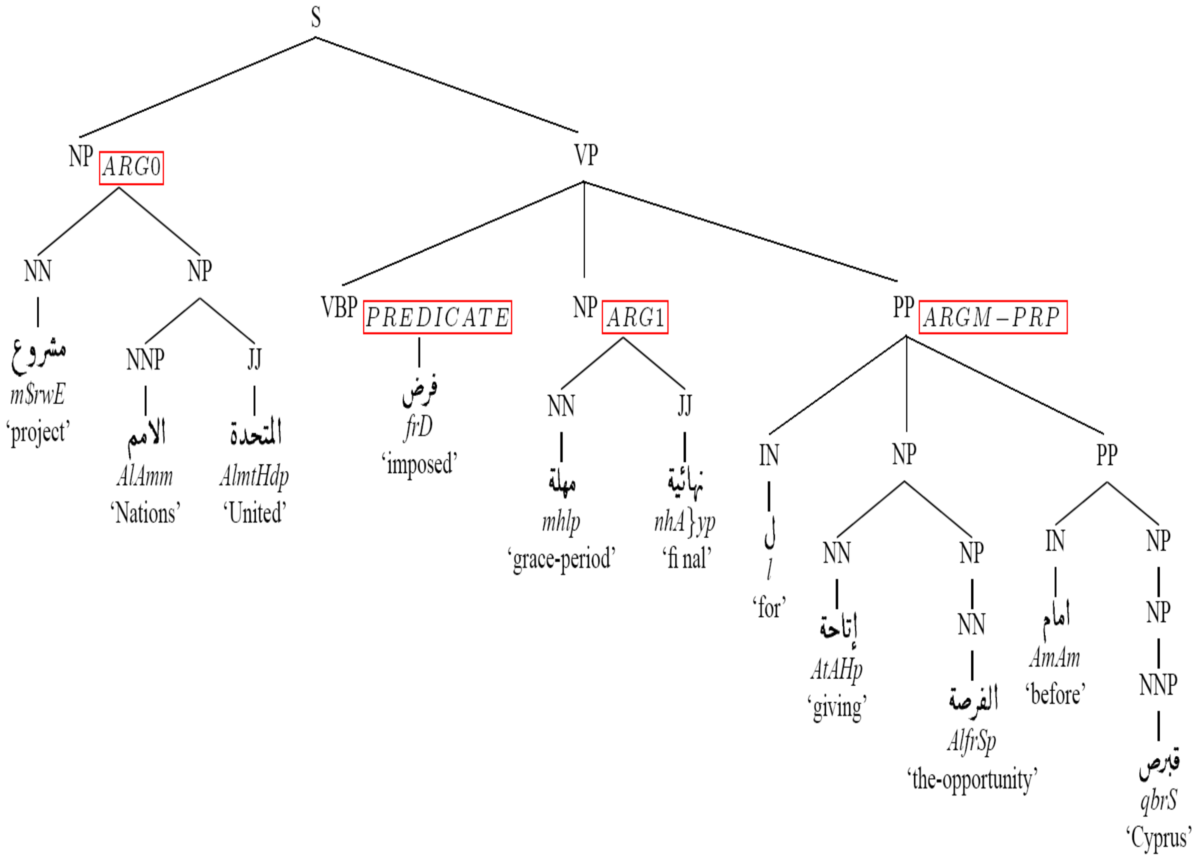
- Annotation of the Holy Quran
  - Morphology, Syntax, Semantic Ontology
- http://corpus.quran.com/



Arabic PropBank

(Palmer et al., 2008)
(Diab et al., 2008)

- Effort to annotate predicate-argument structure on the Penn Arabic Treebank
  - University of Colorado, LDC, Columbia University



Computational Resources

- Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL)
- 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)
    - Dukes' one step hybrid parser (Dukes and Habash, 2011)
  - Base-phrase Chunking
    - (Diab et al, 2004; Diab et al. 2007)
- 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

Marton et al. (2013)

Morphological Features for Arabic Parsing

- Parsing with Rich morphology
  - 😊 – Rich morphology helps morpho-syntactic modeling
    - E.g., agreement and case assignment
  - 😒 – But: Rich morphology increases data sparseness
    - A challenge to statistical parsers
  - 😒 – But: Rich POS tagset can be hard to predict
    - E.g. Arabic case (or state) is usually not explicitly written
  - 😒 – Also: Mapping from form to function is not 1:1
    - E.g. so-called broken plurals, or fem. ending to masc. noun
- Marton et al. (2013) explored the contribution of various Arabic (MSA) morphological features and tagsets to syntactic dependency parsing

Marton et al. (2013)

Morphological Features for Arabic Parsing

- Marton et al. (2013) explored a large space of features
  - Different POS tagsets at different degrees of granularity
  - Different inflectional and lexical morphological features
  - Different combinations of features
  - Gold vs. predicted POS and morphological feature values
  - Form-based vs. functional feature values (gender, number, and rationality)
- CATiB: The Columbia Arabic Treebank
- MALTParser (Nivre et al. 2006)

Marton et al. (2013)

Morphological Features for Arabic Parsing

- POS tagset performance as function of information
  - Approximated by tagset size
  - More informative ➔ better parsing quality (on gold)

Tagset	Size	Gold	Example: Al+xams+ap+u `the-five.fem.sing.nom'
CATIB6	6	81.04	NOM
CATIBEX	44	82.52	Al+NOM+ap
CORE12	12	82.92	ADJ (stripped of any inflectional info)
CORE44	40	82.71	ADJ_NUM
ERTS	134	82.97	DET+ADJ_NUM+FEM_SG
KULICK	32	83.60	DET+ADJ_NUM
BW	430	84.02	DET+ADJ_NUM+FEM_SG+DEF_NOM

Marton et al. (2013)

Morphological Features for Arabic Parsing

- POS tagset performance as function of information
  - Approximated by tagset size
  - More informative ➔ better parsing quality (on gold)
- Gold vs. Predicted POS
  - Lower POS tagset accuracy ➔ worse parsing quality (non-gold)

Tagset	Size	Gold	Predicted	Diff.	Acc.
CATIB6	6	81.04	78.31	-2.73	97.7
CATIBEX	44	82.52	**79.74**	-2.78	97.7
CORE12	12	82.92	78.68	-4.24	96.3
CORE44	40	82.71	78.39	-4.32	96.1
ERTS	134	82.97	78.93	-4.04	95.5
KULICK	32	83.60	79.39	-4.21	95.7
BW	430	**84.02**	72.64	-11.38	81.8

Marton et al. (2013)

CASE and STATE help in gold

PERSON, NUMBER, GENDER and DET help in non-gold

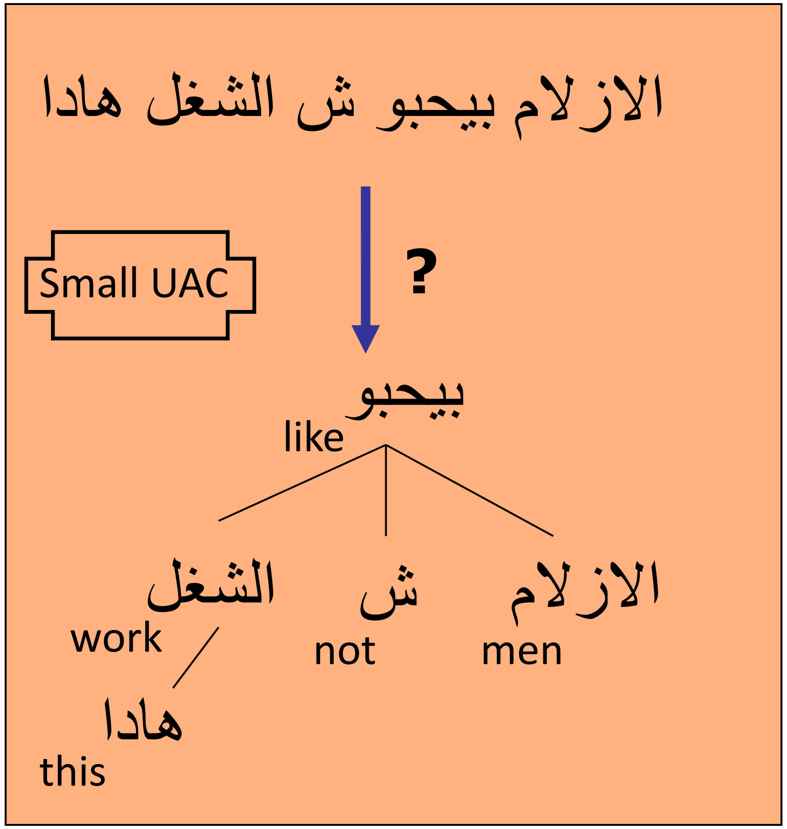
GOLD	**LAS**	***diff***	**PREDICTED**	**LAS**	***diff***
Baseline	82.92		Baseline	78.68	
ALL	85.15	2.23	ALL	77.91	-0.77
CASE	**84.61**	**1.69**	**DET**	**79.82**	**1.14**
STATE	84.15	1.23	STATE	79.34	0.66
DET	83.96	1.04	GEN	78.75	0.07
NUM	83.08	0.16	PER	78.74	0.06
PER	83.07	0.15	NUM	78.66	-0.02
VOICE	83.05	0.13	VOICE	78.64	-0.04
MOOD	83.05	0.13	ASP	78.60	-0.08
ASP	83.01	0.09	MOOD	78.54	-0.14
GEN	82.96	0.04	CASE	75.81	-2.87
CASE+STATE	**85.37**	**0.76**	DET+STATE	79.42	-0.40
CASE+STATE+DET	85.18	-0.19	DET+GEN	79.9	0.08
CASE+STATE+NUM	85.36	-0.01	DET+GEN+PER	79.94	0.04
CASE+STATE+PER	85.27	-0.10	**DET+P.N.G**	**80.11**	**0.17**
CASE+STATE+VOICE	85.25	-0.12	DET+P.N.G+VOICE	79.96	-0.15
CASE+STATE+MOOD	85.23	-0.14	DET+P.N.G+ASPECT	80.01	-0.10
CASE+STATE+ASP	85.23	-0.14	DET+P.N.G+MOOD	80.03	-0.08
CASE+STATE+GEN	85.26	-0.11			

Arabic Dialect Parsing

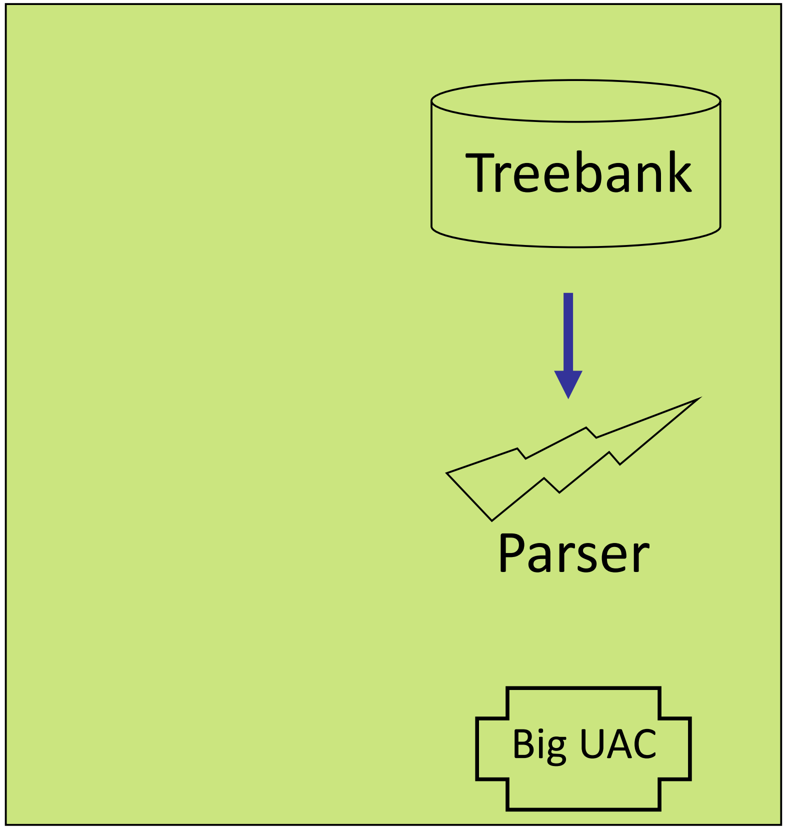
- Possible Approaches
  - Annotate corpora (“Brill Approach”)
    - Too expensive
  - Leverage existing MSA resources
    - Difference MSA/dialect not enormous
    - Linguistic studies of dialects exist
    - Too many dialects: even with dialects annotated, still need leveraging for other dialects

Parsing Arabic Dialects:
The Problem

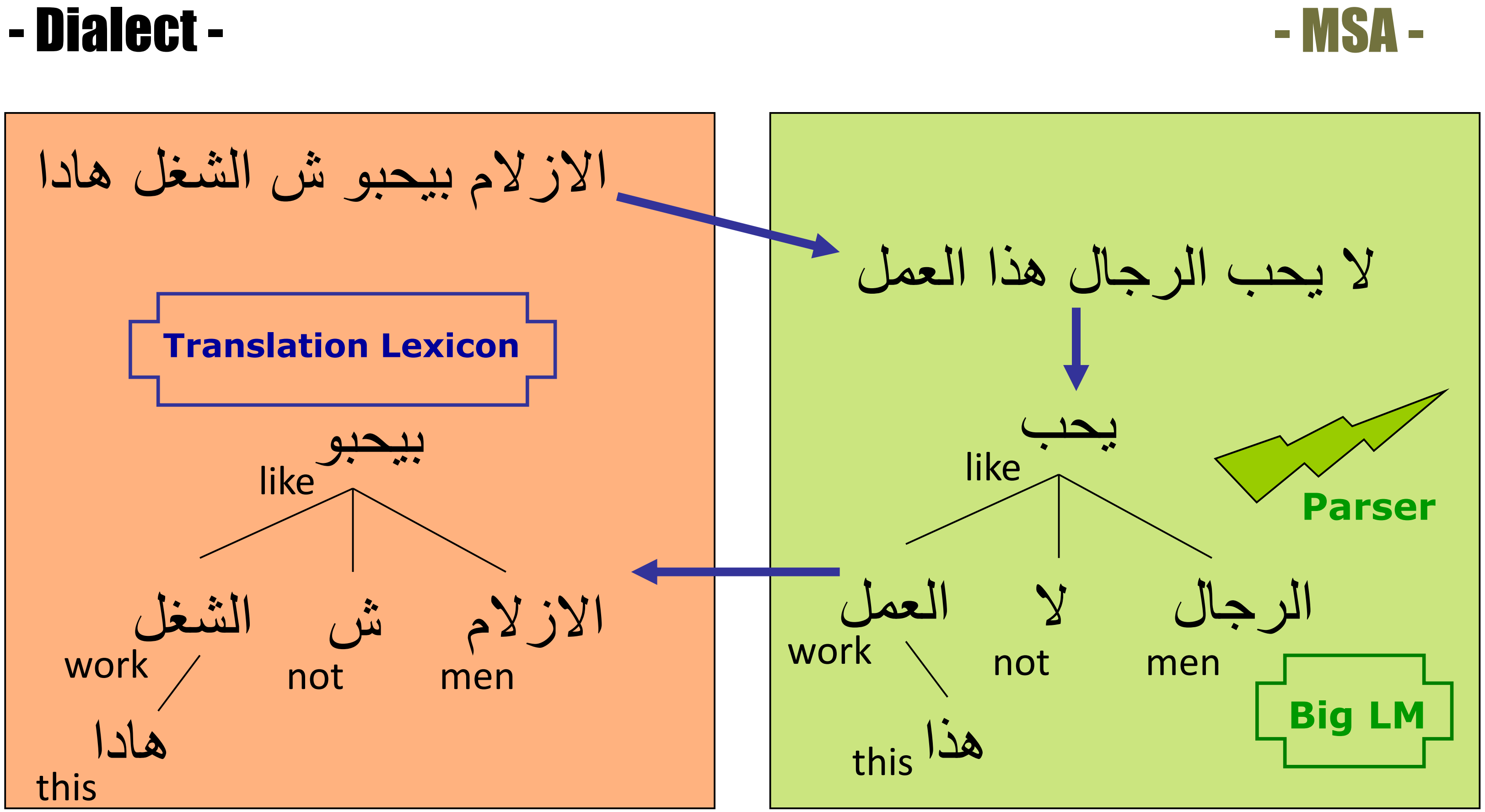
- Dialect -



- MSA -

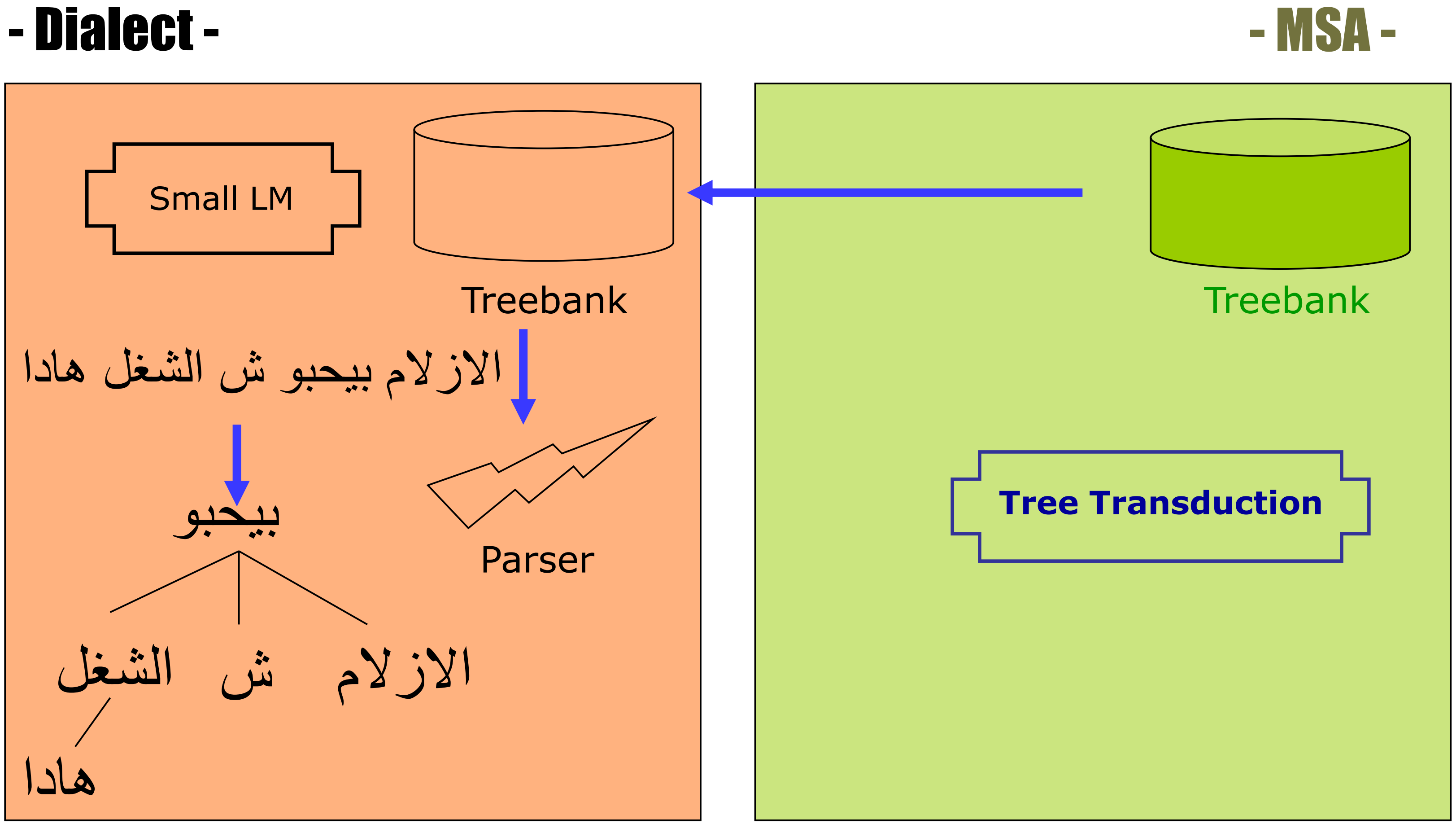


Sentence Transduction Approach



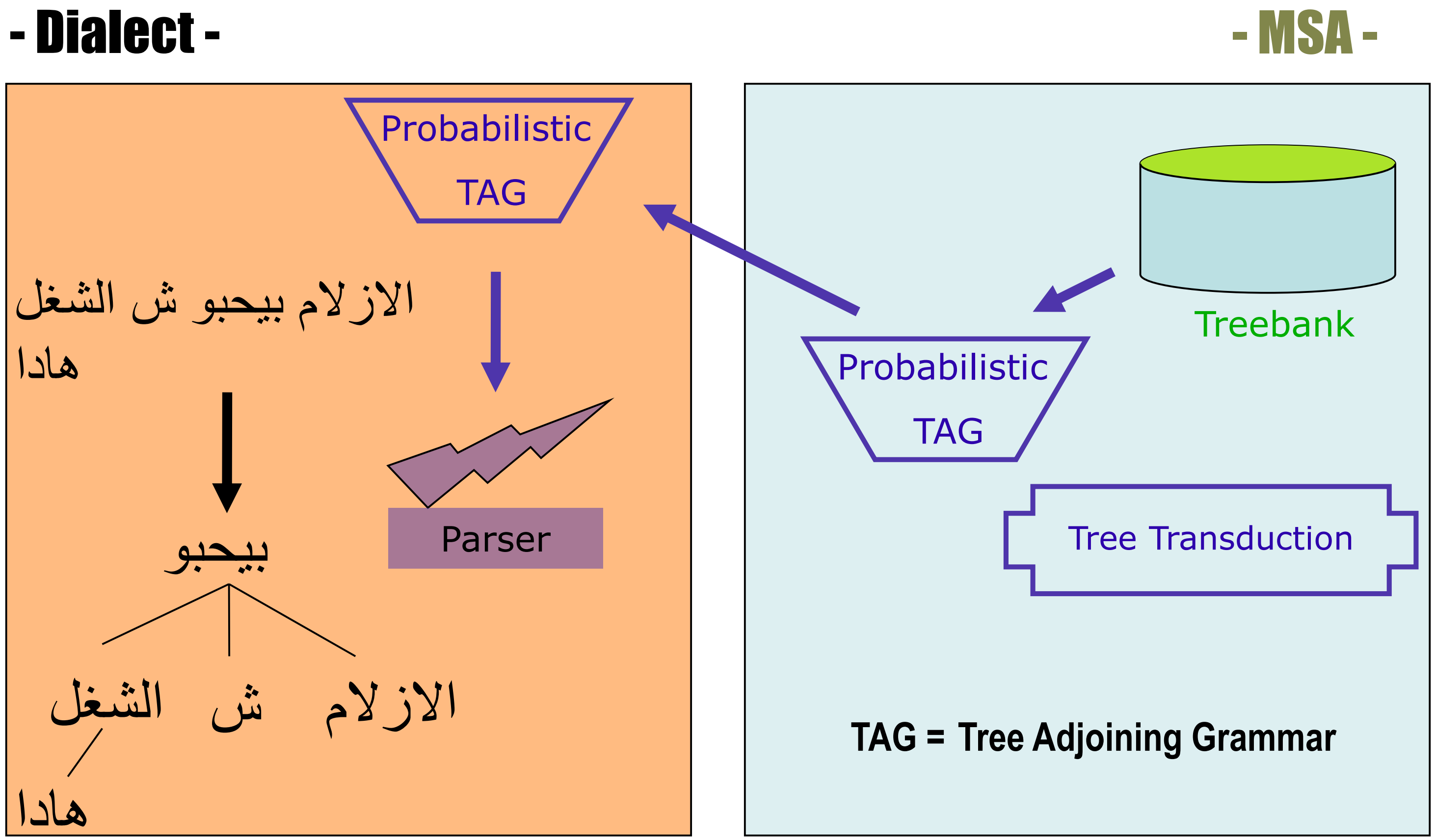
(Rambow et al. 2005; Chiang et al. 2006)

MSA Treebank Transduction



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

Grammar Transduction



- Dialect -

- MSA -

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)

Tutorial Contents

- Introduction
  - The many forms of Arabic
- Orthography
  - Script, phonology and spelling, dialectal variations, spelling inconsistency, automatic spelling correction and conventionalization, automatic transliteration
- Morphology
  - Derivation and inflection, ambiguity, dialectal variations, automatic analysis and disambiguation, tokenization
- Syntax
  - Arabic syntax basics, dialectal variations, treebanks, parsing Arabic and its dialects
- Lexical Variation and Code Switching
  - Dialectal variation, lexical resources, code switching, automatic dialect identification
- Machine Translation
  - Tokenization, out-of-vocabulary reduction, translation from and into Arabic, dialect translation

Arabic 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 fi ما في
Iraqi	mēz ميز	bazzūna بزونة	māl مال	‘arīd اريد	aku اكو	māku ما

- Arabic orthography allows consolidating some variations

Arabic 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

Dialect-MSA Dictionary

- Problem: lack of Dialect-MSA resources
  - No Dialect-MSA parallel text
  - No paper dictionaries for Dialect-MSA
- Dictionary is required for many NLP applications exploiting MSA resources
  - MT and CLIR
  - Parsing with the lack of DA parsers, one would need to translate dialect sentences to MSA before parsing them with an MSA parser
  - Dialect Identification especially with the problem of linguistic code switching and pervasive presence of faux amis (homographs with different meanings in DA and MSA)

Levantine-MSA Dictionary

[Maamouri et al. 2006]

- **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)

THARWA

A Multi-dialectal Dictionary

- What:
  - A three way dictionary for Egyptian Arabic (DA), MSA and English equivalents
  - Predominantly lemma entries
  - All DA entries are in CODA
  - POS tag information provided
  - All Arabic entries are diacritized
  - DA and MSA lemmas are aligned with SAMA and CALIMA databases
  - Manually created and semi automatically consistency checked
- **Dictionary Size:**
  - 65,237 complete unique records

- Example:

Egyptian	MSA	POS	English
شَيِّل $ay~il	حَمَّل Ham~al	verb	carry; blame; impose; charge
ذنّب *an~ib	عاقب EAqab	verb	Punish
أباجورة >abAjawrap	مصباح miSobAH	noun	lamp
أفيونجي >afiyuwnojiy	مدمن mudomin	adj	Opium addict
ظاهرة ZAhirap	ظاهرة ZAhirap	noun	phenomenon

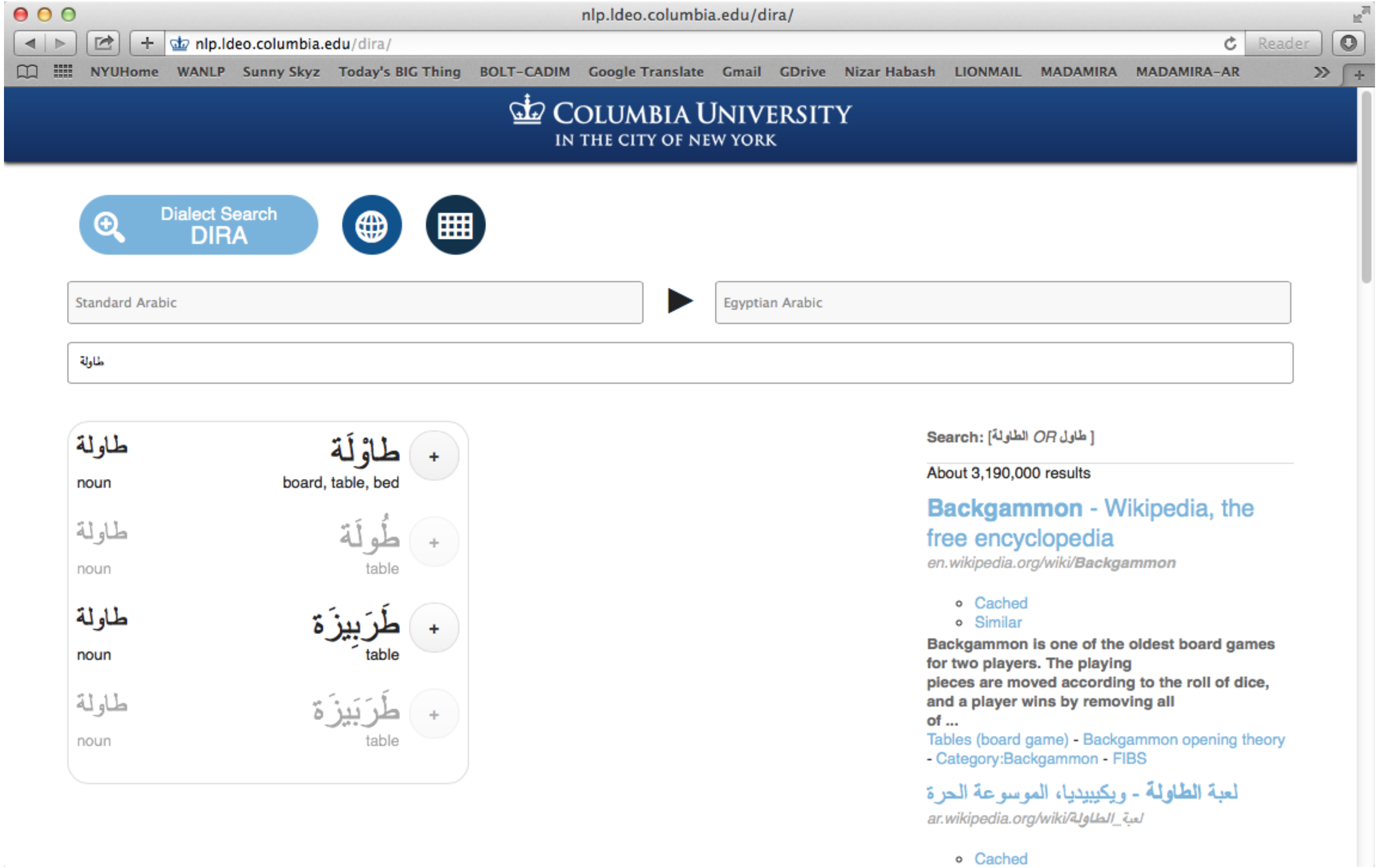
- Used in: DIRA, AIDA, ELISSA
- (Diab et al., 2014 LREC)

DIRA: Dialectal (Arabic) Information Retrieval Assistant

[Diab et al., 2010]

- DIRA is a query expansion application
- Accepts MSA short queries as input and expands them to a dialect(s) of choice
- Multiple MSA expansion modes
  - Expand input MSA with MSA morphology
    - ASbH `he became' >> tSbH, nSbH, ySbHwn, etc.
  - Expand input MSA with DA morphology
    - ASbH `he became' >> **H**tSbH, **H**nSbH, **H**ySbHwA, etc.
  - Translate MSA lemma to DA lemma and expand using DA morphology
    - ASbH `he became' >> tbqY, nbqY, HtbqY, HnbqY, etc.
- Online demo: http://nlp.ldeo.columbia.edu/dira/

DIRA Demo



Standard Arabic ▶ Egyptian Arabic

طاولة

طاولة noun	طاوْلَة board, table, bed	+
طاولة noun	طُولَة table	+
طاولة noun	طَرَبِيزَة table	+
طاولة noun	طَرَبَيزَة table	+

Search: [الطاولة OR طاول]

About 3,190,000 results

Backgammon - Wikipedia, the free encyclopedia

*en.wikipedia.org/wiki/**Backgammon***

- Cached
- Similar

Backgammon is one of the oldest board games for two players. The playing pieces are moved according to the roll of dice, and a player wins by removing all of ...

Tables (board game) - Backgammon opening theory - Category:Backgammon - FIBS

لعبة الطاولة - ويكيبيديا، الموسوعة الحرة

ar.wikipedia.org/wiki/لعبة_الطاولة

- Cached

Lexical Reality of Arabic Data

Data Source	Example
Newswire MSA only	واكد لليوم الثانى ان "الجهود مستمره الى الامام" من اجل مواصلته الحوار الوطنى بخصوص عملية السلام. *And he emphasized for the second day that "efforts are continuing forward" to resume the national dialogue on the peace process.*
Broadcast MSA+some DA	علشان كده هي بتتفاعل مع ما يحدث وتجد إلزاما عليها أن تنبه الشعب العربي إلى حقيقة ما يدور بالمفاوضات *'cause o' this it's interactin' with what is happening and it finds it necessary to awaken the Arab people to the truth of what is happening in the negotiations*
CTS, news groups & blogs more DA	بالعاكس عادي بس لأني متأكد إني بعرفكيش عشان هيك بحكي لك إنتي مخربطة *no problem, but since I am sure I don't know you, that's why I am telling you you're confused.*

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

MSA-LIKE LEV

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

- 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’

Dialect Identification & Classification

- Speech Data
  - State of the art system – 18.6% WER within dialect and 35.1% across dialects (Biadsy et al.,2012)
- Textual Data
  - Sentence Level Dialect ID
    - Zaidan and Callison-Burch (2013)
    - AIDA (Elfardy & Diab, 2012)
  - Token Level Dialect ID and Classification
    - AIDA (Elfardy & Diab, 2012)

Word Level Annotation
[Habash et al., 2008]

- **Word Level 0** ***pure MSA*** **words**
  - *MSA lexemes / MSA morphology / MSA orthography*
  - يكتبون *yaktubuwn* ‘they write’, اعيادكم *AçyAdukum* ‘your holidays’
- **Word Level 1** ***MSA with non-standard orthography***
  - *MSA lexemes / MSA morphology / non-standard orthography*
  - Dialectal spelling: فسطان *fusTAn* (vs. فستان *fustAn* ‘dress’)
  - Spelling error: مساجذ *masAjið* (vs. مساجد *masAjid* ‘mosques’)
- **Word Level 2** ***MSA word with dialect morphology***
  - *MSA lexemes / dialect morphology*
  - بيكتب *byiktib* (Egyptian ‘he writes’)
    - Present tense prefix +ب *b+* (LEV/EGY), +د *da+* (IRQ), +ك *ka+* (MOR)
- **Word Level 3** ***Dialect lexeme***
  - *Dialect lexeme*: never written or spoken when producing MSA
  - The negation marker مش *miš* ‘no/not’
  - عافية *çAfyaħ* (Moroccan for ‘**fire**/health’ but MSA for ‘health’)

Elfardy & Diab (2012, 2013)

AIDA System

- Objectives
  - contextual token and sentence level DA identification and classification with confidence scores
  - As a side effect, AIDA produces linearized gisted MSA and English equivalent text
- Approach
  - Statistical approach combining large scale DA-MSA-ENG dictionaries: Egyptian, Levantine, Iraqi (~63K entries) with language models based on MSA (AGW) and DA corpora (Egy ~6M Tokens/~650K Types, Lev ~7M Tokens/~500K Types)
- Evaluation data
  - Manually annotated 15K Egyptian and 15K Levantine words [Elfardy & Diab, 2012]
  - Manually annotated 20K words for dialect ID [Habash et al., 2008]
- Performance
  - Token Level identification/classification F=81.2 Egyptian, F=75.3 Levantine
- Online demo: http://nlp.ldeo.columbia.edu/aida/

AIDA Example

MSA EGY

هنا رقد الراجل على فراشه يغالب الغيبوبة وكلما افاق يلاقي
مراته جنبه فقلها: لما شركتي فلست كنتي جنبي، ولما بيتنا
إتحرق ، شكلك كده نحس عليا.

Transliteration

hnA rqd AlrAjl ElY frA$h ygAlb Algybwbp wklmA AfAq ylAqy mrAth jnbh fqlhA: lmA $rkty flst knty jnby, wlmA bytnA AtHrq, $klk kdh nHs ElyA.

Tutorial Contents

- Introduction
  - The many forms of Arabic
- Orthography
  - Script, phonology and spelling, dialectal variations, spelling inconsistency, automatic spelling correction and conventionalization, automatic transliteration
- Morphology
  - Derivation and inflection, ambiguity, dialectal variations, automatic analysis and disambiguation, tokenization
- Syntax
  - Arabic syntax basics, dialectal variations, treebanks, parsing Arabic and its dialects
- Lexical Variation and Code Switching
  - Dialectal variation, lexical resources, code switching, automatic dialect identification
- Machine Translation
  - Tokenization, out-of-vocabulary reduction, translation from and into Arabic, dialect translation

Tokenization for Machine Translation

- Tokenization and normalization have been shown repeatedly to help Statistical MT(Habash & Sadat, 2006; Zollmann et al., 2006; Badr et al., 2008; El Kholy & Habash, 2010; Al-Haj & Lavie, 2010; Singh & Habash, 2012; Habash et al., 2013)
- Habash & Sadat 2006
  - Arabic to English Statistical MT
  - *Bleu Metric (Papineni et al. 2002)*

Scheme	40K wd Train	4M wd Train
ST	11.16	37.83
ON	12.59	37.93
WA	15.03	37.79
D1	14.86	37.30
TB	15.94	37.81
D2	16.32	**38.56**
D3	17.72	36.02
EN	**18.25**	36.02

(Habash&Sadat, 2006)

Preprocessing Schemes

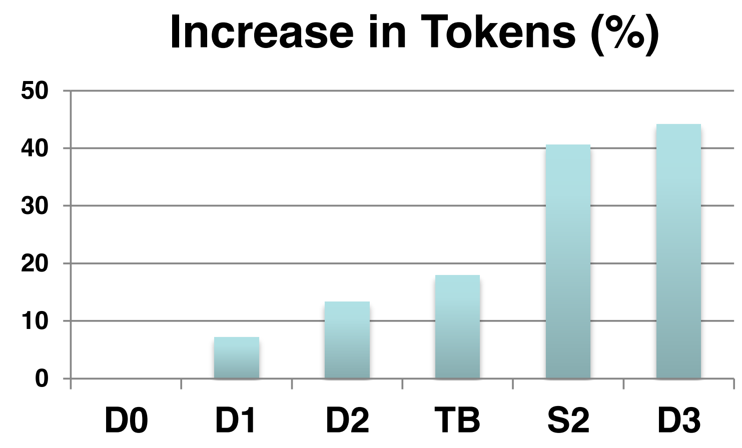
ST	Simple Tokenization
D1	Decliticize CONJ+
D2	Decliticize CONJ+, PART+
D3	Decliticize all clitics
BW	Morphological stem and affixes
EN	D3, Lemmatize, English-like POS tags, Subj

Input:		wsyktbhA?	‘and he will write it?’
	ST	wsyktbhA ?	
	D1	w+ syktbhA ?	
	D2	w+ s+ yktbhA ?	
	D3	w+ s+ yktb +hA ?	
	BW	w+ s+ y+ ktb +hA ?	
	EN	w+ s+ ktb/VBZ S:3MS +hA ?	

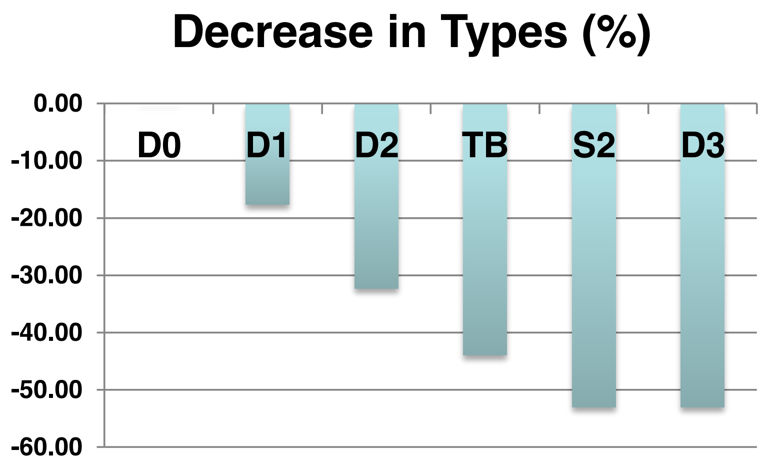
Preprocessing Schemes

• ST	Simple Tokenization
• D1	Decliticize CONJ+
• D2	Decliticize CONJ+, PART+
• D3	Decliticize all clitics
• BW	Morphological stem and affixes
• EN	D3, Lemmatize, English-like POS tags, Subj
• ON	Orthographic Normalization
• WA	wa+ decliticization
• TB	Arabic Treebank
• L1	Lemmatize, Arabic POS tags
• L2	Lemmatize, English-like POS tags

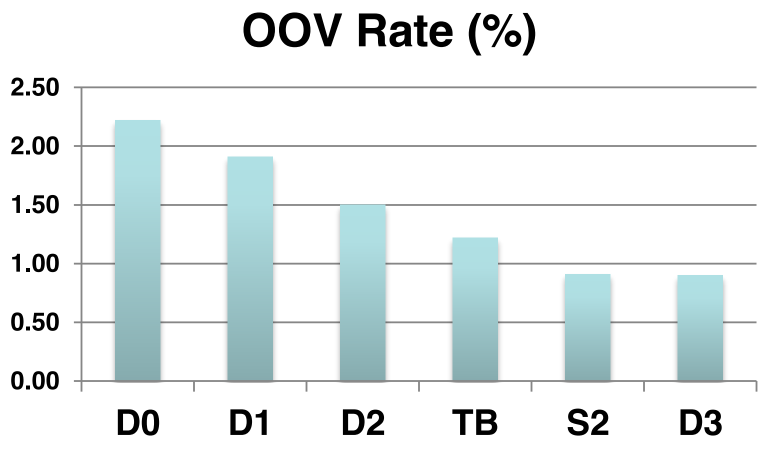
Preprocessing Schemes



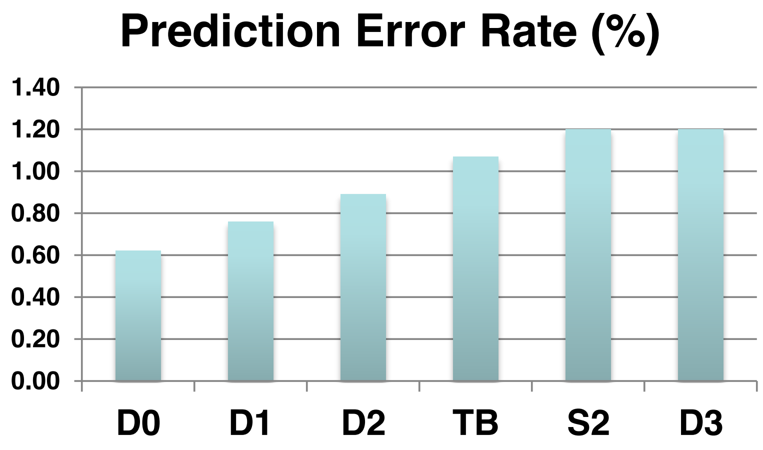
Increase in Tokens (%)



Decrease in Types (%)



OOV Rate (%)



Prediction Error Rate (%)

Tokenization for Machine Translation

- Tokenization and normalization have been shown repeatedly to help Statistical MT(Habash & Sadat, 2006; Zollmann et al., 2006; Badr et al., 2008; El Kholy & Habash, 2010; Al-Haj & Lavie, 2010; Singh & Habash, 2012; Habash et al., 2013)
- Habash & Sadat 2006
  - Arabic to English Statistical MT
  - Different data sizes require different tokenization schemes
  - As size increases, tokenization help decreases
  - In NIST Open MT Evaluation, 9 out of 12 participants in Arabic-English track used MADA

Scheme	40K wd Train	4M wd Train
ST	11.16	37.83
ON	12.59	37.93
WA	15.03	37.79
D1	14.86	37.30
TB	15.94	37.81
D2	16.32	**38.56**
D3	17.72	36.02
EN	**18.25**	36.02

Arabic-to-English VS English-to-Arabic

- Arabic-to-English SMT
  - Tokenization and normalization help

    (Lee, 2004; Habash & Sadat, 2006; Zollmann et al., 2006)
- English-to-Arabic SMT
  - What tokenization scheme?

    (Badr et al., 2008; Al Kholy & Habash, 2010; Al-Haj & Lavie, 2010)
  - Output Detokenization and Denormalization (Enriched/True Form)
    - Anything less is comparable to all lower-cased English or uncliticized and undiacritized French

Normalization	**Example**	**% Words diff. from RAW/ENR**
Reduced (RED)	Âqwý /أقوى/ →Aqwy /اقوي/	**12.1% / 16.2%**
Enriched (ENR) / **TrueForm**	Aqwy /اقوي/ → Âqwý /أقوى/	**7.4 % / 0.0%**

Tokenization for Machine Translation

- Tokenization and normalization have been shown repeatedly to help Statistical MT(Habash & Sadat, 2006; Zollmann et al., 2006; Badr et al., 2008; El Kholy & Habash, 2010; Al-Haj & Lavie, 2010; Singh & Habash, 2012; Habash et al., 2013)
- El Kholy & Habash 2010
  - English to Arabic Statistical MT
  - Funded by a Google award

	Baseline **no tokenization**	**MADA-MSA** **ATB Tokenization**
4 M words	26.00	**27.25**
60 M words	31.30	**32.24**

(Habash, 2008)

REMOOV

- Out-Of-Vocabulary (OOV)
  - Test words that are not modeled in training
  - May be in training data but not in phrase table
  - May be in phrase table but not matchable
- A persistent problem
  - Arabic in ATB tokenization with orthographic normalization:

  Increasing the training data by 12 times

  → 66% reduction in Token/Type OOV

  → 55% reduction in Sentence OOV (sentences with at least 1 OOV word)

	Medium			**Large**		
Word count	**4.1M**			**47M**		
	MT03	**MT 04**	**MT 05**	**MT03**	**MT 04**	**MT 05**
Token OOV	**2.5%**	**3.2%**	**3.0%**	**0.8%**	**1.1%**	**1.1%**
Type OOV	**8.4%**	**13.32%**	**11.4%**	**2.7%**	**4.6%**	**4.0%**
Sentence OOV	**40.1%**	**54.47%**	**48.3%**	**16.9%**	**25.6%**	**22.8%**

Profile of OOVs in Arabic

- Proper nouns (40%)
  - Different origins: Arabic, Hebrew, English, French, Italian, and Chinese
- Other parts-of-speech (60%)
  - Nouns (26.4%), Verbs (19.3%) and Adjectives (14.3%)
  - Less common morphological forms such as the dual form of a noun or a verb
- Orthogonally, spelling errors appear in (6%) of cases and tokenization errors appear in (7%) of cases

Proper Noun	**40%**	**روثبين، جفعاتايم، هوكايدو**
Noun/Adjective	**41%**	**قريتين، مدرستا**
Verb	**19%**	**سيلتقيان، تر، مررنا**
Spelling Error	**13%**	**اشحاض، باكتسان، لروثبين**

OOV Reduction Techniques

- Two strategies for online handling of OOVs by phrase table extension
  - Recycle Phrases
    - Expand the phrase table online with recycled phrases
      - Relate OOV word to INV (in-vocabulary) word
      - Copy INV phrases and replace INV word with OOV word
      - Example: add misspelled variant of a word in phrase table
        - » كناب *knAb* → book
      - Using unigram and bigram phrases was optimal for BLEU
  - Novel Phrases
    - Expand the phrase table online with new phrases
      - Example: باستور *bAstwr* is OOV
      - Use transliteration software to produce possible translations
        - » Pasteur, Pastor, Pastory, Bostrom, etc.

REMOOV Techniques

- MorphEx (morphological expansion)
- DictEx (dictionary expansion)
- SpellEx (spelling expansion)
- TransEx (name transliteration)

	Morphology	**No Morphology**
Recycled Phrases	*MorphEx*	*SpellEx*
Novel Phrases	*Dictex*	*TransEx*

***REMOOV** Toolkit is available for research*
Contact nizar.habash@nyu.edu

Morphology Expansion

- Model target-irrelevant source morphological variations
  - Cluster Arabic translations of English words
    - book ← (كتاب, الكتاب, كتابا)
    - write ← (يكتب تكتب نكتب يكتبون يكتبن سيكتبن ...)
  - Learn mappings of morphological features for words sharing lexemes in the same cluster
    - [POS:V +S:3MS] == [POS:V +S:3FS]
    - [POS:N Al+ +PL] == [POS:N +PL]
    - [POS:N +DU] == [POS:N +PL]
- Map OOV word to INV word using a morphology rule:
- الجماعتين → [POS:N Al+ +DU] == [POS:N +PL] → جماعات

Spelling Expansion

- Relate an OOV word to an INV word through:
  - Letter deletion فلسطني → فلسطيني
  - Letter Insertion فليسطيني → فلسطيني
  - Letter inversion فلسيطني → فلسطيني
  - Letter substitution قلسطيني → فلسطيني
  - Substitution in Arabic was limited to 90 cases (as opposed to 1260)
    - Shape alternations ر <> ز
    - Phonological alternations س <> ص
    - Dialectal variations أ <> ق
- *No modification of the probabilities in the recycled phrases*

Transliteration Expansion

- Use a similarity metric (Freeman et al 2006) to match Arabic spelling to English spelling of proper names
  - – Expand forms by mapping to Double Metaphones (Philips, 2000)
- Assign very low probabilities that are adjusted to reflect similarity metric score

المتنبي	→	MTNP	→	Al-Mutannabi Al-Mutanabi
باستور	→	PSTR	→	Pasteur Pastor Pastory Pasturk Bistrot Bostrom
شوارزنغر شوارزنيجر شوارتزنجر	→	XFRTSNKR	→	Schwarzenegger
قذافي	→	KTF	→	Qadhafi Gadafi Gaddafi Kadafi Ghaddafi Qaddafi Katif Qatif

Dictionary Expansion

- OOV word is analyzable by BAMA (Buckwalter 2004)
- Add phrase table entries for OOV translating to all inflected forms of the BAMA English gloss
- Assign equal very low probabilities to all entries

الموسيقيون	→	موسيقي	→	musical	→	musical musicals
			→	musician	→	musician musicians
المخطئة	→	مخطئ	→	mistaken	→	mistaken
			→	at fault	→	at fault at faults
جلستم	→	جلس	→	sit	→	sit sits sat sitting

REMOOV Evaluation

- Medium Set
  - 4.1 M words
  - Average token OOV is 2.9%
- All techniques improve on baseline
  - TransEx < MorphEx < DictEx < SpellEx
- Combinations improve on combined techniques
  - Least improving combination (on average): MorphEx+DictEx
  - Most improving combination (on average): DictEx+TransEx
- Combining all improves most

BLEU Scores

	MT03	MT04	MT05
BASELINE	44.20	40.60	42.86
TRANSEX	44.83	40.90	43.25
MORPHEX	44.79	41.18	43.37
DICTEX	44.88	41.24	43.46
SPELLEX	45.09	41.11	43.47
MORPHEX+DICTEX	45.00	41.38	43.54
SPELLEX+dMORPHEX	45.28	41.40	43.64
SPELLEX+TRANSEX	45.43	41.24	43.75
DICTEX+TRANSEX	45.30	41.43	43.72
ALL	**45.60**	**41.56**	**43.95**
Absolute improvement	*1.4*	*0.96*	*1.09*
Relative improvement	*3.17*	*2.36*	*2.54*

REMOOV Evaluation

- Learning Curve Evaluation
  - – Different techniques do better under different size conditions
  - – Even with 10 times data, OOV handling techniques still help
- Error Analysis
  - – Hardest cases are Names
  - – 60% of time, OOV handling is acceptable

MT04 BLEU Scores

	1%	10%	100%	1000%
Baseline	13.40	31.07	40.60	42.06
TransEX	13.80	31.78	40.90	42.10
SpellEX	14.02	31.85	41.11	42.25
MorphEX	15.06	32.29	41.18	42.16
DictEx	**20.09**	**33.56**	41.24	42.14
ALL	18.17	33.41	**41.56**	**42.29**
Best Absolute	6.69	2.49	0.96	0.23
Best Relative	49.93	8.01	2.36	0.55

	PN	NOM	V	
Good	26 (40%)	41 (73%)	17 (85%)	*60%*
Bad	39 (60%)	15 (27%)	3 (15%)	*40%*
	46%	*40%*	*14%*	*100%*

OOV Handling Examples

- Foreign name
  - Before: … and president of ecuador lwt$yw gwtyryz .
  - After: … and president of ecuador lucio gutierrez .
- Dual noun
  - Before: … headed the mission to qrytyn in the north .
  - After: … headed the mission to villages in the north .
- Dual verb
  - Before: … baghdad and riyadh , which qTEtA their diplomatic relations …
  - After: … baghdad and riyadh , which sever their diplomatic relations …
- Spelling error
  - Before: … but mHAdtAt between palestinian factions …
  - After: … but talks between palestinian factions …

Arabic Dialect Machine Translation

- **BOLT: Broad Operational Language Translation**
  - Egyptian Arabic → English MT
  - Iraqi <-> English speech-to-speech MT
- **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)

Challenges to processing Arabic dialects: Machine Translation

Arabic Variant	Arabic Source Text	Google Translate
MSA	يوجد كهرباء، ماذا حدث؟ لا	Does not have electricity, what happened?
EGY	الكهربا اتقطعت، ليه كده بس؟	Atqtat electrical wires, Why are Posted?
LEV	شكلو مفيش كهربا، ليش هيك؟	Cklo Mafeesh كهربا, Lech heck?
IRQ	شو ماكو كهرباء، خير؟	Xu MACON electricity, good?

Arabic Dialect Machine Translation

- Problems
  - Limited resources
    - Small Dialect-English corpora & no Dialect-MSA corpora
  - Non-standard orthography
  - Morphological complexity
- Solutions
  - Rule-based segmentation (Riesa et al. 2006)
  - Minimally supervised segmentation (Riesa and Yarowsky 2006)
  - Dialect-MSA lexicons (Chiang et al. 2006, Maamouri et al. 2006)
  - Pivoting on MSA (Sawaf 2010, Salloum and Habash, 2011)
    - Elissa 1.0 (Salloum & Habash, 2012)
  - Crowdsourcing Dialect-English corpora (Zbib et al., 2012)

MSA-pivoting for DA to English MT

[Salloum & Habash, 2011, 2012, 2013]

- Challenge: There is almost no MSA-DA parallel corpora to train a DA-to-MSA SMT
- Solution: use a rule-based approach to
  - produce MSA paraphrases of DA words
  - create a lattice for each sentence
  - pass the lattice to an MSA-English SMT system
- The rule-based approach needs:
  - A dialectal morphological analyzer
  - Rules to transfer from DA analyses to MSA analyses
- Elissa 1.0

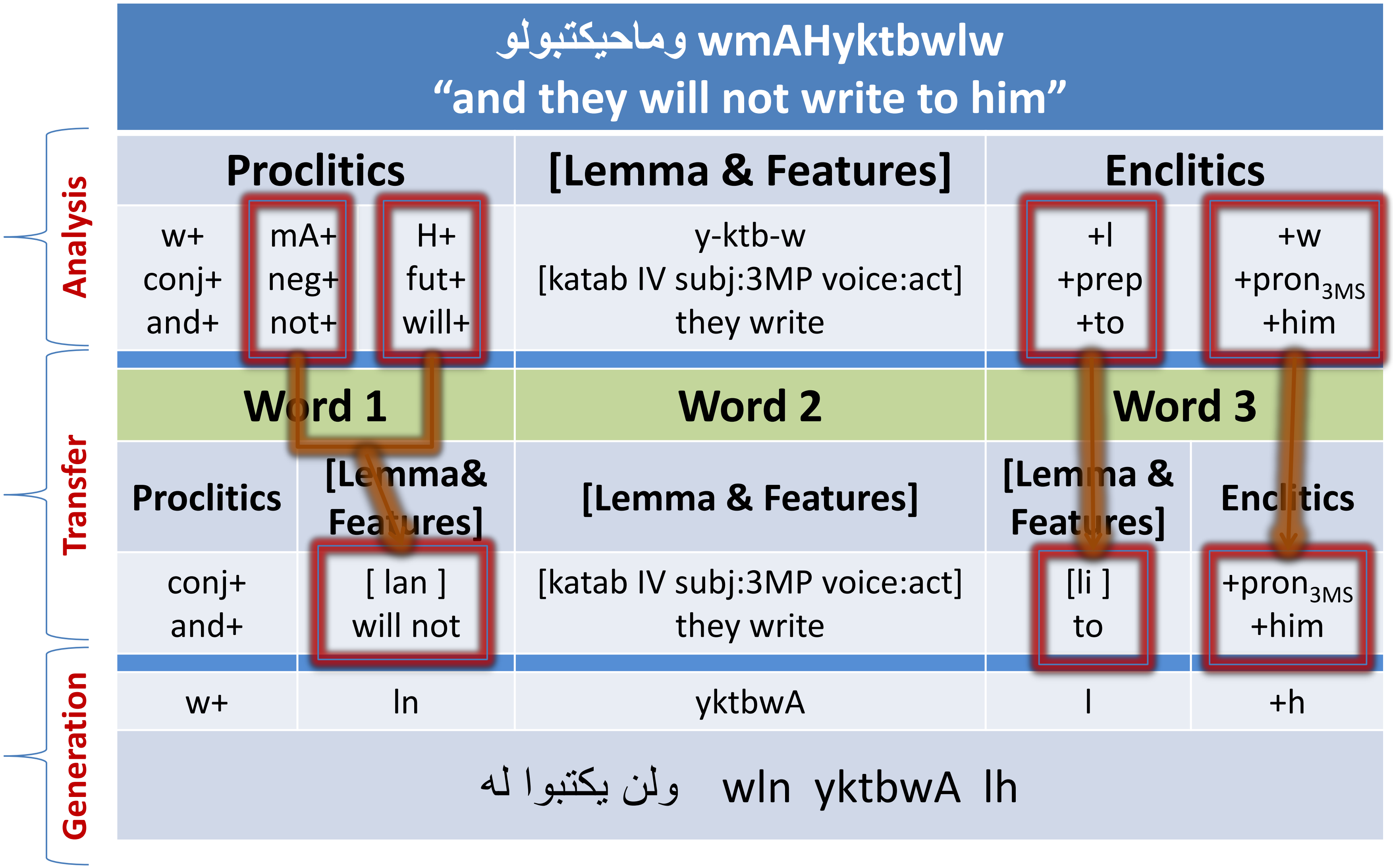
[Salloum & Habash, 2011, 2012, 2013]

Elissa 1.0

- Dialectal Arabic to MSA MT System
- Output
  - MSA top-1 choice, n-best list or map file
- Components
  - Dialectal morphological analyzer (ADAM) (Salloum and Habash, 2011)
  - Hand-written morphological transfer rules & dictionaries
  - MSA language model
- Evaluation (DA-English MT)
  - MADA preprocessing (ATB scheme)
  - Moses trained for MSA-English MT
  - 64 M words training data
  - Best system only processes MT OOVs and ADAM dialect-only words
  - Top-1 choice of MSA
  - Results in BLEU

System	Dev. Set	Blind Test
Baseline	37.20	38.18
Elissa + Baseline	37.86	38.80

Example



وماحيكتبولو wmAHyktbwlw

“and they will not write to him”

Analysis

Proclitics			**[Lemma & Features]**	**Enclitics**	
w+ conj+ and+	mA+ neg+ not+	H+ fut+ will+	y-ktb-w [katab IV subj:3MP voice:act] they write	+l +prep +to	+w +pron$_{3MS}$ +him

Transfer

Word 1		**Word 2**	**Word 3**	
Proclitics	**[Lemma& Features]**	**[Lemma & Features]**	**[Lemma & Features]**	**Enclitics**
conj+ and+	[lan] will not	[katab IV subj:3MP voice:act] they write	[li] to	+pron$_{3MS}$ +him

Generation

w+	ln	yktbwA	l	+h

ولن يكتبوا له   wln yktbwA lh

Elissa 1.0: DA to MSA translation

Direct Translation of Dialectal Arabic (DA)	
Dialectal Arabic	بهالحالة ماحيكتبولو شي عحيط صفحتو لأنو ماخبرهن يوم اللي وصل عالبلد
DA-English Human Transaltion	In this case, they will not write on his page wall because he did not tell them the day he arrived to the country.
Arabic-English Google Translate	Bhalhalh Mahiketbolo Shi Ahat Cefhto to Anu Mabrhen day who arrived Aalbuld.

Pivoting on Modern Standard Arabic (MSA) using Elissa	
DA-MSA Elissa Translation	في هذه الحالة لن يكتبوا شي علي حائط صفحته لانه لم يخبرهم يوم الذي وصل الي البلد
Arabic-English Google Translate	In this case it would not write something on the wall yet because he did not tell them the day arrived in the country.

General References

- ACL Anthology (search for Arabic)
  - http://www.aclweb.org/anthology/
- Machine Translation Archive (search for Arabic)
  - http://www.mt-archive.info
- Zitouni, I. ed., Natural Language Processing of Semitic Languages. Springer. 2014.
- Soudi, A., S. Vogel, G. Neumann and A. Farghaly, eds. Challenges for Arabic Machine Translation. John Benjamins. 2012.
- Habash, N. and H. Hassan, eds. Machine Translation for Arabic. Special Issue of MT Journal. 2012.
- Habash, N. Introduction to Arabic Natural Language Processing. Synthesis Lectures on Human Language Technologies. Morgan & Claypool. 2010.
- Farghaly, A. ed. Arabic Computational Linguistics. CSLI Publications. 2010
- Soudi, A., A. van den Bosch, and G. Neumann, eds. Arabic Computational Morphology. Springer, 2007.
- Holes, C. Modern Arabic: Structures, Functions, and Varieties. Georgetown University Press. 2004.
- Bateson, M. Arabic Language Handbook. Georgetown University Press. 2003.
- Brustad, K. The Syntax of Spoken Arabic: A Comparative Study of Moroccan, Egyptian, Syrian, and Kuwaiti Dialects. Georgetown University Press. 2000.

Thank you!

Natural Language Processing of Arabic and its Dialects

Mona Diab

The George Washington University
mtdiab@gwu.edu

Nizar Habash

New York University
Abu Dhabi
nizar.habash@nyu.edu