

AthNLP 2026
September 5, 2026

Introduction to Arabic NLP

مقدمة في المعالجة الآلية للغة العربية

Prof. Nizar Habash
New York University Abu Dhabi
United Arab Emirates

أ. د. نزار حبش
جامعة نيويورك أبوظبي
الإمارات العربية المتحدة

New York University Abu Dhabi

جامعة نيويورك أبوظبي



New York University Abu Dhabi

جامعة نيويورك أبوظبي

- A research university with a fully integrated liberal arts and science college
- Students from all over the world
 - 2,200 students
 - 115 nationalities

• جامعة بحثية مع كلية علوم وفنون
ليبرالية متكاملة

• طلاب من جميع أنحاء العالم
• ٢,٢٠٠ طالب
• ١١٥ جنسية



أبوظبي

Abu Dhabi

***Father of
Gazelles***



Abu Dhabi

أبو ظبي





MOHAMED BIN ZAYED
UNIVERSITY OF
ARTIFICIAL INTELLIGENCE





EMNLP 2022
Dec 7-11, Abu Dhabi
United Arab Emirates





Who I am

- Professor Nizar Habash د. نزار حبش
- Computer Science @ NYUAD
- Research Scientist at Columbia University
- Ph.D. in Computer Science
University of Maryland College Park
- B.S. Computer Engineering
B.A. Linguistics and Languages
Old Dominion University



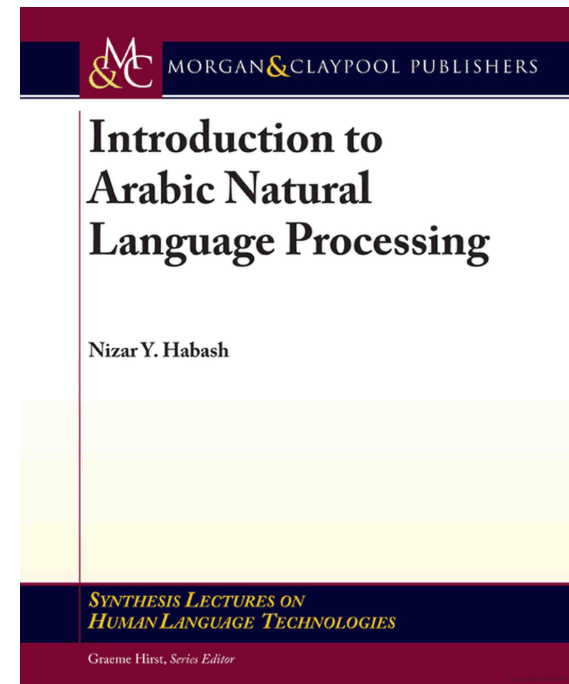
Who I am

- Professor Nizar Habash د. نزار حبش
- Computer Science @ NYUAD
- Specialty area
 - Artificial Intelligence
 - Natural Language Processing
 - Computational Linguistics
 - Language Technologies
- Director of CAMEL Lab
 - Computational Approaches to Modeling Language



Who I am

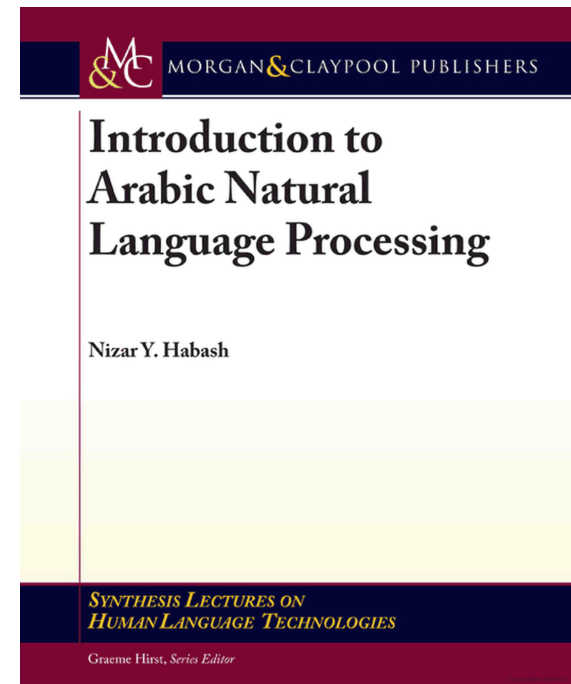
- Professor Nizar Habash د. نزار حبش
- Computer Science @ NYUAD
- Specialty area
 - Artificial Intelligence
 - **Arabic** Natural Language Processing
 - **Arabic** Computational Linguistics
 - **Arabic** Language Technologies
- Director of CAMEL Lab
 - Computational Approaches to Modeling Language





Who I am

- Professor Nizar Habash د. نزار حبش
- Computer Science @ NYUAD
- Awards
 - King Salman Academy Award for the Arabic Language (2022)
 - Antonio Zampolli Prize (2024) for Outstanding Contributions... to Human Language Technologies
 - Fellow of the Association for Computational Linguistics (2025)



CAMEL Lab

- **Computational Approaches to Modeling Language Lab**

- Established in September 2014

- **Research Areas**

- Core Arabic & Arabic dialect NLP
- Resource and tool development
- Machine translation
- Pedagogical applications

- **240+ publications/30 resources**

مختبر «كامل»

- **مختبر الأساليب الحاسوبية لنمذجة اللغة**

- أسس في شهر أيلول/سبتمبر ٢٠١٤

- **نطاقات البحث**

- أنظمة المعالجة الآلية للغة العربية ولهجاتها
- تطوير موارد البيانات والأدوات
- الترجمة الآلية
- التطبيقات التربوية

- **أكثر من ٢٤٠ ورقة علمية و ٣٠ مورداً**

جامعة نيويورك أبوظبي



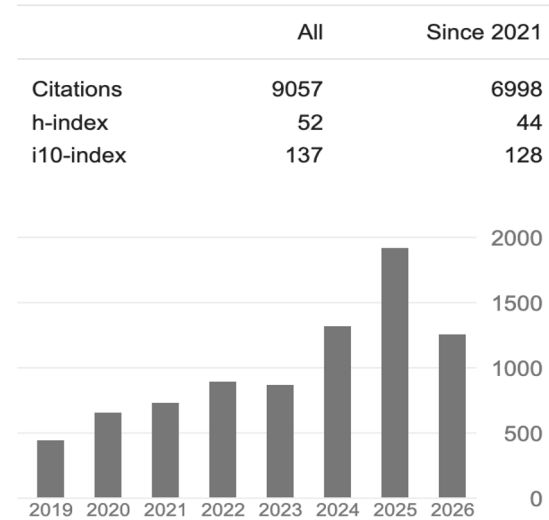
CAMeL Lab

مختبر «كامل»

- Follow us on X
- Publications on Google Scholar
- تابعونا على X
- أوراقنا على الباحث العلمي من جوجل



@CamelNlp



scholar.camel-lab.com

Roadmap

خارطة الطريق

- Introduction
- Arabic AI Challenges
- Solutions & Outlook
- Questions

- مقدمة
- تحديات اللغة العربية
- حلول ورؤى مستقبلية
- أسئلة

Why Arabic NLP?

- **Arabic is one of the top 10 used languages in the world**
 - The national language of ~400M people from 23 states
 - The liturgical language of ~1.8B Muslims
 - One of the six official languages of the United Nations
- **Arabic presents many interesting challenges to NLP/CL**
 - Complex and rich morphology
 - A high degree of orthographic ambiguity
 - Many co-existing geographical and historical variants (dialects)
 - Arabic is written mainly in Arabic Script... but there are other scripts
 - e.g. Arabizi, Maltese, Judeo-Arabic, Garshuni, ...
 - And Arabic Script is used to write many other languages
 - e.g. Azeri, Kurdish, Pashto, Persian, Sindhi, Urdu, Uyghur, ...

Language Specific

Why ~~Arabic~~ NLP?

- **Modern NLP... (as of 2026)**

- Language-independent ML algorithms

- *Bias in algorithms and assumptions*

- Multilingual models

- *Distributional bias*

- Language-specific data sets and tasks

- *Language-specific standards, challenges, and quirks*
 - *Data preparation*
 - *Error Analysis*
 - *Classification vs (controllable) generation tasks*
 - *Explainability*

The Main Challenges for Arabic Language Processing

التحديات الرئيسية للمعالجة الآلية للغة العربية

- Orthographic ambiguity
- Morphological richness
- Dialectal variation
- Orthographic inconsistency
- Resource poverty (data & tools)
- Limited research

- الإبهام الإملائي
- الغنى الصرفي
- تعدد اللهجات
- الأخطاء الإملائية
- فقر موارد البيانات والأدوات
- البحث العلمي المحدود

Arabic Script

الخط العربي

العربية



العربية



- Written right-to-left
- Letters have contextual variants

- مكتوب من اليمين إلى اليسار
- حروف أشكالها متغيرة سياقيا

Arabic Script

الخط العربي

العربية ←

العرابية ←

- Written right-to-left
- Letters have contextual variants

- مكتوب من اليمين إلى اليسار
- حروف أشكالها متغيرة سياقيا

The good, the bad, and the ugly

- Regrettable Arabic tattoos
- Can you sport the problems?



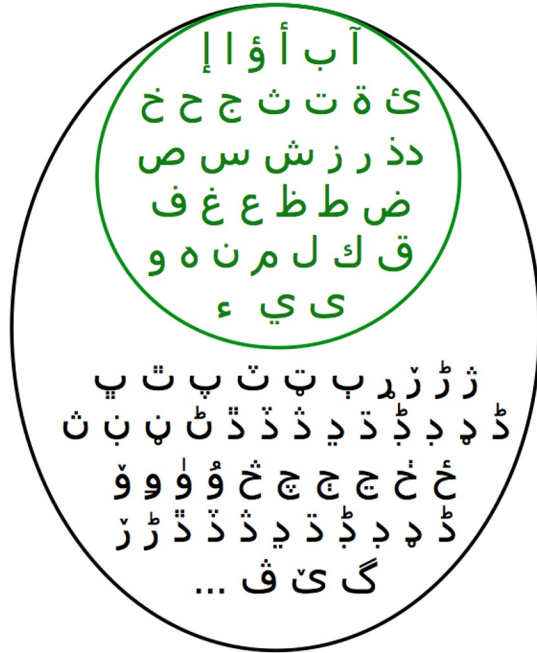
ماريا Maria



مقاتل Fighter

Arabic Script

الخط العربي



العربية ←

العرابية ←

- Written right-to-left
- Letters have contextual variants
- Used to write many languages besides Arabic: Persian, Kurdish, Urdu, Pashto, etc.

- مكتوب من اليمين إلى اليسار
- حروف أشكالها متغيرة سياقيا
- يستخدم لكتابة العديد من اللغات إلى جانب العربية: الفارسية والكردية والأردية والباشتو ، إلخ.

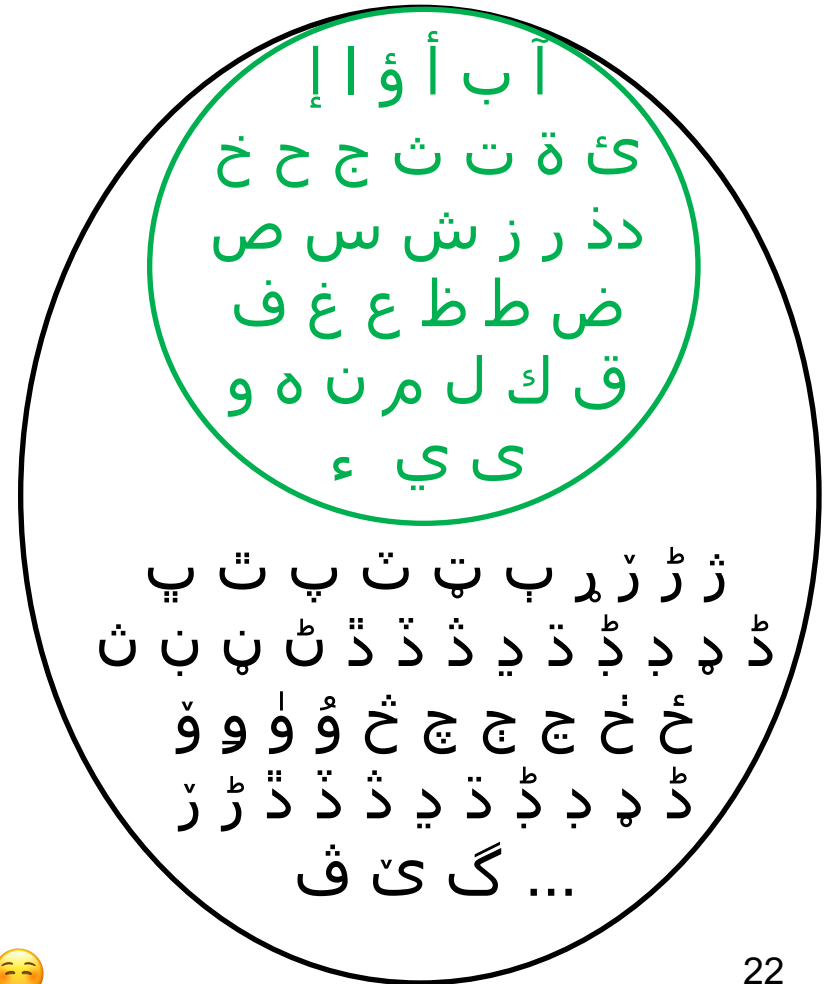
Arabic Script
Other languages

Arabic

- No more than 3 dots
- Dots either above or below
- Marks are 1/2/3 dots, hamza (ء) or madda (~) only
- Rare borrowing for foreign words
 - پ/p/, ف/v/, گ/g/, چ/tʃ/
 - regionally variable

Not Arabic

- Extra marks: haft (v), ring (o), taa (ط), four dots (::), vertical dots (:)
- Some Numerals (٤,٥,٦)



Once you learn the alphabet, it is easier 😊

News

People Are Using an Ancient Method of Writing Arabic to Combat AI Censors

After Palestinian content was restricted or removed from Instagram and Facebook, social media users developed a crafty method of altering Arabic script.



Hakim Bishara May 19, 2021

217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132

الانسان ضد الآلة: بوره الباط العربيه

217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132
217 131 216 179 217 133 32 216 165 216 179 216 177 216 167 216 166 217 138 217 132

Social media users have been using an ancient dotless Arabic script to elude censorship of Palestinian content (courtesy Mada Masr)

Disambiguating Arabic without dots

الإنسانُ ضدُّ الآلةِ

الإنسانُ ضدُّ الآلةِ

الإنسان ضد الآلة

الانسان ضد الاله

الاسان صد الاله

human against machine

Transliteration

Buckwalter's Scheme

- Romanization
 - One-to-one orthographic mapping
 - Not a phonological map
 - Left-to-right
 - Easy to learn and use
- Commonly used in Arabic NLP

AlslAm Elykm

AlslAm Elykm السلام عليكم

/as-salāmu ‘alaykum/

/as-salāmu ‘alaykum/

Peace be upon you! = Hello!

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

Orthographic Ambiguity

الإبهام الإملائي

- Arabic script uses optional diacritical marks
 - 1.5% of newspaper words have some diacritical marks
 - Standard Arabic has 6.8 diacritizations and 2.7 lemmas per word

- الإملاء بالخط العربي يستخدم التشكيل الاختياري
 - ١,٥ ٪ من كلمات نصوص الصحف تحتوي على علامات تشكيل
 - لكلمات العربية الفصحى ٦,٨ تشكيلات و ٢,٧ مداخل معجمية بالمتوسط

Vowel				Nunation			Gemination
بَ	بُ	بِ	بْ	بَّ	بُّ	بْ	بّ
/ba/	/bu/	/bi/	/b/	/ban/	/bun/	/bin/	/bb/

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 صحفي ان التعاون بين اسبانيا والمغرب لم يتوقف ابدا ولم يجمد.

إِسْبَانِيَا تَنْفِي تَجْمِيدَ الْمُسَاعَدَةِ الْمَمْنُوحَةِ لِلْمَغْرِبِ

مَدْرِيد 1 - 11 (إِف ب)- أَكَّدَ رَئِيسُ الْحُكُومَةِ الْإِسْبَانِيَّةِ خُوسِيَّة مَارِيَا أَثْنَارَ الْيَوْمَ الْخَمِيسَ أَنَّ إِسْبَانِيَا لَمْ تَوَقِّفَ الْمُسَاعَدَةَ الَّتِي تُقَدِّمُهَا لِلْمَغْرِبِ خِلَافًا لِمَا أَكَّدَهُ أَمْسَ الْأَرْبَعَاءَ وَزِيرَ الشُّؤُونِ الْخَارِجِيَّةِ وَالتَّعَاوُنِ الْمَغْرِبِيِّ مُحَمَّدُ بْنُ عَيْسَى أَمَامَ مَجْلِسِ النُّوَابِ الْمَغْرِبِيِّ . وَقَالَ رَئِيسُ الْحُكُومَةِ الْإِسْبَانِيَّةِ فِي مُؤْتَمَرٍ صَحَافِيٍّ أَنَّ التَّعَاوُنَ بَيْنَ إِسْبَانِيَا وَالْمَغْرِبِ لَمْ يَتَوَقَّفَ أَبَدًا وَلَمْ يُجَمَّدْ.

Orthographic Ambiguity

الإبهام الإملائي

- Arabic script uses optional diacritical marks
 - 1.5% of newspaper words have some diacritical marks
 - Standard Arabic has 6.8 diacritizations and 2.7 lemmas per word

- الإملاء بالخط العربي يستخدم التشكيل الاختياري
 - ١،٥ ٪ من كلمات نصوص الصحف تحتوي على علامات تشكيل
 - لكلمات العربية الفصحى ٦،٨ تشكيلات و ٢،٧ مداخل معجمية بالمتوسط

ولعين

Orthographic Ambiguity

الإبهام الإملائي

- Arabic script uses optional diacritical marks
 - 1.5% of newspaper words have some diacritical marks
 - Standard Arabic has 6.8 diacritizations and 2.7 lemmas per word

- الإملاء بالخط العربي يستخدم التشكيل الاختياري
 - ١,٥ ٪ من كلمات نصوص الصحف تحتوي على علامات تشكيل
 - لكلمات العربية الفصحى ٦,٨ تشكيلات و٢,٧ مداخل معجمية بالمتوسط

ولعين

وَلَعَيْنُ

wa+laEiynN

and+cursed

وَلِعين

wa+li+EaynK

and+for+an_eye

وَلِعينَ

waliEiyna

Infatuated (m.pl)

Orthographic Ambiguity

- Arabic words can be very ambiguous due to optional diacritics
- But how ambiguous?
- 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 l ‘Alif’
- Some final short vowels are deterministically inferable

ths is wht an Arbc txt lks lik wth no vwls

- For a computer ...
 - A word on average has 12.3 analyses, **6.8 diacritizations**, and 2.7 lemmas (core meanings)
 - Not all of this ambiguity is due to orthography! More on this later.

Morphological Richness

الغنى الصرفي

- Arabic has a very rich inflectional system
 - For example, Arabic verbs have 5,400 inflected forms
 - Whereas English verbs have 6 and Chinese verbs have 1!

- اللغة العربية نظام صرفي غني جداً
ينتج عنه تراكيب كثيرة
 - فمثلاً، للفعل العربي حوالي ٥,٤٠٠ تصريف
 - بينما للفعل الإنجليزي ٦ تصاريف،
وللفعل الصيني تصريف واحد!

وسنقولها
/wasanaqūluhā/
و + س + ن + قول + ها
wa+sa+na+qūl+u+hā
and+will+we+say+it
And we will say it

قال، قالت، قالاً، قالوا، قلت،
قلت، قلتما، قلتهم، قلتن،
يقول، يقول، يقل، تقول، تقول،
تقل، تقولين، تقولي،
...فقال، فقالت، فقالا...
...وسأقولها، وسنقولها ...

Morphological Complexity

- English is **not** morphologically rich.
 - The number of inflected forms is small
 - The verb paradigm is limited to 6

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

- The complete English part-of-speech tag set has 48 tags
- The complete Arabic part-of-speech tag set has 22,400 tags

Analysis vs. Disambiguation

byn

بين



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

Morphological Analysis

Morphological Disambiguation

is out-of-context

is in-context

Analysis vs. Disambiguation

Will Ben Affleck be a good Batman?

هل سينجح بين أفليك في دور باتمان؟



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

Morphological Analysis

is out-of-context

Morphological Disambiguation

is in-context

Dialectal Variation

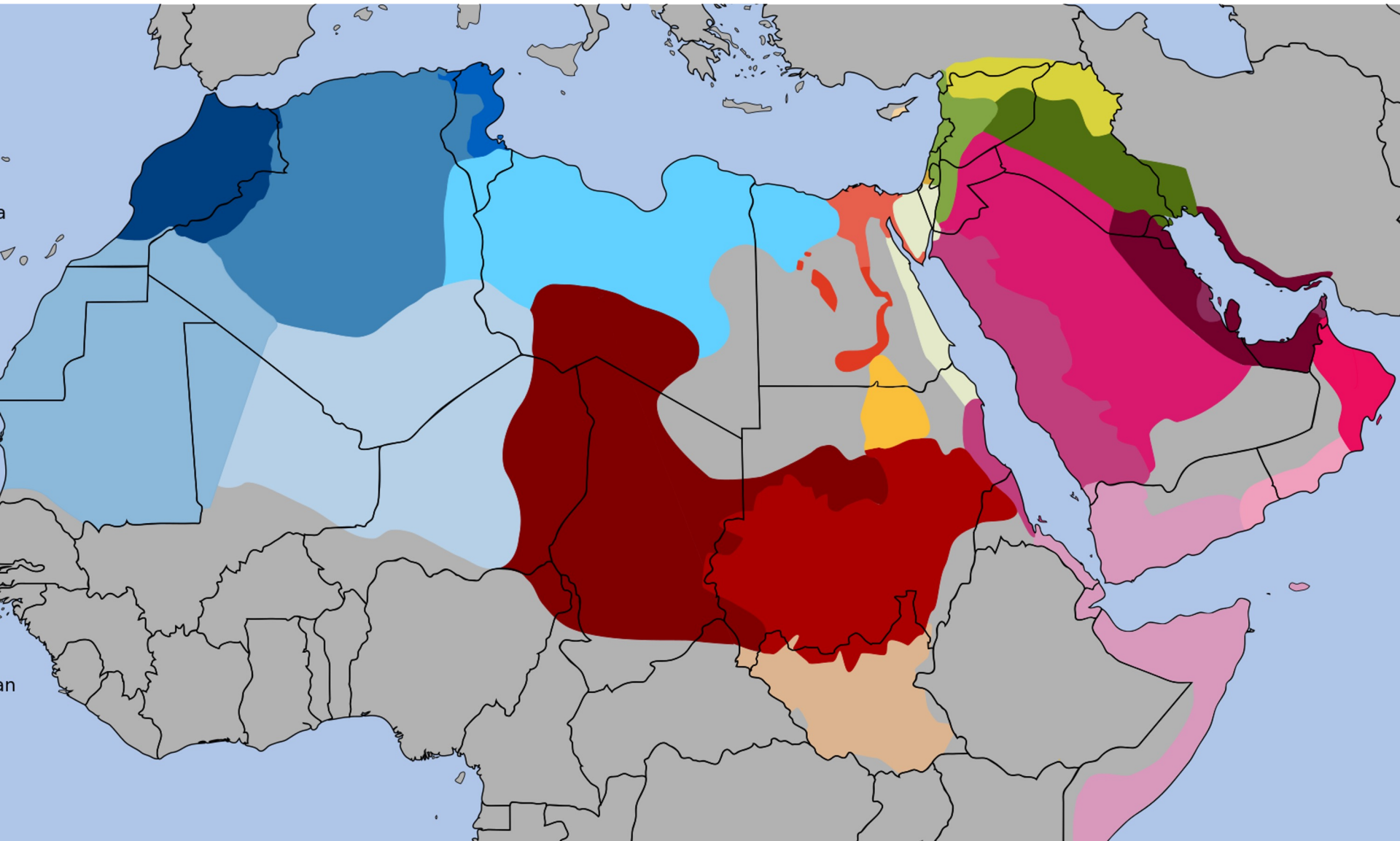
- Classical Arabic
 - Quranic Arabic
 - Historical texts
- Modern Standard Arabic
 - Official language
 - Language of news & media
 - Standard writing & grammar
 - The National Language
- Dialectal Arabic
 - Predominantly spoken
 - No official standardization
 - The Mother Tongue
 - Increasing use on social media

تعدد اللهجات

- فصحي التراث
 - لغة القرآن الكريم
 - نصوص تاريخية
- فصحي العصر
 - اللغة الرسمية في البلدان العربية
 - لغة الصحف والإعلام
 - لها إملاء قياسي وقواعد قياسية
 - لغة الأمة
- اللهجات العربية
 - لغة التحدث الرئيسية
 - بدون إملاء أو قواعد قياسية رسمية
 - لغة الأم
 - وجود متزايد على وسائل التواصل الاجتماعية

Dialectal Variation

تعدد اللهجات



Dialectal Variation

تعدد اللهجات

- Coarse Classification

- Gulf Arabic
- Levantine Arabic
- Egyptian Arabic
- Maghrebi Arabic

- Classification Problems

- Countries with multiple varieties
- Sub-dialectal varieties:
e.g., Urban, Rural, and Bedouin

- Diglossia and Multilinguality

- التصنيف العام

- اللهجة الخليجية
- اللهجة الشامية
- اللهجة المصرية
- اللهجة المغاربية

- مشاكل تصنيفية

- تعدد اللهجات في نفس البلد
- تعدد اللهجات الفرعية :
المدنية، والفلاحية والبدوية، مثلا

- الازدواجية والتعددية اللغوية

Arabic Diglossia

الازدواجية اللغوية في العربية

	Formal	Informal
MSA	Typical MSA	<i>Cartoon Arabic</i> <i>Telenovela Arabic</i> <i>MSA L2</i>
Dialect	Formal Spoken Arabic	Typical Dialect

Dialectal Variation

تعدد اللهجات

• Coarse Classification

- Gulf Arabic
- Levantine Arabic
- Egyptian Arabic
- Maghrebi Arabic

• Classification Problems

- Countries with multiple varieties
- Sub-dialectal varieties:
e.g., Urban, Rural, and Bedouin

• Diglossia and Multilinguality

• Dialects or Languages?

- Irrelevant question
- Can we exploit dialect linguistic relationships to build better AI?

• التصنيف العام

- اللهجة الخليجية
- اللهجة الشامية
- اللهجة المصرية
- اللهجة المغاربية

• مشاكل تصنيفية

- تعدد اللهجات في نفس البلد
- تعدد اللهجات الفرعية :
المدنية، والفلاحية والبدوية، مثلا

• الازدواجية والتعددية اللغوية

• لغات أو لهجات أو لغجات؟

- سؤال غير مهم من ناحية تقنية
- السؤال الأهم: هل يمكن استغلال علاقة الصلة بين اللهجات لتطوير أنظمة أذكى؟

The Case of Maltese

- A language that is very closely related to Arabic.
 - “A language is a dialect with an army and navy”
Max Weinreich
- Standardized Latin-based orthography

Maltese	Il-bnedmin	kollha	jitwiellu	hielsa	u ugwali	fid-dinjità	u d-drittijiet
Arabic	البنادمين <i>AlbnAdmyn</i>	كلها <i>klhA</i>	يتوالدوا <i>ytwAldwA</i>	خالصة <i>xAlSp</i>	واجوالي <i>wAjwAly</i>	فالدنيّتا <i>fAldnytA</i>	والدريّتيّات <i>wAldrytyAt</i>
Gloss	the-humans	all	they-are-born	free	and equal	in-the-dignity	and the-rights
Origin	AR	AR	AR*	AR*	AR IT	AR IT	AR IT*
English	All human beings are born free and equal in dignity and rights						

Kurt Micallef, Fadhl Eryani, Nizar Habash, Houda Bouamor, and Claudia Borg. 2023. [Exploring the Impact of Transliteration on NLP Performance: Treating Maltese as an Arabic Dialect](#). In *Proceedings of the Workshop on Computation and Written Language (CAWL 2023)*, Toronto, Canada.

Phonological Variation

الاختلافات الصوتية

- First noticed differences
 - Not the most important

- الاختلافات الصوتية تُلاحظ فوراً
 - ليست الأكثر أهمية

الفصحى MSA		اللهجات Dialects	
ق	/q/	ق، ك، ء، گ، ج	/q/, /k/, /ʔ/, /g/, /dʒ/
ث	/θ/	ث، ت، س	/θ/, /t/, /s/
ذ	/ð/	ذ، د، ز	/ð/, /d/, /z/
ج	/dʒ/	ج، گ، ج	/dʒ/, /g/, /ʒ/

/qalb/, /galb/, /ʔalb/ ⇔ heart q/b قلب

fjr dawn فجر ، fqr poverty فقر ⇔ /fagr/

Morphological Variation

الاختلافات الصرفية

- Compared to Standard Arabic, dialectal morphology can be simpler or more complex
 - No case
 - New complex clitics

• مقارنة بصرف الفصحى، فإن صرف اللهجات أحياناً أبسط وأحياناً أكثر تعقيداً

- اختفاء الحركات الإعرابية
- تعدد التراكيب اللواصقية

MSA $\Leftrightarrow$ Dialect

كتاب $\Leftrightarrow$ كتابٌ، كتابٌ، كتاباً، كتابٍ، كتابِ

ktAbu, ktAbN, ktAba ktAbAF... $\Leftrightarrow$ ktAb

book (different case endings) $\Leftrightarrow$ book

Morphological Variation

- Compared to Standard Arabic, dialectal morphology can be simpler or more complex
 - No case
 - New complex clitics

الاختلافات الصرفية

- مقارنة بصرف الفصحى، فإن صرف اللهجات أحياناً أبسط وأحياناً أكثر تعقيداً
 - اختفاء الحركات الإعرابية
 - تعدد التراكيب اللواصقية

MSA $\Leftrightarrow$ Dialect

وماكتبوها لوش $\Leftrightarrow$ ولم تكتبوها له

w+lm t+ktb+w+hA lh $\Leftrightarrow$ w+mA+ktb+tw+hA+l+w+\$

and you did not write it for him

Morphological Variation

الاختلافات الصرفية

- Dialects vary systematically among themselves

- هناك العديد من الاختلافات المنهجية الصرفية بين اللهجات

	Past الماضي	Future المستقبل
فصحى Standard	كتب /kataba/	سيكتب /sayaktubu/
شامي Levantine	كتب /katab/	حيكتب /hayiktob/
خليجي Gulf	كتب /kitab/	بيكتب /byiktib/
مصري Egyptian	كتب /katab/	هيكتب /hayiktib/
مغربي Moroccan	كتب /kteb/	غيكب /ayekteb/

Morphological Variation

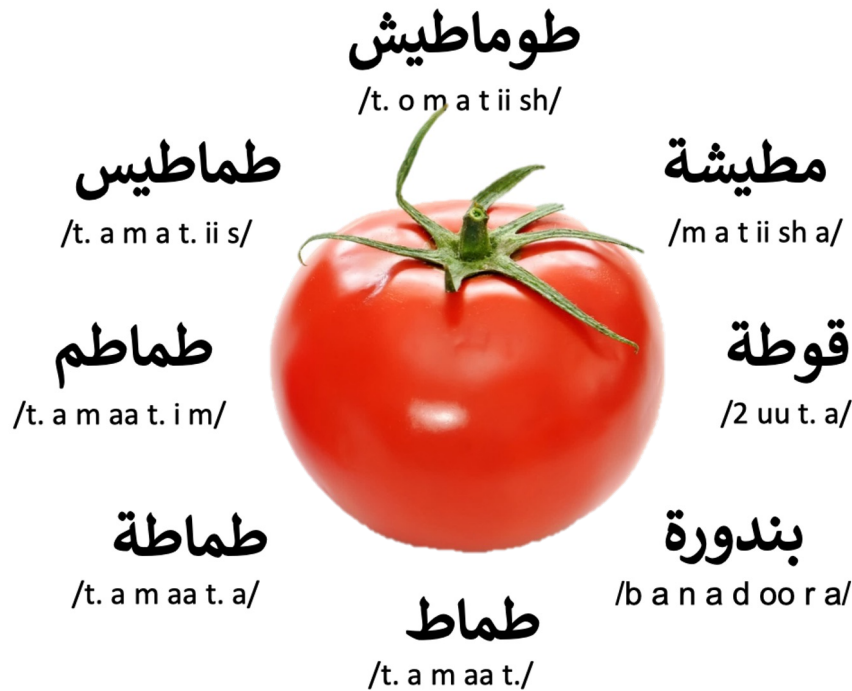
الاختلافات الصرفية

- Dialects vary systematically among themselves

- هناك العديد من الاختلافات المنهجية الصرفية بين اللهجات

	Pastالماضي			Presentالمضارع	
	المتكلم 1S	المخاطب 2S	المخاطبة 2S	المتكلم 1S	المتكلمون 1P
شامي Levantine	كتبت /katab t /	كتبت /katab ti /		اكتب /a k tob/	نكتب /no k tob/
مغربي Moroccan	كتبت /kteb t /	كتبت /kteb ti /		نكتب /ne k teb/	نكتبوا /ne k tebu/

Lexical Variation



*You say **to-MAY-to**,
I say **to-MAH-to**!*

الاختلافات المعجمية

brAd؟ براد؟



/b a r r a a d/



/b a r r a a d/



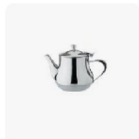
أنا أشهر
براد بيت!

"House Fridge"

/b r a a d # b e t/



عصاف صيف
... تسوق افضل افریق ككثيره متيل ...



دافية ستر
براد شاي استيل 0.50 لتر - دافو...



زودع دنیا التوفیر
الصيف براد شاي اصفر صناعة ك...



ويكيبيديا
براد مخربي - ويكيبيديا



Eval El-Torky - In stock
براد استنسل بيد بكليت - key...



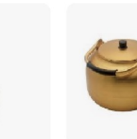
Craft Electronics
براد ذو بابين رمادي والسود



مخبر شركة باصرة للآرائي الشراعية
16-93082 - براد شاي استيل ...



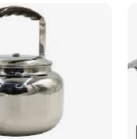
الصيف كيتش
براد شاي علاقي فخسي



شركة العارس سوتر التجارية
براد اصفر مقاس 1.5 لتر



الصيف كيتش
براد يد بني استيل مستحد الاحجام



in stock - سارون مسر
براد استنسل مثيل 2 لتر - ميلق...



شركة نوازم لغات
براد اصفر صيفي 1.5 لتر ...



شركة نوازم لغات
براد اصفر صيفي 1.5 لتر ...



دافية صيف
براد شاي استيل 0.9 لتر - دافو...



مخبر شركة باصرة للآرائي الشراعية
16-93050 - براد شاي كلاسي...



480degree
الشر براد شاي استنسل مثيل 1.5...



مخبر نوازم لغات
براد شاي الكليوم خاني (مخ)



مخبر افراسي اري الشراعية
براد شاي استيل - مخبر افراسي اري...



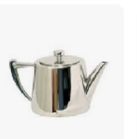
in stock - سارون مسر
براد من الالاستن استنسل 1.5 لتر...



ALHAFIDH
Panasonic Free Stand...



كارفور
تسوق هوم براد شاي سنكليل ستر...



مخبر شركة لسانا للآرائي الشراعية
براد شاي استيل مخبر 0.60 لتر...



Mogambo Store
براد ماء هيفاني هيفاني...



تيمبل فورد
براد ماء ارجين - كوري - خلوا...



Orindi
براد ماء Denka water...

Related searches

براد ثلاثة

براد قديم

براد شاي



مخبر شركة باصرة للآرائي الشراعية
16-90069 - براد شاي عرود...



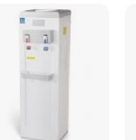
Jum3a.com
براد صودي



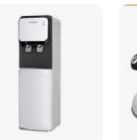
Elyran
الروان: براد ماء موندكس ...



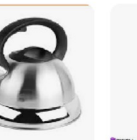
umbrella-egypt.com
براد شاي سنكليل



... شركة خبثات على السلسل والفران
براد مهندا جالي باراد 1932S ...



Miwag
... oler Dispenser ...



IRAK Store
براد صفاة افراسي...



ديمي شي
تسوق براد 18 قدم فريد هرات...



ALHAFIDH
براد ماء ماضتي - AFIDH ...



شركة اركا الشدة لشراعية
براد مافري 1.4 لتر فخسي



Alsaif Home
براد نوازي 0.30 لتر - ايفس زجاج مثيرة من ال...



Mogambo Store
براد ماء هيفاني STORE ...



umbrella-egypti.com
براد شاي سنكليل تركي 1 لتر

Related searches

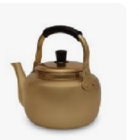
براد مغربي

براد ماء

براد شاي قديم



Beylina Homeware
براد شاي جويانيه تركي لوي...



شركة الصومعي للآرائي الشراعية
براد شاي اصفر كوري - مخبر اله...



هولي
براد شاي استيل 0.9 لتر فخسي مث...



ويكيبيديا
براد بيت - ويكيبيديا



زنان كور
تسوق من افكر لشركة براد سنكليل...



Moroccan Home
براد فاني - Moroccan Home



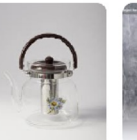
مئة شمالي توافرات
براد شاي - مئة شمالي توافرات



... مطهي الحارة زيت الزيتون
براد شاي مغربي - حجم ك...



Jumia
براد شاي سنكليل - مزود بصفاة ...



مخبر نوازي ارفي الصيغة
براد زجاج مع صفاة - مخ...



www.ico.ae
براد ماء لمسجد - هيئة الاصل الخيرة العالمية



rawcon.com
براد شاي مغربي ايق مثيل مث...



مخبر صوبنج
براد شاي استنسل 1.5 لتر ال...



2 - سارون مسر
براد شاي تركي- لايال - 2 لتر : on.eg



زنان كور
تسوق من افكر لشركة براد سن...



Beylina Homeware
براد شاي جويانيه تركي لوي...



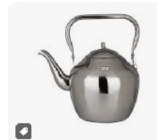
480degree
افضل براد شاي سنكليل مثيل ...



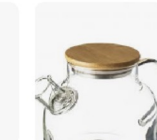
براد بيت
براد بيت: رجلة افراسي هوليود - المية وال...



مخبر لافان
براد شاي الكليوم مقوش - مخبر لافان بيت ال...



in stock - مخبر افراسي
براد سيل 3 لتر هدي - مخبر افراسي



Mini Outlet
براد شاي - آل الاسمار والت...



لوازم كور
براد فلهو 4 لتر



مخبر صوبنج
براد شاي استنسل 1 لتر ال...

Foreign Borrowings

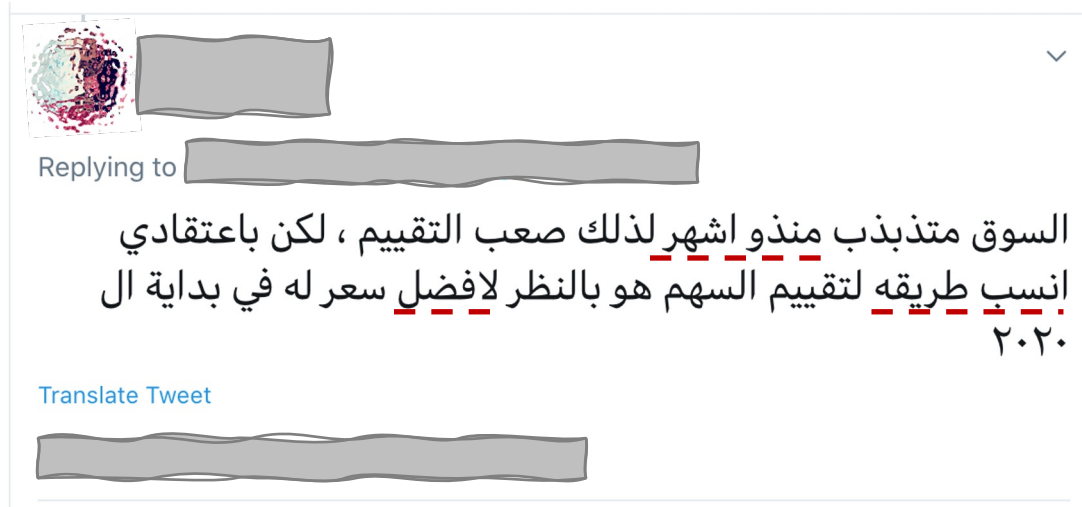
○ فلسفة	falsafah	philosophy
○ فيلسوف	faylasuwf	philosopher
○ بندورة	banaduwhah	pomodoro (tomato)
○ بيرا	biyrA	birra (beer)
○ أوكي	<uwky	okay
○ مرسي	mirsiy	merci
○ تلفون	talifawn	telephone
○ تلفن	talfan	to phone
○ فرمت	farmat	to format
○ فرمته	farmatah	formatting

Common Spelling Errors and Inconsistency

كثرة الأخطاء الإملائية والتضارب الإملائي

- Unedited standard Arabic text on news commentaries has an error rate of ~30% (QALB project)

- نسبة الخطأ في التعليقات الغير محررة المكتوبة بالعربية الفصحى على صفحات بعض مواقع الأخبار تصل إلى ٣٠٪ (مشروع «قلب»)



Common Spelling Errors and Inconsistency

كثرة الأخطاء الإملائية والتضارب الإملائي

- But in the dialects, there is no official orthography! So, there are many free forms including Arabizi.

- ولكن في اللهجات، لا يوجد معيار إملائي قياسي! لذا، هناك تنوع في التهجئات بما في ذلك الإملاء المرومن (العريزي).

وجدنا على الإنترنت ٢٧ تهجئة لهذه الكلمة باللهجة المصرية

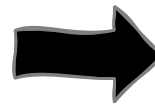
We found 27 spellings online of the Egyptian Arabic word

ما يقولهاش

/mabiʔulhāʃ/

“he does not say it”

(Habash et al., 2018)



Arabic Orthography	Arabic Transliteration	Frequency
مبيقولهاش	<i>mbyqwlhAš</i>	≈ 26,000
ما بيقولهاش	<i>mA byqwlhAš</i>	≈ 13,000
ما بقلهاش، مبقولهاش، مبقلهاش، ما بقلهاش، مابقولهاش	<i>mAbqlhAš, mbqwlhAš, mbqlhAš, mA bqlhAš, mAbbyqwlhAš</i>	≤ 10,000
ما بقلهاش، ما بقلهاش، مبيقلهاش، ما بيقلهاش، مابقلهاش	<i>mAbqwlhAš, mA bqwlhAš, mbyqlhAš, mA byqlhAš, mAbbyqlhAš</i>	≤ 1,000
مبئلهاش، ما بيئولهاش، ما بيئولهاش، ما بيئولهاش	<i>mbġlhAš, mAbġwlhAš, mA byġwlhAš, mAbbyġwlhAš</i>	≤ 100
ما بيؤلهاش، ما بيؤلهاش، مبيؤلهاش، ما بيؤلهاش، ما بؤلهاش، ما بؤلهاش، مبيؤلهاش، ما بؤلهاش	<i>mA byŵlhAš, mAbġlhAš, mbyġwlhAš, mA byġlhAš, mAbġwlhAš, mA bġlhAš, mA bŵlhAš, mbġwlhAš, mbyŵlhAš, mAbŵlhAš, mbŵlhAš</i>	≤ 10

Arabizi variants: mabi2ulhash, mabiquhash, mabi'ulhash,... بضعة تهجئات عريزية :

Common Spelling Errors and Inconsistency

كثرة الأخطاء الإملائية والتضارب الإملائي

- But in the dialects, there is no official orthography! So, there are many free forms including Arabizi.

- ولكن في اللهجات، لا يوجد معيار إملائي قياسي! لذا، هناك تنوع في التهجئات بما في ذلك الإملاء المرومن (العريزي).



كثرة الأخطاء الإملائية والتضارب الإملائي

- ولكن في اللهجات، لا يوجد معيار إملائي قياسي! لذا، هناك تنوع في التهجئات بما في ذلك الإملاء المرومن (العريزي).
- But in the dialects, there is no official orthography! So, there are many free forms including Arabizi.

[illegible]

Arabizi - 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



Akl 1961

Ḥ ḥaleef	أ	F fe
B be		V ve
P pe		Q qaaf
T te		L laam
Ṭ 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
Z zahh		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)		

عربي	IPA	Latin	عربي	IPA	Latin
أ إ آ ء ؤ ئ	/ʔ/	‘ 2 ∅	ث	/θ/	th
ة	/a/,/t/	a t	ط	/t/	t T 6
ح	ħ	H h 7	ع	/ʕ/	‘ 3 ∅
خ	/x/	kh 7’ x 8	غ	/ɣ/	g gh 3’
ذ	/ð/	th	ق	/q/	q
ش	/ʃ/	sh ch	ي	/y/ /ay/ /ī/ /ē/	y,i,e, ai,ei,...

Duolingo Romanization

Other important symbols

أ silent	آ 2aa	ء silent	ؤ silent
إن in	ون un	ان an	اا aa

Special letter combinations

لـ l+m	لا laa
-----------	-----------

Short vowel markings

ي i	و u	ا a
--------	--------	--------

ف f	غ gh	ع 3	ظ DH
م m	ل l	ك k	ق q
ي ii / y	و uu / w	ه h	ن n
	ة a/at	ى aa	

Placements of hamza

أ 2	ؤ 2	إ 2	أ 2	ء 2
--------	--------	--------	--------	--------



Let's learn Arabic!

Get to know the Arabic writing system

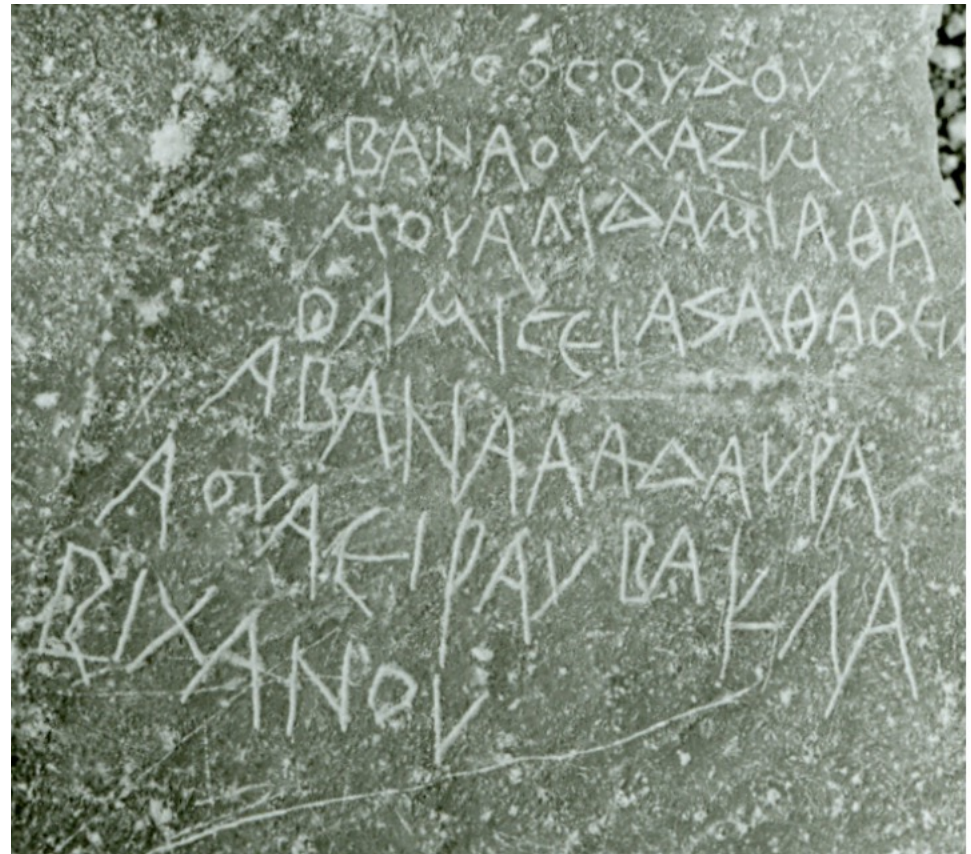
LEARN THE CHARACTERS

ث th	ت t	ب b	ا aa
د d	خ kh	ح H	ج j
س s	ز z	ر r	ذ dh
ط T	ض D	ص S	ش sh

New Epigraphica from Jordan I: a pre-Islamic Arabic inscription in Greek letters and a Greek inscription from north-eastern Jordan

A. Al-Jallad (Leiden University)

A. al-Manaser (University of Oxford)



Reading 1

- 1) Αυσος Ουδου
- 2) Βαναου Χαζιμ
- 3) μου αλ-Ιδαμι αθα
- 4) οα μι- Σιεα ζαθαοε ω
- 5) α Βαναα α-δαυρα
- 6) αουα ειραυ βακλα
- 7) βι-Χανου[v]

Translation

- 1) 'Aws son of 'ūd (?)
- 2) son of Bannā' son of Kazim
- 3) the 'idāmite came
- 4) from Sī' to spend the winter
- 5) with Bannā' in this region
- 6) and they pastured on fresh herbage
- 7) during Kānūn

Greek Aljamiado

- https://en.wikipedia.org/wiki/Greek_Aljamiado

Original text

افندی آپو کرڈیا بندہ ٹیلو
- تٹیرو سو تخوما نادفیلو
دی دتذریمی اغو نکسفره توتو
- ابتریمو یسینا نیدو فیلو
اسی فیلس امینا یاتی زویتی
- اغو ایسینا افندی دنسی فیلو
اکینون بومسس ایسی نمیسو
- اکینون بو تو ٹیلس نیمی فیلو
شلیانزس ایشیشاش خوما ایینان
- یشیلی بوقتسن ستوسون تو شیلو

Transcription into Modern Greek

Αφέντη, από καρδιά πάντα θέλω,
του θυρού σου το χώμα ναν το φιλώ·
τι δεντρί 'μαι εγώ, να 'ξευρα τούτο,
όπου τρέμω για σέναν γοιον το φύλλο.
Εσύ φιλείς εμένα για τη ζωή <τού>τη·
εγώ εσένα αφέντη δεν σε φιλώ.
Εκείνον που μισείς εσύ να μισώ·
εκείνον που το θέλεις να μη φιλώ.
Χιλιάδες οι ψυχές, χώμα έγιναν·
οι χίλιοι π<ρ>όφτασαν στο σον το χείλο.

English translation

Master, from my heart I always want
to kiss the ground at your door.
What kind of a tree am I, I'd like to know,
to tremble for you like a leaf.
You kiss me for [throughout] this life;
I do not kiss you, master.
Whom you hate, I will hate.
Whom you desire, I will not kiss.
There are thousands of souls, they have become dust.
The thousand managed to get to your lips.

Problems with Dialect Translation

Facebook MT – 2017

- 'يصبحهم' may [God] grant them [good] morning' is translated as 'לפגוע בהם' 'to hurt them'
- Character-based models suspected...

HAARETZ

Home | Israel News

Israel Arrests Palestinian Because Facebook Translated 'Good Morning' to 'Attack Them'

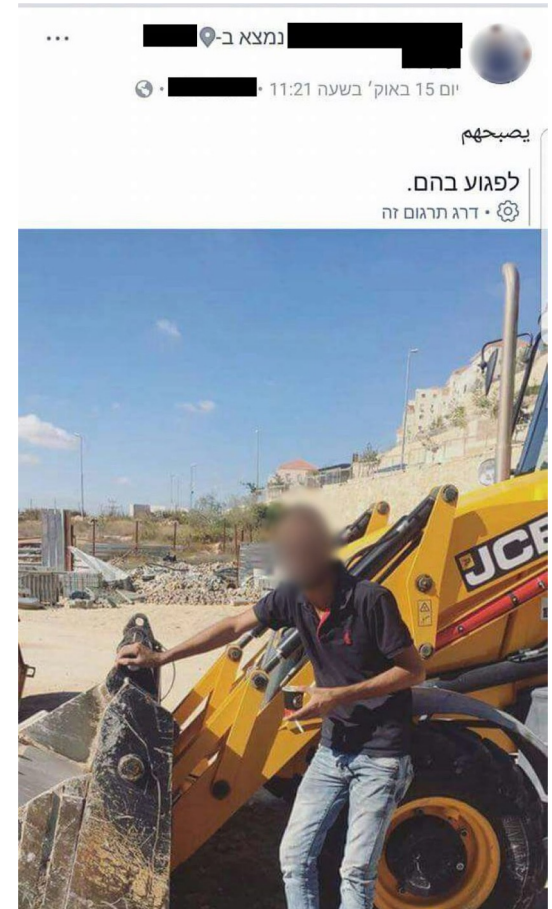
No Arabic-speaking police officer read the post before arresting the man, who works at a construction site in a West Bank settlement

Save Zen Read

Yotam Berger
Follow
Oct. 22, 2017

Listen to this article now
01:49
Powered by Trinity Audio

The Israel Police mistakenly arrested a Palestinian worker last week because they relied on automatic translation software to translate a post he wrote on his Facebook page. The Palestinian



לפגוע בהם *can* یصبحهم mean

Excellent question — and you're thinking in exactly the right direction linguistically.

The Arabic *يُصْبِحُهُم* (*yusbiḥuhum*) *can* sometimes **imply** harm — but not **literally mean** "to harm them" (**לפגוע בהם**) on its own.

Let's unpack this carefully:

◆ Contextual meaning

In **classical or Quranic Arabic**, this phrase is often used in contexts like:

"فَأَصْبَحَهُمُ الْعَذَابُ" — *the punishment befell them in the morning.*

Here, *يُصْبِحُهُمُ الْعَذَابُ* literally means "the punishment overtook them in the morning" — but **implies** that **harm or destruction struck them**.

So in **context**, yes — it can **imply** *לפגוע בהם* (to harm them), but **the verb itself** doesn't inherently mean *to harm*; rather, it describes *the time or manner* something happens.

what is the Quranic phrase that makes this usage?

Excellent — that’s a very insightful follow-up.

Yes — the Qur’anic expression يُصْبِحُهُم (yuṣbiḥuhum) appears in a few places where it carries this **“befalls them / overtakes them in the morning”** meaning, often describing divine punishment.

Here are the key examples 📌

3. Surah Al-Hijr (15:83) explicitly uses the form:

فَأَصْبَحَهُمُ الْعَذَابُ مُسْتَقَرًّا
fa-aṣḥaḥumu al-‘adhābu mustaqirran
→ “So the punishment overtook them in the morning, firmly established.”

Here you can clearly see:

- أَصْبَحَهُم = “befell them in the morning”
- الْعَذَاب (the punishment) = the subject that causes harm

So, the harm comes from the context (the punishment), not from the verb itself.

15:83



فَأَخَذَتْهُمُ الصَّيْحَةُ مُصْبِحِينَ ۝٨٣

But the ‘mihṭv’ blast overtook them in the mornina.

54:38



وَلَقَدْ صَبَحَهُمْ بُكْرَةً عَذَابٌ مُسْتَقَرٌّ ۝٣٨

And indeed, by the early morning they were overwhelmed by an unrelenting torment.

Hallucinated/Confabulated
Quranic Verse

Roadmap

خارطة الطريق

- Introduction
- Arabic AI Challenges
- Solutions & Outlook
- Questions

- مقدمة
- تحديات اللغة العربية
- حلول ورؤى مستقبلية
- أسئلة

Solution Outlines

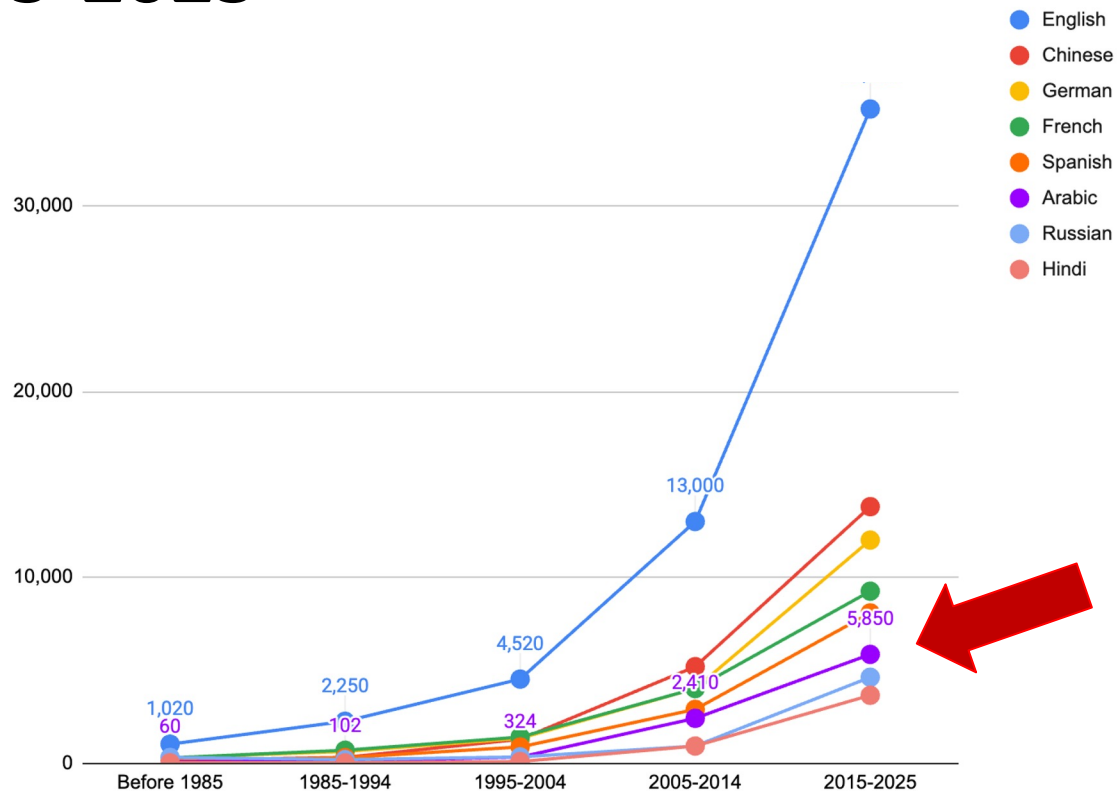
- **Open source data and tools**
 - Closed resources slow progress
 - Intellectual property issues
- **Developing shared standards**
 - But variety has great value
- **Continuous technique updates**
 - Accuracy, efficiency & coolness
- **Nurturing strong collaborations**
 - Teams working on Arabic in many locations
 - Government, Industry & Academia
 - Conferences & Competitions
 - Train a new generation of researchers and developers

الخطوط العريضة للحلول

- **البيانات والأدوات مفتوحة المصدر**
 - الموارد المغلقة تعيق التقدم
 - قضايا الملكية الفكرية
- **وضع معايير مشتركة**
 - لكن للتنوع قيمة كبيرة
- **مواكبة التحديث في التقنيات**
 - الدقة والكفاءة والروعة
- **رعاية التعاون البحثي**
 - فرق تعمل على اللغة العربية في كل مكان في العالم
 - الحكومات والشركات والجامعات
 - المؤتمرات والمسابقات
 - تدريب جيل جديد من الباحثين والمطورين

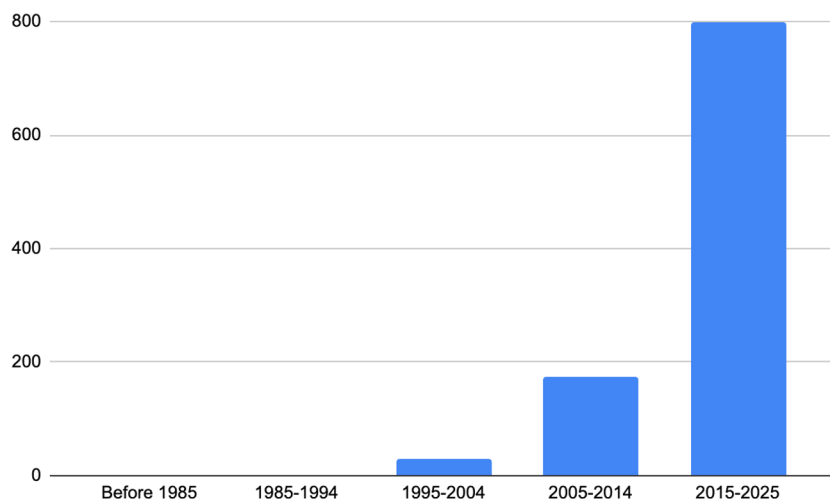
Publications 1985-2025

- Papers mentioning some specific languages in the **ACL Anthology**.



Resources 1985-2025

- Resources in the MASADER Catalog.
- arbml.github.io/masader



Masader+

Stats

Contribute

Search through datasets

ARABIC NLP DATA CATALOGUE

The catalogue currently has 1002 datasets added by 63 contributors

Check the DART dataset

12345...101

PAPER																				
NO.	T.	NAME	T.	LINK	T.	YEAR	T.	DIALECT	T.	VOLUME	T.	UNIT	T.	LINK	T.	ACCESS	T.	TASKS		
1	WDC	GitHub open in ↗	2014	MSA	6000000.0	Tokens												Automatic Creation of Arabic Named Entity Annotated Corpus Using Wikipedia	Free	named-entl recognition
2	Arabic Wiki data Dump 2018	kaggle	2018	MSA	0.0	Documents												Not Available	Free	language-modeling
3	AASL	kaggle open in ↗	2023	MSA	7857.0	Images												RGB Arabic Alphabets Sign Language Dataset	Free	sign-langua recognition
4	Religious Hate Speech	GitHub open in ↗	2018	Mixed	6136.0	Sentences												Are they Our Brothers? Analysis and Detection of Religious Hate Speech in the Arabic Twittersphere	Free	hate-speech detection
5	Arabic Wikipedia Talk Pages	GitHub open in ↗	2017	Mixed	5259.0	Sentences												Code-Switching as a Social Act: The Case of Arabic Wikipedia Talk Pages	Free	codeswitch-detection

جامعة نيويورك أبو ظبي



Selected Projects

مشاريع مختارة



أدوات كامل

CAMeL Tools

- An Open-source Python Toolkit for Arabic Natural Language Processing
 - Arabic Specific
 - Flexibility in use and reuse
 - Modularity
 - High Performance
 - Ease of use for beginner
- مجموعة أدوات مفتوحة المصدر بلغة بايثون لمعالجة اللغة العربية
 - التركيز على اللغة العربية
 - مرونة الاستخدام وإعادة الاستخدام
 - وحدات مستقلة سهلة التجميع
 - الأداء عالي الجودة
 - سهولة الاستخدام للمبتدئين

https://github.com/CAMeL-Lab/camel_tools



Sign in

CAMeL-Lab / camel_toolsPublic

NotificationsFork 95Star 578

<> CodeIssues 30Pull requests 5DiscussionsActionsProjects...

READMEContributingMIT license



أدوات كامل CAMeL Tools

Introduction

CAMeL Tools is suite of Arabic natural language processing tools developed by the [CAMeL Lab](#) at [New York University Abu Dhabi](#).

```
pip install camel-tools
```

Camelira

نظام كامليرا

- **An enabling technology for Arabic**
- **A hybrid solution**
 - Rule-based morphology analyzer
 - Machine learning tagger trained on a large annotated news corpus
 - Select part-of-speech, diacritics, lemma and many features in context in one fell swoop
- **Multi dialectal system**
 - Standard Arabic and Egyptian, Levantine and Gulf Arabic dialects

• تقنية مُمَكِّنة لخدمة اللغة العربية

• حل هجين أسلوباً

- محلل صرفي قواعدي الأسس
- مُوسِّم يعتمد على التعلم الآلي وموارد بيانات تدريبية من نصوص صحفية
- يعين قسم الكلام والتشكيل والمدخل المعجمي ومعظم السمات اللغوية في السياق بضرية واحدة

• نظام متعدد اللهجات

- العربية الفصحى واللهجات المصرية والشامية والخليجية

<http://camelira.camel-lab.com/>

Example of Annotated Data

وقد كاتبته فتحية لمدة سنتين.

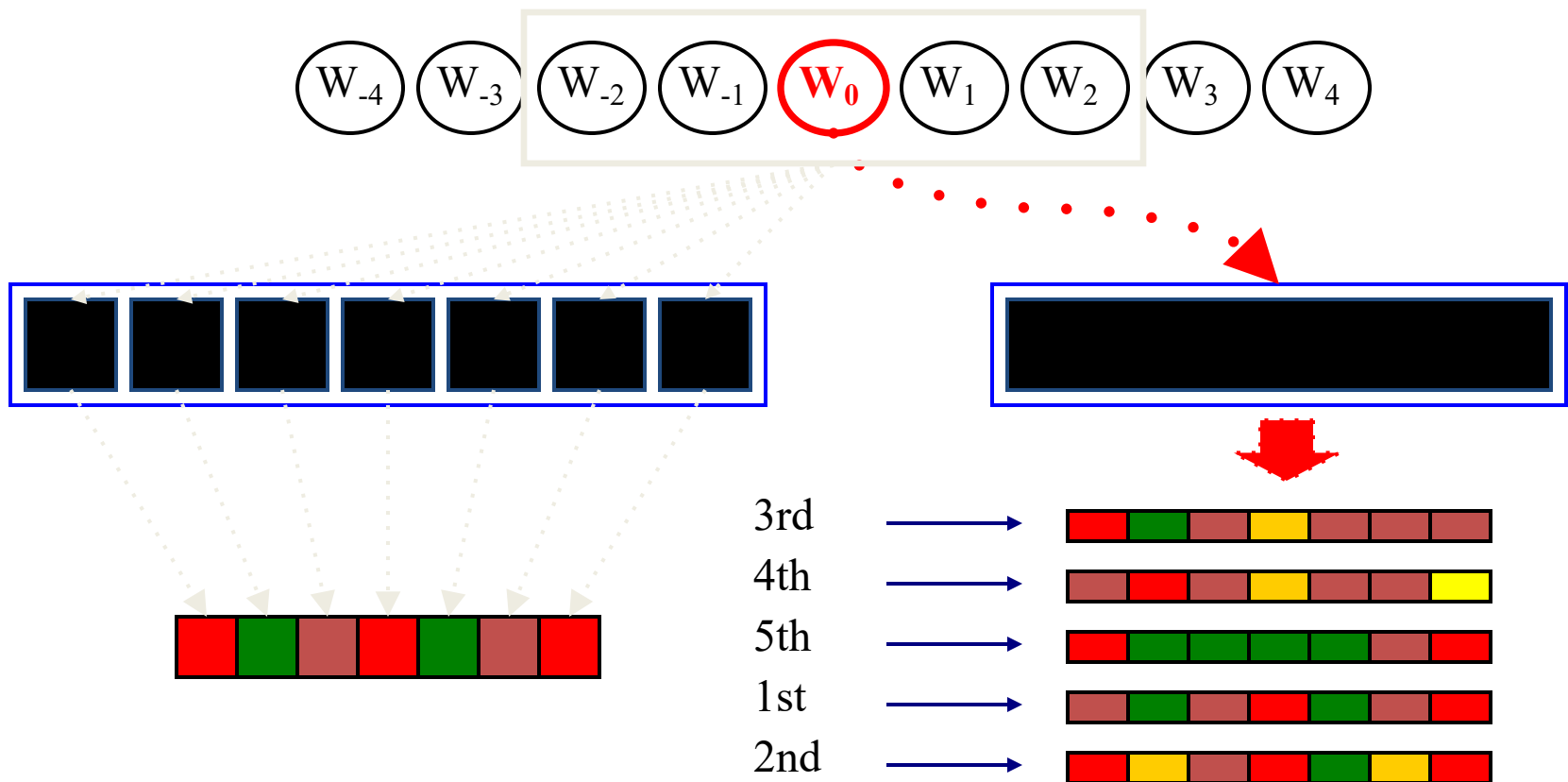
Word	Lemma		POS	Features	Gloss
وقد	قَدَّ	و+قد	Noun	Masc Sg	size/physique
	قَد	و+قد	Particle		may/might/has/have
	وَقَدَّ	وقد	Noun	Masc Sg	fuel/burning
	وَقَدَّ	وقد	Verb	Perf 3rd Masc Sg	kindle/ignite
	وَقَدَّ	وقد	Verb	Perf 3rd Masc Sg	ignite/burn
كاتِبته	كَاتِب	كاتبة+هـ	Noun	Fem Sg	author/writer/clerk
	كَاتَب	كاتبت+هـ	Verb	Perf 3rd Fem Sg	correspond_with
	كَاتَب	كاتبت+هـ	Verb	Perf 1st Sg	correspond_with
	كَاتَب	كاتبت+هـ	Verb	Perf 2nd Fem Sg	correspond_with
	كَاتَب	كاتبت+هـ	Verb	Perf 2nd Masc Sg	correspond_with
فتحية	تَحِيَّة	ف+تحية	Noun	Fem Sg	greeting/salute
	فَتْحِيَّة	فتحية	Proper	Fem Sg	Fathia
لمدة	مُدَّة	ل+مدة	Noun	Fem Sg	interval/period
سنتين	سِنْتُ	سنتين	Noun	Masc Sg	cent
	سَنَة	سنتين	Noun	Fem Sg	year
.	.	.	Punc		.

Example of Annotated Data

وقد كاتبته فتحية لمدة سنتين.

Word	Lemma		POS	Features	Gloss
وقد	قَدَ	و+قد	Noun	Masc Sg	size/physique
	قَدْ	و+قد	Particle		may/might/has/have
	وَقَدَ	وقد	Noun	Masc Sg	fuel/burning
	وَقَدَ	وقد	Verb	Perf 3rd Masc Sg	kindle/ignite
	وَقَدَ	وقد	Verb	Perf 3rd Masc Sg	ignite/burn
كاتِبته	كَاتِبَ	كاتِبة+هـ	Noun	Fem Sg	author/writer/clerk
	كَاتَبَ	كاتبت+هـ	Verb	Perf 3rd Fem Sg	correspond_with
	كَاتَبَ	كاتبت+هـ	Verb	Perf 1st Sg	correspond_with
	كَاتَبَ	كاتبت+هـ	Verb	Perf 2nd Fem Sg	correspond_with
	كَاتَبَ	كاتبت+هـ	Verb	Perf 2nd Masc Sg	correspond_with
فتحية	فَتْحِيَّةَ	ف+تحية	Noun	Fem Sg	greeting/salute
	فَتْحِيَّةَ	فتحية	Proper	Fem Sg	Fathia
لمدة	مُدَّةَ	ل+مدة	Noun	Fem Sg	interval/period
سنتين	سِنَتَ	سنتين	Noun	Masc Sg	cent
	سَنَةً	سنتين	Noun	Fem Sg	year
.	.	.	Punc		.

diac	وَسَيَكْتُبُونَهَا	التشكيل
lex	كُتِبَ1-u	المدخل المعجمي
root	ك.ت.ب	الجذر
pattern	وَسَيَ123ُونَهَا	الوزن
caphi	w_a_s_a_y_a_k_t_u_b_uu_n_a_h_aa	اللفظ
gloss	and+_will+_they_(people)+write+it;them;her	الترجمة
pos	verb	قسم الكلام
catib6	PRT+PRT+VRB+NOM	قسم الكلام/كاتب
ud	CONJ+AUX+VERB+PRON	قسم الكلام/يو دي
prc2	wa_conj	سابقة ٢
prc1	sa_fut	سابقة ١
per	3	الشخص
gen	m	الجنس
num	p	العدد
asp	i	الزمن
vox	a	البناء
mod	i	الصيغة
enc0	3fs_dobj	لاحقة ١
d1tok	وَسَيَكْتُبُونَهَا	تقطيع د١
d2tok	وَسْ+يَكْتُبُونَهَا	تقطيع د٢
atbtok	وَسْ+يَكْتُبُونَ+هَا	تقطيع بنك بن الشجري
d3tok	وَسْ+يَكْتُبُونَ+هَا	تقطيع د٣
bwtok	وَسْ+يَكْتُبُونَ+وَسْ+يَكْتُبُونَ+هَا	تقطيع باكوالتر
bw	وَسْ+CONJ/وَسْ+FUT_PART/ي/IV3MP/كُتِبَ/وَسْ+I+MOOD:SUBJ:MP/وَسْ+DO:3FS/IVSUFF	تحليل باكوالتر
pos_logprob	-1.023208	احتمالية قسم الكلام
lex_logprob	-3.648503	احتمالية المدخل المعجمي



MORPHOLOGICAL CLASSIFIERS

- Multiple independent classifiers
- Corpus-trained

RANKER

- Heuristic or corpus-trained

MORPHOLOGICAL ANALYZER

- Rule-based
- Human-created

Camelira

An Arabic Multi-Dialect Morphological Disambiguator

محلل آلي متعدد اللهجات للغة العربية

<http://camelira.camel-lab.com/>

حصدت المهرة «دريم» لتيري وسولتاو إيتون، وبإشراف المدرب تيري إيتون وقيادة أدریان راموس، ثمار جهودها، حين توجت بطلاً لسباق كأس الوثبة ستاليونز، والذي أقيم بمضمار أرابهو بارك الأميركي، برعاية النسخة الـ

Submit

MSA

Analyzing as Modern Standard Arabic

Diacritized/POS

Tokenized

Lemmatized

حَصَدَتْ الْمُهْرَةُ «دَرِيم» لِتِيرِي وَسُولْتَاوِ إِيْتُونِ ، وَبِإِشْرَافِ الْمُدَرَّبِ تِيرِي إِيْتُونِ وَقِيَادَةِ أَدْرِيانِ رَامُوسِ ، ثَمَارَ جُھُودِهَا ،
حَيْنَ تَوَجَّتْ بِطَلَّةً لِسَبَاقِ كَاسِ الْوُثْبَةِ سْتَالِيُونزِ ، وَالَّذِي أُقِيمَ بِمُضْمَارِ أَرَابُھُو بَارَكِ الْأَمِيرِكِيِّ ، بِرِعَايَةِ النُّسَخَةِ آلِ 14 لِمَهْرَجَانِ
سَبَاقَاتِ سُمُو الشَّيْخِ مَنصُورِ بِنِ زَايِدِ آلِ نَهْيَانِ .

حَصَدَتْ

POS: Verb
Person: 3rd
Gender: Feminine
Number: Singular
Aspect: Perfective
Voice: Active
Mood: Indicative
Gloss: harvest; mow

1.000 - diac: حَصَدَتْ lex: حصدت pos: verb bw: حصد/PV+ت/PVSUFF_SUBJ:3FS gloss: harvest;mow+it;they;she_
0.928 - diac: حَصَدَتْ lex: حصدت pos: verb bw: حصد/PV+ت/PVSUFF_SUBJ:2FS gloss: harvest;mow+you_[fem.sg
0.857 - diac: حَصَدَتْ lex: حصدت pos: verb bw: حصد/PV+ت/PVSUFF_SUBJ:1S gloss: harvest;mow+I_<verb>
0.857 - diac: حَصَدَتْ lex: حصدت pos: verb bw: حصد/PV+ت/PVSUFF_SUBJ:2MS gloss: harvest;mow+you_[masc.:
0.357 - diac: حصدت lex: حصدت pos: noun_prop bw: حصدت/NOUN_PROP gloss: NO_ANALYSIS

Camelira

An Arabic Multi-Dialect Morphological Disambiguator

محلل آلي متعدد اللهجات للغة العربية

<http://camelira.camel-lab.com/>

طلعي لفرقتج ولا عاد اسمعج تكلمين بهالموضوع

Submit Gulf

Analyzing as Gulf

POS Tokenized Lemmatized

طلع ي ولا عاد ا+سمع+ج ت+تكلم+ين

ب+ه+ال+موضوع

بهاالموضوع

POS: Noun
Gender: Masculine
Number: Singular
Proclitic 1: Preposition
Proclitic 0: Demonstrative Particle + Determiner

1.000 - diac: بهالموضوع lex: مَوْضُوع pos: noun bw: ب-/PREP+*/DEM_

أغنية جامدة جدا هتندم لو مشفتهاش

Submit Auto-detect

Analyzing as Egyptian (auto-detected)

Diacritized/POS Tokenized Lemmatized

أَغْنِيَة جَامِدَة جَدًّا خ ت نَدَمَ لَوْ مَا شَفْتُهَاش

مَا شَفْتُهَاش

POS: Verb
Person: 2nd
Gender: Masculine
Number: Singular
Aspect: Perfective
Voice: Active
Mood: Indicative
Proclitic 0: Negative Particle

1.000 - diac: مَا شَفْتُهَاش lex: شَافَ pos: verb bw: مَا/NEG_PART+شَفْ/

MADAR Project

مشروع «مدار»

• Multi-Arabic Dialect

Applications and Resources

- 25 Arab cities (Rabat to Muscat)
- Lexicon, parallel data, dialect id, and machine translation
- More than 100,000 sentences
- 47,000 lexical entries
- 55 individuals worked on MADAR
- with Houda Bouamor and Owen Rambow

• مشروع التطبيقات والموارد متعددة

اللهجات

- ٢٥ مدينة عربية (رباط لمسقط)
- معجم، مكنز متوازي، نظام تعرف على اللهجات، وترجمة آلية للهجات
- أكثر من ١٠٠ ألف جملة
- ٤٧ ألف مدخل معجمي
- ٥٥ شخصاً شاركوا بالعمل
- بالتعاون مع هدى بوعمر وأوون رامبو

• Dialect Id shared task

- 21 teams, 15 countries

• مسابقة عالمية لتحديد اللهجات

- ٢١ فريق من ١٥ دولة

جامعة نيويورك أبو ظبي



Carnegie Mellon University

COLUMBIA UNIVERSITY
IN THE CITY OF NEW YORK

<http://madar.camel-lab.com>

The Cities in MADAR

مدن «مدار»



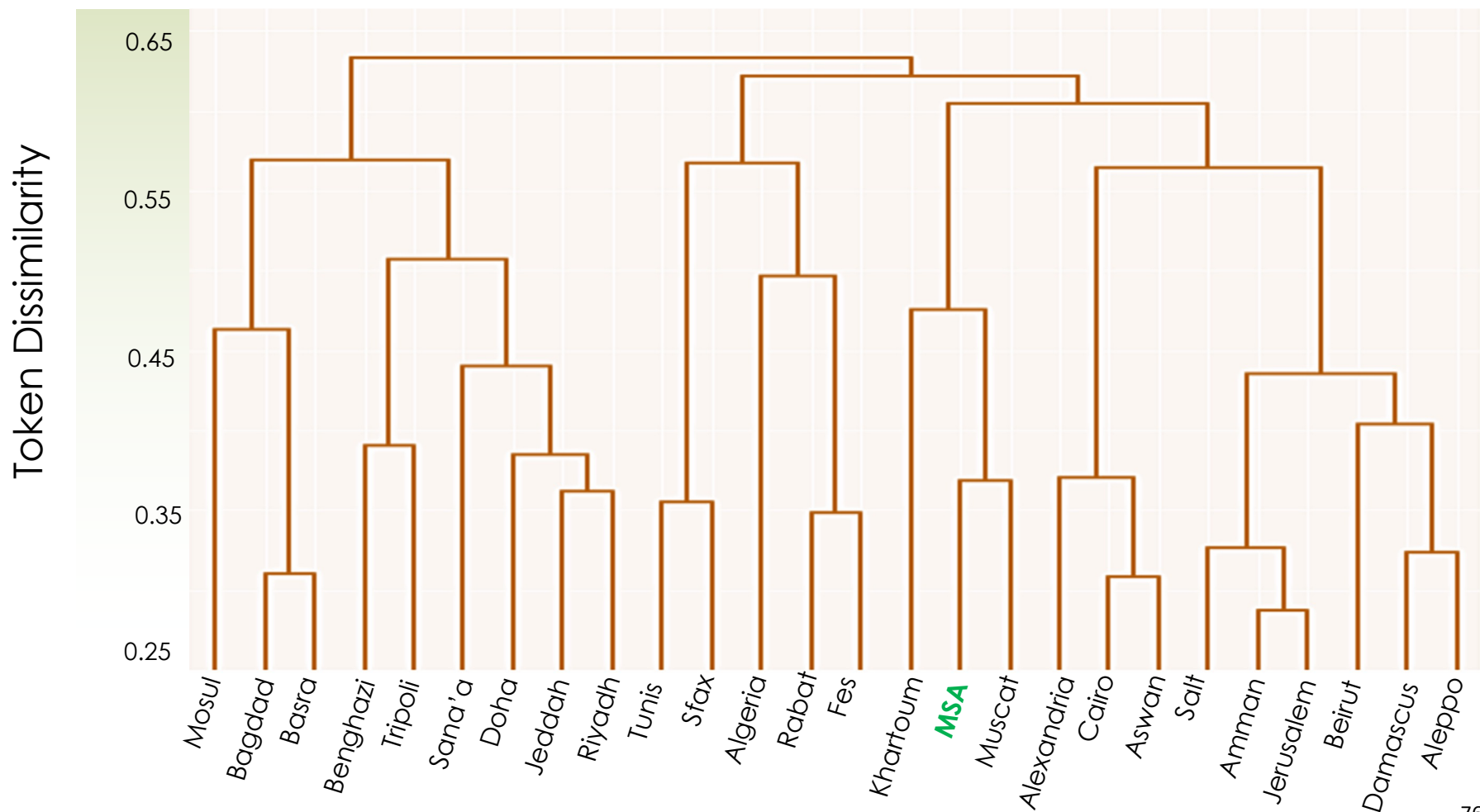
MADAR Corpus Example

مثال من مكنز «مدار»

MSA:	هذه الغرفة صغيرة جداً .		English:	This is a small room.	
	Beirut	Lebanon	الأوضة دي صغيرة أوي .	Cairo	Egypt
	Damascus	Syria	الأوضة دي صغيرة خالص .	Aswan	
	Aleppo		الأوضة دي صغيرة جدا .	Alexandria	
	Jerusalem	Palestine	الغرفة دي صغيرة شديد .	Khartoum	Sudan
	Salt	Jordan			
	Amman				
	Baghdad	Iraq	الغرفة صغيرة قوي .	Sana'a	Yemen
	Basra				
	Mosul				
	Muscat	Oman			
	Doha	Qatar			
	Riyadh	KSA			
	Jeddah				
	Baghdad	Iraq	هاد الغرفة صغيرة بزاف .	Fes	Morocco
	Basra		هاد الغرفة صغيرة بزاف .	Rabat	
	Mosul	Iraq	لبيت هذي صغيرة .	Tunis	Tunisia
	Muscat		البيت هذه ياسر صغيرة .	Sfax	
	Doha	Qatar	هاذ الغرفة صغيرة بزاف .	Algeria	Algeria
	Riyadh	KSA	الدار صغيرة بكل .	Benghazi	Libya
	Jeddah		الدار هادي صغيره هلبه .	Tripoli	

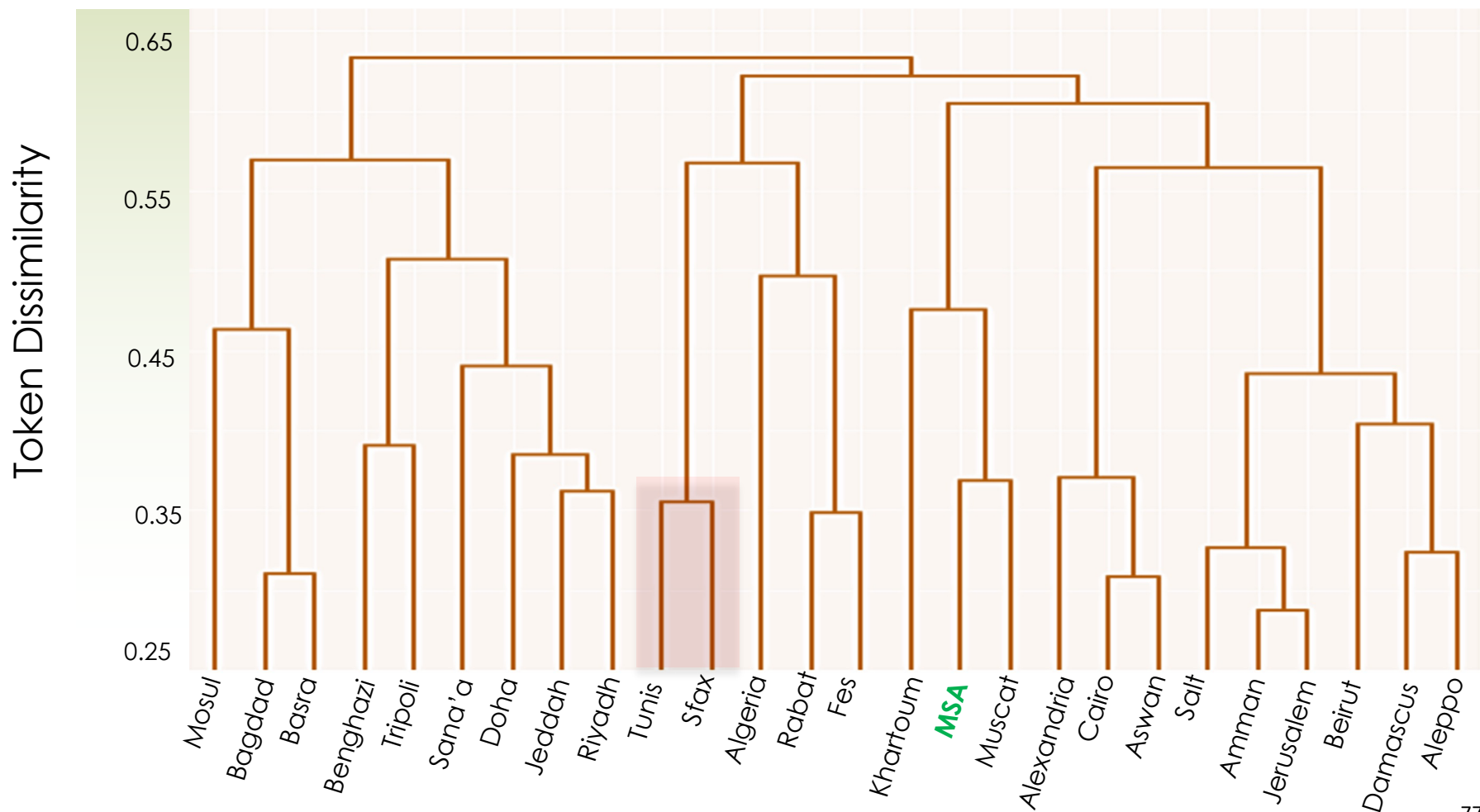
Dialectal Similarity

درجة تشابه اللهجات



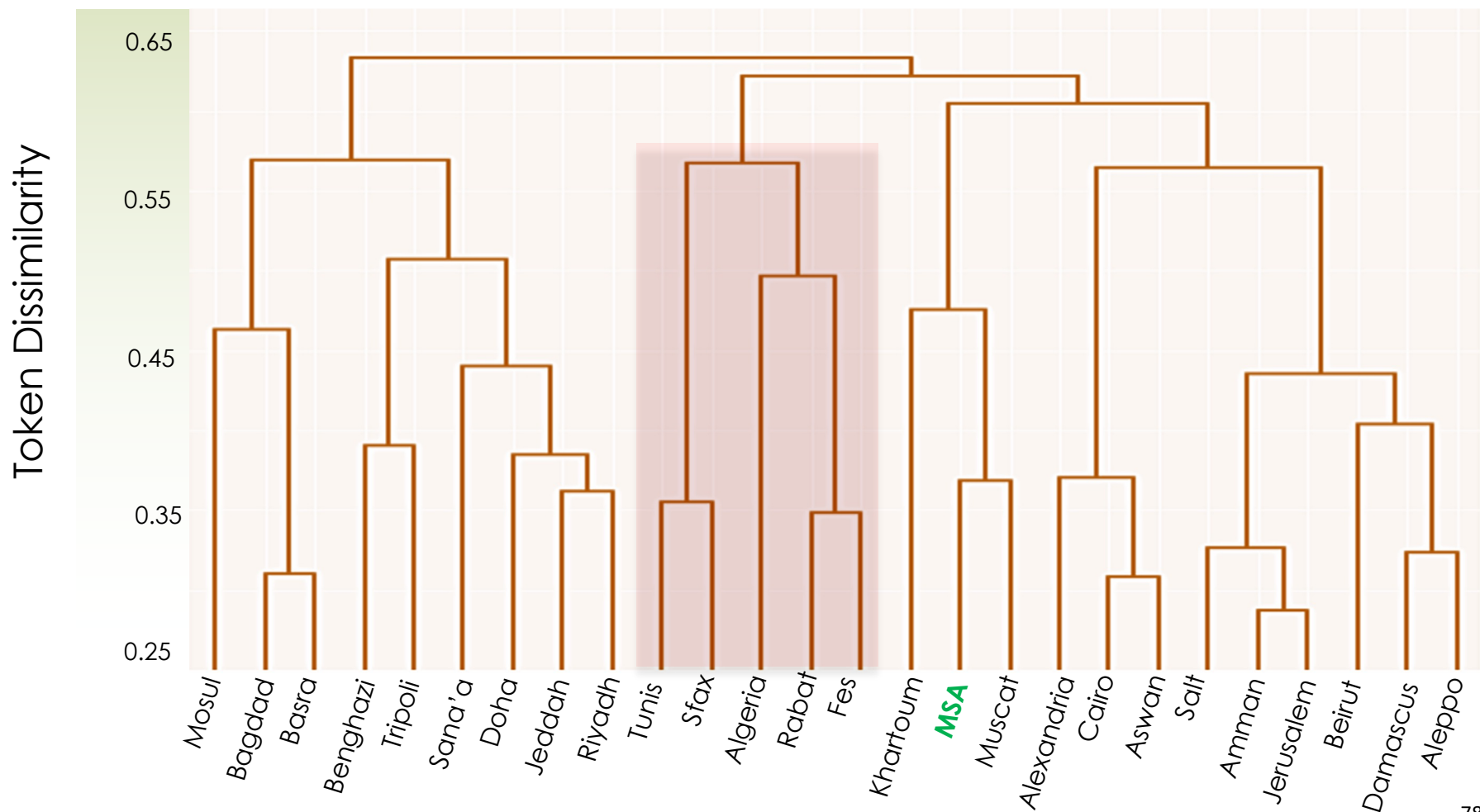
Dialectal Similarity

درجة تشابه اللهجات



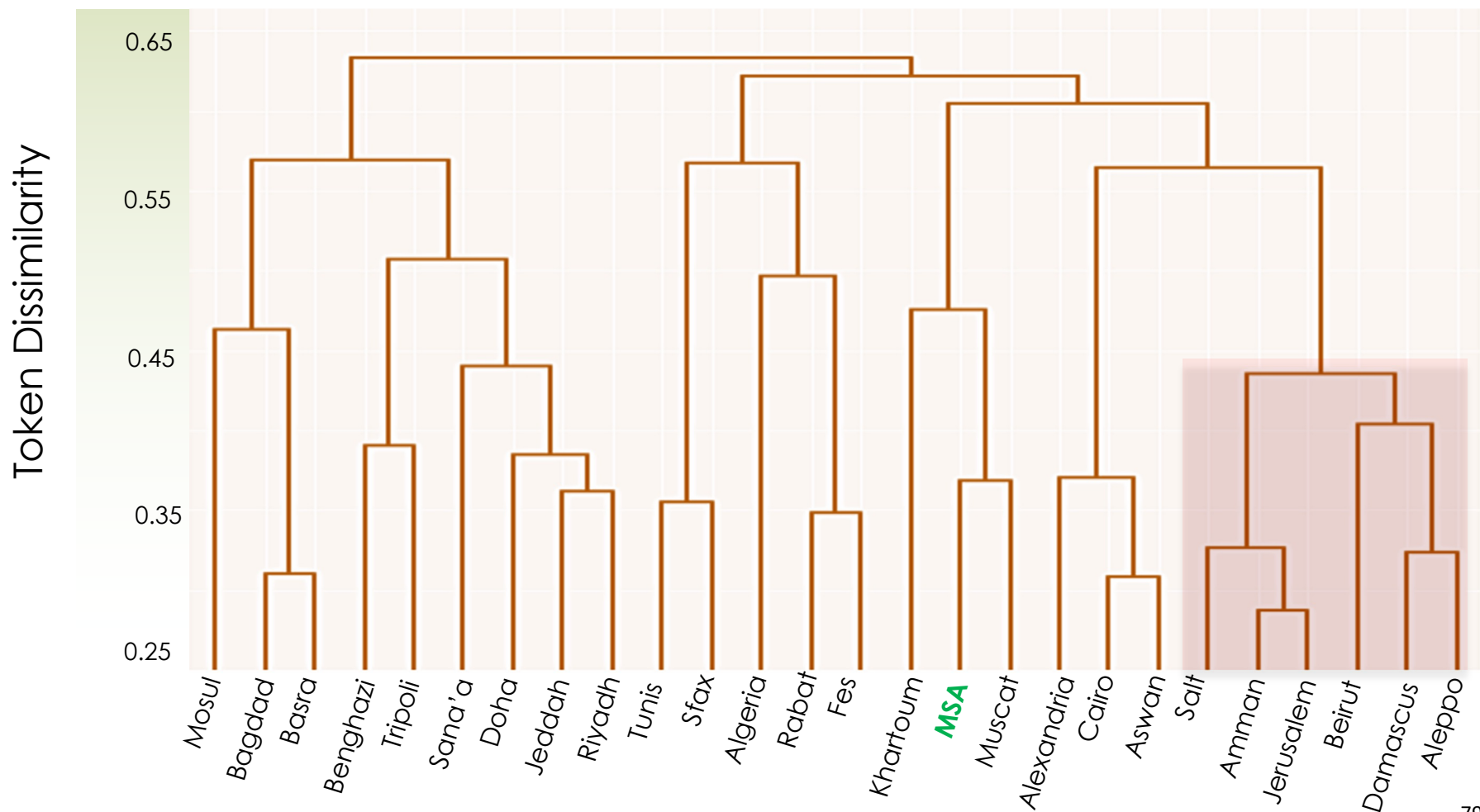
Dialectal Similarity

درجة تشابه اللهجات



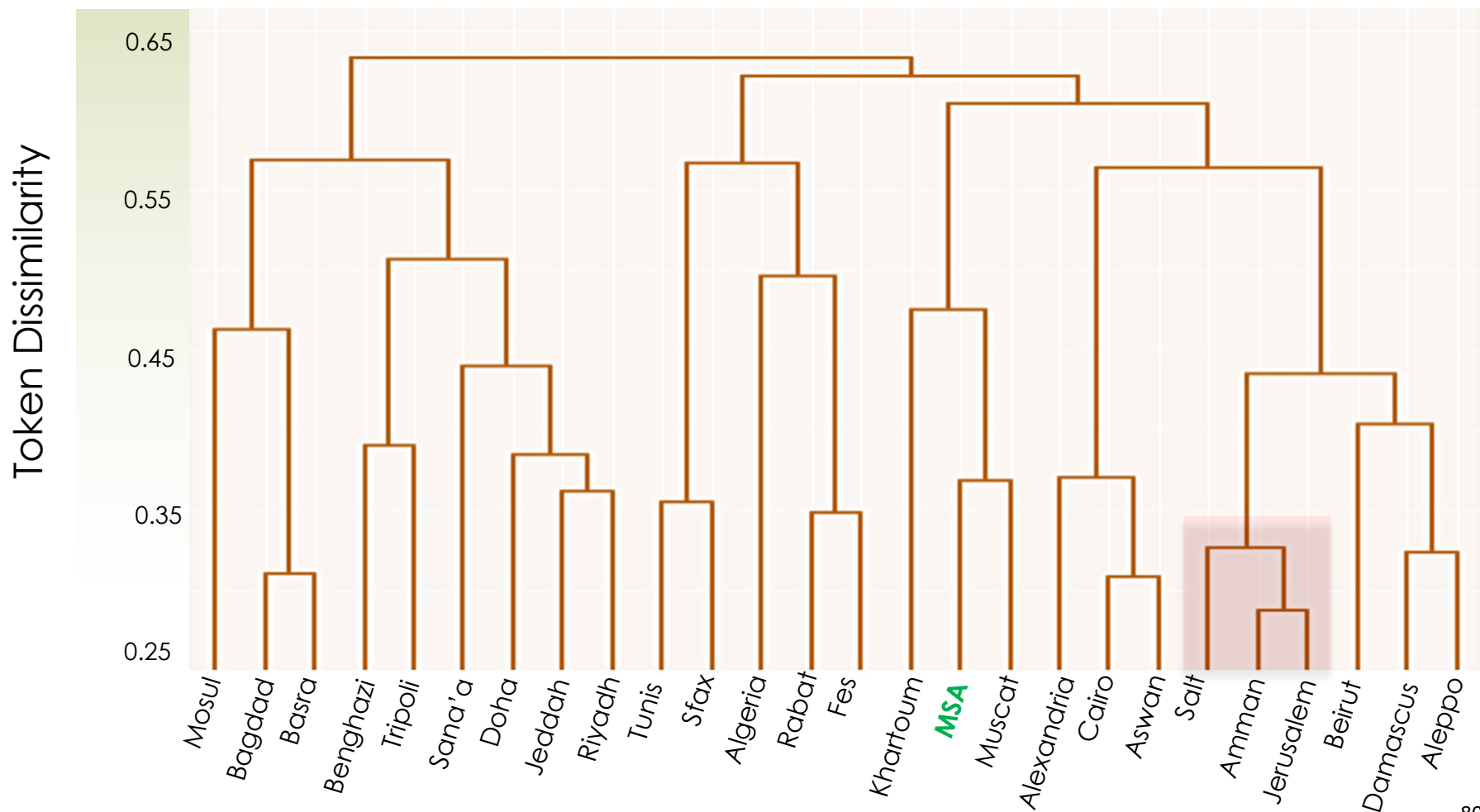
Dialectal Similarity

درجة تشابه اللهجات



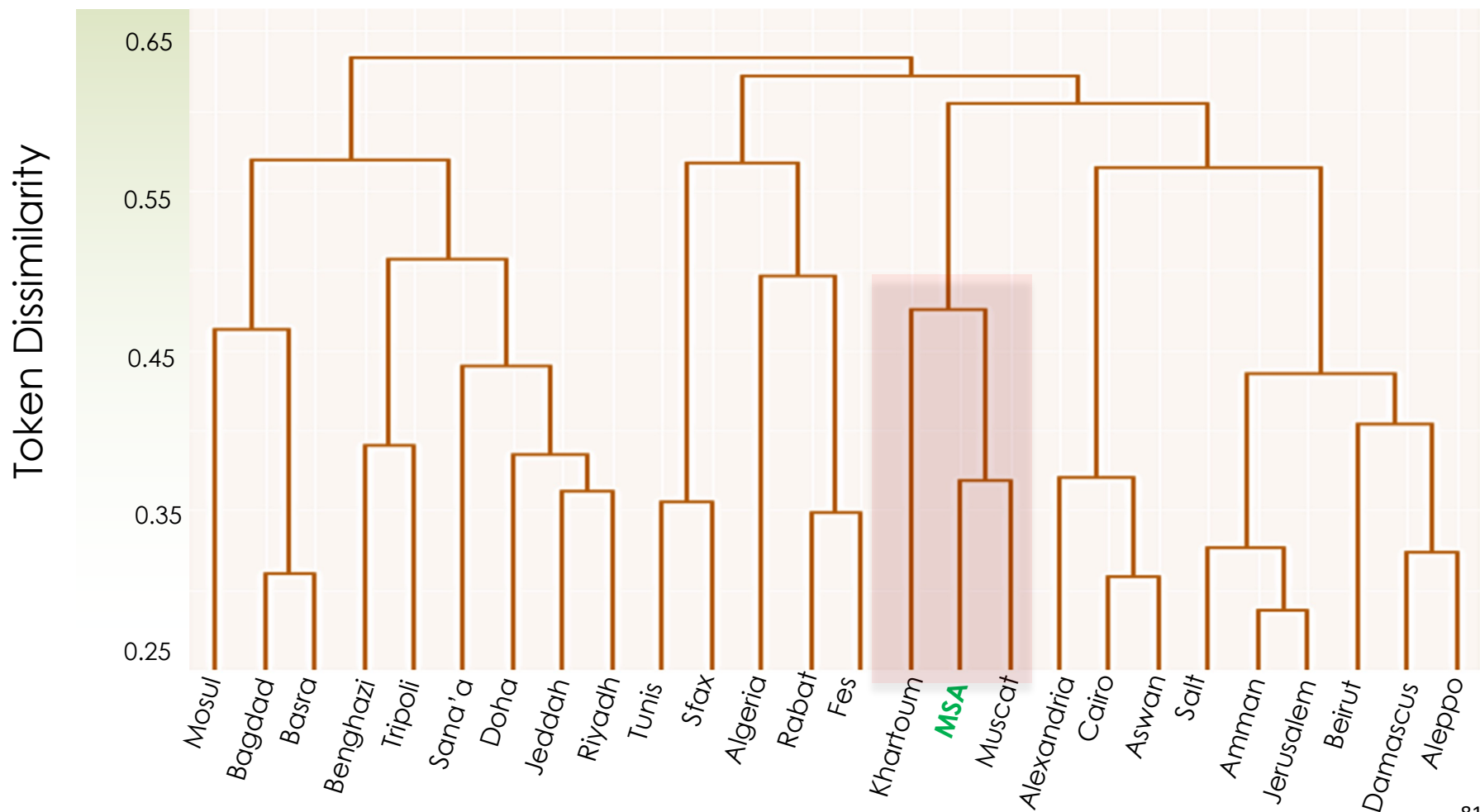
Dialectal Similarity

درجة تشابه اللهجات



Dialectal Similarity

درجة تشابه اللهجات



MADAR Project

مشروع «مدار»

بدي دوب قلي قلي بجنون بجنك انا مجنون ما بنسى حبك يوم

Identify

Output

Beirut: 96.0% Aleppo: 2.4% Damascus: 0.7% Salt: 0.5% Jerusalem: 0.4% Other: 0.1% MSA: 0.0 %



- ADIDA: Automatic Dialect Identification for Arabic
 - Classify sentences into one of 25 Arabic city dialects and MSA
 - 68% for 7-word sentences
 - 90% for 16-word sentences

• «عديدة»: نظام التعرف التلقائي على اللهجات

- يحدد لهجة الجملة المدخلة من بين ٢٥ مدينة عربية والفصحى
- الدقة: ٦٨٪ لجمل من ٧ كلمات
- الدقة: ٩٠٪ لجمل من ١٦ كلمة

<https://adida.abudhabi.nyu.edu/>

MADAR Project

مشروع «مدار»

CODA:	الامتحانات	قبل	الي	الفترة	صُحابي	ما شفتش	كودا:
	الإمتحانات	أبل	اللى	الفترة	صحابي	ما شفتش	
	المتحانات	ابل	إلي	الفطرة	صوحاي	مشفتش	
	الامتحنات	abl	إلى	الفطره	صوحابى	ما شوفتش	
	ilimti7anat	qbl	الي	ilftra	Su7abi	mashoftish	
	limtihanaat	qabl	illi		sohaby		

• CODA: Conventional Orthography for Dialectal Arabic

- Designed for all Arabic dialects
- Minimizes differences in writing
- Uses Arabic script

• «كودا»: كتابة واحدة للدارجة

- نظام مصمم لكل اللهجات سوية
- يقلل من الاختلافات في الكتابة
- يستخدم الخط العربي

Pre-trained and Large Language Models for Arabic

AraBERT

AraGPT



MARBERT

ARBERT



AraT5



CAMeLBERT



noor



... and many more!



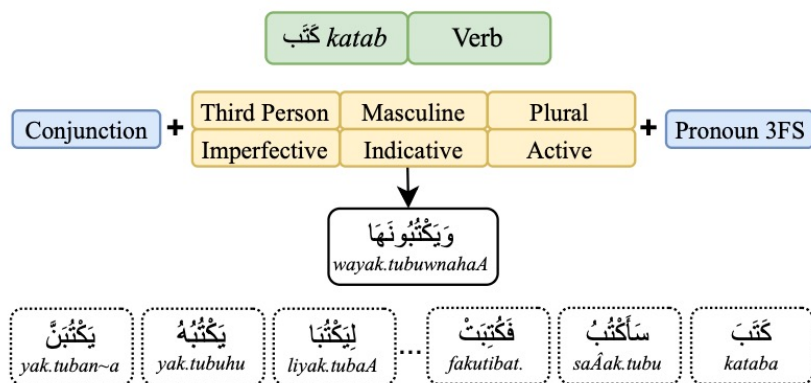
YallaMorph: A Benchmark for Evaluating Arabic Morphological Generation in Large Language Models

Mahmoud Reda,¹ Salam Khalifa,^{1,2} Reham Marzouk,³ Nizar Habash¹

Computational Approaches to Modeling Language (CAMEL) Lab

¹New York University Abu Dhabi,

²Stony Brook University, ³Mohamed bin Zayed University of Artificial Intelligence



EMNLP 2026

Model	10-Shot					
	Any Match Accuracy			F1 Score		
	Diac	Undiac	Norm	Diac	Undiac	Norm
GPT_{en}	51.9	67.0	67.7	54.4	70.5	71.3
Gemini_{en}	49.7	60.7	61.4	52.4	64.2	64.9
Fanar_{en}	21.5	36.5	37.2	22.5	38.3	39.1
Fanar_{ar}	15.8	26.8	27.2	16.6	28.1	28.5
Jais-70B_{en}	9.1	20.9	21.2	9.5	21.7	22.1
Jais-70B_{ar}	8.2	17.5	17.8	8.6	18.3	18.6
ALLaM_{en}	4.7	10.7	11.0	4.7	10.9	11.2
ALLaM_{ar}	4.5	10.0	10.3	4.1	9.3	9.5
Qwen_{en}	4.5	13.0	13.6	4.9	14.4	15.0
Jais-8B_{en}	3.4	10.1	10.2	3.3	9.9	10.0
Jais-8B_{ar}	2.3	8.2	8.5	2.1	7.6	7.9

Arabic Morphosyntactic Tagging and Dependency Parsing with Large Language Models

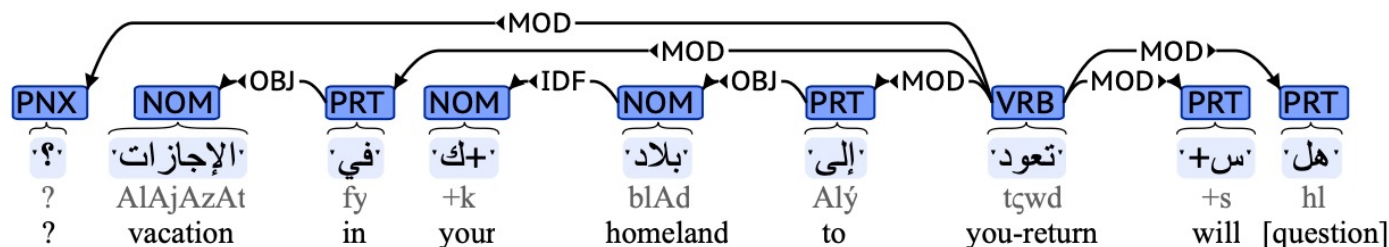
EMNLP 2026

Mohamed Adel¹ Bashar Alhafni² Nizar Habash¹

Computational Approaches to Modeling Language Lab

¹New York University Abu Dhabi

²Mohamed bin Zayed University of Artificial Intelligence



Model	Zero-shot				Best-ICL			
	Tok F ₁	LS	UAS	LAS	Tok F ₁	LS	UAS	LAS
Llama4	49.5	42.7	37.6	23.8	87.8	78.6	71.9	66.0
Qwen3	51.2	49.1	37.4	24.1	84.9	80.6	74.1	68.8
GPT5.2	93.7	70.5	69.7	60.1	98.2	90.9	87.4	84.1
Gemini3	92.8	78.1	75.0	68.4	96.2	91.9	88.1	85.3
CamelParser2.0	98.0	91.5	87.3	84.8	–	–	–	–
CamelParser2.1	98.0	92.5	88.5	86.0	–	–	–	–

ACL Findings 2024



CIDAR: Culturally Relevant Instruction Dataset For Arabic

Zaid Alyafeai ^{1,*} **Khalid Almubarak** ^{2,*} **Ahmed Ashraf** ^{3,*} **Deema Alnuhait** ^{4,*}
Saied Alshahrani ^{5,6} **Gubran A. Q. Abdulrahman** ¹ **Gamil Ahmed** ^{1,7}
Qais Gawah ¹ **Zead Saleh** ¹ **Mustafa Ghaleb** ^{1,8}
Yousef Ali ¹ **Maged S. Al-Shaibani** ¹

¹ King Fahd University of Petroleum and Minerals (KFUPM) ² Prince Sattam bin Abdulaziz University (PSAU)

³ ARBML ⁴ University of Illinois Urbana-Champaign ⁵ Clarkson University ⁶ University of Bisha

⁷ Interdisciplinary Research Center for Smart Mobility and Logistics (IRC-SML), KFUPM

⁸ Interdisciplinary Research Center for Intelligent Secure Systems (IRC-ISS), KFUPM

ACL 2025 – best resource paper



Palm:
A Culturally Inclusive and Linguistically Diverse Dataset for Arabic LLMs

Muhammad Abdul-Mageed
Canada Research Chair in NLP and ML
The University of British Columbia
muhammad.mageed@ubc.ca



ArabicNLP 2025

BALSAM: A Platform for Benchmarking Arabic Large Language Models

Rawan Al-Matham, Kareem Darwish, Raghad Al-Rasheed, Waad Alshammari, Muneera Alhoshan, Amal Almazrui, Asma Al Wazrah, Mais Alheraki, Firoj Alam, Preslav Nakov, Norah Alzahrani, Eman alBilali, Nizar Habash, Abdelrahman El-Sheikh, Muhammad Elmallah, Haonan Li, Hamdy Mubarak, Mohamed Anwar, Zaid Alyafeai, Ahmed Abdelali, Nora Altwairesh, Maram Hasanain, Abdulmohsen Al Thubaity, Shady Shehata, Bashar Alhafni, Injy Hamed, Go Inoue, Khalid Elmadani, Ossama Obeid, Fatima Haouari, Tamer Elsayed, Emad Alghamdi, Khalid Almubarak, Saied Alshahrani, Ola Aljarrah, Safa Alajlan, Areej Alshaqarawi, Maryam Alshihri, Sultana Alghurabi, Atikah Alzeghayer, Afrah Altamimi, Abdullah Alfaifi, Abdulrahman AlOsaimy

EMNLP 2026



Bulbul: A Dataset for Dialectal Arabic Speech Recognition

Ahmed Ashraf¹, Aisha Alansari¹, Fadel Al Abbas¹, Nada Almarwani², Samah Aloufi², Saad Ezzini¹
Maged S. Al-Shaibani¹, Doaa Dalaq¹, AbdelRahim A. Elmadany³, Muhammad Abdul-Mageed³
Mohamed Mehdi Trigui¹, Dania Refai¹, Layan Refai⁴, Mohamed Akrou¹, Mustafa Jarrar^{5,6}
Wasfi G. Al-Khatib¹, Alaa Dalaq¹, Darin El-Nakla¹, Samir Abdaljalil⁷, Abdulrahman Al-Fakih¹
Nour El Imane Zeghib¹, Moussa REDAH¹, Salmane Chafik⁸, Mohamed El-Attar⁹, Rima Grati⁹
Sarah Kohail⁹, Malak Alkhorasani¹⁰, Khadijah Al Safwan¹, Ismail M. Mudhaffar¹, Ali Altam¹¹
Ahmed Al-Shaikh¹, Adnan Saeed¹², Hamzah Luqman¹

¹KFUPM, ²Taibah University, ³UBC, ⁴PSUT, ⁵Birzeit University, ⁶HBKU
⁷Texas A&M University, ⁸Mohammed VI Polytechnic University, ⁹Zayed University, ¹⁰IAU
¹¹Symbiosis International University, ¹²Taiz University
{g202411740,aisha.ansari,hluqman}@kfupm.edu.sa



- **SIGARAB** = ACL Special Interest Group in Arabic Language Processing.
- We celebrated our four-year anniversary earlier this year (January 20, 2026).
- We have **1,100+ members**.
- We have an active mailing list that has been enriching to the community
- Sponsoring a conference, three workshops and a yearly NLP School.
- Check out the website: www.sigarab.org >>>>>



The Second Arabic NLP School

Co-located with EACL 2026

Rabat, Morocco

March 24, 2026

Organized by SIGARAB



✳️ AI-generated content

The Fourth Arabic Natural Language Processing Conference (ArabicNLP 2026)

<https://arabnlp2026.sigarab.org>

Co-located with EMNLP 2026, Budapest, Hungary, October 24-29, 2026

- With EMNLP 2026
- October 28-29, 2026
- 11 Shared Tasks
- Acceptance Rate 22.6%



General Chair



Muhammad
Abdul-Mageed

UBC

Program Chairs



Houda
Bouamor

CMU Qatar



Samhaa El-
Beltagy

Nawy



Hend Al-
Khalifa
King Saud
University

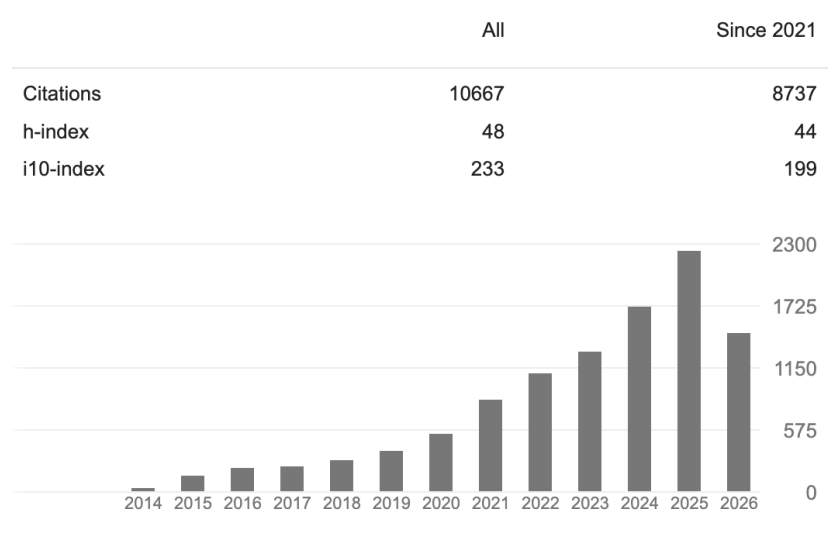


Tamer
ElSayed
Qatar
University

ArabicNLP is #17 in Computational Linguistics

ArabicNLP is #17 among the top publication venues in “Computational Linguistics” according to Google’s 2025 Scholar Metrics.

The ranking is calculated based on the citation counts of the papers published within the last 5 years, (2020-2025).



<http://scholar.arabic-nlp.net/>

Summary

- **Arabic poses important challenges to NLP/CL/AI**
 - Complex orthography and morphology, widespread variations
 - Limited research and resources
- **Arabic CL is a large and growing field**
 - Great improvements in quality
 - More problems to solve
- **CAMeL Lab is one of many ongoing efforts**
 - Focus on Arabic NLP
 - Many resources and tools

ملخص

- **اللغة العربية تطرح تحديات مهمة للذكاء الاصطناعي**
 - تحديات إملائية وصرفية واختلافات لهجوية واسعة النطاق
 - الفقر في البحث والموارد
- **اللسانيات الحاسوبية العربية مجال كبير ومتنامي**
 - شهد تحسناً عظيماً في الجودة
 - وفتح الأبواب لمشاكل جديدة
- **مختبر كامل واحد من الجهود العديدة**
 - التركيز على اللغة العربية
 - العديد من الموارد والأدوات

Summary

ملخص

- **The future is bright!**

- More work is being done on Arabic Artificial Intelligence!
- We need more people who can help!

- **If you are interested in Arabic NLP, join SIGARAB**



<http://www.sigarab.org/>

- **لكن المستقبل مشرق!**

- الأبحاث والتطبيقات عن اللغة العربية والذكاء الاصطناعي في تزايد!
- نحن بحاجة إلى المزيد من الأشخاص الذين يمكنهم المساعدة!

- **للمهتمين بالتحليل الآلي للغة العربية، ننصح بالانضمام إلى SIGARAB**

The background image shows a modern university building with a courtyard. The building has multiple stories with large windows and balconies. The courtyard is paved and has some trees and plants. The sky is cloudy and colorful, suggesting a sunset or sunrise. The text "شكراً على المتابعة ! Thank You" is overlaid on the image.

شكراً على المتابعة ! Thank You

www.camel-lab.com
nizar.habash@nyu.edu