Arabic script

36=28 consonants,6-(ؤأإآءئ), Ta ة, A̱lif ى	NLP Task - Orthographic Transliteration:
19 letters shape & ك والهمزة والنقاط	Backwlater Transliteration model
Lettershape: Initial, medial, final and isolate	NLP Task - Orthographic Normalization:
Cursive connected style	encoding cleanup->complex ligature, ك
Ligtures	Tatweel and diacritics remove
Diacritic (Vowel, Nunation, Shadda, Dagger)	Letter normalization: l <--أإآ
Digits (Westren and eastren) - LTR	Hand writing - MADCAT
Punctuation [: . !"] [؟،] Tatweel	Automatic Diacritization - Only 1.5%
Typography, growing font library	Other Languages Letter: Persian, Kurdish
Encoding: Unicode, ISO-8859, CP-1256	

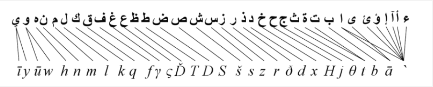
Challenges to Arabic NLP

- Orthographic ambiguity
- Orthographic inconsistency
- Morphological Complexity
- Dialect Variation
- Annotated resource poverty

Arabic Phonolgy

Phonem:	Minimal pair: /k//g/ & قلب كلب
MSA - ق /q/	/q/, /k/, /ʔ/, /g/, /ʤ/, /ɢ/
ث /θ/	/θ/, /t/, /s/
ذ /δ/	/δ/, /d/, /z/
ج /ʤ/	/ʤ/, /g/, /ʒ/

Orthography Connected to Phoneme


ي و ه ن م ل ك ق ف غ ع ظ ط ض ص ش س ز ر ذ د خ ح ج ث ة ت ب ا ى ئ ؤ إ آ أ ء
ī y ū w h n m l k q f γ ç Ď T D S š s z r ð d x H j θ t b ā `

Orthography - Ambiguity

Optional Diacritization	Complex = No vowels? long vowels, initial
Arabic words has on average:	12.3 Analysis & 6.8 Diacritics & 2.7 lama
Morpho-Phonemic Spelling issue 1:	الشمسية والقمرية || ة -> ه || عصا -> عصى
Morpho-Phonemic Spelling issue 2:	التنوين ينطق نون - ألف واو الجماعة لاينطق
Standardization Issue	سوريا وسورية || فيلم وفلم || أفريقية وأفريقيا
NLP Task:	Proper Name Transliteration
Qaddafi problem (Kadafi, Qadafi.....)	Schwarzenegger Problem (, شوارزينجر شوارزينغر)
Hassan Problem	Marie problem ---> Los it to ماري

Arabic Spelling

Hamzated Alif and Alif maqsura 11%	Penn Arabic bank tree
(30%) of words have errors/out of 2M words	Qatar Arabic Language Bank
Arabic spelling errors are a big challenge	GIGO: Garbage In Garbage Out
Inconsistencies in Dialectal Arabic !=standard	مابيقولهاش
و+ ب+أدلة+هاَ	و+بَادَلت+هاَ
العَيـن = (eye, water spring, Alain city)	Spelling variants

Arabic Morphology

Morphological Complexity	A core word has many inflected forms
Gender(2), Number(3), Person(3), Aspect(3)	Tense particle (2), Mood(3), Voice(2),
Pronominal clitic-(12), Conjunction clitic(3)	وسنقولها =/wasanaqūluhā/= و+ س+ ن+ قول + ها
go went going gone go goes	VB VBD VBG VBN VBP VBZ
Arabic POS tags: 22,400 tags	English POS tags: 48 tags
12.3 analyses and 2.7 lemmas per word	Functional Morphology جمع تكسير
Form based morphology التصريف الطبيعي	علم العروض. الفراهيدي - الكتابة العروضية
الكتاب الرمزية - 00//0/	التنوين, الشدة, الاشباع والألف :هندن أخلل يحرمي هاذا

Tools and Papers

MADAMIRA	State-of-the-art Arabic and Arabic Dialect processing
MADAR	Multi-Arabic Dialect Applications and resources
SAMER	Simplification of Arabic Masterpieces for Extensive
CODA	A conventional orthography for Dialectal Arabic
tashaphyne	Arabic lite Stemmer

POS For Arabic

Stanford Arabic parser tagset	Morphological annotation Quranic Arabic corpus
Arabic MADA system tagset	POS tag set for Modern Standard Arabic

POS For Arabic (cont)

The APT tagger (Khoja, 2001) / hybrid	The Qutuf (Altabba et al., 2010) tagger
Al-Dahdah (1989),	TreeTagger by Schmid (1995)) - 20+ lang

Noun (اِسْم)<Aisom>, Verb (فِعْل)<fiEol>, and Particle (حَرْف)<Harof>.

Each one of these categories has many subcategories.