Introduction to Arabic Natural Language Processing (part 1)

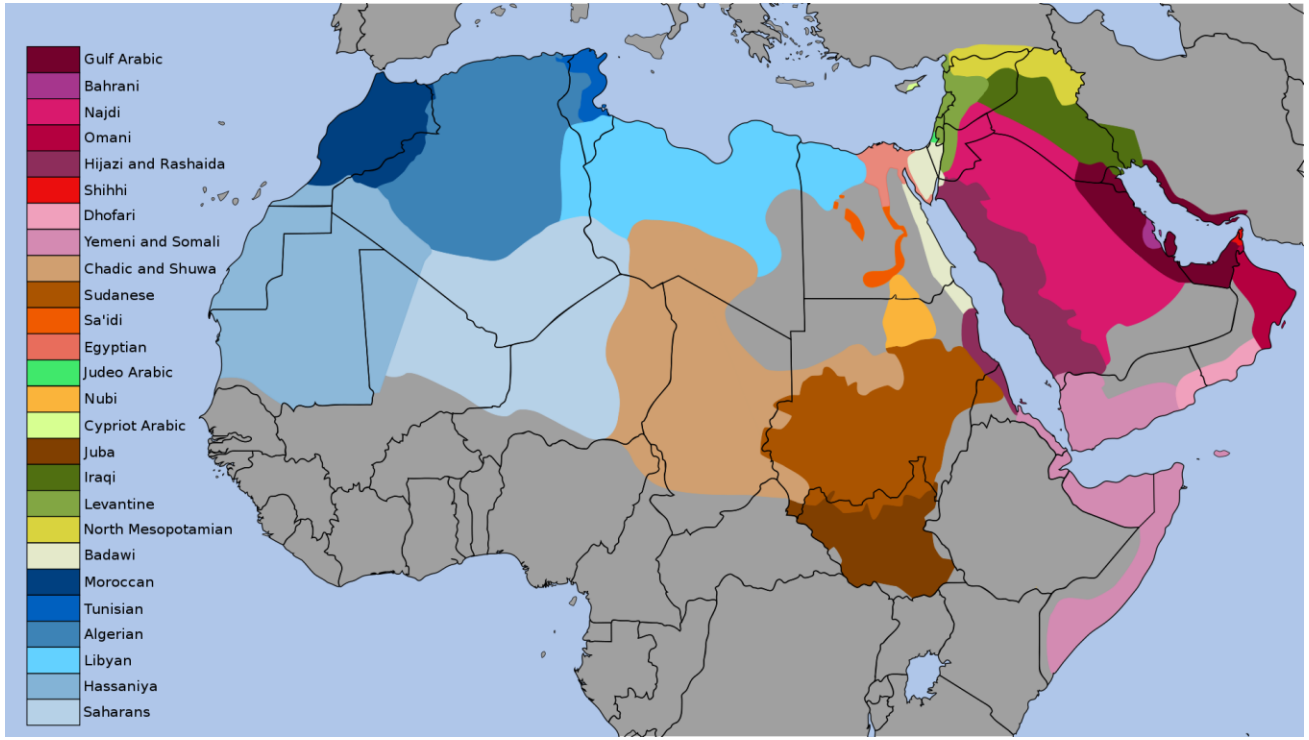
Kareem Darwish

aixplain Inc.

مقدمة في معالجة اللغة العربية آليا

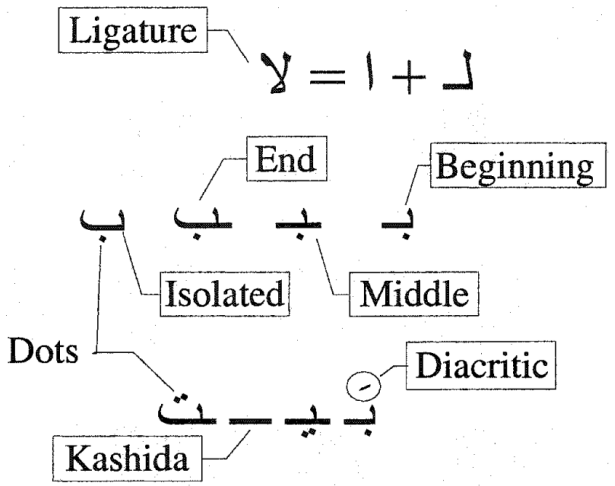
What is Arabic?

- Arabic is a Semitic language that is being spoken by more than 450 million people.
- Arabic was the language of science and culture for hundreds of years.
- Arabic script is being used by many languages (ex. Urud, Farsi, Kurdi, Hausa, etc.), and used to the script for many more languages (ex. Turkish, Russian, etc.)
- Arabic has fundamental influence on many world languages (ex. Urdu, Maltese, Hebrew, Turkish, Farsi, Somali, Mahri, etc.)
- Arabic is a family of languages (ex. Arabic, Maltese, Mahri, Suqutri, Shehri, etc.) and dialects (ex. Maghrebi, Hassaniya, etc.).



How is Arabic Different – Orthography 1

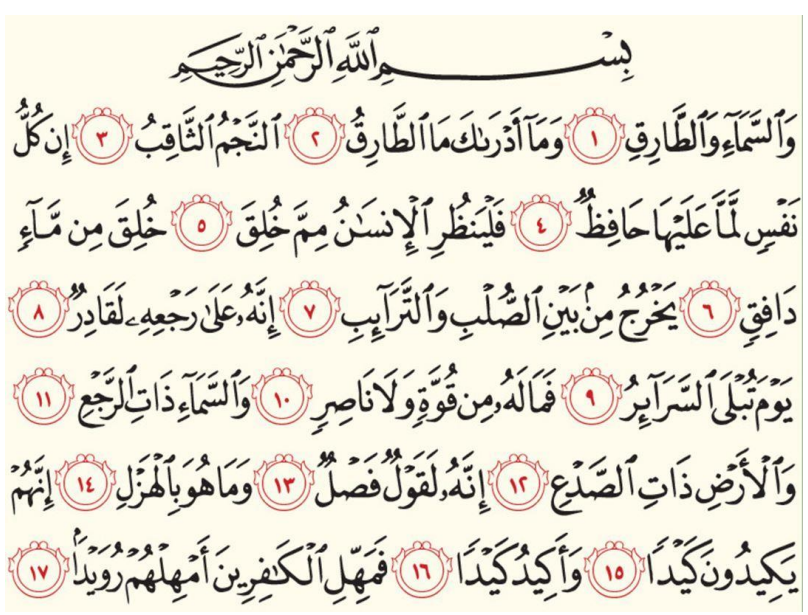
- Orthography refers to how a language is written
- Arabic:
  - Is written from right-to-left
  - Uses connected characters that change shape depending on position
  - Is typically written without diacritics – diacritics are in fact part of words (ex. رَجُل، رِجْل، رَجْل)

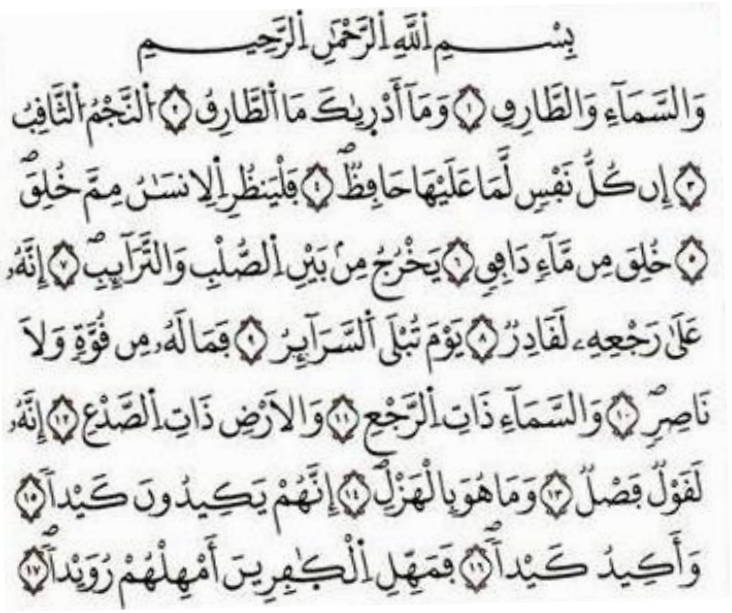


قَالَ بَلَىٰ وَلَٰكِن لِّيَطْمَئِنَّ قَلْبِي قَالَ فَخُذْ أَرْبَعَةً مِّنَ ٱلطَّيْرِ فَصُرْهُنَّ إِلَيْكَ ثُمَّ ٱجْعَلْ عَلَىٰ كُلِّ جَبَلٍ
مِّنْهُنَّ جُزْءًا ثُمَّ ٱدْعُهُنَّ يَأْتِينَكَ سَعْيًا ۚ وَٱعْلَمْ أَنَّ ٱللَّهَ عَزِيزٌ حَكِيمٌ

How is Arabic Different – Orthography 2

- Arabic:
  - Has multiple scripts:
    - Othmani script
    - Maghrebi script
    - Arabizi
  - Has multiple encoding:
    - Buckwalter
    - UTF-8
    - Windows CP1256





How is Arabic Different – Orthography 2

- Arabic:
  - Has different writing conventions:
    - Ex. Terminal: ي (ex. علي) vs ى (ex. على)
    - Ex. Default diacritic: قَال vs قال

How is Arabic Different – Orthography 3

Arabic letters	ا	ب	ت	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	ى[4]
DIN 31635	ʾ / *ā*	b	t	*ṯ*	*ǧ*	*ḥ*	*ḫ*	d	*ḏ*	r	z	s	*š*	*ṣ*	*ḍ*	*ṭ*	*ẓ*	ʿ	*ġ*	f	q	k	l	m	n	h	*w* / *ū*	*y*	*ī*
Buckwalter	A	b	t	v	j	H	x	d	*	r	z	s	$	S	D	T	Z	E	g	f	q	k	l	m	n	h	w	y	Y
Qalam	' / aa			th			kh		dh				sh					`	gh									y	
BATR	A / aa			c			K		z'				x					E	g								w / uu	y	ii
IPA (MSA)	ʔ, aː	b	t	θ	dʒ g ʒ	ħ	x	d	ð	r	z	s	ʃ	sˤ	dˤ	tˤ	ðˤ zˤ	ʕ	ɣ	f	q	k	l	m	n	h	w, uː	j, iː	

- **Arabic text**
  - يُولَدُ جَمِيعُ ٱلنَّاسِ أَحْرَارًا مُتَسَاوِينَ فِي ٱلْكَرَامَةِ وَٱلْحُقُوقِ.

- **Buckwalter transliteration**
  - yuwladu jamiyEu {ln~aAsi >aHoraArFA mutasaAwiyna fiy {lokaraAmapi wa{loHuquwqi.

hamza
- lone hamza: '
- hamza on alif: >
- hamza below alif: <
- hamza on wa: &
- hamza on ya: }

alif
- madda on alif: |
- alif al-wasla: {
- dagger alif: `
- alif maqsura: Y

harakat
- fatha: a
- damma: u
- kasra: i
- fathatayn: F
- dammatayn: N
- kasratayn K
- shadda: ~
- sukun: o

ta marbouta: p

How is Arabic Different – Morphology 1

- Morphology refers to how a word is constructed
- Arabic has a root-based morphological system:
  - Start with root (ex. كتب – ktb) – about 10,000 roots
  - Fit into *stem template*, ex.:
    - CCAC: كتاب – kt***A***b
    - CACC: كاتب – k***A***tb
    - mCCwb: مكتوب – ***m***kt***w***b
  - Add diacritics: كِتَاب – k***i***t***a***Ab, كُتَّاب – k***u***t~***a***Ab
  - Add prefixes and suffixes:
    - Ex. Prefixes: coordinating conjunctions (و، ف); determiners (ال); future (س); etc.
    - Ex. Suffixes: pronouns (هم، كم، هـ، ها، ك، كما، كن، هن); plural/dual markers (ين، ون، ان، ي، و، ا); etc.
    - Ex. وَكِتَابهُنَّ – w***a***k***i***t***a***Abh***u***n~***a***

* ك ت ب — (كَتَبَ) من بابِ نصَر و (كَتَابًا) أيضا و (كِتَابةً) . و (الكِتابُ) أيضا الفَرْضُ والحُكْمُ والقَدَرُ. و (الكاتِبُ) عنـدَ العربِ العَالِمُ ومنـه قولُهُ تعـالى : « أَمْ عِنْـدَهُمُ الغَيْبُ فَهُمْ يَكْتُبُونَ » و (الكُتَّاب) بالضمِّ والتشديدِ (الكَتَبَةُ) . و (الكُتَّابُ) أيضا و (المَكْتَبُ) (٢) واحدٌ والجَمْـعُ (الكَتَاتِيبُ) و (المَكَاتِبُ) . و (الكَتِيبَةُ) الجَيْشُ . و (اكْتَتَبَ) أي

من مختار الصحاح لمحمد بن أبي بكر الرازي

How is Arabic Different – Morphology 2

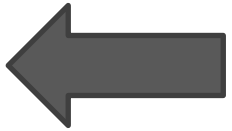
- Arabic morphology is ambiguous:
- Ex. وليد:
  - وَلِيد – w***a***l***i***yd
  - وَلِيَد – w***a***+l***i***+y***a***d

- Ex. ايمان:
  - إيمَان -- <***i***ym***a***An – أ م ن
  - أَيْمَان -- >***a***y***o***m***a***An – ي م ن
  - أَيِّمَان -- >***a***y~***i***m***a***+An – أ ي ن
  - أَيَمْأَن -- >***a***+y***a***m***o***>***a***n – م أ ن

How is Arabic Different – Morphology 3

- One of the fundamental problems in Arabic is word segmentation:

ويكيبيديا هي موسوعة رقمية، متعددة اللغات، حرة المحتوى. يستطيع أي شخص التحرير فيها بدون تسجيل، ويستطيع أي شخص الاستفادة من المحتوى، واستغلاله بهدف تجاري أو غيره وفقًا لترخيص الموسوعة.



ويكيبيديا هي موسوع+ة رقمي+ة ، متعدد+ة ال+لغ+ات ، حر+ة ال+محتوى . يستطيع أي شخص ال+تحرير في+ها ب+دون تسجيل ، و+يستطيع أي شخص ال+استفاد+ة من ال+محتوى ، و+استغلال+ه ب+هدف تجاري أو غير+ه وفق+ا ل+ترخيص ال+موسوع+ة .

How is Arabic Different – Morphology 3

- Segmentation is important for many downstream applications such as:
  - Information retrieval (search)
    - كتاب ➔ والكتاب، الكتاب، كتابهم، وكتابهن
  - Lemmatization (الرجوع لأصل الكلمة):
    - وليد ➔ وليد; يد
    - الرجال ➔ رجل
  - Syntactic processing:
    - وليد ➔ وليد: noun; و+ل+يد: coordinating_conjunction + preposition + noun
  - Diacritic recovery:
    - وليد ➔ وَلِيد، وَ+لِ+يَد

ويكيبيديا هي موسوع+ة رقمي+ة ، متعدد+ة ال+لغ+ات ، حر+ة ال+محتوى . يستطيع أي شخص ال+تحرير في+ها ب+دون تسجيل ، و+يستطيع أي شخص ال+استفاد+ة من ال+محتوى ، و+استغلال+ه ب+هدف تجاري أو غير+ه وفق+ا ل+ترخيص ال+موسوع+ة .

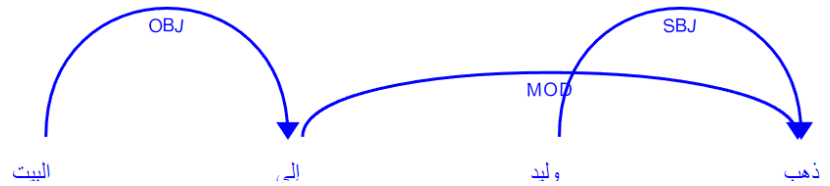
ويكيبيديا هي موسوعة رقمي ، متعدد لغة ، حر محتوى . استطاع أي شخص تحرير في دون تسجيل ، استطاع أي شخص استفادة من محتوى ، استغلال هدف تجاري أو غير وفق ترخيص موسوعة .

NOUN PRON NOUN+NSUFF ADJ+NSUFF PUNC ADJ+NSUFF
DET+NOUN+NSUFF PUNC ADJ+NSUFF DET+NOUN PUNC V NOUN NOUN
DET+NOUN PREP PRON PREP NOUN NOUN PUNC CONJ V NOUN NOUN
DET+NOUN+NSUFF PREP DET+NOUN PUNC CONJ NOUN PRON PREP
NOUN ADJ CONJ NOUN PRON NOUN CASE PREP NOUN
DET+NOUN+NSUFF PUNC

وِيكِيبِيدْيَا هِيَ مَوْسوعَةٌ رَقْمِيَّةٌ ، مُتَعَدِّدَةُ اللُّغَاتِ ، حُرَّةُ المُحْتَوَى . يَسْتَطِيعُ أَيُّ شَخْصٍ التَّحْرِيرِ فِيهَا بِدُونِ تَسْجيلٍ ، وَيَسْتَطِيعُ أَيَّ شَخْصٍ الِاسْتِفادَةَ مِنْ المُحْتَوَى ، واسْتِغْلالِهِ بِهَدَفٍ تِجارِيٍّ أَوْ غَيْرِهِ وَفْقًا لِتَرْخيصِ المَوْسوعَةِ .

How is Arabic Different – Morphology 4

- Segmentation is important for many downstream applications such as:
  - Parsing (إعراب):
    - ذهب وليد إلى البيت



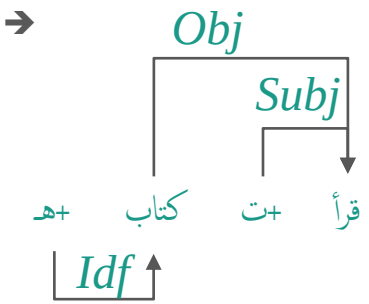
Dependency Relation	HEAD ID	FORM	ID
---	0	ذهب	1
SBJ	1	وليد	2
MOD	1	إلى	3
OBJ	3	البيت	4

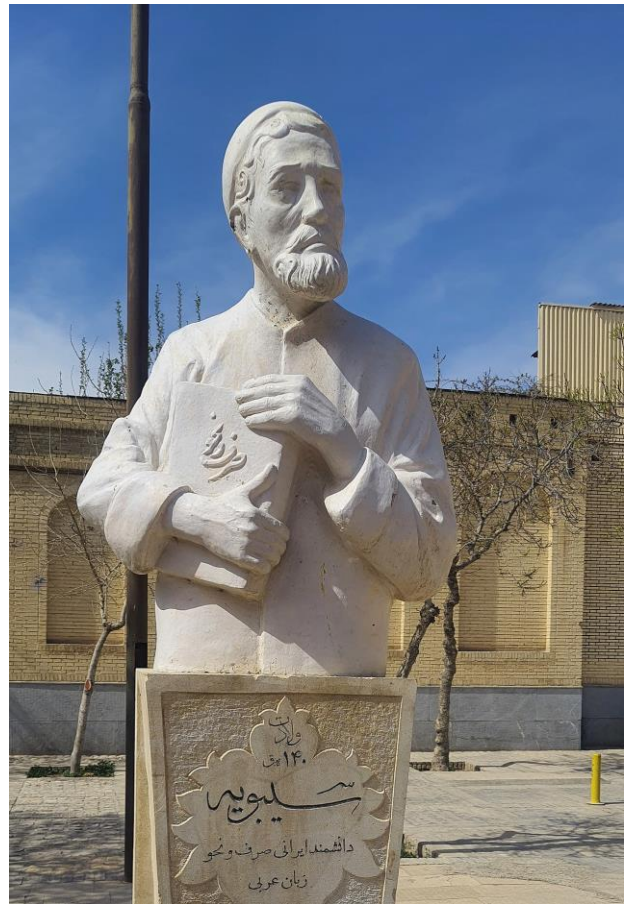
How is Arabic Different – Syntax 1

- Syntax refers to how a sentence is constructed
- Arabic is generally free form:
  - Verb-Subject-Object (VSO): أكل الأسد الأرنب
  - Subject-Verb-Object (SVO): الأسد أكل الأرنب
  - Verb-Object-Subject (VOS): أكل الأرنب الأسد
  - Subject-Predicate (SP - جملة أسمية): الأسد مفترس
- Two central problems are associated with syntax, namely:
  - Part of speech (POS) tagging:
    - Ex. قرأت كتابه ➔

Pronoun	هـ+	Noun	كتاب	Pronoun	ت+	Noun	قرأ

  - Parsing:
    - Ex. قرأت كتابه ➔





Diacritization Convention

Diacrtization is NOT Always Easy

فكاهة: من نظر إلى أخيه نظرة ...

مثلث قطرب!

خلا أخ لك فلا تخله ... من لك يوما بأخيك كله!

إن الدنيا أقبلت بصرم وأدبرت حذاء، ولم يبق منها إلا كصبابة الإناء يشربها أحدكم

Diacrtization is NOT Always Easy

بدا فحيا بالسَلام ~ رمى عذولي بالسِلامْ

أشار نحوي بالسُلام ~ بكفه المخضَبِ

بالفتح لفظُ المبتدي ~ والكسرِ صخرُ الجَلمَدي

والضمِ عِرقٌ في اليدي ~ قد جاء في قول النبي

Diacrtization is NOT Always Deterministic

سَاعَدَ الرَّجُلُ الْوَلَدَ vs سَاعَدَ الرَّجَلَ الوَلَدُ

أَطْعَمْتُ الْهِرَّةَ vs أُطْعِمَتُ الْهِرَّةُ vs أَطْعَمَتْ الْهِرَّةَ

Which form is correct?

بَرَزَ الثَّعْلَبُ يَوْمًا فِي ثِيَابِ الْعَارِفِينَ ... وَمَشَى فِي الْأَرْضِ يَهْذِي وَيَسُبُّ الْمَاكِرِينَ

بَرَزَ الثَّعْلَبُ يَوْمًا فِي ثِيَابِ الْعارِفِينَ ... وَمَشى في الْأَرْضِ يَهْذي وَيَسُبُّ الْماكِرِينَ

قَالَتْ: دَعِينِي وَهُزَالِي وَالزَّمَنْ ... وَأَجِيبِي صَاحِبَ السِّكِينِ يَا ذَاتَ الثَّمَنْ

قَالَتْ: دَعِينِي وَهُزَالِي وَالزَّمَنِ ... وَأَجِيبِي صَاحِبَ السِّكِينِ يَا ذَاتَ الثَّمَنِ

Which form is correct?

ثَوَى طَاهِرُ الأَرْدَانِ . لَمْ تَبْقَ بُقْعَةٌ ... غَدَاةَ ثَوَى إِلاَّ اشْتَهَتْ أَنَّهَا قَبْرُ

ثَوى طاهِرُ الأردانِ . لَم تَبقَ بُقعَةٌ ... غَداةَ ثَوى إِلاَّ اشْتَهَتْ أَنَّها قَبرُ

اشْتَقْتُ لِرُؤْيَتِهِ

اِشْتَقْتُ لِرُؤْيَتِهْ

Conventions R Important 4 Consistency

Should we explicitly put default diacritics? Ex. قالَ vs قَالَ

Should we put explicit *sukun* on letters? Ex. مِثلُ vs. مِثْلُ

Should we put diacritics on different forms of *hamza*? أَنْتَ vs. أَنْتَ

Should we put diacritics on *hamzat-ul-wasl*? اضْرِبْ vs. اِضْرِبْ

Should we put *sukun* on *lam* in *Al Alqamariyya*? القَرْيَةُ vs الْقَرْيَةُ

Formulating a Convention for Gulf

Let's look at some examples:

لوى راسه بسرعة وقام يصيح: ش تقول إنت ش تقول!

ابتسمت بخبث ورمست: شو سالفتج

اتصل وخل الباقي على الله

وربع هزاع صوبه ومط الفون منو وسكره بوجهه

Dialectal Diacritization NEEDS Conventions

In Gulf dialect, which diacritized form is correct:

تْقُولْ vs تِقُولْ vs تُقُولْ

Can we start a word with *sukun*?

Can we have two letters with *sukun* back-to-back? Ex. عِنْدْها

Are we going to put ANY case endings (حركة إعراب)? Ex. عَبْدَ الله vs عَبْد الله

How do we diacritize the coordinating conjunction و?

Sample Convention

- الحرف الأول يجب تسكينه إذا كان يسكن مع إضافة حرف الواو قبله. مثال: كلمة "يْقُول" يسكن أولها لأنه ما إضافة "و" قبلها تشكل كالآتي: "وِيْقُول". إذا حرك مع وضع "و" قبل، إذا توضع الحركة المناسبة. مثال: "بِتْقُول
- عند الشك بين الكسر أو تشكيل آخر، يغلب الكسر. مثال: الواو في "وِالْحِين" تكسر "و" الإضافة في أغلب الأحيان (مثال: "وِهِي") إلا إذا كان هناك حركة أخرى ظاهرة
- التقاء الساكنين مسموح به في حدود ضيقة. مثل: "بَعْدf"
- يترك الحرف الأخير دون تشكيل حيث إن اللهجات ليس بها حركة إعراب، ولكنه في بعض الأحيان يحتاج إلى شدة (مثل: "رِدّ")

Sample Convention

- عند التشكيل بالفتح يوضع التنوين قبل الألف. مثال: "طَبْعًا"
- عند المد بالألف أو الواو أو الياء، يجب وضع حركة على الحرف التي قبلها. مثال: "قَال"
- يجب تشكيل جميع الحروف إلا حروف المد وهمزة الوصل: مثال: "اسْتَحَت"
- يجب التشكيل حسب طريقة النطق في اللهجة
- تسكن واو الجماعة في أغلب الأحيان (مثال: "شَافَوْا")، إلا إذا استلزم وضع ضمة على الحرف التي قبلها (مثال: رِدُّوا)

Sample Convention

هذه بعض الجمل المشكلة:

- هَزَّاعْ : الْرِيم قَالَت كِل شَيْ
- مَانِعْ : قُولِي لَا خَلِّيت مِنِّج اِنْ شَاءْ اَلّلَّه
- وَقَّفْنِي صُوت عَمّي خَالِدْ
- نَاصِرْ : اِنْتِي مَا لِجْ خَصْ جِزْوِي سِيرِي وِفِكّينِي
- الْرِيم : عَبَالِي ^ ^
- مِيرَةْ : يَعْنِي عُقُب سَاعَة
- فَزّ خَلِيفَةْ : هَلَا جِدُّوه . .