

The Arabic Language اللغة العربية



- Is the largest member of the Semitic branch of the Afro-Asiatic language family
- It is spoken throughout the Arab world and is well studied throughout the Islamic world.
- 206 million native speakers according to the Ethnologue.
- It's the 5th most widely spoken language in the world





Challenges of working with Arabic

- Arabic has a diglossia situation. This refers to the situation where there is a High language and a Low language.
- The Arabic writing system carries some ambiguity because Arabic doesn't represent the short vowel. They are represented through diacritics.

كُتِبَتْ

ا س = ع ه و -
= -

- A lot of variation in the orthography.



Overview

In this presentation we will focus on the different kinds of Arabic orthographic issues that we have encountered and handled while building our various linguistic software solutions for Arabic such as:

- *Transliteration of Arabic names*
- *Arabic Name Matching*
- *Arabic Name Cleaner*
- *Transliteration of Romanized Arabic used in chat rooms*
- *Arabic editor*
- *Arabic part of speech tagging*



Types of Arabic Corpora

- The Quran
- Modern standard Arabic in Arabic script
 - news articles, e-books
- Colloquial Arabic in Arabic script
 - Arabic telephone speech, Arabic forums or chat rooms
- Colloquial Arabic in Roman script
 - Arabic forums or chat room



What is the Issue of Orthographic Arabic Variants?





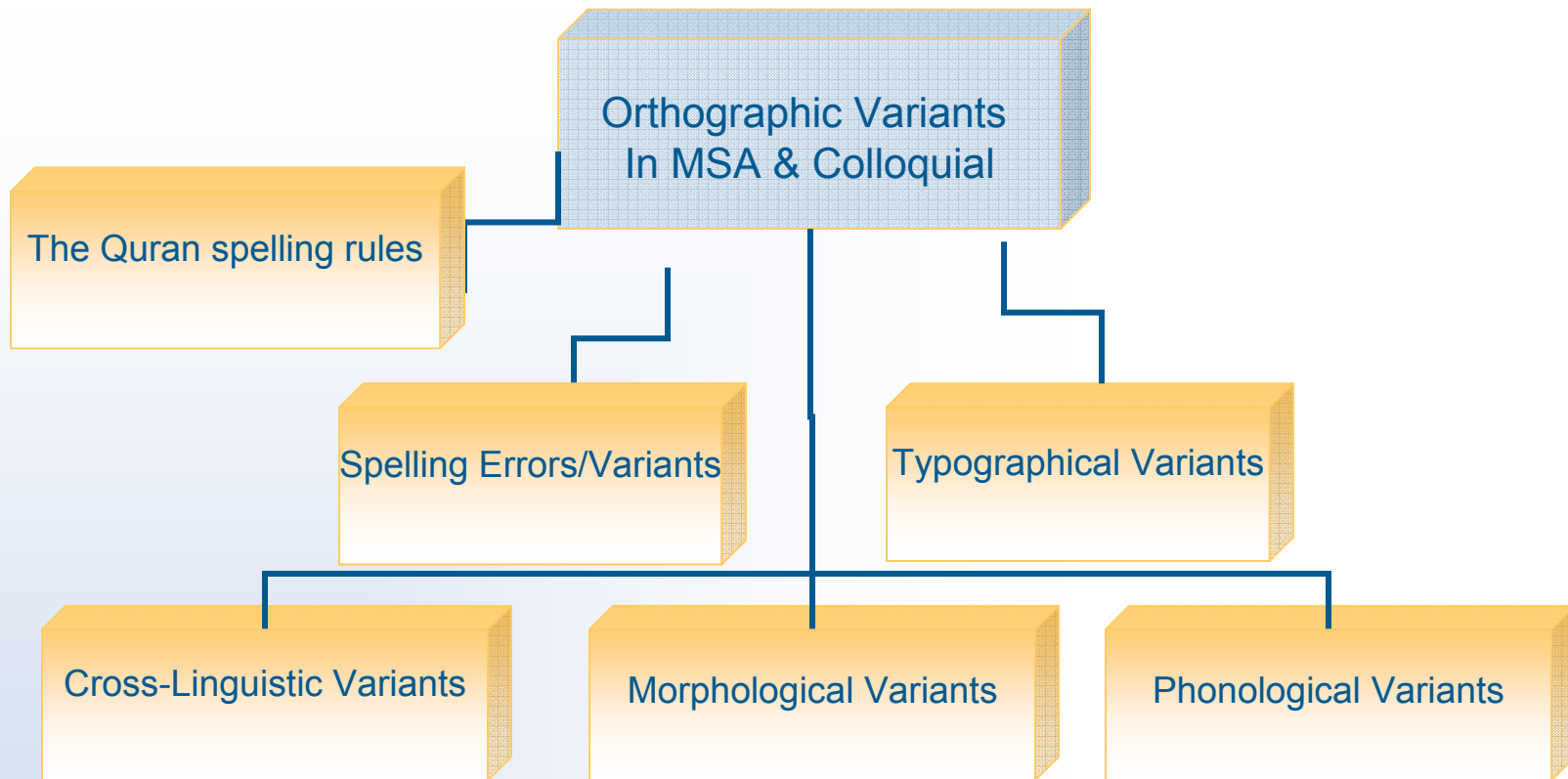
How to Handle it?

- Software Solution
 - (word, gloss) => unique spelling
 - *speech recognition, transliteration, machine translation, or information extraction*

- Lexicon Solution
 - Free of variants
 - Unique normalized form
 - Variants generating rules
 - Linking underlying form with surface forms



Types of Variants



Spelling in the Quran

- The orthography of the Quran was singled out as a separate branch of study known as `ilm al-rasm. Suyūṭī (909/1503) summarized the rules of Qur'ānic rasm to 6 as follows:

- 1. The rule of deletion, hadhf
- 2. The rule of addition, ziyādah
- 3. The rule of substitution, badal
- 4. The rule of the hamza
- 5. The rule of joining and separating, al-wasl wa-l-fasl
- 6. The rule of cases where there are two canonical readings but the text is written according to one of them, ma fīhi qirā'ātan

(ما فيه قرائتان فكتب على إحداهما) fakutiba `alā ihdāhumā



Examples for the rule of the deletion of Aleph

- is deleted after vocative yā' as in يا أيها الناس vs يا أيها الناس “oh people”
- after na of the plural as in أنجيناكم vs أنجيناكم “rescue (we [verb] + you)”
- after laam as خلف vs خلاف “dispute”
- in regular masculine and feminine plural as in الصّديقين – الصّداقات vs الصادقين – الصادقات “the truthful”

All of these rules are specific to the Quran. They are not found in MSA.



Typographic Variants

- Buckwalter, 2004, mentions the following variation:
 - *The drop of hamza initially, medially, and finally*

MSA	Variant	Gloss
إحسان	احسان	charity
أريج	اريج	fragrance
رؤوف	رووف	merciful
هناء	هنا	happiness

- *Two dots inserted on aleph maqsura, and two dots removed from yaa*

MSA	Variant	Gloss
سامي	سامى	exalted
موسى	موسى	Moses
متى	متى	Matthew



Typographic Variants (cont)

- Dropping the madda from the aleph
- Hamza insertion below vs. above aleph
- Two dots inserted on final haa, and two dots removed from taa marbouta

MSA	Variant	Gloss
آل صباح	ال صباح	Al Sabah
الشيخ آل خليفة	الشيخ ال خليفة	Al-Sheikh Al Khalifah
أمي	إمي	my mother
ليلة	ليله	night
طه	طة	Taha

- Diacritics: partial, full, or none

Word	Normalized	Google Hits
الجُمْهُورِيَّة	الجمهورية	1
الجمهورية	الجمهورية	33,200
الجمهورية	الجمهورية	2,120,000



Typographic Variants (cont)

- Use of aleph wasla vs. hamza or none.

إبتسام ابتسام أبتسام

- Typing hamza + aleph maqsura separately vs. together

- طواریء

- طواریئ

- Typing two diacritics on top of each other.

- Scheherazade vs. Arial

كُتَبَ vs كَتَبَ

Morphological Errors/Variants

- Omitting final aleph after the *waw* from the 3rd masculine plural suffix

MSA	Variant	Gloss
كتبوا	كتبو	they wrote

- Unclear morpheme boundaries upon colloquial transliteration

Colloquial	Variant	Gloss
أحكي لكم	أحكيكم	I tell you (masc pl)
على بالي	علبالي	on my mind
أشكي لك	أشكيلك	I complain to you (sg)

Morphological Errors/Variants (cont)

- Run on words

MSA	Variant	Gloss
سي محمد	سيمحمد	Si Muhammed
أبو أحمد	أبوأحمد	Father of Ahmad
ذو الفقار	ذوالفقار	Dhu Alfaqar
نور الدين	نورالدين	Noureddin

Cross-Linguistic Variants

- Final taa representation in Persian vs. Arabic
- Initial hamza representation in Berber transcription vs. MSA
 - Example:*
 - مُوَحمَد وُمَادِي - Libyan
- Initial hamza representation in Kurdish transcription vs. MSA
 - Examples:*
 - نَارَاس هُوَشِيَار اِبْرَاهِيم رَسُول
 - نَالَا بَارَام مُحَمَّد

MSA	Variant	Gloss
نشأة	نشأت	growth
صفوة	صفوت	pure

MSA	Variant	Gloss
أومادي	ؤمادي	son of Madi
أويحيى	ؤيحيى	son of Yahia

MSA	Kurdish	Gloss
الآن	ئالان	now
أشنا	ئاشنا	friend

Cross-Linguistic Variants (Cont)

- Variants resulting from transliterating foreign words

MSA	Gloss
طوني توني	Tony
ماري ميري	Mary
أنجليكا أنجيليكا أنجليكا أنجليكة	Angelica

- Transliteration vs. transcription or transliteration vs. translation

MSA	Gloss
عمر برادلي	Omar Bradley
اومار برادلي	Omar Bradley
ديفيد / داؤود / داوود	David

Spelling Errors/ Variants

- Wrong hamza insertion
 - Corpus specific error

- Aleph maqsura representation as aleph mamduda or taa marbouta

- Dropping laam after prefixing the definite article al-

- Laam is present phonetically but must be deleted orthographically

MSA	Variants	Gloss
إيمان أحلام	أيمان أحلام	Iman Ahlam
نسأل	نسل	We ask
خطأ	خطاء	mistake

MSA	Variants	Gloss
بشرى	بشرة/بشرا	good news
ليلة	ليلى	night
ليلى	ليلة	ecstasy

MSA	Variants	Gloss
الليل	اليل	night
التي	اللتى	which/whom (fem sg)
الذي	اللذي	which/whom (masc sg)



Spelling Errors/Variants (cont)

- Slight different pronunciation in a dialect (from MSA) results in misspelling of MSA words

MSA	Variants	Gloss
واحدة	وحدة	one/single
لكن	لاكن	however

- Conventions lacking for spelling colloquial words or affixes

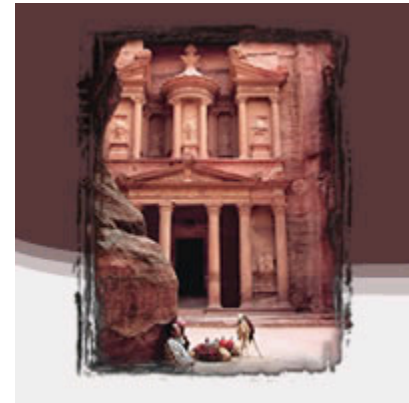
MSA	Variants	Gloss
هذا الولد	ها الولد	this boy
هذا الولد	هالولد	this boy
هذا الولد	هلولد	this boy



Newspaper Example

عمان- بترا - استقبل جلالة الملك عبدالله الثاني امس الفريق اول ميشيل جاكسون رئيس اركان الجيش البريطاني الذي يزور الاردن حاليا وجرى خلال اللقاء الذي حضره رئيس هيئة الاركان المشتركة الفريق اول الركن خالد جميل الصرايره والسفير البريطاني في عمان كريستوفر برينتس بحث اوجه التعاون بين البلدين وخاصة في المجالات العسكرية والدفاعيه

- **Typographical Variants**
 - *Lacking hamza on aleph*
 - اوجه, الاركان, الاردن, اركان, استقبل
 - *Dropping final hamza*
 - بترا
 - *Lacking two dots on taa marbouta*
 - والدفاعيه, الصرايره
- **Morphological Variants**
 - *Run on word*
 - عبدالله





Phonological Impact on Orthography

- The uvular /q/ could be pronounced as a glottal stop /ʔ/, /k/, /g/, [ɣ].

MSA	Variants	Gloss
قاسم	كاسم / آسم گاسم / غاسم	Qasem

- The voiceless interdental fricative “th” could be pronounced as a “t”
 - تكرم عينك بدل ال ثلاثة منجيب 10

MSA	Variants	Gloss
كثير	كتير	A lot
ثلاثة	تلاثة	three

- The voiceless interdental fricative “th” could be pronounced as an [s]
 - قال والله سعلب مكار كل شي يسوى

MSA	Variants	Gloss
ثمار	سمار	fruits
ثعلب	سعلب	fox

Phonological Impact on Orthography (cont)

- the voiced interdental fricative as in the first sound in “the” would be pronounced as a [d]

- يا أم الجدائل ذهب والدهب ده لون قلبك

MSA	Variants	Gloss
ذهب	دهب	gold

- the [ض] would be pronounced as a [ظ]

MSA	Variants	Gloss
ضرار	ظرار	penalty

- voiced emphatic interdental fricative [ظ] would be pronounced as a [ض] or [ز] (emphatic [Z])

MSA	Variants	Gloss
ظاهر	ضاھر	apparent
ظاهر	زاهر	apparent

Phonological Impact on Orthography (cont)

- spelling/pronouncing [k] as [tch]

- *علشان خاطر عيونتش*
- *اعتبرين نفسستش منا وفيينا*

MSA	Variants	Gloss
عيونك	عيونتش	your eyes

- Spelling/pronouncing [ج] as [ي]

- *ليش انتو ميانين للحين؟*

MSA	Variants	Gloss
مجانين	ميانين	crazy

- spelling [ج] as [دج] in Maghrebi, and Sudanese

- *كما طلبت المحامية سماع شهادة التونسي نزار والمتهم الجزائري دجميل*

MSA	Variants	Gloss
جميل	دجميل	nice

- [ذ] could be pronounced as a [ز]

انا شو زني؟

MSA	Variants	Gloss
ذنب	زنب	fault

Variations in Romanized Arabic



Basis Technology dealt with the issue of orthographic variations in Romanized Arabic in the following issues:

- Transliteration of Arabic names
- Transliteration of romanized Arabic that is used in chatrooms

Reason why there are orthographic variations in Romanized Arabic:

1. There is no one to one correspondence between the Arabic letters and the Roman letters. Arabic has the guttural sounds.
2. English spelling is chaotic.
3. Different languages use different letters in the Roman script to express the same sounds.

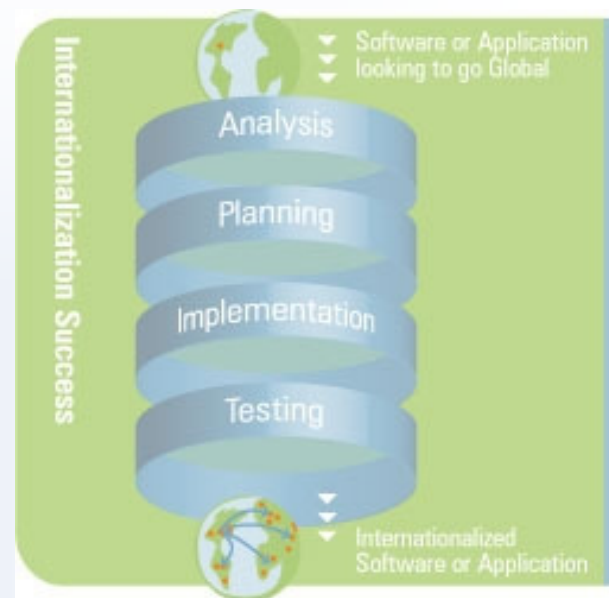


Example: Romanization of قاسم

- Qasim
- Gasim
- 8asim → 8asim El9abunchi
- Kasim
- Asim → could also be عاصم
- 2asim → 2asim ya 7ramy
- 'asim
- Qasem
- Gasem
- 8asem
- Kasem
- Asem
- 2asem
- 'asem



These issues of Arabic variants represent a small set of Arabic Linguistic challenges that we encountered and we have provided tools to handle them in our Arabic product solutions.





References

Tim Buckwalter (2004)

Issues in Arabic Orthography and Morphology Analysis

Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages, COLING 2004, Geneva, August 28, 2004.

<http://www.islamicawareness.org/Quran/Text/Scribal/haleem.html>



Thank You!

bushraz@basistech.com

zinas@basistech.com