Issues in Arabic Grammar: from Tokenization to Transfer

Mohammed A. Attia
Mohammed.attia@postgrad.manchester.ac.uk
School of Informatics,
The University of Manchester

Tokenization

Concatenations in Arabic nouns

Proclitics			Stem	Suffix	Enclitic
Conjunction/ question article	Preposition	Definite article	Noun	Gender/Number	Genitive pronoun
Conjunctions و "wa" (and) or ف "fa" (then)	ب "bi" (with), ك "ka" (as) or ل "li" (to)	ال "al" (the)	Stem	Masc Dual (4)	First person (2)
				Fem Dual (4)	
Question word أ "a" (does or did)				Masculine regular plural (4)	Second person (5)
				Feminine regular plural (1)	Third person (5)
				Feminine Mark (1)	

Possible concatenations in Arabic nouns

Tokenization

Concatenations in Arabic verbs

Proclitics		Prefix	Stem	Suffix	Enclitic
Conjunction/ question article	Complementizer	Tense/mood – number/gender	Verb	Tense/mood – number/gender	Object pronoun
Conjunctions و "wa" (and) or ف "fa" (then)	ل "li" (to)	Imperfective tense (5)	Stem	Imperfective tense (10)	First person (2)
Question word أ "a" (does or did)	س "sa" (will)	Perfective tense (1)		Perfective tense (12)	Second person (5)
	ل "la" (then)	Imperative (2)		Imperative (5)	Third person (5)

Possible concatenations in Arabic verbs

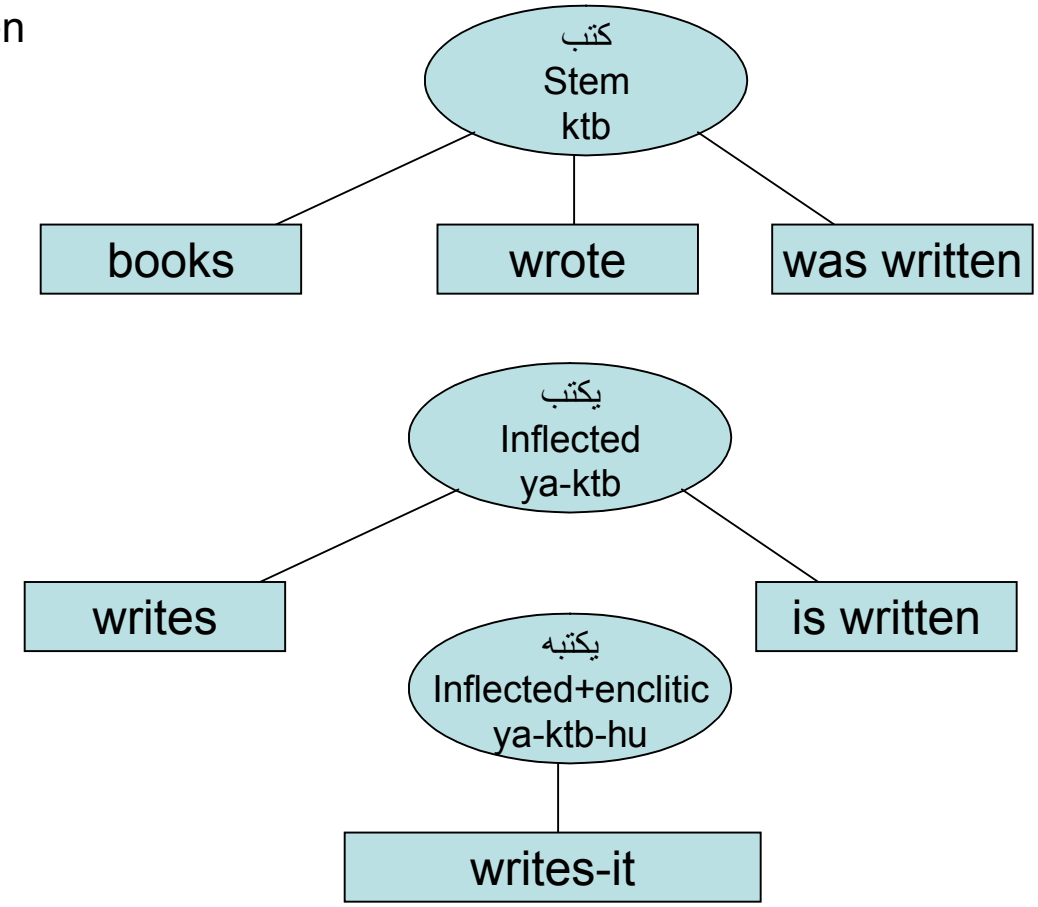
Tokenization

The Ambiguity Pyramid Hypothesis

كتب ktb books / wrote / was-written

يكتب ya-ktb writes / is-written

يكتبه ya-ktb-hu [he]-writes-it



Tokenization

The Ambiguity Pyramid Hypothesis

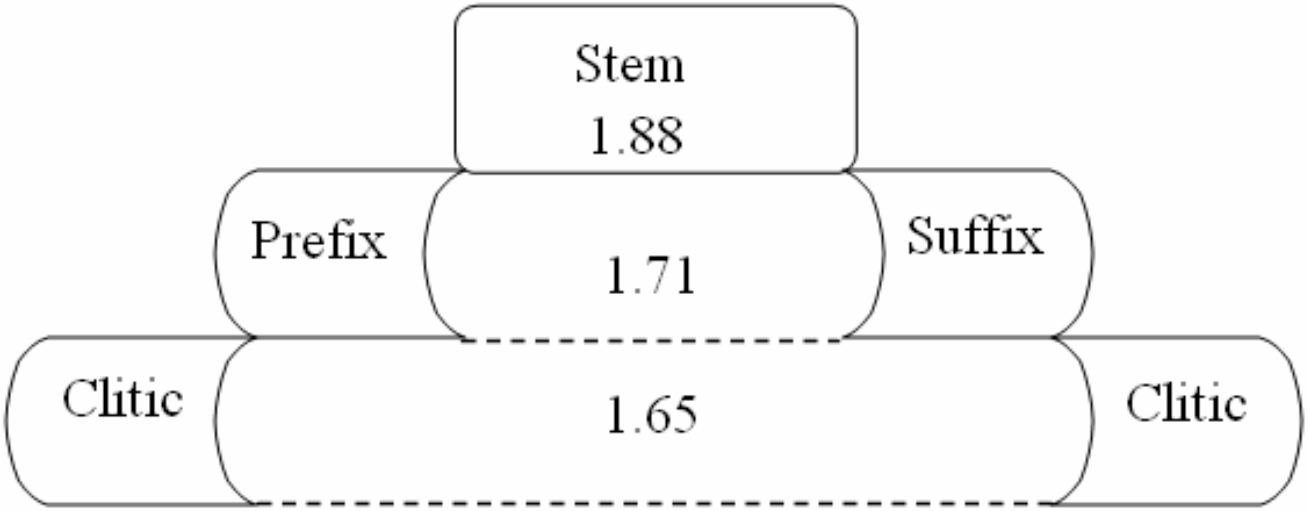
In Arabic, tokenization creates extra ambiguities.

- Form: كتابهم kitabu-hum (their book) – Not ambiguous
- Tokenized: كتاب@هم kitab@hum, three different readings
  - كتاب هم kitab hum (book they)
  - كتابهم kitabu-hum (their book)
  - كتاب هم ktabu-hamm (book of worry)

Tokenization

The Ambiguity Pyramid Hypothesis

Ambiguity rates decrease, on average, with the increase in word buildup complexity



The ambiguity pyramid hypothesis

Tokenization and Tagging

- Proposed solution - Tagging
- يكتبهya-ktb-hu ([he]-writes-it)
  - يكتب/+verb+active@ه/+pron+acc
- كتابهم kitaba-hum (their book)
  - كتاب/+noun@هم/+pron+gen

Arabic Morphology

- Reducing the ambiguity rate per word
- Choosing the most probable solution

```
OPTIMALITYORDER
        LeaveOut "Imperative"
        NOGOOD
        uncommon "uncommon morphological forms  مثل قاضي قاض ومعاني معان"
        STOPPOINT
         FirstPerson
         SecondPerson
         Passive
       +PreferFunc "prefer function word analysis over other words"
       +propernoun
        .
```

Grammar

- Current State
  - Rules: 49
  - Disjuncts: 1788
  - Test sentences: 482
  - Real sentences: 40
  - Longest sentence: 43

Grammar

- Cascade of STOPPOINT Optimality Mark

uncommonSubcat “on Bush to help the Palestinians على بوش أن يساعد الفلسطينيين "

STOPPOINT

uncommon "uncommon morphological forms مثل قاضي قاض ومعاني معان"

STOPPOINT

SubjNoArtRel "Arab commentators (who) neglected him of ignoring the Israeli occupation معلقين عرب اتهموه بتجاهل احتلال إسرائيل"

STOPPOINT

AdjAsNoun "the latter (person) clarified the point قام الأخير بتوضيح المسألة"

STOPPOINT

AdjAsNounDispref

Grammar

- Increased ambiguity with the increased number of words
  - (384+435,107,008) (35 words)
  - (384+44,851,840) (32 words)
  - (384+44,851,840) (32 words)
  - (120+16,074,008) (31 words)
  - (120+15,273,560) (30 words)
  - (24+2,789,560) (29 words)
  - (128+5,606,400) (25 words)
  - (400+299,328) (22 words)
  - (128+720) (12 words)

Grammar

- Clitics are %35 of the number of forms
  - 43 words => 192 tokenizations => 63 tokens (Longest analyzed sentence)
  - 21 words => 8 tokenizations => 33 tokens
  - 14 words => 4 tokenizations => 23 tokens
  - 26 words => 128 tokenizations => 43 tokens

Grammar

- Current approach: 5 articles, 200 sentences
- Limitation of the current approach to grammar development
  - Too limited portion of the corpus
  - Contains all varieties: long sentences, rare constructions, rare morphological forms, etc.
  - It does not allow us to make general judgments about the language, such as:
    - Average length
    - Frequency of certain syntactic structures

Grammar

- Future Plan
- Developing a sentence segmenter
  - Arabic currently segmented by hand
  - Segmentation in English revolves around punctuation marks which signal the end of a sentence [. ?], you need only to avoid abbreviations and acronyms.
  - In Arabic, a comma can signal a new sentence and also certain words can signal a new sentence

(almesryoon.com 10 September 2006)

تاريخ يجب أن يوثق

جمال سلطان : بتاريخ 10 - 9 - 2006

المرحلة التي أعقبت أحداث أكتوبر 1981 عرفت فصولا مأساوية على صعيد حقوق الإنسان ، وخاصة المعتقلين السياسيين من التيار الإسلامي ، والمشاهد التي رويتها في مقال الامس هي بكل أمانة أقل كثيرا مما عايشه آخرون تحملوا أهوالا كبيرة كانت تتم خارج نطاق القانون والأخلاق وأي شيئ يمت إلى الإنسانية بصلة ، كثيرون من أبناء التيار الإسلامي يملكون شهادات أكثر خطورة وأكثر تفصيلا ، وبالأسماء ، هناك قيادات أمنية أصبحوا وزراء للداخلية بعد ذلك ومارسوا التعذيب بأنفسهم ، أحدهم كان يشرف بنفسه على التعذيب في القلعة ، وكان يتدخل عندما يشرف المعتقل على الموت وينهمر الدم من أنفه وفمه ووجهه ويدخل في إغماءة ، فيضع معاليه حذاءه في فمه ويقلب وجهه بقدمه ثم يشير إلى زبانيته : لسه صاحي ، وصاحب هذه الشهادة حي يرزق الآن في إحدى مدن الدلتا ، وآخرون من مختلف شرائح التيار الإسلامي لديهم شهادات مروعة ، بعض زملاء الاستقبال على سبيل المثال يملكون ـ إذا تكلموا ـ ما يملأ مكتبة كاملة من الشهادات والأحوال والأهوال ، ومعظم هؤلاء ما زالوا أحياء يرزقون بعضهم داخل مصر وآخرون خارجها ، أذكر على سبيل المثال عبد الحميد شكر وحمدي عبد الرحمن وتوفيق علوان وعبد المجيد الشاذلي وعبد الآخر حماد ، وهؤلاء كانوا جيران عنبر ألف في الاستقبال ، وهناك آخرون من نزلاء باقي السجون وخاصة القلعة في عزها ، وأنا أناشد هؤلاء جميعا أن يسجلوا شهاداتهم ويوثقوها كتابة و أن تكون هناك جهة مؤسسية تنسق هذا الجهد ، وبغض النظر عن القيمة القانونية لهذه الشهادات ، لأن الظروف الحالية يستحيل فيها تحقيق نصر قضائي يثأر من الجلادين والقتلة ، ولكن الأهم هي تسجيل هذه الشهادة للتاريخ وللأجيال القادمة ، كما أنه إبقاء للأمل في أن يأتي اليوم التي يتمكن فيه قضاء مصر العظيم من أن يثأر لأصحاب تلك المآسي التي طواها النسيان ، في الوقت الذي يتبختر الزبانية في الفضائيات والمنتديات بوصفهم حكماء سياسيون وخبراء أمنيون في مكافحة الإرهاب ، كما أنه يمثل كشفا للغطاء عن الجلادين القدامى والجدد ويحاصر منتهكي كرامة البشر ويفزع منامهم ، ثقيلة هي هذه الشهادات على النفس ، وثقيلة على المشاعر ، لأنه بدون شك عندما تستحضرها فأنت تستحضر قدرا من عذابها وآلامها ، خاصة عندما تدرك أنك صارخ في البرية ، وأن كلامك لن يفلح في أن يعيد حقا لأصحابه ، أو حتى يوقف مسلسل التعذيب المستمر حتى الآن ، وأنا أوافق بعض الأصدقاء الذين علقوا على مقال الأمس بأن التعذيب استمر بعد ذلك بصورة أبشع وأكثر هولا ، هذا صحيح ، وأنا لدي شهادات عديدة من رموز وشخصيات عايشت هذه المحنة ، ولعل الله ييسر الفرصة والوقت والصبر على إنجازها وتحريرها .

Grammar

- Future Plan
- Developing a sentence segmenter
  - Use of punctuation marks has not been regular in Arabic for a long time. A sentence is a stream of ideas with or without a full stop at the end = English paragraph
  - Effect of translation appears:
    والمشتبه فيهم هم عمر سعيد عمر، وسلمان محمد خميسي، ومحمد كوبوا، ونجله محمد كوبوا سيف، وسعيد ساغار أحمد، وعبود روغو محمد.
    "The suspects are Umar, and Sulaiman, and Mohammed, and Said, and Abbud"
  - وبهذا and hence
  - وعلى ذلك and therefore
  - وهكذا and thus
  - وعادة and usually
  - وأضاف and [he] added
  - وكان ومن ولكن ويعتبر وذلك وقال وأكد
  - و ، comma + and (productive but not safe واصفا وفق)

Transfer

- Where to place the bilingual lexicon?
- In Transfer?

Duplicate lexicon

- @noun_noun(ولد, boy).
- @noun_noun(موزة, banana).
- @noun_noun(تفاحة, apple).
- PRED(%X, عالم), +NTYPE(%X, %%), +HUMAN(%X, +) ==> PRED(%X, scientist).
- PRED(%X, عالم), +NTYPE(%X, %%), +HUMAN(%X, -)==> PRED(%X, world).

Duplicate Subcats

- @verb_subj_obj(%X, أكل, %Subj, %Obj), +HUMAN(%Subj, +) ==> @verb_subj_obj(%X, eat, %Subj,%Obj).
- @verb_subj(%X, أكل, %Subj), +HUMAN(%Subj, +) ==> @verb_subj(%X, eat, %Subj).
- @verb_subj(%X, أكل, %Subj), HUMAN(%Subj, -) ==> 0.

Transfer

- Where to place the bilingual lexicon?
- One default gloss in morphology (+gloss-???+gloss-)
  - upper language: كتاب+sg+nonhuman+gloss-book+gloss-
  - lower language: كتاب

- Modified in lexicon according to subcats and mwes
  - اعتمد V XLE {@(V-Subj-Obl %stem على) GLOSS=rely` on
    |@(V-Subj-Obj %stem) GLOSS=accredit.

- Refined in transfer