Arabic Language: Nature and Challenges

Mohammed Attia

The British University in Dubai

May 29, 2012
Outline

• Introduction
• Lexicons and Corpus Linguistics
• Morphology
• Syntactic Parsing
• Tokenization
• Multiword Expressions
• Statistical Parsing
• Which is better?
• Spelling Checking and Correction
• Integration with Applications
Introduction
Introduction

• Living language are just ... living
Introduction

• Living language are just ... living
• They grow
Introduction

- Living language are just ... living
- Leaves fall off
Introduction

- Living language are just ... living
- New leaves appear
- ... constantly changing
- They may grow old and die
- They maybe reborn
- New languages may appear
Introduction

• Language reveals everything about us ...
Introduction

• Language tell everything about us ....
• How rich or poor
Introduction

• Language tell everything about us ....
• How well-educated
Introduction

- Language tells everything about us.
- Where we come from.
Introduction

• Language tell everything about us ....
• what kind of work we do
Introduction

• Language tell everything about us ....
• our feelings and our sentiments
Introduction

• Language is the key to business expansion: Translation and localization

World translation business in 2011 = $30 billion
Introduction

• And a repository of knowledge and information
Lexicons and Corpus Linguistics
Definition of a dictionary

- A description of the vocabulary ( حصيلة لغوية) used by members of a speech community (مجتمع يتحدث نفس اللغة). A dictionary deals with:
  - Conventions عرفي not idiosyncrasies شخصي
  - norms سائد not rarities نادر
  - Probable واقع not possible ممكن نظريا

- Lexical evidence
  - Subjective evidence
    - Introspection الاستبطان
    - informant-testing المعرفة أصحاب استشارة
  - Objective evidence
    - A corpus ( ذخيرة النصوص) provides typifications (تصنيف للأشكال) of the language
      - A typical lexical entry means it is both “frequent” or “recurrent” and “well-dispersed” in a corpus.
      - A typical lexical entry belongs to the stable “core” of the language.

Principles of Lexicography

• Corpora and Dictionaries
  – Brown Corpus, 1 million words, 1960s, → Citations for American Heritage Dictionary
  – British National Corpus (BNC), 100 million words, 1990s set the standard (balance, encoding)
  – The Oxford English Corpus, one billion words, 2000s → Oxford English Dictionary
  – Longman Corpus Network, 330 million word → Longman Dictionaries
Principles of Lexicography

• Dictionaries before Corpora
  – Citation banks مراجع اقتباسية استشهادية
    • A citation is a short extract providing evidence for a word usage or meaning in authentic use.
  – Disadvantages
    • labour-intensive
    • instances of usage are authentic, but there is a big subjective element in their selection.
      – People tend to notice what is remarkable and ignore what is typical
      – bias towards the novel or idiosyncratic usages
Principles of Lexicography

• **Characteristics of a reliable corpus** (مواصفات ذخيرة النصوص)
  
  – The corpus does not favour high class language
  – The Corpus should be large and diverse
  – The corpus should be either synchronic or diachronic
  – The corpus should be well-balanced using “stratified sampling” (أخذ عينات بشكل نسبي)
  – The corpus should avoid skewing (الانحراف أو التحيز)
Principles of Lexicography

- **Lexical Profiling**
  - Word POS
    - v, n, adj, adv, conj, det, interj, prep, pron
  - Valency Information
    - subcat frames, other obligatory or optional syntactic constructions
  - Collocations
    - commit a crime, sky blue, lame duck
  - Colligational preferences
    - was acquitted, trials (difficult experiences)
Principles of Lexicography

– Lexical Profiling Software
– Concordancers
– Word Sketch (Sketch Engine) - Adam Kilgarriff
Concordancer
## Sketch Engine

### Part of the Word Sketch for the noun *bargain*

<table>
<thead>
<tr>
<th>object_of</th>
<th>object</th>
<th>modifier</th>
<th>object</th>
<th>modifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>strike</td>
<td>264</td>
<td>2.7</td>
<td>hard</td>
<td>251</td>
</tr>
<tr>
<td>drive</td>
<td>61</td>
<td>43.38</td>
<td>real</td>
<td>20</td>
</tr>
<tr>
<td>get</td>
<td>26</td>
<td>27.56</td>
<td>best</td>
<td>14</td>
</tr>
<tr>
<td>seal</td>
<td>27</td>
<td>16.38</td>
<td>good</td>
<td>19</td>
</tr>
<tr>
<td>make</td>
<td>5</td>
<td>14.82</td>
<td>bad</td>
<td>8</td>
</tr>
<tr>
<td>find</td>
<td>26</td>
<td>13.6</td>
<td>better</td>
<td>8</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>modifies</th>
<th>modifies</th>
<th>modifier</th>
<th>number</th>
<th>modifier</th>
</tr>
</thead>
<tbody>
<tr>
<td>basement</td>
<td>221</td>
<td>0.9</td>
<td>plea</td>
<td>26</td>
</tr>
<tr>
<td>hunter</td>
<td>22</td>
<td>38.62</td>
<td>wage</td>
<td>6</td>
</tr>
<tr>
<td>price</td>
<td>54</td>
<td>33.65</td>
<td>credit</td>
<td>6</td>
</tr>
<tr>
<td>bookshop</td>
<td>11</td>
<td>26.73</td>
<td>sale</td>
<td>5</td>
</tr>
</tbody>
</table>

Part of the Word Sketch for the noun *bargain*
1973 مستويات العربية المعاصرة
للسعيد محمد بدوي

• فصحي التراث
• فصحي العصر
• عامية المثقفين
• عامية المتنورين
• عامية الأميين
Modern Standard Arabic vs. Classical Arabic vs. Colloquial Arabic

• Modern Standard Arabic
  – The language of modern writing, prepared speeches and the language of the news

• Classical Arabic
  – The language of Arabia before Islam and after Islam until the Medieval Times
  – Present religious teaching, poetry and scholarly literature.

• Colloquial Arabic
  – Variety of Arabic spoken regionally and which differs from one country or area to another. They are to a certain extent mutually intelligible.

Code Shifting – Code Switching – Diglossia – multi-layered diglossia
Modern Standard Arabic vs. Classical Arabic vs. Colloquial Arabic

• Modern Standard Arabic
  – Tendency for simplification
    • Some CA structures to die out
    • Structures marginal in CA started to have more salience
    • no strict abidance by case ending rules
  – A subset of the full range of structures, inflections and derivations available in CA
  – MSA conforms to the general rules of CA
  – How “big” or how “small” the difference (on morphological, lexical or syntactic levels) need more research and investigation
Review of Arabic lexicographic work

- *Kitab al-'Ain* by al-Khalil bin Ahmed al-Farahidi (died 789)  
  (refinement/expansion/organizational Improvement)
  ▼
- *Tahzib al-Lughah* by Abu Mansour al-Azhari (died 980)
- *al-Muheet* by al-Sahib bin 'Abbad (died 995)
- *Lisan al-'Arab* by ibn Manzour (died 1311)
- *al-Qamous al-Muheet* by al-Fairouzabadi (died 1414)
- *Taj al-Arous* by Muhammad Murtada al-Zabidi (died 1791)
- *Muheet al-Muheet* (1869) by Butrus al-Bustani
- *al-Mu'jam al-Waseet* (1960)
Review of Arabic lexicographic work

• Bilingual Dictionaries
  – Edward William Lane's *Arabic-English Lexicon* (1876) indebted to *Taj al-Arous* by al-Zabidi
  – Hans Wehr's Dictionary of Modern Written Arabic (1961)
    • Size: 45,000 entries
    • Aim: Using scientific descriptive principles to describe present-day vocabulary through wide reading in literature of every kind
• Application
  – Selection of works by high flying poets and literary critics such as Taha Husain, Taufiq al-Hakim, Mahmoud Taimur, al-Manfalauti, Jubran Khalil Jubran
  – Use of secondary sources (dictionaries) for expansion
  – Inclusion of rarities and classicisms that no longer formed a part of the living lexicon
Review of Arabic lexicographic work

• Bilingual Dictionaries
  – Landau and Brill (1959) *A Word Count of Modern Arabic Prose*
    • A word count based on 270,000 words based on: 136,000 from the news (Moshe Brill, 1940) and 136,000 from 60 contemporary books on: fiction, literary criticism, history, biography, political science, religion, social studies, economics, travels and historical novels
    • 6,000 words in the news
    • 11,000 words in literature
    • 12,400 words in the combined list (does not include proper nouns)
Review of Arabic lexicographic work

• Bilingual Dictionaries
    • COBUILD-style, Corpus-based (3 million words)
    • Manually constructed
    • Covers the whole range of the actual vocabulary in the corpus with 17,000 entries compared to 45,000 entries in Hans Wehr
    • 5% of frequent new words not found in Hans Wehr
Review of Arabic lexicographic work

- Buckwalter Arabic Morphological Analyzer (2002)
  - Size: 40,222 lemmas (including 2,034 proper nouns)
  - Includes many obsolete lexical items
    (But how many?)

<table>
<thead>
<tr>
<th>#</th>
<th>Meaning</th>
<th>Classical Word</th>
<th>Google</th>
<th>MSA Word</th>
<th>Google</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>sully</td>
<td>قلعط qal`at</td>
<td>8</td>
<td>لطخ laṭṭaḥa</td>
<td>29,600</td>
</tr>
<tr>
<td>2</td>
<td>caulk</td>
<td>قلفط qalfaṭ</td>
<td>9</td>
<td>أفسد `afsada</td>
<td>205,000</td>
</tr>
<tr>
<td>3</td>
<td>wear</td>
<td>استكد `istakadda</td>
<td>4</td>
<td>أنهك `anhaka</td>
<td>37,100</td>
</tr>
<tr>
<td>4</td>
<td>fickle</td>
<td>غملج ġamlağ</td>
<td>7</td>
<td>متقلب mutaqqallib</td>
<td>189,000</td>
</tr>
<tr>
<td>5</td>
<td>erosion</td>
<td>انتكال `i’tikāl</td>
<td>7</td>
<td>تأكل ta‘ākul</td>
<td>1,700,000</td>
</tr>
</tbody>
</table>

Google score for Classical vs. MSA entries
Corpus-based Lexicon

Largest corpus of modern Arabic to date
Arabic gigaword 1,200,000,000
  = 16,000 large books
  = 800 meters of bookshelves
Burj Khalifah is 830m

Avr reader reads 200 wpm
With 60% comprehension.

You will need 11 years 24/7
to read the Gigaword corpus
Review of Arabic lexicographic work

Buckwalter obsolete words: 8,400 obsolete words

 صحراea: فِيِّفِاء فِدْفِد فِوَاء مَوْماة مِتْلُف سِبْسِب
رِمل: هِيَلَان وَعْس مِيعاِس عِصِّير

سيرج: حِداِجَة مَخْلُوَّة
حمِل: ظِعْيِنَة حِدْج ظِعْوَن وَفْر
لحِجام: فِدْام كَعْم كَعَم آَرَبِيَّة
شَكْم غمَامَة
راَكِب: حِدَّاء
جمال: هَجِيَّة
رداء: دَفْيَةٌ بِشَتَة
حِذَاء: مِيدِّل بِشْمِق زِربَول زِرْبَون
صَرْمَة قُيِّفَاب
Review of Arabic lexicographic work

Not in Dictionaries: about 10,000 need to be added

تكنولوجيا:
رقمية، آمنة، مكتبة
فيسبوك، تويتر، تغريدة
هاتف، جوال، تليفون محمول
لاب توب
الهواتف الذكية
حوسبة
بريد إلكتروني
آي فون، دي في دي، سي دي
سبام، فيروس
ملتي ميديا
كمبيوتر، لوحى، نسخة
شيفرة

اقتصاد:
خصوصية ريمى، بورصة تعليم داو، جونز تضخم أسهم، قيمة، دفترية، مليار، ترليون، تجارة، إلكترونية
Morphological Lexicon AraComLex

• How our lexical database will be different from Buckwalter’s. We include
  – only entries attested in a corpus
  – subcategorization frames
  – +/-human semantic information for nouns
  – Information on allowing passive and imperative inflection for verbs
  – Information on diptotes
  – detailed information about derived nouns/adjectives (active or passive participle or a verbal noun, *masdar*)
  – multi-word expressions
  – classification of proper nouns: person, place, organization, etc.
  – Frequency information
  – Citation in real examples
Corpus-based Lexicon

Lexicons as a truthful representation of the language as evidenced in a corpus

The Arabic Gigaword corpus
AraComLex

Our morphological analyser – based on a lexical database automatically derived from the Arabic Gigaword Corpus
Morphology
Arabic Morphology

• Arabic Morphotactics

<table>
<thead>
<tr>
<th>Root</th>
<th>درس</th>
<th>dars</th>
</tr>
</thead>
<tbody>
<tr>
<td>Template</td>
<td>R₁aR₂aR₃a</td>
<td>R₁aR₂R₃a</td>
</tr>
<tr>
<td>POS</td>
<td>V</td>
<td>V</td>
</tr>
<tr>
<td>Stem</td>
<td>d ar as a</td>
<td>d ar r as a</td>
</tr>
<tr>
<td></td>
<td>‘study’</td>
<td>‘teach’</td>
</tr>
</tbody>
</table>
Arabic Morphology

- Design Approach: Three approaches
  1. Root-based Morphology
     Xerox Arabic FTM
  2. Stem-based morphology
     Buckwalter
     $kr$ $akar$ PV thank;give thanks
     $kr$ $okurIV$ thank;give thanks
  3. Lemma-based morphology
Morphological Lexicon - AraComLex

- AraComLex Lexicon Writing Application
Syntactic Parsing
Algorithms and Data Structure

Output

Parsing

Training

Bikel Parser
Algorithms and Data Structure

1. **Output**
2. **Tokenizer**
3. **Morphological Analyser**
4. **Treebank**
5. **Parsing**
6. **Training**
7. **Bikel Parser**

**Linguistics**
Why Linguistics

• Linguistic Data is a naughty blackbox:
  – You get non-deterministic answers
  – You can get wrong answers
  – For the same question, you can get a set of inconsistent answers

• We need to make the algorithms suite the data structure, and we also need to make sure that the data is structured properly.
Handcrafted Grammar: A Quick Overview

Sentence

ساعدة الهيئة الفلسطينية
sā‘adat al-hai’atu al-filistīniyyīn/ al-filistīniyyain
helped the-agency the-Palestinian.pl/ the-Palestinian.dual
‘The agency helped the Palestinians/ the two Palestinians.’

Tokenization

ساعدت@الـ@هيئة@الـ@فلسطينيين
helped@the@agency@the@Palestinians
Handcrafted Grammar: A Quick Overview

Morphological analysis

ساعدت  + verb+past+active+1pers
helped  + verb+past+active+3pers+sg+fem
       + verb+past+active+2pers+sg+fem
       + verb+past+active+2pers+sg+masc

الـ  + defArt
the

 الهيئة  + noun+nonhuman+fem+sg
agency

Filipina  + adj+masc+dual+accgen
Filipinos  + adj+masc+pl+accgen
       + noun+human+filipina+masc+dual+accgen
       + noun+human+filipino+masc+pl+accgen
Handcrafted Grammar: A Quick Overview

Lexicon (Lexical properties/subcategorization frames)

helped

\[
V 
\text{XLE} (\text{^GLOSS}) = \text{help } "\text{This verb has three different subcat frames}" \\
\{ (\text{^PRED}) = '%\text{stem}(\text{^SUBJ}(\text{^OBJ}(\text{^COMP})))' \\
\text{^COMP COMP-FORM} = c \text{ (^COMP COMP-TYPE) = c verbal} \\
\mid (\text{^PRED}) = '%\text{stem}(\text{^SUBJ}(\text{^OBJ}(\text{^OBL})))' (\text{^OBL OBJ PCASE}) = c \text{ على} \\
\mid (\text{^PRED}) = '%\text{stem}(\text{^SUBJ}(\text{^OBJ})))'.
\]

agency

\[
N 
\text{XLE} (\text{^GLOSS}) = \text{agency } (\text{^PRED}) = '%\text{stem}', (\text{^PERS}) = 3 \\
\{ (\text{^NUM}) (\text{^NUM}) \sim= sg \mid (\text{^NUM}) = sg \} \text{ "the default number is singular"}.
\]

Palestinian

\[
N 
\text{XLE} (\text{^GLOSS}) = \text{Palestinian } (\text{^PRED}) = '%\text{stem}', (\text{^PERS}) = 3 \\
\{ (\text{^NUM}) (\text{^NUM}) \sim= sg \mid (\text{^NUM}) = sg \} \text{ "the default number is singular"}; \\
\text{ADJ XLE} (\text{^PRED}) = '%\text{stem}', (\text{^GLOSS}) = 'Palestinian' \\
\{ (\text{^ATYPE}) = c \text{ predicative} \mid (\text{^ATYPE}) = \text{ attributive}\}.
\]
Handcrafted Grammar:  
A Quick Overview

Grammar Rules: PS-rules and functional equations

MT ARABIC RULES (1.0)

S_Nonequational --> "There are three word orders permitted in Arabic: VSO, SVO and VOS"
  { VSO
    | SVO
    | VOS.}

VSO --> V: ^=! @DefSTense (^ VTYPE)= copular (^ COMP-TYPE)= verbal
  { (^ SUBJ PRED)=c 'pro' (^ SUBJ NUM) = (^ AGR NUM)
    | (^ SUBJ PRED)= 'pro' (^ AGR NUM)= sg)
  (^ AGR GEND)=(^ SUBJ GEND) (^ AGR PERS)=(^ SUBJ PERS);
  { NP: (^ SBJ)=! (! FIRST-CONJ)=+
    (! CASE)=nom (! PRON-TYPE) =~ pers
    | e: (^ SUBJ PRED)= 'pro' "ProDrop"
      (^ AGR PERS)=(! PERS) (^ AGR NUM)=(! NUM) (^ AGR GEND)=(! GEND) }
  (NP: (^ OBJ)=! (! CASE)= acc).
Handcrafted Grammar: A Quick Overview

Output: c-structures and f-structures

helped

the agency

the Palestinians
Tokenization
Tokenization in XLE

wasayashkurunahu
wa@sa@yashkuruna@hu
and@will@thank[they]@him

wa@li@al@rajuli
and@to@the@man

Proclitics

Enclitic

Conjunction

Comp./Tense Marker

Stem with Affixes

Object Pronoun

Conjunction

Preposition

Definite Article

Stem with Affixes

Genitive Pronoun

Proclitics

Enclitic
Tokenization in XLE

Deterministic Tokenizer

(والرجل: and to the man)
@وال@ال@رجل wa@li@al@ragul@ and@to@the@man@

Non-Deterministic Tokenizer

(والرجل: and to the man)
@وال@ال@رجل wa@li@al@ragul@ and@to@the@man@
@وال@ال@رجل
@وال@ال@رجل
@وال@ال@رجل
@وال@ال@رجل
Tokenization in Bikel

• English parser
  – Input sentence:
    The President led his country in reform.
  – Formatted sentence:
    (The President led his country in reform.)

(VBZ has) (RB n't) (NNP Chicago) (POS 's)
Tokenization in Bikel

• English parser
  – Output:
    
    \((S \ (NP \ (DT \ The) \ (NNP \ President))) \ (VP \ (VBD \ led) \ (NP \ (PRP$ \ his) \ (NN \ country))) \ (PP \ (IN \ in) \ (NP \ (NNP \ reform.))))\)

  – Tree
Tokenization in Bikel

• Arabic parser

Diagram showing the tokenization process for verbs and nouns in Arabic, including conjunction, comp/tense marker, stem with affixes, object pronoun, preposition, definite article, stem with affixes, and genitive pronoun.
Tokenization in Bikel

• Arabic parser
  – Input sentence:
    الرئيس قاد بلده في الإصلاح
    The President let his country in reform.
  – Formatted sentence:
    • Alra\(\text{iysu}\) q\(\text{Ada}\) balad\(\text{ahu}\) fiy Al\(<\text{iSlaAHi}\)
    • Alra\(\text{iysu}\) q\(\text{Ada}\) balada- -hu fiy Al\(<\text{iSlaAHi}\)
    • Al\(\text{+ra}\)iys\(\text{+u}\) qAd\(\text{+a}\) balad\(\text{+a- -hu}\) fiy Al\(+<\text{iSlaAH+i}\)
    • \((\text{Al+ra}\text{iys+u qAd+a balad+a- -hu fiy Al+<iSlaAH+i})\)
Tokenization in Bikel

• Arabic parser
  – Output:
    $(S (NP (NN Al+ra{j}iys+u)) (VP (VBD qAd+a) (NP (NN balad+a-) (PRP$ -hu)) (PP (IN fiy) (NP (NN Al+<iSlaAH+i)))))$
  – Tree
Multiword Expressions
Multiword Expressions in XLE

- Three types of MWEs
  - Fixed Expressions: Lexically, morphologically and syntactically rigid. A word with spaces.
    - **New York**
    - **United Nations**
  - Semi-Fixed Expressions: Lexically, or morphologically flexible
    - **Sweep somebody under the rug/carpet**
    - **Transitional period(s)**
  - Syntactically-flexible Expressions
    - **to let the cat out of the bag**
    - **The cat was let out of the bag.**
Multiword Expressions

• MWEs are important
  – High frequency in natural language (30-40%)
  – Important for MT, literal translation is usually wrong
  – When taken as a block, they relieve the parser from the burden of processing component words
  – We collected 34,658 MWEs in addition to 45,202 Named Entities
The United States looks for Bin Laden.
Multiword Expressions in Bikel

• Compositional, yet detectable in the English treebank

(NP (DT the) (NNP United) (NNP Kingdom) )

(NP (NNP New) (NNP York) )

(NP (DT the) (NNP Middle) (NNP East) )

(NP (NNP Saudi) (NNP Arabia) )

(NP (NNP Las) (NNP Vegas) )

(NP (NNP Los) (NNP Angeles) )

(CONJP (IN in) (NN addition) (TO to) )
Multitword Expressions in Bikel

- Compositional, undetectable, sometimes inconsistent, in Arabic treebank

Los Angeles
(NP (NOUN_PROP luws)
  (NOUN_PROP >anojiliys))

United States
الولايات المتحدة
(NP (DET+NOUN+NSUFF_FEM_PL+CASE_DEF_NOM Al+wilAy+At+u)
  (DET+ADJ+NSUFF_FEM_SG+CASE_DEF_NOM Al+mut~aHid+ap+u))

The Middle East
الشرق الأوسط
(NP (DET+NOUN+CASE_DEF_GEN Al+$aroq+i)
  (DET+ADJ+CASE_DEF_GEN Al+>awosaT+i))

in addition to
((CONJP (NOUN+NSUFF_FEM_SG+CASE_INDEF_ACC <iDAf+ap+F) (PREP <ilaY))

(NP-ADV (NP (NOUN+NSUFF_FEM_SG+CASE_INDEF_ACC -*iDAf+ap+F)) (PP (PREP <ilaY)
(NP (NP (NOUN_PROP EarafAt)))}
Multiword Expressions in Bikel

• Example
The United States looks for Bin Laden.

الولايات المتحدة تبحث عن بن Laden

(S (NP (NNS Al+wilAy+At+u) (JJ Al+mut~aHid+ap+u)) (VP (VBP ta+boHav +u) (PP (IN Ean) (NP (NNP bin) (NNP IAdin))))))

[Diagram]

S
  /   |
 NP    VP
   /     /
 NNS JJ  VBP
   /     /
 Al+wilAy+At+u Al+mut~aHid+ap+u ta+boHav+u

States United looks

IN NP
  /   /
  Ean NNP
       /   /
      NNP NNP
        /   /
       Bin Laden

for
Statistical Parsing
Bikel Arabic Parser Evaluation

• Coverage of the statistical parser on sentence <= 40 words:
  – Arabic: 75.4%
  – Chinese: 81%
  – English: 87.4%

  (Bikel, 2004)

  – Arabic is “far below” the required standard.

  (Kulick et al., 2006)
Bikel Arabic Parser Evaluation

• Why Arabic performs poorly? (Kulick et al. 2006)
  – The ATB tag set is very large and dynamic, this is why they are mapped into 20 PTB tags. The tagset reduction is extreme and important information is lost.
  
  – Verb
    » IV3FS+IV+IVSUFF_MOOD:I
    » IV3MS+IV+IVSUFF_MOOD:J
    » PV+PVSUFF_SUBJ:3MS
    » IVSUFF_DO:3MP
  
  – Noun
    » NOUN+CASE_DEF_ACC
    » DET+NOUN+NSUFF_FEM_PL+CASE_DEF_GEN
    » NOUN+NSUFF_FEM_SG+CASE_DEF_GEN
Bikel Arabic Parser Evaluation

• Why Arabic performs poorly? (Kulick et al. 2006)
  – Average sentence length in Arabic is 32 compared to 23 in English
  – Significant number of POS tag inconsistencies, for example /ys is tagged as NEG_PART and PV
  – 5% of VP in Arabic have non-verbal heads
  – Base Noun Phrases (NPB) are 30% in English compared to 12% in Arabic.
  – Construct states in Arabic *roughly* correspond to possession constructions in English
Bikel Arabic Parser Evaluation

• Why Arabic performs poorly? (Kulick et al. 2006)
  – Arabic has a much greater variance in sentence structure than English.

<table>
<thead>
<tr>
<th>Sentence Type</th>
<th>Arabic %</th>
<th>English %</th>
</tr>
</thead>
<tbody>
<tr>
<td>VSO</td>
<td>62</td>
<td>0</td>
</tr>
<tr>
<td>SVO</td>
<td>17</td>
<td>90</td>
</tr>
<tr>
<td>No VP</td>
<td>19</td>
<td>11</td>
</tr>
<tr>
<td>Subjectless VP</td>
<td>2</td>
<td>0</td>
</tr>
</tbody>
</table>

• Major revision of Arabic treebank guidelines 08
Which is better?
Which is better?

• Common wisdom: statistical parsers are:
  – Shallow: They do not mark syntactic and semantic dependencies needed for meaning-sensitive applications

(Kaplan et al., 2004)
Which is better?

- XLE: “We parse the web.”
Which is better?

- Common wisdom is not entirely true.
- DCU: “We can also parse the web.”
Which is better?

- **Summary**
  - Handcrafted grammars are built on assumptions and intuitions. They depend on how good these assumptions are.
  - Handcrafted grammar can be improved by:
    - Effectively managing the development project
    - Making use of statistical facts (treebanks, and TIGERSearch)
Which is better?

– Statistical grammars are built on facts. They depend on how true these facts are.

– Statistical grammar can be improved by:
  • Improving the quality and size of treebanks.
Which is better?

• Statistical grammars are more efficient because:
  – there is a clear separation between the algorithm and the data structure
  – there is a clear division of labour, the linguists fight their battle, and the engineers fight their own battle
Spelling Checking and Correction
Spelling Checking and Correction

How frequent is spelling errors in news web sites?

Spelling errors in News Web Sites

- Agence France-Presse
- Xinhua News Agency
- Al Nahar
- Al Hayat
- Al-Quds Al-Arabi
- Al-Ahram
- Assabah
- Asharg Al-Awsat
- Ummah Press
- Al-Jazeera

Ratio %
Creating a wordlist

English: 708,125, French: 338,989, Polish: 3,024,852

<table>
<thead>
<tr>
<th></th>
<th>No. of Words</th>
<th>MS Accepted</th>
<th>MS Rejected</th>
</tr>
</thead>
<tbody>
<tr>
<td>AraComLex⁴</td>
<td>12,951,042</td>
<td>8,783,858</td>
<td>4,167,186</td>
</tr>
<tr>
<td>Arabic-Spell for Aspell (using Buckwalter)</td>
<td>938,977</td>
<td>673,875</td>
<td>265,103</td>
</tr>
<tr>
<td>1 billion-word corpus (Gigaword⁵ and Al-Jazeera)</td>
<td>2,662,780</td>
<td>1,202,481</td>
<td>1460,447</td>
</tr>
<tr>
<td>Ayaspell for Hunspell</td>
<td>292,464</td>
<td>230,506</td>
<td>61,958</td>
</tr>
<tr>
<td>Total*</td>
<td>15,147,199</td>
<td>9,306,138</td>
<td>5,841,061</td>
</tr>
</tbody>
</table>
Spelling Checking and Correction

Spelling error detection
1. Matching against a word list  
2. Character-based LM
Spelling Checking and Correction

Automatic correction of spelling errors

<table>
<thead>
<tr>
<th>Spell Checker</th>
<th>First order ranking</th>
</tr>
</thead>
<tbody>
<tr>
<td>MS Spell Checker</td>
<td>80.54%</td>
</tr>
<tr>
<td>Hunspell using Ayaspell</td>
<td>45.64%</td>
</tr>
<tr>
<td>Approach 1: Edit distance &amp; Noisy Channel</td>
<td>68.20%</td>
</tr>
<tr>
<td>Approach 2: Adding heuristics to Edit distance</td>
<td>71.3%</td>
</tr>
<tr>
<td>Approach 2 with post-processing</td>
<td>75%</td>
</tr>
</tbody>
</table>
Integration with Applications
Applications of Arabic Language Technologies

Structured Data
- Annotated, classified, information easily obtained

Unstructured Data
- Round-the-clock streaming, raw data, un-annotated

Language Resources
- Lexicons for single words, Named Entities, Multiword Expressions

Future Information Technologies
- Machine Translation
- Information Retrieval
- Question Answering

Focus on Arabic

> Heuristics
> Algorithms
> Testing
> Evaluation
> Validation