Workshop on Arabic Corpus Linguistics
11-12 April 2011, Lancaster University

Corpus linguistics resources and tools for Arabic lexicography

Majdi Sawalha and Eric Atwell
School of Computing,
University of Leeds, Leeds, LS2 9JT, UK
http://www.comp.leeds.ac.uk/sawalha
http://www.comp.leeds.ac.uk/eric
Outline

• Introduction
  – Oxford University Dictionaries
    • Monolingual Dictionaries
    • Bilingual Dictionaries
  – Arabic & Arabic NLP
    • Difficulties
    • Morphological analysis
  – Traditional Arabic lexicography
  – Constructing the Arabic broad-coverage lexical resource
  – Key notes
  – Conclusions
  – References
Oxford Dictionaries

• Searching for a word

• Dictionary entries are the lemmas of the word.
• Lemmatising of input words is needed to direct the search for
  the input word to the correct dictionary entry.

• What happens if the user entered a miss-spelled (unknown) word?

• Gives suggestions of similar words

No exact results found for "univercity" in the dictionary.

Did you mean university?
Did you mean inveracity?
Did you mean intercity?
Dictionary Entry

- Dictionary entry: Information provided

university

Pronunciation: /ˈjuːnəvɜrsetɪ/ noun (plural universities)

a high-level educational institution in which students study for degrees and academic research is done.

I went to university at the Sorbonne
his daughter is at university

Examples

the university of life
the experience of life regarded as a means of instruction

- That is acquiring the knowledge which you will do well to put into practice as you travel to the next stage of your journey through the university of life.

Phrases

Origin

Middle English: from Old French universite, from Latin universitas 'the whole', in late Latin 'society, guild', from universus (see UNIVERSE)

Origin

12/4/2011

Pronunciation

POS + Plural form

Meaning

Examples

Phrases

Meaning

Examples

Other Dictionaries

412/4/2011
What information can be provided by the Arabic dictionaries?

Dictionary entry

Position in dictionary

Meaning

Pronunciation

Lemma

Root

Pattern

Plural form

Examples

Using a suitable font that shows the letters and the diacritics
Using colours to illustrate clitics attached to the word

Phrases, Collocations, Idioms
(meaning and examples)

Origin (relation to traditional Arabic Dictionaries)

Morphological features
Detailed description of the morphological features of the word's morphemes

Related words list
A list of the words derived from the same root or words that have the same lemma

More on this word

university in other languages:
- German: Universität (f)
- Italian: università (f)
- Portuguese: universidade (f)
- Spanish: universidad (f)
- Japanese: 大学 (daigaku)
- Korean: 대학 (daehak)
- Polish: uniwersytet (m)
- Russian: университет (m)
- Chinese: 大学 (daxue)
- Thai: มหาวิทยาลัย (mahaviyadhi)
- Arabic: دانشگاه (danshghah)

See Language Web display
Find this word in Oxford Reference Online
university, n.

1. n. The whole body of teachers and scholars engaged, at a particular place, in giving and receiving instruction in the higher branches of learning; such persons associated together as a society or corporate body, with definite organization and acknowledged powers and privileges (esp. that of conferring degrees), and forming an institution for the promotion of education in the higher or more important branches of learning; also, the colleges, buildings, etc., belonging to such a body. In recent use, const. without article: at (or to) university, etc.

Sometimes, especially in former use, synonymous with college: see COLLEGE n. 4c.

St. Edmund in S. Eng. Leg. I. 438/256 So that he began at Oxenford of diuinite, So noble aed ... he vnuersite. Ibid. 439/278 He began so depe desputi of the trinite. Ita gret wonder me hadde [surfe] al the vnuersite. c1384 YWCLIF Wks. (1880) 157 He[ne] mensus lavis and worldly cleris statuts ben red in vnuersiteis. c1400 Rom. Rose 6/79 At Paris... he had. The accorde of the vnuersite, And of the puple as semeth me. c1423- (see COLLEGE n. 4). c1450 Godstowl Reg. 438 The house... that Robert of Stauntome held of the vnuersite of Oxenford. 1509 FISHER Funeral Sermon. Cress Richmond Wks. (1876) 301 The studendes of bothe the vnuersites... 1579 W. WILKINSON Confut. Fam. Love 40 They labour to put out the eyes of this land (the Vnuersites I meanes). 1644 MILTON Educ. 3 This place should be at once both Scholl and Universtite. 1663 LAWRENCE in Extr. St. Peters Friends Ser. ii. (1912) 126 University, Town and Country are far more active and vigilant then before. 1792 LITTREL Brief Rel. (1857) V. 149 A patent... for founding an university..., to be called king Williams universtite. 1745 BALLOT Examen. Collug. (1732) 259 Are you going to Louvain to see the University? 1785 J. ADAIR Wks. (1854) IX. 550 He an is anxious to study some time at your university [= Harvard College] before he begins the study of law. 1840 CARLYLE Heroes v. (1838) 305 Universities are a notable, respectable product of the modern ages. 1856 STANLEY Sina & Pal. x. 364 The great Jewish university...
Traditional Arabic Lexicography

• Arabic lexicography is one of the original and deep-rooted arts of Arabic literature.

• The first lexicon constructed was *kitāb al-‘ayn* كتاب العين 'ayn lexicon’ by al-farāḥīdī (died in 791).

• Over the past 1200 years, many different kinds of Arabic language lexicons were constructed; these lexicons are different in ordering, size and aim or goal of construction.

• Many Arabic language linguists and lexicographers studied the construction, development and the different methodologies used to construct these lexicons.
Ordering lexical entries in the Arabic lexicons

- **al-
halīl** methodology [5 lexicons]
  - Listed the lexical entries based on the pronunciation of the letters starting from the farthest in the mouth to the nearest

- **abī‘ubayd** methodology [3 lexicons]
  - Listed the lexical entries based on similarity in meaning.

- **al-ḡawharī** methodology [4 lexicons]
  - Listed the lexical entries based on last letter of the word.

- **al-barmakī** methodology [11 lexicons]
  - Listed the lexical entries alphabetically.

- **Modern dictionaries** uses a combination of root/word as lexical entries arranged alphabetically.
تاَجُ العُرُوس من جِوُاهِرِ الْقَافُوس

الجَّزيرة

تَجَلَّىُ الفَتْرَةُ، وَفِي الْفَرْقُوتَاتِ لِلْمَرْأَبِ وَتَيَّاَةِ الْمَسَطْحَةِ في الْمَفْتَرَةِ.

الجَّزيرةُ، ضَمَّ الْعَلَمِ، وَتَقْرِيبُ بَعْضِهِ مِنْ بَعْضِهِ.

تَجَلَّىُ الفَتْرَةُ، وَفِي الْفَرْقُوتَاتِ لِلْمَرْأَبِ وَتَيَّاَةِ الْمَسَطْحَةِ في الْمَفْتَرَةِ.
The use of Corpora in building dictionaries – Example 1

Selecting examples from concordance lines

Lexicographers
### Lemmatized Arabic Internet Corpus

[http://corpus.leeds.ac.uk/query-ar.html](http://corpus.leeds.ac.uk/query-ar.html)

#### Querying Arabic Corpora

<table>
<thead>
<tr>
<th>Arabic Internet</th>
<th>Al Hayat News</th>
<th>Arabic Wikipedia</th>
<th>Corpus of Contemporary Arabic</th>
<th>Computer Science corpus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Arabic legal texts</td>
<td>v2</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Submit Query**  **Reset**  **CQP syntax only (Examples)**  **Click here for getting help on the query interface**

**Back to the query window**

#### Set parameters of your query

- **Concordance**
  - Context: 50 (`c for characters, w for words`)
  - Sort by: Document  Frequency  lemma  word
  - Then by: left  right
  - Output: 100 lines

- **Collocations**
  - Collocation scores: Mutual Information  t-score  Loglikelihood
  - Context: 5 words on the left  5 words on the right
  - POS tag of the collocate: POS tags

I also created the lists of the most frequent word forms in Internet, LDC, W

After lemmatisation done by [Mr. Sohail Salama](http://www.corpus.leeds.ac.uk) there is also the frequency list c

The corpora are:

1. The Internet corpus was compiled using the procedure described in
3. The Wikipedia corpus — from the public Wiki data retrieved on July
4. The Arabic Legal Corpus — from keywords collected by Hanem El
5. The Arabic Legal corpus of Arabic — from keywords collected by

The interface was developed by Serge Sharoff, contact me at s.sharoff@leeds.ac.uk

12/4/2011

#### See 100 examples of 'جامعة' cut 100' in I-AR-LEMMAT

```
الجامعة من المؤكد أن تكون ذات معنى مع وحيدة جميرة، في نفسها.
الجامعة من المؤكد أن تكون ذات معنى مع وحيدة جميرة، في نفسها.
الجامعة من المؤكد أن تكون ذات معنى مع وحيدة جميرة، في نفسها.
الجامعة من المؤكد أن تكون ذات معنى مع وحيدة جميرة، في نفسها.
```

I also corrected the lists of the most frequent word forms in Internet, LDC, W

After lemmatisation done by navigation| Mr. Sohail Salama| there is also the frequency list c

The corpora are:

1. The Internet corpus was compiled using the procedure described in
3. The Wikipedia corpus — from the public Wiki data retrieved on July
4. The Arabic Legal Corpus — from keywords collected by Hanem El
5. The Arabic Legal corpus of Arabic — from keywords collected by

The interface was developed by Serge Sharoff, contact me at s.sharoff@leeds.ac.uk

12/4/2011

#### See 100 examples of 'جامعة' cut 100' in I-AR-LEMMAT

```
الجامعة من المؤكد أن تكون ذات معنى مع وحيدة جميرة، في نفسها.
الجامعة من المؤكد أن تكون ذات معنى مع وحيدة جميرة، في نفسها.
الجامعة من المؤكد أن تكون ذات معنى مع وحيدة جميرة، في نفسها.
الجامعة من المؤكد أن تكون ذات معنى مع وحيدة جميرة، في نفسها.
```

I also corrected the lists of the most frequent word forms in Internet, LDC, W

After lemmatisation done by [Mr. Sohail Salama](http://www.corpus.leeds.ac.uk) there is also the frequency list c

The corpora are:

1. The Internet corpus was compiled using the procedure described in
3. The Wikipedia corpus — from the public Wiki data retrieved on July
4. The Arabic Legal Corpus — from keywords collected by Hanem El
5. The Arabic Legal corpus of Arabic — from keywords collected by

The interface was developed by Serge Sharoff, contact me at s.sharoff@leeds.ac.uk

12/4/2011
The use of Corpora in building dictionaries – Example 2

university

college, academy, educational establishment/institution, institute, varsity

* North American school
* historical polytechnic

View dictionary entry »

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Score</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>college</td>
<td>0.504</td>
<td>207476</td>
</tr>
<tr>
<td>institution</td>
<td>0.416</td>
<td>190572</td>
</tr>
<tr>
<td>school</td>
<td>0.41</td>
<td>981325</td>
</tr>
<tr>
<td>University</td>
<td>0.371</td>
<td>461863</td>
</tr>
<tr>
<td>organization</td>
<td>0.34</td>
<td>416383</td>
</tr>
<tr>
<td>agency</td>
<td>0.339</td>
<td>215584</td>
</tr>
<tr>
<td>hospital</td>
<td>0.326</td>
<td>276588</td>
</tr>
<tr>
<td>department</td>
<td>0.324</td>
<td>203022</td>
</tr>
<tr>
<td>education</td>
<td>0.32</td>
<td>357278</td>
</tr>
<tr>
<td>community</td>
<td>0.319</td>
<td>672366</td>
</tr>
<tr>
<td>chub</td>
<td>0.313</td>
<td>360545</td>
</tr>
<tr>
<td>council</td>
<td>0.312</td>
<td>263624</td>
</tr>
<tr>
<td>church</td>
<td>0.31</td>
<td>279459</td>
</tr>
<tr>
<td>firm</td>
<td>0.309</td>
<td>212234</td>
</tr>
<tr>
<td>industry</td>
<td>0.302</td>
<td>434801</td>
</tr>
<tr>
<td>business</td>
<td>0.298</td>
<td>845818</td>
</tr>
<tr>
<td>union</td>
<td>0.295</td>
<td>223911</td>
</tr>
<tr>
<td>society</td>
<td>0.293</td>
<td>346017</td>
</tr>
<tr>
<td>government</td>
<td>0.293</td>
<td>1199449</td>
</tr>
<tr>
<td>company</td>
<td>0.293</td>
<td>1185039</td>
</tr>
<tr>
<td>bank</td>
<td>0.291</td>
<td>278006</td>
</tr>
<tr>
<td>student</td>
<td>0.287</td>
<td>775851</td>
</tr>
<tr>
<td>authority</td>
<td>0.287</td>
<td>347240</td>
</tr>
<tr>
<td>sector</td>
<td>0.284</td>
<td>197731</td>
</tr>
</tbody>
</table>
The use of Corpora in building dictionaries – Example 2, cont.

- Frequency lists /measurements can compare between two words, and find how these words are related to each other.
Collocations example from the Arabic Internet Corpus

**Corpus: I-AR-LEMMA; Tokens: 193842936**

| Query: [word="المصطلحات"]  |
| Colloc: left=0, right=1; Filter: |

<table>
<thead>
<tr>
<th>LL score</th>
<th>LL</th>
<th>Collocation</th>
<th>Joint Freq</th>
<th>Freq2</th>
<th>LL score</th>
<th>Concordance</th>
<th>Examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>جامع محلة</td>
<td>154 77355</td>
<td>9982 473.49</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>219 77355</td>
<td>152191 286.27</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>209 77355</td>
<td>140791 276.40</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>133 77355</td>
<td>22059 267.59</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>94 77355</td>
<td>15693 188.70</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>105 77355</td>
<td>31312 180.57</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>119 77355</td>
<td>57162 176.73</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>131 77355</td>
<td>111277 158.22</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>89 77355</td>
<td>24380 165.75</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>50 77355</td>
<td>1501 143.41</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>76 77355</td>
<td>30729 110.21</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>24 77355</td>
<td>45 105.88</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>95 77355</td>
<td>102740 103.58</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>61 77355</td>
<td>21642 99.58</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>38 77355</td>
<td>2264 95.90</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>41 77355</td>
<td>3827 94.14</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>42 77355</td>
<td>38066 84.03</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>59 77355</td>
<td>47107 72.89</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>48 77355</td>
<td>38441 59.21</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>47 77355</td>
<td>36262 58.82</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>41 77355</td>
<td>342 54.11</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>39 77355</td>
<td>16253 42.70</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>جامع محلة</td>
<td>25 77355</td>
<td>8201 41.75</td>
<td>Examples</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

See 94 examples of 'MU(meet [lemma='جامع' [lemma='متحدة' cut 100' in I-AR-LEMMA
Oxford bilingual dictionaries

Examples from parallel corpora

- university
  - A. n
    - POS: English
    - université (f)
  - B. modif
    - [lecturer] d’université; [degree / town] universitaire; [place] à l’universite
    - university
      - entrance = entrée (f) à l’université
      - university
        - education = formation (f) universitaire

Other languages:
- German: Universität
- Italian: università
- Portuguese: universidade
- Spanish: universidad
- Japanese: 大学
- Korean: (종합) 대학교
- Polish: uniwersytet
- Russian: universitet
- Chinese: 综合性大学
- Thai: มหาวิทยาลัย
- Arabic: جامعة

See Language Web display
Find this word in Oxford Reference Online
Google translation

University

Dictionary

Related phrases

Google translation
Oxford bilingual dictionaries

• Do users need to know the meaning of word in many different languages?
• Users need to translate from one language to another.
• Connecting terms with a central language (English) can connect two languages together.
• Specific linguistic information to each word from their monolingual dictionaries can be provided, in addition to information of the central language (English)
Arabic & Arabic NLP

• 200 million people speaking Arabic as first language.
• More than 1 billion Muslims need Arabic to recite the Quran (the holy book of Muslims).
• One of the UN official languages.
• Increased potential for learning Arabic recently.
• *Many commercial software companies invest in Arabic NLP.*
Why is Arabic NLP difficult?

- **Complex morphology**
  - Words consist of multi-morphemes of 5 kinds
  - Proclitics, prefixes, stem/root, suffixes, enclitics.

- "وَسَيْ كُتُبْ عَنْهَا [wasayaktubūnhā]" (And they will write it)

<table>
<thead>
<tr>
<th>Conjunction</th>
<th>Particle of futurity</th>
<th>Progressive letter</th>
<th>Root / Stem</th>
<th>Relative Pronoun (Plural/Subject)</th>
<th>Relative Pronoun (Object)</th>
</tr>
</thead>
<tbody>
<tr>
<td>وَ</td>
<td>سَ</td>
<td>يَ</td>
<td>كُتُبْ</td>
<td>وُنَ</td>
<td>هَ</td>
</tr>
</tbody>
</table>

- **Vowels & Diacritic Marks**
  - 3 long vowels (ا alif, و wāw, ي yā’)
  - 3 short vowels (فتا'ha', داممαha', كسرαh.)
  - Other diaractics: sukūn ـ، Šadda ـّ، Tanwīn (ٌ، ٍ، ُ)
  - ta‘īvīl character (ـ).
  - همزةٜ(إ) tā’ marbūtaٞ(ء) and هاٞ(ه) ، ياٞ(ي) and alif maqṣūraٞ(ى) ， and maddaٞ(ا)
Morphological Analyses of Arabic text

- Morphological analysis is essential for processing text corpora and building dictionaries.
- Existing Arabic morphological analyzers failed to achieve accuracy rates more than 75%. (Sawalha & Atwell, 2008)
- We can not rely on such analyzers for further analysis such as part-of-speech tagging and parsing.

Morphological Analyzer for Arabic text

Step 1: Tokenization
- Different text types
- Spell-checking

Step 2: Function words

Step 3: Clitics, Affixes & Stems

Step 4: Root/Lemma extraction

Step 5: Pattern generation

Step 6: Vowelization

Step 7: Assigning detailed morphological features tags for each of the word’s morphemes
Example of Analyzed word

*wa al-ğāmiʿāt* (And the universities)

<table>
<thead>
<tr>
<th>Feminine plural letters</th>
<th>stem</th>
<th>Definite article</th>
<th>conjunction</th>
</tr>
</thead>
<tbody>
<tr>
<td>جَامِعَاتُ</td>
<td>جَامِعَ</td>
<td>الْ</td>
<td>وَ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Lemma</th>
<th>Pronunciation</th>
<th>Root</th>
<th>Meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>جَامِعَةٌ</td>
<td>جَامِعَ</td>
<td>جَامِعَةٌ</td>
<td>جَامِعَةٌ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Pattern</th>
<th>Plural form</th>
<th>Examples</th>
<th>Phrases, collocations, idioms</th>
<th>Origin (links to traditional Arabic lexicons)</th>
</tr>
</thead>
<tbody>
<tr>
<td>فَاعِلَةٌ</td>
<td>جَامِعَاتِ</td>
<td>جَامِعَةٍ جَامِعَةٍ جَامِعَةٍ جَامِعَةٍ</td>
<td>جَامِعَةٍ جَامِعَةٍ جَامِعَةٍ</td>
<td>جَامِعَةٍ جَامِعَةٍ جَامِعَةٍ</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word’s morphemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>وَ</td>
</tr>
<tr>
<td>الْ</td>
</tr>
<tr>
<td>جَامِعَ</td>
</tr>
<tr>
<td>اتُ</td>
</tr>
</tbody>
</table>

http://www.comp.leeds.ac.uk/sawalha/tagset.html
Samples of traditional Arabic lexicons

كتب: الكتاب: معروف، والجمع كتب. كَتَبَ الشيء، يَكُتِبَ كُتِبًاً وَكِتَابًا. كَتَبَهُ رَجُلٌ، كَتَبَهُ الشيء، يَكُتِبَهُ كُتِبًّاً وَكِتَابًّا وَكِتَابًا. كَتَبَهُ خَطًّا، قال أبو النجم: أُقبلْتُ من عِندَ زِبَادٍ كَاحِرٍ، تَخْطُ رَجُلاً خَطًّا مُخْلِفٍ. تَكُتِبُانِ في الْطَرِيقِ لَمَّا أَلَفَ فَقُولُون: وَرَأَيْتُ في بَعْضِ النَسْخَ تَكُتِبَانِ، كَبْسُ النَّاطِرِ، وَهِيْ لَغةُ بَهْرَاءٍ، يُكُسَرُونَ النَّاطِرَ، فَيَفْقُولُون: تَعْلَمُونَنَّ، ثُمَّ أَنْبُثُ الْكَافِ كَسْرَةُ النَّاطِر. والكَتَابُ أَيْضاً: الاسم، عن الْحَيَاةَ الأُزْهَرِيَّةِ: الْكَتَابُ أَسْمَ أَلْمَا كَتِبَ مَجْمُوعًاٍ، والكَتَابُ مَصْدَرٌ، والكِتَابَةُ لَمْ يَكُنْ لهَ صِنَاعَةٌ، مثل الْسِّيَاغَةَ والْخَيَاطَةِ. والكِتَابَةُ: اكْتَبَكَانِ كِتَابًاٍ تَسْخَحَهُ. وَيَقُولُ: اكْتَبَ فَلَانًا فَلَا أُأَيْ سَأَلَهُ أَنْ يَكُتِبَ لهَ كِتَابٌ فِي حَاجَةٍ. وَإِلَّا كَتَبَ لَهُ، قُولُ: كَتَبَ كُتَابَهُ، وَيَقُولُ: كَتَبَهُ حُجُّهُ، وَاكْتَبَهُ: اكْتَبَهُ. وَكَذِلِكَ اكْتَبَ. وَكاَتِبَهُ كَتَابًا، وَاكْتَبَهُ كَتَابًا. وَفِي التَّزِيلِ العَزِيزِ: اكْتَبَهُا فِي تَمْلَى عَلَيْهِ بَكْرَةً وأَصْيَلًاٍ: أَيْ اكْتَبَهُا. وَيَقُولُ: اكْتَبَ الْرَجُلُ إِذَا كَتَبَ نَفْسَهُ فِي دِوَانِ السَّلَطَانِ ...
Constructing the Arabic broad-coverage lexical resource

Analyzing Lexicon’s Text Separately

- Convert each lexicon text into a unified format.

A bag of words is extracted from the definition text.

- A normalization analysis that verifies the word-root pairs by applying linguistic knowledge.
The first 60 lexical entries of the root كتاب k-t-b ‘wrote’ stored in the broad-coverage lexical resource.
Combining into One Broad-Coverage Lexical Resource

- After analyzing each lexicon, a combination algorithm is applied to construct the broad-coverage lexicon.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><em>lisūn al-‘rab</em></td>
<td>207,992</td>
<td>207,992</td>
<td>100.00%</td>
<td>47.80%</td>
</tr>
<tr>
<td>2</td>
<td><em>mu‘gām al-muḥtar al-luğaṭ</em></td>
<td>74,507</td>
<td>61,113</td>
<td>82.02%</td>
<td>14.04%</td>
</tr>
<tr>
<td>3</td>
<td><em>tağ al-‘arūs min ḡawāhir al-qāmūs</em></td>
<td>128,119</td>
<td>95,415</td>
<td>74.47%</td>
<td>21.93%</td>
</tr>
<tr>
<td>4</td>
<td><em>muḥtar al-ṣ-ṣāhā</em></td>
<td>19,540</td>
<td>16,573</td>
<td>84.82%</td>
<td>3.81%</td>
</tr>
<tr>
<td>5</td>
<td><em>al-muğrab ṣīḥartub al-mu‘rab</em></td>
<td>12,396</td>
<td>9,805</td>
<td>79.10%</td>
<td>2.25%</td>
</tr>
<tr>
<td>6</td>
<td><em>kitābu al-‘ayn</em></td>
<td>30,292</td>
<td>18,878</td>
<td>62.32%</td>
<td>4.34%</td>
</tr>
<tr>
<td>7</td>
<td><em>al-mu‘gām al-wasīf</em></td>
<td>36,660</td>
<td>25,364</td>
<td>69.19%</td>
<td>5.83%</td>
</tr>
<tr>
<td></td>
<td><strong>Totals</strong></td>
<td><strong>509,506</strong></td>
<td><strong>435,140 [C]</strong></td>
<td><strong>85.40%</strong></td>
<td><strong>100.00%</strong></td>
</tr>
</tbody>
</table>
The Corpus of Traditional Arabic Lexicons

- Lexicons’ text can be used as a corpus of traditional Arabic lexicons.
- **Different domain** than existing corpora.
- The Arabic corpus of dictionaries **covers a period of more than 1200 years.**
- Consists of **large number of words and word types.**
- Has both **vowelized and non-vowelized text.**

<table>
<thead>
<tr>
<th>Number of files</th>
<th>247</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Size</strong></td>
<td>178.32 MB</td>
</tr>
<tr>
<td><strong>Vowelized words analysis</strong></td>
<td></td>
</tr>
<tr>
<td># of words</td>
<td>14,369,570</td>
</tr>
<tr>
<td># of word types</td>
<td>2,184,315</td>
</tr>
<tr>
<td><strong>Non-vowelized word analysis</strong></td>
<td></td>
</tr>
<tr>
<td># of words</td>
<td>14,369,570</td>
</tr>
<tr>
<td># of word types</td>
<td>569,412</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Partially-vowelized</th>
<th>Non-vowelized</th>
</tr>
</thead>
<tbody>
<tr>
<td>Word</td>
<td>Frequency</td>
</tr>
<tr>
<td>في</td>
<td>292,396</td>
</tr>
<tr>
<td>من</td>
<td>269,200</td>
</tr>
<tr>
<td>قال</td>
<td>172,631</td>
</tr>
<tr>
<td>و</td>
<td>120,060</td>
</tr>
<tr>
<td>على</td>
<td>108,252</td>
</tr>
<tr>
<td>ما</td>
<td>89,195</td>
</tr>
<tr>
<td>وقال</td>
<td>88,233</td>
</tr>
<tr>
<td>عن</td>
<td>82,027</td>
</tr>
<tr>
<td>إذا</td>
<td>81,479</td>
</tr>
<tr>
<td>أي</td>
<td>78,622</td>
</tr>
<tr>
<td>وهو</td>
<td>75,149</td>
</tr>
</tbody>
</table>

12/4/2011
Key notes

• When do users start using dictionaries?
• What are the users’ needs of monolingual or bilingual dictionaries?
• What is the best arrangement methodology for Arabic dictionaries?
• What are the technologies needed to search for a word in dictionary?
• What linguistic information that can be provided in a dictionary for each dictionary entry?
• How is dictionary entry connected to other dictionary entries?
• How are examples of the dictionary entry selected?
• What are the phrases, collocations, idioms that are related to dictionary entry?
Conclusions

• Morphological analysis is essential technology for constructing dictionaries.
• The Arabic morphological analyzer provides detailed information about the processed words.
• Traditional Arabic lexicons are important because they provide the correct linguistic information about the words, examples from the Qur’an and poetry, collocation used in the past 1200 years, etc...
• Arrangement of Arabic dictionary entries is a challenge for Arabic lexicography.
• Three types of dictionaries can be provided according the user needs; Modern Arabic dictionary, traditional Arabic dictionary, and a dictionary of dialects.
References


Thank you