Assessing Word-form based Search for Information in Arabic: Towards a New Type of Lexical Resource

Mouna Anizi¹, Joseph Dichy²

¹ Université Lumière-Lyon 2 & ICAR (UMR 5191-CNRS/Lyon 2); École Nationale Supérieure des Sciences de l’Information et des Bibliothèques (ENSSIB), Lyon, France
anizi@enssib.fr

² Université Lumière-Lyon 2 & ICAR (UMR 5191-CNRS/Lyon 2), Lyon, France
joseph.dichy@univ-lyon2.fr

Abstract

Albeit real progress has been made during the last two decades, finding or retrieving information in Arabic with the help of a search engine remains difficult, owing to the high level of ambiguity entailed by the structure of ‘unvowelled’ Arabic writing. These language-specific difficulties are brought to a peak in the case of queries based on single words. The contribution analyses the results of search queries on Google, which are compared to the results of word-form analyses obtained both on the ArabiCorpus site (http://arabicorpus.byu.edu/) and with analysers based on the DIINAR.1 lexical resource (references at http://diinar.univ-lyon2.fr). An assessment protocol is proposed. Clearly, it aims at evaluating neither the analysers of ArabiCorpus and DIINAR.1, nor the Google search engine. Examining the latter is quite another question, related to Google ranking, and speed. The aim of the paper, instead, is to explore and assess the possibilities and limitations of word-form based queries in Arabic, i.e. the result of queries obtained with word-form based analysers and language resources (what can be obtained and what strictly speaking cannot). The protocol includes (a) comparing results obtained through Google with the often numerous word-forms obtained through the two other sources, (b) considering a number of semantic aspects related to the contexts query words appear in, and (c) taking into account word-before/word-after collocations and set phrases. It eventually introduces essential features of a new type of lexical resource for future Arabic search engines, which needs to contain, among other components: (a) a compact and comprehensive database operating at word-form level, such as DIINAR.1, and (b) an extended lexical resource that includes semantic relations, collocations and set or semi-set expressions.

1- Introduction

This paper is both a continuation of previous works on the assessment of Arabic language resources and software [Dichy, 2004; 2005], and a contribution to the BLARK (Basic Language Resources Kit) concept in Arabic [Krauwer et al., 2004]. We focus here on resources needed in retrieving, or searching for information in Arabic. The task remains difficult, owing to the fact (a) that high level language resources similar to those found, for instance, in English or French, are still missing in Arabic, in spite of recent advances [Nikhou & Choukri, 2004], and (b) that the structure of Arabic writing is both agglutinative and ‘unvowelled’ [Dichy, 1997]. Traditionally ‘unvowelled’ Arabic script is known to generate a high level of ambiguities [Dichy, 1990]; additional ambiguities are associated with the Arabic language in newspapers ([Buckwalter, 2004], [Abbes, 2004], [Abbes & Dichy, 2008a]) and on the Web [Hassoun, Dichy & Abbes, 2008].

In order to tackle the question of how these difficulties affect search results, we present a short experiment of information retrieval in Arabic, conducted on the current Google search engine. The protocol of the experiment includes comparing the output of Google searches with results obtained with lexical queries using the ArabiCorpus site (http://arabicorpus.byu.edu/) and the DIINAR.1 lexical resource (http://diinar.univ-lyon2.fr [Dichy, Braham, Ghazali & Hassoun, 2002], [Dichy & Hassoun, 2005]).

The reader should note, for the sake of clarity, that, in this work, we do not endeavour to assess any of these tools, i.e. neither the Google Arabic search engine, nor the underlying lexical database and morphological analyser of ArabiCorpus or DIINAR.1. This is, in fact quite another discussion, which involves comparing Google ranking and statistic approaches on the one side, and, on the other, considering the frequency of occurrences in a given corpus as well as morphological analysis based on rules and on grammar-lexis information drawn from a lexical resource. This could be the matter of another work, which would include speed parameters, discussions around the Google ranking approach, and other linguistic and semantic aspects. The question of the optimisation of morphological analysers drawing on lexical resources such as DIINAR.1 is no easy matter when it comes to such results as the hundreds of thousand answers obtained in split seconds through Google. The aim of the paper is, instead, to explore and evaluate the possibilities and limitations of word-form based information queries in Arabic. The object of the assessment is what can be done with single word queries, and what, strictly speaking, cannot. The contribution eventually aims at highlighting the need for a new type of lexical resource.

First, we introduce in section 2 the protocol of the experimental procedure followed. The section includes a short recall of the structure of the Google Arabic search.
engine, as well as the ArabiCorpus site, and the DIINAR.1 resource. Second, we present, in section 3, a small number of actual queries, and analyse the results obtained. Third, we outline, in section 4, some features of the new lexical resource that is needed.

2- The Assessment Protocol

The testing protocol consists of the following procedural steps:

Step 1: Google Arabic single word search
– Perform a simple search on Google. The query only contains Arabic strings.
– Observe the results obtained and detect difficulties (flaws, if any, and misses). Try and categorize them according to the problem encountered. On the opposite, note effective results.

Step 2: Considering word-form variation related to a given lemma (with ArabiCorpus and DIINAR.1)
– Perform another search using the same queries (words) on the ArabiCorpus site (BYU) and also in the DIINAR.1 resource.
– Compare results found in step 2 with those obtained with Google. The frequency of Arabic word-forms obtained through ArabiCorpus should be taken into account.

Step 3: Word-before/word-after contextualization
– Identify “frozen” or set expressions or terms (that include two words or more). On the ArabiCorpus site, consider very basic collocations and phrases, using the ‘word before/word after’ function, only taking into account the higher frequency collocations. Identify, in the step 1 results of the Google search (single word queries), the most salient set phrases including two words or more.
– Go back to Google and perform a new set of searches based on frequent ‘word before’ or ‘word after’ collocations obtained through the ArabiCorpus site (using the Google quotation marks convention). These new queries focus on contexts.
– Evaluate results. List, whenever found, lacks in the contextualized research on Google and propose an analysis. Also consider results that appear on Google and that are not found using ArabiCorpus or the DIINAR related tools.

2.1- Google, a short recall

The Google search engine is based on a method called “pagerank calculation”. The ‘pagerank’, as one knows, is a quotation from 0 to 10, reflecting the popularity of a site. The more a site receives links from other sites, the higher its quotation. The calculation of the ‘pagerank’ value is conducted through what is called the ‘Google Dance’. During this phase, a computer robot roams all web pages in Google's indexes, counting and comparing the number of links pointing to each site [Peyronnet, 2007]. The diagram and the descriptive recall below show how the search engines manage to answer queries with both speed and precision.

<table>
<thead>
<tr>
<th>Phase 1</th>
<th>Phase 2</th>
<th>Phase 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Web crawl and storage of pages found</td>
<td>Indexation and feeding of the Google databases</td>
<td>Sorting and presentation of the results of the search queries</td>
</tr>
</tbody>
</table>

Table 1: Overall structure of the Google search engine.

In the first phase a robot, called "bot" or "spider" operates. Its program roams the web continuously, in order to feed and update a database that includes:
– the address of every page found;
– a description of the contents of pages (title, text, meta-tags, names of pictures, images, etc.);
– a list of links between every page and other pages.

The Google engine is powered by thousands of robots operating continuously on thousands of computers around the world. Every time a page contains a link to another page, the robot, once users have finished reading, jumps to the linked page and continues its work.

The second phase is the building of the index. Google's computers permanently process the content of pages found, in order to generate indexes that will enable it to find almost instantly the result of a given query among billions of pages.

The third and final phase is the website opened by users connecting, for instance, to www.google.com or www.google.fr. There are over a hundred sites (also called ‘DataCenters’) throughout the world. Each contains a copy of the index of all the pages liable to respond to a given query. This allows each distribution centre to remain successful despite the large number of users of the system (see http://www.rankspirit.com).

2.2- ArabiCorpus, short presentation

ArabiCorpus is a site (http://arabicorpus.byu.edu/) offering query words in Arabic for the purpose of lexical analyses. It belongs to the Brigham Young University (BYU) in the state of Utah. The corpus includes over 68 million words, belonging, mainly, to newspapers from different Arabic countries and to contemporary literature (novels, essays). It also includes the Koran, a few medieval science treatises, A Thousand and One Nights, etc. The site presents users, on a freeware basis, with the results of a concordance software operating with a language resource, “the lookup items” of which are “based loosely on (but are quite different from) the dictionary files in Buckwalter's Morphological Analyser” (D. Parkinson, site information pages).

The entry of queries can be either words, for which the user is asked to specify the lexical category (name, adjective, verb, adverb), or strings of characters, which allows entering set phrases (variation in set phrases require repeated queries).

The result of a given query appears on a number of pages, which include on the whole:
– the number of occurrences per 100,000 words in the part of the corpus selected by the user (newspaper, novel, etc.);
– the KWIC ('Key Words In Context') concordance related to each entry, with about 15 words before and after, and corpus location information;
– the sub-sections of the selected corpus where the results were found;
– the word-forms in which the subject of the request is accompanied by a frequency index;
– the ‘word/word after’ of the entry, associated with their number of occurrences.

Links to either the concordance results or to the paragraph the query word appears in are offered. There are other features, but we need here to cut short.

2.3- The DIINAR.1 lexical database and the related word-level analysers, a quick outline

DIINAR.1 (‘Diccionnaire Informatisé de l’ARabe, version 1’), Arabic acronym for mu’jam al-arabiyiya l-’alîyya (www.elda.org) encompassing 19,457 verbs, 70,702 deverbal entries (verbal nouns, active and passive participles, ‘analogous’ adjectives, nouns of time and place), 39,099 nominal stems, around 150 tool-words and a prototype of 1,384 proper names. Each entry is associated with word-level morpho-syntactic specifiers ensuring ‘legal’ grammar-lexis relations between the lexical basis of a given word-form and other word-formatives [Dichy, 1997]. This means that the lexica generated by combining lemmas and affixes and/or clitics are constrained, and only include forms that effectively exist in the language. For instance, the lemma kawkab ‘celestial object’ (latin astrum), to write’, can be followed by the clitic complement pronoun hu ‘it’ (or ‘him’, in a metaphoric use, meaning ‘he wrote his name’), as opposed to nazala ‘to go down’, which cannot be associated with a complement pronoun (kataba is transitive and nazala is not). Verbal entries can be conjugated, and nominal ones, subjected to declension.

The total amount of minimal words (i.e. of lemmas with their prefix and suffixes) generated from the database is 7,774,938 [Abbes, Dichy & Hassoun, 2004; 2005]. All these forms are existing words (as indicated above, the resource is not based on the generation of purely virtual forms; these would amount to over 65 or 70 million forms!) DIINAR.1 is available at ELRA/ELDA (www.elda.org).

One of the main tools based on the DIINAR.1 lexical resource is the AraConc concordance software, the output of which can be considered as a triple: the word-form, its analysis and its position in the corpus, and the MorphArab word-form analyser [Abbes, 2004], which shares with the other morphological analysers and generators based on DIINAR.1 ([Zaafrani, 2002], [Ouersighni, 2001]) the functions of:
– segmentation and analysis of word-forms into lower-level formats, and
– identifying the word-forms that belong to the language.

The results of Google queries and the interrogation of the ArabicCorpus site will also be compared to the contents of the generated lexic of DIINAR.1.

3- The Results of Queries

The protocol presented at the beginning of section 2 is applied below to queries based on the single word, kawkab ‘celestial object’, kawkab ‘star’, ‘planet’. This noun is related to a group of notions, and features a high level of polysemy, including a number of metaphorical uses that appear to be lexically coded. Its plural form is built through ‘internal derivation’ (i.e. a change in the morpho-syllabic pattern, known as ‘broken plural’ – [Zaafrani, 2002], i.e., respectively: kawwakib and ‘amwal ‘of time and place’. In addition, various types of collocations and set phrases can be found. These set phrases are – on the whole – of the <noun+adjective> or <noun+noun> (‘construct state’, ‘addition’) structure, and present interesting semantic features.

3.1- Step 1: Google Arabic single word search

The results for the kawkab query on Google yielded 39,700,000 results¹. In the first most relevant 443 results, kawkab appears as an isolated word, except for a very few occurrences of al-kawkab (with the clitic definite article –al).

Here are the most prominent search results obtained through Google. Meanings – which have been checked in their context whenever needed – are exemplified with a significant excerpt or two:

First results:

kawkab = ‘celestial object’ (latin astrum).

“The sun”

“kawkab al-anam”

Second results:

kawkab = the ‘planet’ a person is said to originate from (metaphorical use).

“The player who came from another planet.”

The phrase refers to the football player Maradona, considered in this context as talented enough to be described as extraterrestrial.

Similar result:

“She has no feather on her head, neither is she from another planet.”

The two phrases included in the sentence above mean that the person in consideration is just ordinary (with nothing special about her).

Third results:

kawkab = a ‘planet’, meaning ‘a world’ or ‘a universe’ (metaphorical use).

“The planet heart”

The expression, which is borrowed from a magazine, refers to the world of feelings and love.

Fourth results:

kawkab = a ‘star’, metaphorical use referring to a ‘radiant beauty’ (actress, singer, etc.).

“He has the face of a star”

This comparison refers to beauty (in English, the phrasing would go: “his/her face is radiantly handsome” or “beautiful”).

¹ All the figures for the Google queries refer to the first week of April 2009.
Fifth results: kawkab = ‘a star’, metaphorical use referring to a ‘star’ in the ‘Hollywood’ meaning of the term. Several answers include figurative senses relating to human referents.

“The star of Orient”

This set expression is almost a nickname, which traditionally refers to the Egyptian singer Umm Kulthum, and to her unique, and now legendary status. We found no other use on Google.

“The star of cooking”

The phrase refers, among others, to ‘Chef Ramzi’, a star cook, who presents cooking recipes on one of the Lebanese television companies.

“film star”

Another phrasing of the same meaning is: نجمة سينمائية, najma sinama‘iyah. Similar phrases are found in English or French (“a movie star”, “une star du cinema”). The cliché, both in French and Arabic is obviously borrowed from English (and Hollywood). In Arabic, though, the immediate word for ‘star’ is najma. Kawkab is used here in its hyperonymic meaning.

Sixth result: kuwaakib, ‘asteroid’.

“The luminous flares and gases emitted by the explosion of an asteroid…”

As already indicated, the word kawkab refers, literally, to a “celestial object” (compare to French astre or Latin astra). It appears here in the diminutive form kuwaakib, which allows construing, in Arabic terminology, a word for “asteroid”.

Seventh result: kawaakib, plural of kawkab, used as a feminine proper name.

“But Kawaakib became the star of cooking...”

The result shows the form of the ‘broken’ plural (جمع (تكسر) of kawkab, used as a feminine proper name (a number of plural nouns of the mafaa‘il/fawaa‘il pattern are re-used as feminine proper names, e.g. jawaahir, ‘awaaTīf...)

The last two results feature ‘internal derivation’ (a change in pattern, the radical consonants remaining unaltered). It suggests that the Google search engine may retain, although obviously not on a systematic basis, what we identify, linguistically, as morphologically related forms.

3.2. Step 2: Word-form variation, with ArabiCorpus and DIINAR.1

I) DIINAR.1 potential word-form variation related to the lemma kawkab

The results obtained from the DIINAR.1 lexical resource include fully developed word-form generation, based on potentially existing forms (combinations excluded by grammar or by grammar-lexis relations, are filtered out). To make the presentation shorter, the only results given below are associated with the singular form kawkab. The plural kawaakib, and the relative nominal/adjectival form kawkab-iyy have not been included, for lack of space. The proclitic forms have also been reduced, for short: the proclitic prepositions (bi-, li-) or the article ‘al-, for instance, are not included below. Case-endings have not been submitted to variation.

The high figure, 4,970,000, is not relevant for us here, because other hand, the result for ‘one enclitic’ with the 3rd person encountered combinations of clitic formatives. On the last line of the table obviously features rarely results. Two proclitics one enclitic

<table>
<thead>
<tr>
<th>Proclitic formatives</th>
<th>Word-forms</th>
<th>Enclitic formatives (pronouns)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ل التوكلك</td>
<td>ECL, 3P, MF, D,</td>
<td>هما</td>
</tr>
<tr>
<td>و وكلك</td>
<td>ECL, 3P, MF, D,</td>
<td>هما</td>
</tr>
<tr>
<td>ا لきます</td>
<td>ECL, 3P, MF, D,</td>
<td>هما</td>
</tr>
<tr>
<td>تكلك</td>
<td>ECL, 3P, F, S,</td>
<td>ها</td>
</tr>
<tr>
<td>ل التوكلك</td>
<td>ECL, 3P, M, P,</td>
<td>هنا</td>
</tr>
<tr>
<td>فكلك</td>
<td>ECL, 3P, M, P,</td>
<td>هنا</td>
</tr>
<tr>
<td>ل التوكلك</td>
<td>ECL, 3P, M, P,</td>
<td>هنا</td>
</tr>
<tr>
<td>انكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>ل التوكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>فكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>ل التوكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>انكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>فكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>ل التوكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>انكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>فكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>ل التوكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>انكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
<tr>
<td>فكلك</td>
<td>ECL, 3P, F, S,</td>
<td>كا</td>
</tr>
</tbody>
</table>

Table 2: A subset of the potentially existing word-forms including the lemma kawkab (from DIINAR.1)

The above word-forms do not appear in the results obtained on Google with the kawkab کوكب query. Some of them, though, are bound to be relevant, e.g.:

<table>
<thead>
<tr>
<th>Number of clitic formatives</th>
<th>Word-form</th>
<th>Google query results</th>
</tr>
</thead>
<tbody>
<tr>
<td>One proclitic</td>
<td>کوكب</td>
<td>54,900</td>
</tr>
<tr>
<td></td>
<td>کوكب</td>
<td>99,800</td>
</tr>
<tr>
<td></td>
<td>الكوكب</td>
<td>488,000</td>
</tr>
<tr>
<td></td>
<td>الكوكب</td>
<td>1,520,000</td>
</tr>
<tr>
<td>One enclitic</td>
<td>کوكب</td>
<td>7,920</td>
</tr>
<tr>
<td></td>
<td>کوكب</td>
<td>9,770</td>
</tr>
<tr>
<td></td>
<td>الكوكب</td>
<td>25,100</td>
</tr>
<tr>
<td></td>
<td>الكوكب</td>
<td>228,000</td>
</tr>
<tr>
<td>Two proclitics</td>
<td>بالکوكب</td>
<td>122,000</td>
</tr>
<tr>
<td></td>
<td>والکوكب</td>
<td>279,000</td>
</tr>
<tr>
<td>One proclitic and one enclitic</td>
<td>کوكب</td>
<td>456</td>
</tr>
<tr>
<td></td>
<td>الكوكب</td>
<td>1,010</td>
</tr>
<tr>
<td>Two proclitics and one enclitic</td>
<td>الكوكب</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>والکوكب</td>
<td>76</td>
</tr>
</tbody>
</table>

Table 3: Additional queries on Google, based on word-forms from the DIINAR.1 language resource

The last line of the table obviously features rarely encountered combinations of clitic formatives. On the other hand, the result for ‘one enclitic’ with the 3rd person sing. masculine pronoun –hu, albeit it amounts to a very high figure, 4,970,000, is not relevant for us here, because kawkahu-hu کوكبة “his” or “its planet, star” (“celestial object”) and kawkaba& کوكبة “constellation” are not currently distinguished by the Google search engine.

2) ArabiCorpus results: corpus-based word-form variation related to the lemma kawkab

The number of occurrences of the lexical entry kawkab in the Egyptian newspaper Al-Ahram (year 1999) is 428, with an average of 2.6 kawkab per 100,000 words.

Word-forms based on kawkab appearing in the above corpus are the following:
- al-kawkab, which includes the proclitic article al-: 123 occurrences;
- kawkab-V-naa, including an undetermined case-ending suffix, conventionally transcribed here with a “V” (for either -u-, -a or -i, respectively nominative, accusative or genitive) and the enclitic pronoun -naa (1st person, plural): 37 occurrences;
- kawkab-an, including the suffix -an (case-ending = accusative in indefinite nouns): 20 occurrences
- li-kawkab-in, with the proclitic li- (preposition, roughly here: ‘for’, or ‘to’): 14 occurrences;
- wa-l-kawkab-V, with the proclitic coordination marker wa-, the article -al- and an undetermined case-ending suffix (noted with a “V”): 8 occurrences;
- kawkabi, with the enclitic pronoun –ii (1st pers. sing.): 8 occurrences.

These forms did not appear in the Google search. On the other hand, the plural form kawakabih, which appeared in the results of Google kawkab query (albeit in the limited way mentioned above) and is included in DIINAR.1, could not be found with ArabiCorpus.

When starting a new Google query with the word-form kawkabha-naa, one finds as many as 228,000 occurrences, which were not included at all in the query based on kawkab presented above. Another query, adding a letter y (ョ) after kawkab gave 65,000 responses. These nevertheless divide into (a) the adjectival form kawkab-iyy “star-like” on the one hand, and (b) the noun followed by the pronoun of the first person (ii – ﻓ) on the other. Two overall remarks can be made:

(1) The Google search requires the user to consider by himself the various types of word-form variation, and then launch as many new queries as he can think of. The large number of results found with the two examples above (228,000 and 65,000), as well as the adding up of figures than can be obtained on Google using potentially existing word-forms from Table 2, also demonstrates the importance of this gap, due to the lack of an underlying word-level lexical resource such as DIINAR.1.

(2) Coming to the assessment of word-level queries, one must not forget that it is extremely difficult for any analyser (including morpho-syntactic analysers [Ouersighni, 2001]) to distinguish, e.g., between the two word-forms behind the unvowelled graphic word kwkby, referred to by (a) and (b) above. The limitation is neither that of Google, ArabiCorpus or DIINAR.1, but pertains,

2 The few examples given in Table 3 already amount to about 2.8 million occurrences.
rather, to the category of what, strictly speaking, cannot be done at word-level.

3.3- Step 3: Word-before/word-after contextualisation

The following results come from the ArabiCorpus site. The most recurrent word before/word after combination are given in the table below:

<table>
<thead>
<tr>
<th>Word immediately after</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>الأرض</td>
<td>59</td>
</tr>
<tr>
<td>Al-'ArD</td>
<td>59</td>
</tr>
<tr>
<td>(without hamza)</td>
<td></td>
</tr>
<tr>
<td>الشرق</td>
<td>51</td>
</tr>
<tr>
<td>Al-sharq</td>
<td>51</td>
</tr>
<tr>
<td>المريخ</td>
<td>25</td>
</tr>
<tr>
<td>Al-Marriix</td>
<td>25</td>
</tr>
<tr>
<td>المشرق</td>
<td>13</td>
</tr>
<tr>
<td>Al-Musharti</td>
<td>13</td>
</tr>
<tr>
<td>الآخر</td>
<td>11</td>
</tr>
<tr>
<td>'aaxar</td>
<td>11</td>
</tr>
<tr>
<td>الأرض</td>
<td>11</td>
</tr>
<tr>
<td>Al-'ArD</td>
<td>11</td>
</tr>
<tr>
<td>(with hamza)</td>
<td></td>
</tr>
<tr>
<td>الأرضي</td>
<td>10</td>
</tr>
<tr>
<td>Al-'arDiyy</td>
<td>10</td>
</tr>
<tr>
<td>(with hamza)</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Word immediately after</th>
<th>Occurrences</th>
</tr>
</thead>
<tbody>
<tr>
<td>سطح</td>
<td>22</td>
</tr>
<tr>
<td>saTH</td>
<td>22</td>
</tr>
<tr>
<td>سكان</td>
<td>14</td>
</tr>
<tr>
<td>Sukkaan</td>
<td>14</td>
</tr>
</tbody>
</table>

Table 4: Words immediately following kawkab in the Al-Ahram (1999) newspaper – ArabiCorpus consultation.

Note that the first query on Google (see § 3.1) included all the occurrences of the ‘word-after’ function above, except for the phrase al-Kawkbab al-'arDiyy, “the planet Earth”. Regarding the latter set expression, a query launched on Google using the “quotes” convention yielded as many as 14,300 results, which did not appear in the first query. We obtained 4,680 results for sukkaa al-kawkab, “the inhabitants of the planet” and 23,500 for SaTH al-kawkab, “the surface of the planet”, 8,900 of which are included in the syntagn SaTH al-kawkab al-'aHar, “the surface of the red planet”, i.e. Mars.

4.2- Information that cannot be accessed at word-form level

On the other hand, there does not seem to be a way in which the high level of ambiguity of Arabic script can be neutralised at word-form level. If the query is, for instance, the written form Twl طول , the user can be asked (as on the ArabiCorpus site) to specify the lexical category (noun, adjective, verb...). This may allow the search engine, if the object of the query is a verb, to look for all the conjugated forms (which can be generated with a resource such as DIINAR.1), but will not allow eliminating the noun Twal، ‘length’ from the search, since it is supported by the same graphic form Twl as the verb Tawwal طول، ‘to make long’, ‘to lengthen’, ‘to protract’, at the 3rd pers. masc. sing. of the perfective; in addition, one finds the noun Tawl طول، ‘might, power’.

We have also seen in § 3.2- 2) that the graphic word kwkby كوكبي is ambiguous because the last letter (y) can be either the 1st person sing. clitic pronoun (the meaning is: “my star” or “planet”), or the relative adjective suffix (the meaning being “star-like”, “planet-like” or “planetary”). The search noises caused by these ambiguities cannot be rubbed out at word-level.

4.3- The need for a lexical resource that includes collocations and set expressions

The example of kawkab is also interesting because it underlines another type of ambiguity, which is related to the various meaning of a given word. As seen in § 3.1 and 3.3, kawkab can mean:

(a) a “planet” or a “star”, in the proper sense of “celestial object”, e.g. kawkab al-arD, “the planet Earth” (other planets are mentioned in Table 4), or kawkab ash-shams, “the sun”;

(b) a “planet” in the metaphorical meaning of “the world of”, as in the third results found with Google (§ 3.1): kawkab al-gulb, word-for-word, “the planet heart”, i.e., “the world of feelings and love”;

(c) a “star”, in another metaphorical meaning, related to radiance (and the glittering of ‘Hollywood stars”), e.g. kawkab sinamaa iyy, “movies” or “film star”, or in kawkab ash-shaq, “the star of Orient” (Unm Kuhlhum).

Each of these meanings is associated with a word before/word after context:

In (a), kawkab is associated with a named entity, referring to a planet or a star. In Arabic, one does not say “Paris”, “the Thames”, or “Mars”, but madinatu Baariis, “the city of Paris”, nahr at-taamz, “the river Thames”, kawkab al-marriix, “the planet Mars”. It is therefore possible to list the named entities that are liable to follow kawkab in that sense.

In (b), the context is that of a magazine. The phrase refers to a given heading.

In (c), the examples are set expressions.

All three types contexts can be entered in a language resource. The work presented here paves the way for
corpus-based listing and description of contexts and collocations.

4.4- Conclusion: Structure of the language resource needed for Arabic search engines

On the basis of the above analyses, one can outline the structure of the language resource needed for the optimisation of Arabic search engines:

– On the one hand, one needs searches to be as comprehensive as possible. This requires a good level of morphological analysis. The language resource should therefore include generated lexica of the same level of comprehensiveness and efficiency as those of DIINAR.1.

– On the other hand, one needs to restrict the search according to the actual aims of the user. A language resource including the type of contextual information outlined in the previous paragraph will allow presenting the users of Arabic searches with semantic and contextual choices (see, for instance, the types of results listed in § 3.1). Searches results easily amount to millions of choices (see, for instance, the types of results listed in the users of Arabic searches with semantic and contextual outlined in the previous paragraph will allow presenting

References


– (2005). The crucial role of language resources (LR-s) in the assessment of Arabic NLP applications including recognition, Key-note address at the Special session on “Linguistic Information Integration in Arabic Character and Text Recognition” – International Conference on Machine Intelligence, Tozeur, Tunisia, Nov. 5-7, 2005.


