



## White Paper

### The Deficiency of Traditional (non-Arabic) Search Systems

With the vast increase in the IT industry, in recent years, there has been an expansion of textual content, both on the web and in companies' intranets. This has created a dire need for search tools, in order to extract information from largely unstructured text.

Traditional search engines used for English, for example, at best use wildcards to search for words; this may be good enough for English, but it is very deficient in searching through Arabic text. The reason for this stems from the richness of Arabic morphology (which governs the inflected forms of Arabic words); it was, thus, necessary to develop a search engine that suited the nature of Arabic. Therefore, a search engine, that takes into account Arabic morphology in the first place, has been developed.

But Arabic morphological rules alone do not completely solve the problems of searching Arabic text; therefore, an engine has been developed, that, in addition to morphology, allows the selection of the meaning of each word in the search query. This helps to reduce the number of unnecessary words (redundant words) in the search results.

Traditional search engines, for non-Arabic languages, depend upon either exact string matching or the use of wildcards in the search query. If we were to try to use wildcard search in order to obtain all the inflected forms of an Arabic word such as اجتماع(meeting), for example, we might try entering the query word \*اجتماع\* to obtain words such as اجتماعات(meeting[s]), الاجتماعان([the][two] meeting[s]), etc.

However, we also need to find inflections of the verb اجتمع([he] met) and its various inflections, by also attempting to enter the query \*اجتمع\*. In addition, we would need to find inflections of the noun مجتمع(meeting {assembled}), by entering \*مجتمع\*, etc. This means that the user should be aware of all derivations of the word he wishes to search for! This means that wildcard search is highly incomprehensive for Arabic, meaning that it cannot obtain all the required results.

At the same time, if the user tries to trick the system by attempting to obtain all the search results using the single query \*ع\*د\*, he will run into further problems. Supposing that a user wishes to obtain all the inflected forms of "عاد" ([he] returned), in order to retrieve العودة([the] return), عائد(returning), عُد(return {Imperative form}), etc., he would also retrieve words that bear no relation to what he set out to search for; he might retrieve words such as بعيد(far), وعود(promises), etc. This would render that type of search redundant for Arabic, meaning that it produces results that are not required.

So we can clearly see that traditional search engines, when used to search for Arabic words, lose the two major features that should be present in any search engine, namely, comprehensiveness and accuracy (the opposite of redundant).

### Proposed Solution

#### First Step - Using Morphological Rules

Using morphological rules enables the user to retrieve all the inflected forms that belong to the same morphological root as the query word. For example, if the user were to search for the word اجتمع([he] met), then words such as الاجتماع([the] meeting), يجتمعان([they - dual] met), مجتمعون([they are] assembled), etc.

This way, the user may enter the query phrase "اجتماع اللجنة العامة" (the meeting of the general committee) and be able to retrieve results like ".....وقد اجتمعت أمس الأول اللجنة العمومية لمنظمة الوحدة الإفريقية....." (... the **general committee** of the Organization of African Unity **met** ..."), as well as "إن اللجان الفرعية ..... والعامة سوف تقوم بالاجتماع فى الشهر القادم" (".... the **general** and sub **committees** will **meet** next month...).

As we can see, it is possible to specify retrieving the resultant search words in the same order of the query words, or not. This type of search is more suited to the nature of the Arabic language and it achieves the feature of comprehensiveness discussed above.

A fast Arabic morphological analyzer was developed and is integrated in KSearch, to allow searching for all Arabic word inflections, using morphological rules ("Morphological Search"). We have established that morphological search produces comprehensive search results, but what about the accuracy of these results? To answer this question, let us take the trilateral morphological root ع-م-ج and look at some derivatives of it: اجتماع([he] met), جمعية(society {as in "dead poets' society"}), إجماع(unanimity or consensus), اجتماعية(social), etc.

Here we can see that, using morphology alone, we might search for "اجتماع" (meeting) and obtain results such as:

- " .. and after the following council **meetings** .."
- " .. and the council members had agreed **unanimously** to elect ..."
- " ... the ministry of **social** affairs decided to grant the **society** a period of grace ..."

We can see here, from these results, that the feature of accuracy is not complied with, even using morphology. So, there has to be a way to improve the accuracy of the search results and reduce the number of words that bear no relation to the required results; this is what we shall consider in the next section.

## Second Step - Finding a Better Solution for Arabic Search - Using Meaning

First of all, let us think about what we mean by required results. When a user searches, he searches, in his mind, for a meaning. Therefore, there needs to be a means of specifying that meaning, if a word's orthographic form has more than one meaning or "sense". For example, in the search query "المتحف الحديث" ("the modern museum"), the Arabic word الحديث has several senses (it is "ambiguous"); it might mean <modern> or <conversation>. In this case, the user should specify the required meaning, in order to obtain the required results.

If the meaning <modern> is selected, then the search results might contain "المتاحف الحديثة فى العالم ... العربى" ("... the **modern museums** in the Arab world..."), in addition to other results that contain inflections of the word "الحديث" (modern) and that belong to the meaning <modern>.

However, if the meaning <conversation> is selected, then the result in the previous paragraph will not be retrieved. The results may contain a sentence like "....وفى المتحف قال المتحدثون الرسميون من مجلس الإدارة ..." ("... At the **museum**, the official **speakers** of the board of directors said that ..."), as well as other results that contain inflections of the word "الحديث" (modern) and that belong to the meaning <conversation>.

It can be seen from the previous paragraph that the redundancy in the search results has been reduced to results that are closer to the user's required meaning, thereby rendering the results more accurate than those obtained using morphological rules alone. KSearch displays the meanings of the query words entered, so that the user can select the meaning he requires, for each word.

The component which achieves this is a lexical semantic analyzer that is linked to the morphological analyzer. This complete morpho-conceptual system is the at the heart of KSearch.