

## ASPI Arabic Search Plug-In

The richness and variety of Arabic makes it one of the world's most difficult languages for comprehensive and effective searching. As one of the oldest languages, it is also one of the few with fundamentally different versions – classical and Modern Standard – that make it even more difficult to create and qualify a comprehensive lexicon for practical use. This challenge is compounded further by Arabic's complex linguistic structure which makes it easy to commit a wide range of writing errors.

**ASPI: The world's most advanced, comprehensive, and easy-to-use Arabic search plug-in.**

ASPI is the most sophisticated Arabic search tool available today. That's because our Middle Eastern roots and deep understanding of the Arabic language and natural language processing give us an advantage over the competition. It has enabled us to create a unique, innovative methodology for writing Arabic language concepts and rules specifically for computer processing rather than attempting to manipulate traditional written methods that simply do not translate well to the digital world.

ASPI can be integrated with your search application through a simple user interface that's fully customizable with our software developer's kit. Our open platform system makes it compatible with Windows, Linux, and most other standard environments. Once integrated into your search engine, our plug-in performs two primary functions:

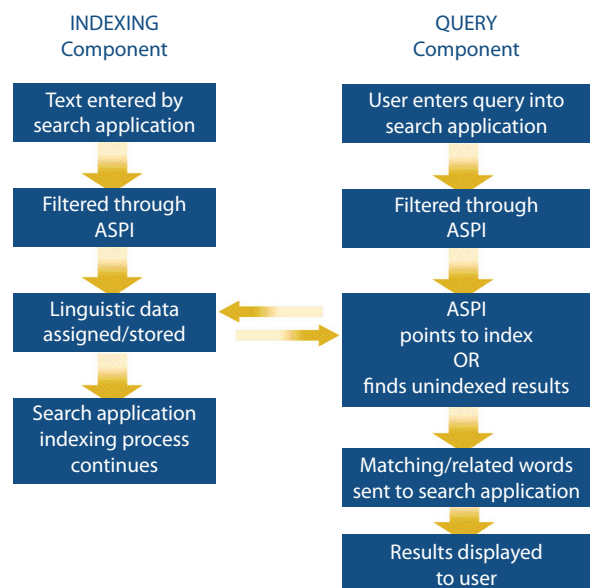
### Index-phase analysis

When your engine indexes pages, ASPI goes to work immediately analyzing all the Arabic words and assigning specific "info", "attributes", or "tags" to each word during indexing. When indexing is complete, our plug-in does all the processing off-line, storing the analyzed and processed data so that when a query is performed, the search engine will immediately retrieve only relevant results.

### Rapid and relevant query results

Despite the complexity of the Arabic language, ASPI is fast processing – at least 400,000 words/sec during index processing in a Pentium 4, 512MB RAM, Windows XP environment – in some cases successfully reaching 800,000 words/sec. Proven to increase both precision and recall of your search engine, ASPI can add tremendous capability and value to your search engine without slowing it down.

Once a query comes in, the search engine first filters it through ASPI which quickly looks at all the previously-stored information that could be associated with the query and provides only appropriate results based on the earlier index processing. Results are either narrowed or expanded based on the user's selected search technique and writing mode.



**ASPI Arabic Search Plug-In**

Our unique language model provides even more advantages.

Simply put, no other Arabic search tool is based on a language model as subtle and sophisticated as ASPI. With over 70% of Arabic words having multiple, often ambiguous, meanings and any single word having an average of 90 valid prefixes, 200 suffixes, and 3000 prefix/suffix combinations, conventional search engines can quickly become overwhelmed.

Our robust language model contains a number of linguistic tools, resulting from over thirty years of R&D, making it very capable of dealing with Arabic-related complexities. Our comprehensive lexical database is based on a fully-tagged Arabic corpus of over 150 million words to enable a complete, relevant word and concept search.

**Complete searching by the way people write.**

ASPI takes into account the unique linguistic features of written Arabic by providing search capabilities for three different modes of writing.

**THREE WRITING MODES**

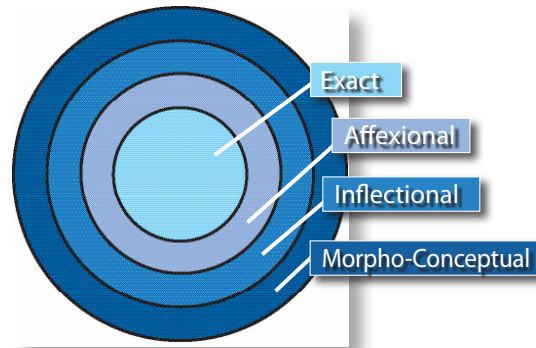
The word "terrorist" is currently written in Arabic as follows:

- |           |   |               |
|-----------|---|---------------|
| 1. إرهابي | → | Correct Form  |
| 2. إرهابي | → | Accepted Form |
| 3. ارهابي | → |               |
| 4. ارهابي | → | Wrong Form    |
| 5. أرهابي | → |               |
| 6. أرهابي | → |               |

Assuming that "إرهابي" is the query word:

- If the writing mode is Exact, the search process will only cover writing form #1.
- If the mode is Relaxed, the search process will cover writing forms #1 – 4 which guarantees more comprehensive search results.
- If the mode is Common Error, the search process will include the correct and the accepted forms as well as the wrong ones. Our strategy is to cover as much as possible the tokens related to the query word ignoring the linguistic correctness, which is not one of the user's search targets.

**SEARCH TECHNIQUE COVERAGE**



In addition to considering the three writing modes, our Arabic plug-in provides four basic linguistic search techniques for even greater search versatility and flexibility:

**Exact** – Considers only the exact query word, limiting the search results.

**Affexional** – (commonly referred to as Intelligent Wild Card) considers not only the exact query word, but also all acceptable prefixes and suffixes related to the query word as well. Conventional Wild Card searching is particularly difficult in Arabic because it can yield results with completely different meanings from the query word. The Affexional technique is ideal for searching on a specific proper noun.

**Inflectional** – In addition to prefixes and suffixes, this technique also searches for changes to the middle of a word for the most complete results on common nouns. This technique searches for words closely related to the query word.

**Morpho-conceptual** – Matching only those derivations based on the Arabic that are conceptually similar to the search word, results will include words sharing the same morphological origin + concept (idea). The Morpho-conceptual technique overcomes the disadvantages of pure root-based searching, while supporting affixation, inflection and derivation. It guarantees comprehensive, accurate, non-redundant search results.

ASPI Arabic Search Plug-In

Examples of search results that might be returned using the various search techniques:

Query/Exceptional	ثروات	بحر	إسلام	جاسوس	جريمة	صحف	بطل
Affexional	ثرواتهم	البحر	والإسلام	للجاسوس	الجريمتان	كصحفها	بطلهم
Inflectional	ثروة	بحار	إسلامي	جواسيس	جرائم	صحيفة	أبطال
Morpho-Conceptual	ثراء	البحارة	مسلمين	التجسس	مجرم	صحافة	بطولة

These are just a sample of the many words that would be included in the search results, including those with related prefixes and suffixes. To limit results to only the most appropriate, prefixes and suffixes NOT related to the query word would not be included.

Searches can be expanded even further using additional linguistic features found in the Advanced Search function such as searching by Synonym – the query word and all its synonyms are searched.

Further analysis for named entities

Intelligent Wild Card Search is the best search technique for named entities because it extracts the proper noun with its related prefixes & suffixes which meets the specific need of the user.

However, some named entities (proper nouns) in Arabic can hold completely different meanings when adding to them prefixes or suffixes; for example: مصر = Egypt, where as المصير = The insistent, making it very important to have an engine that recognizes named entities.

Examples of search results that might be returned using the various search techniques:

Named entities can become even more complicated when they're in the form of noun groups. A conventional search engine does not identify word groups as a named entity, causing the search results to be inappropriate, even if Affexional Search is used.

Noun groups such as "البيت الأبيض" which is "the White House" in English, will never be recognized as such by a search application, leading you to get results such as "البيوت البيضاء" which means "the white houses" if searched on using the Inflectional search – which obviously does not refer to the White House in Washington DC.

For further analysis of named entities an additional tool such as our Entity Extractor is the optimal solution.

For more information, please contact us for a product sheet or visit <http://www.coltec.net/entityextractor>

