

WORDCON: Phonemic-based Word Conversion

A major stumbling block to cross-language content management and data mining is transliteration – the conversion of non-western language characters to Roman characters. This is especially challenging for Arabic-to-English conversion because the Arabic alphabet uses only consonants and diacritics, making it difficult to accurately and fully represent all English versions of Arabic words.

To help non-Arabic speakers, we've developed WORDCON, a software tool that accurately converts Arabic words to Latin characters, whether the source word is entered in Arabic or English. The trick is not just transforming written consonants, but the phonemic short vowels that are created with diacritics as well, since Arabic uses mainly consonants. Our unique methodology is phonemic-based – that is, it uses the sound of the original Arabic word to reproduce phonemes that more accurately represent English vowel usage and generate all the different Latin-based writing forms of an Arabic word – particularly important when using proper names. For example, "Mohamed" is written only one way in Arabic محمد, but has over twenty versions in English. In a process such as entity extraction or search, it is vital that all spellings be considered for truly comprehensive search results.

When integrated into a search application, WORDCON enables English-speaking users to search for an Arabic name using a single English spelling, even if it isn't the common usage – searching with "Muhammad" in English will automatically yield results including Mohamed, Mohammad and Muhamed as well as all the other English variations, significantly increasing coverage of your results.

Our proprietary conversion method features two components - the first using natural language symbols and the second identifying all spelling variations and dominant uses; for example the Arabic name "يوسف" will automatically convert to "Joseph" in English but will also convert to phonetically similar conversions such as "Youssef" and "Youssuf". This unique feature is critical for conversion of words with no phonemic similarity - "Egypt" in English is referred to as "Misr / مصر" in Arabic.

Perfect for a wide range of educational, corporate and government uses.

WORDCON is ideal for English-speaking users who may not be familiar with Arabic, including media professionals, data base managers, librarians, students, and anyone who wants to search Arabic documents and sources for further study. It's also valuable for teaching anyone how to pronounce Arabic words, such as language students, broadcasters, politicians, government officials, and anyone engaged in public speaking on Arabic issues.

WORDCON is also an invaluable tool for corporate or government intelligence gathering and information verification. By automatically returning all spelling variations of a name, WORDCON can be used to verify identity and credit history, cross-check name variations for immigration and travel restriction cases, and uncover possible identity fraud.

Sample English Spelling Variation Chart

أسامة	مصطفى	حسين	عبد الله
Ossama	Mostafa	Hussein	Abd-Allah
Oussama	Moustafa	Husseyin	AbdAllah
Ussama	Mostapha	Hussayn	AbdEllah
Usama	Moustapha	Houssain	Abdalah
Osama	Mustafa	Houssayn	Abdullah
Ousama	Mustapha	Houssein	Abdulah

As seen in this chart a single Arabic word or name can yield many English spellings.

WORDCON: Phonemic-based Word Conversion

Sample English Google® Search Engine Results

•Query •Definition	محمود Arabic Name	No. of Results	محمد Arabic Name	No. of Results	يوسف Arabic Name	No. of Results
Spelling (1)	Mahmoud	12,400,000	Mohammed	26,300,000	Youssef	4,930,000
Spelling (2)	Mahmud	5,060,000	Mohamed	27,800,000	Yousef	2,130,000
Spelling (3)	Mahmood	3,740,000	Mohammad	15,200,000	Yousuf	1,390,000

■ As seen in these sample queries, different English spelling variations of the same Arabic name yield very different results which severely hinders precision and recall. You might not find the information you're seeking simply because of your choice of spelling. WORDCON yields all possible forms of the query with a single click.

Source: www.google.com May, 2007

Users searching with any popular search engine such as Google simply type the word in English and WORDCON automatically converts it accurately into correct Arabic. WORDCON can be provided with all common transliteration schemes such as those used by the U.S. Bureau of Geographic Names, FBIS Arabic Romanization Guidelines, IC Standard for the Transliteration of Arabic, SATTs, and the Buckwalter system. WORDCON is even useful for English-only users, when no Arabic is involved at all. For example, using the English version of Google with WORDCON integrated, a search for an Arabic name in English will automatically return ALL English variations of that name.

WORDCON accommodates three writing modes for users writing in Arabic – exact, relaxed, and common error. This means that even Arabic words written incorrectly will still convert accurately into English.

Text alternative conversion feature

Informal communications such as texting, chatting, and instant messaging (IM) are becoming an increasingly greater percentage of all digital information traffic. Over 25% of Internet users have used, or are currently using, online chat rooms as a means of communication

and a large and growing number of cell phone subscribers – 1.4 billion globally – are generating 350 billion text messages every month.

That's why WORDCON offers a convenient text alternative conversion feature for character substitutions made necessary by incompatible alphabets and keypad restrictions. Now such vital information as proper names that may appear in text messages as a combination of letters and numbers can be quickly and easily identified and converted to their correct alphabetic version.

Use as a stand-alone tool or integrate WORDCON into another application

WORDCON is an ideal stand-alone conversion tool and can also provide value-added benefits to other applications. As with all our products, the WORDCON Phonemic Conversion system can be delivered as a Windows, Macintosh, or Linux SDK for easy integration with other applications. It can also be integrated with ANEE, our entity extractor tool, as well as our Search Plug-in.

Chatting Alternatives

Number	Arabic
7	ح
2	أ
3	ع

These examples show numerical digits used to represent Arabic characters that do not have equivalents in the Latin alphabet.

محمد	عمر	مؤمن
Mohamed	Omar	Moamen
Mo7amed	3omar	Mo2men