Buckwalter-based Lookup Tool as Language Resource for Arabic Language Learners

Jeffrey Micher
Multilingual Computing Branch
Army Research Laboratory
Adelphi, MD 20783 USA
jmicher@arl.army.mil

Clare R. Voss
Multilingual Computing Branch
Army Research Laboratory
Adelphi, MD 20783 USA
voss@arl.army.mil

The morphology of the Arabic language is rich and complex; words are inflected to express variations in tense-aspect, person, number, and gender, while they may also appear with clitics attached to express possession on nouns, objects on verbs and prepositions, and conjunctions. Furthermore, Arabic script allows the omission of short vowel diacritics. For the Arabic language learner trying to understand non-diacritized text, the challenge when reading new vocabulary is first to isolate individual words within text tokens and then to determine the underlying lemma and root forms to look up the word in an Arabic dictionary.

Buckwalter (2005)'s morphological analyzer (BMA) provides an exhaustive enumeration of the possible internal structures for individual Arabic strings in XML, spelling out all possible vocalizations (diacritics added back in), parts of speech on each token identified within the string, lemma ids, and English glosses for each tokenized substring.

The version of our Buckwalter-based Lookup Tool (BBLT) that we describe in this poster provides an interactive interface for language learners to copy and paste, or type in, single or multiple Arabic strings for analysis by BMA (see Fig. 1)

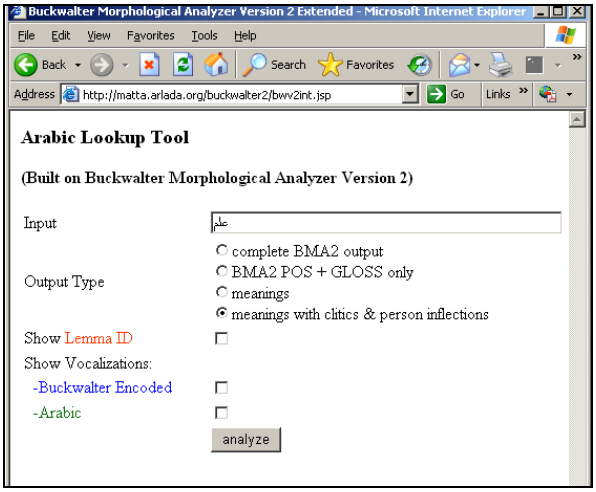

Figure 1. BBLT Input Screen

We originally developed BBLT for ourselves as machine translation (MT) developers and evaluators, to rapidly see the meanings of Arabic strings that were not being translated by our Arabic-English (MT) engines (Voss et al. 2006), while we were also testing synonym lookup capabilities in Arabic WordNet tool (Elkateb et al. 2006). While BBLT allows users to see the "raw" BMA XML (see Fig. 2), the look-up capability that sorts the entries by distinct lemma and presents by English gloss has proved the most useful to English-speaking users who cannot simply lookup Arabic words in the Hans Wehr dictionary (considered the most complete source of Arabic words with about 13,000 entries, but requires the user to be able to "know" the underlying form to search for).

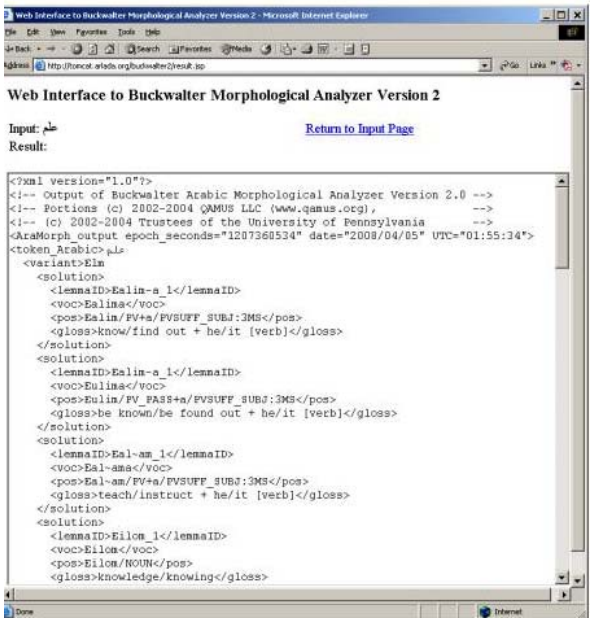

Figure 2. BBLT Output for single token with option "meanings with clitics and person inflections" on

The BBLT user can opt to see the glosses with or without the clitics or inflections, with their diacritized forms either transliterated or rewritten

Software Engineering, Testing, and Quality Assurance for Natural Language Processing, pages 66–67,
Columbus, Ohio, USA, June 2008.

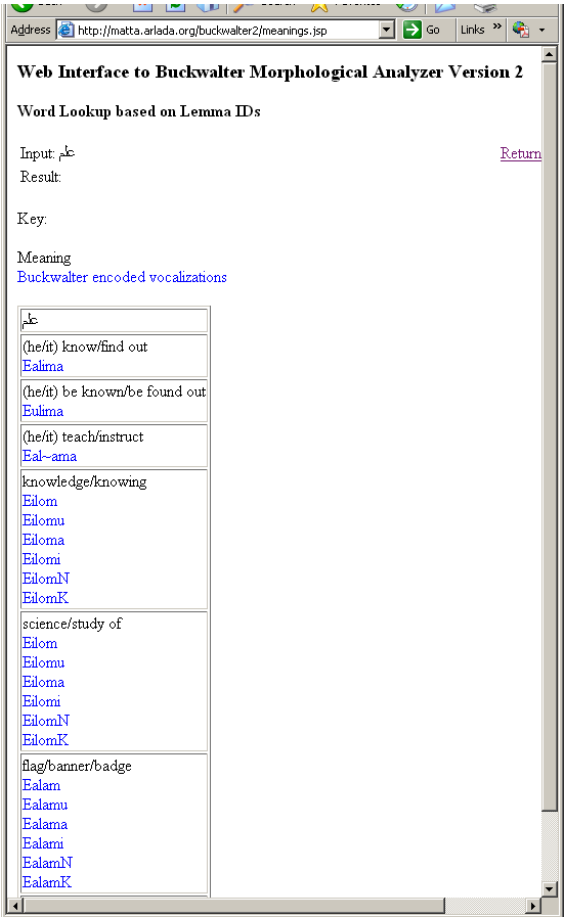

Figure 3. BBLT Output for single token with additional option "Buckwalter encoded vocalizations" on

in Arabic script (see Fig. 3) or in full table form for full sentence glossing (see Fig. 4).

The web application is written as a Java webapp to be run in a tomcat web server. It makes use of wevlets written as both standalone sevlets, extending HttpServlet, and .jsp pages. One servlet handles running BMA as a socket-server process and another servlet handles request from the input .jsp page, retrieves the raw output from the former, process the output according to input page parameters, and redirects the output to the appropriate .jsp page for display.

References

Buckwalter Arabic Morphological Analyzer (BAMA), Version 2.0, LDC Catalog number LDC2004L02, www.ldc.upenn.edu/Catalog.

Buckwalter,T. (2005) www.qamus.org/morphology.htm

Elkateb, S., Black, W., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A. and Fellbaum, C., (2006). Building a WordNet for Arabic, in *Proceedings of The fifth international conference on Language Resources and Evaluation (LREC 2006)*.

Voss, C., J. Micher, J. Laoudi, C. Tate (2006) "Ongoing Machine Translation Evaluation at ARL," Presentation, In Proceedings of the NIST Machine Translation Workshop, Washington, DC.

Wehr, Hans (1979) Arabic-English Dictionary:: The Hans Wehr Dictionary of Modern Written Arabic. Edited by J. M. Cowan. 4th ed..Wiesbaden, Harrassowitz.

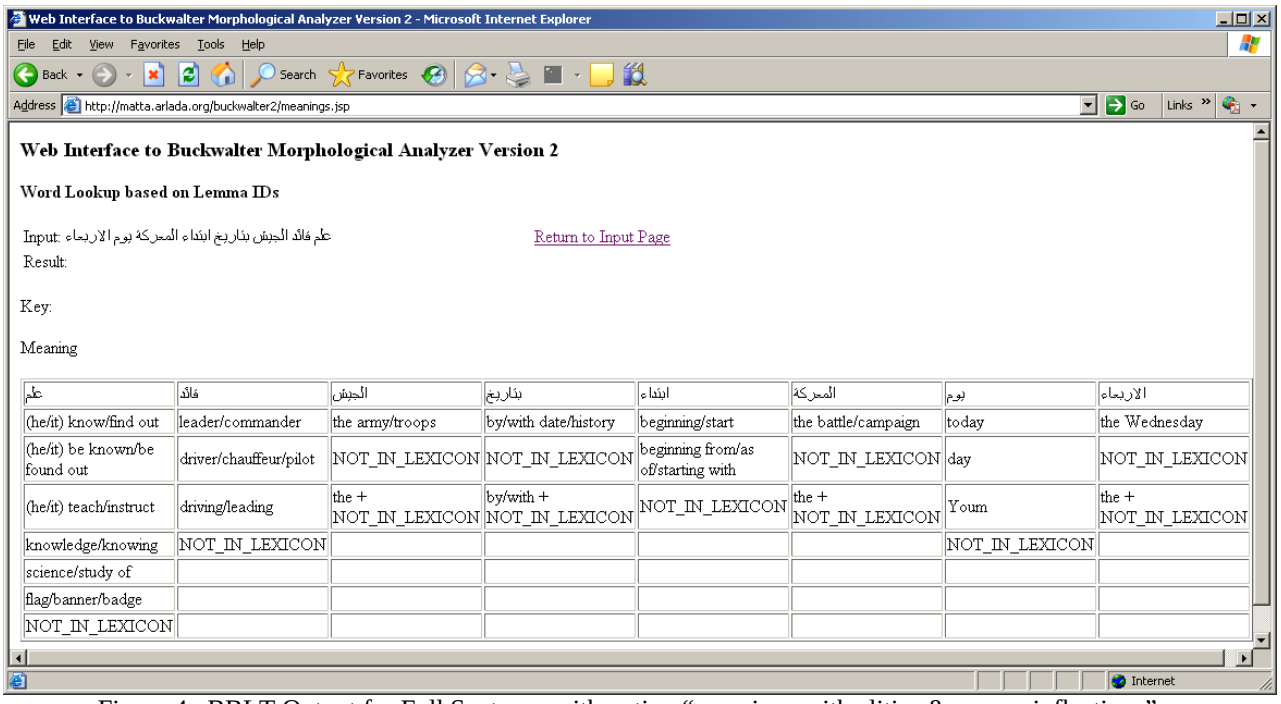

Figure 4. BBLT Output for Full Sentence with option "meanings with clitics & person inflections"