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

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

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

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 Englishspeaking 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).

web Interface to Buckwalter Morphological Analyze	er Version 2 - Microsoft Internet Explorer	_ D X
Die Edit Yoow Figwarites Iaads Heip	Concerning and Million and Concerning	£57
J-Back	(man) ()·3 同·11 []	in Marine
dama 🗿 http://toncet.arlada.org/budiwater2/result.isp	- 8% u	** ** *
Web Interface to Buckwalter I	Morphological Analyzer Version 2	-
Input: علم	Return to Input Page	
Result:		
xml version="1.0"?		-
<pre>(i- (c) 2022-2004 Trustees of AraMorph output epoch seconds <token arabic=""> µis <solution> <lenstruktion> <lenstruktion> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <lenstruktion> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution> <solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></solution></lenstruktion></solution></solution></solution></solution></solution></solution></solution></lenstruktion></lenstruktion></solution></token></pre>	<pre>the University of Feinsylvania> ="120760534" date="2008/04/05" UTC="01:55:34"> aID> UEJ:3MS e/it [verb] aID> UFF 8UB0:3MS out + he/it [verb] ID> 3UB0:3MS </pre>	
 <solution></solution>		
<pre><lennaid>Eilom_1<voc>Eilon</voc> <pos>Eilon/NOUN</pos> <pos>Eilon/NOUN</pos></lennaid></pre>	D>	-
al agroups who will de la constant d	a Arona	
1		



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. ©2008 Association for Computational Linguistics

	/ 📈 🤍 🔛 🤝							
Address a http://matta.arlada.org/buckwa	ilter2/meanings.jsp 💽 ラ Go Links » 🍖 📼							
Web Interface to Buckwalter Morphological Analyzer Version 2								
Word Lookup based on Lemma	IDs							
Input: ﷺ	Return							
Result:								
Key:								
Meaning								
Buckwalter encoded vocalizations								
علم								
(he/it) know/find out								
Ealima								
(he(it) he known/he found out								
Fulima								
Charles the set of the strengt								
(nent) teach/instruct								
Eal~ama								
knowledge/knowing								
Eilom								
Eilomu								
Eiloma								
Eilomi								
EllomIN The IZ								
Ellomk								
science/study of								
Eilom								
Eilomu								
Eiloma								
Eilomi								
Ellomin TalantZ								
LIOUNE								
flag/banner/badge								
Ealam								
Ealamu								
Ealama								
Ealami								
EalamN								
Halamk.								
4	•							

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.

🚳 Web Interface to Buckwalter Morphological Analyzer Version 2 - Microsoft Internet Explorer								_			
Ele Edit View Favorites Iools Help									2		
🕞 Back 🔹 🕥 😰 🙆 🏠 🔎 Search 🤸 Favorites 🚱 🔗 - 🌭 🔳 - 🛄 🎉											
Agdress 🕘 http://matta.arlada.org/buckwalter2/meanings.jsp							-	🔁 Go 🛛 Links » 🍕	<u>1</u> -		
Web Interface to Buckwalter Morphological Analyzer Version 2											
Word Lookup based on Lemma IDs											
Input: لمحركة نوم الاربعاء Result: Key:	علم فائد الجيش بذاريخ ابنداء ا	:	<u>Return to Input</u>	Page							
Meaning											
علم	فائد	الجرش	بئاريخ	ابئداء	المعركة	نوم			الاربياء		
(he/it) know/find out	leader/commander	the army/troops	by/with date/history	beginning/start	the battle/campaign	today			the Wednesday		
(he/it) be known/be found out	driver/chauffeur/pilot	NOT_IN_LEXICON	NOT_IN_LEXICON	beginning from/as of/starting with	NOT_IN_LEXICON	day			NOT_IN_LEXICO	'n	
(he/it) teach/instruct	driving/leading	the + NOT_IN_LEXICON	by/with + NOT_IN_LEXICON	NOT_IN_LEXICON	the + NOT_IN_LEXICON	Youm			the + NOT_IN_LEXICO	N	
knowledge/knowing	NOT_IN_LEXICON					NOT_I	N_LE	XICON			
science/study of											
flag/banner/badge											
NOT_IN_LEXICON											
E									🌍 Internet	_	

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