Behavior Research Methods
2010, 42 (2), 481-487
doi:10.3758/BRM.42.2.481

Aralex: A lexical database for Modern Standard Arabic

Sami Boudelaa and William D. Marslen-Wilson
MRC Cognition and Brain Sciences Unit, Cambridge, England

In this article, we present a new lexical database for Modern Standard Arabic: Aralex. Based on a contemporary text corpus of 40 million words, Aralex provides information about (1) the token frequencies of roots and word patterns, (2) the type frequency, or family size, of roots and word patterns, and (3) the frequency of bigrams, trigrams in orthographic forms, roots, and word patterns. Aralex will be a useful tool for studying the cognitive processing of Arabic through the selection of stimuli on the basis of precise frequency counts. Researchers can use it as a source of information on natural language processing, and it may serve an educational purpose by providing basic vocabulary lists. Aralex is distributed under a GNU-like license, allowing people to interrogate it freely online or to download it from www.mrc-cbu.cam.ac.uk:8081/aralex.online/login.jsp.

Psycholinguistic databases, providing statistical information such as word frequency, length, and imageability, have proved to be invaluable tools for the experimental investigation of the cognitive processes underlying language functions and for the design of language assessment tools for both educational and clinical purposes (Lété, Sprenger-Charolles, & Colé, 2004; Stadthagen-Gonzalez & Davis, 2006). Such databases have long been available for European languages such as English, French, German, and Spanish and have contributed to major advances in basic theoretical and translational research in psycholinguistics (Baayen, Piepenbrock, & van Rijn, 1993; Content, Mousty, & Radeau, 1990; New, Pallier, Brysbaert, & Ferrand, 2004; Sebastián-Gallés, Marti, Cuetos, & Carreiras, 2000). The obverse of these achievements is that much of what we know to date about human language understanding and representation is based on the study of a select few languages, with a single language (English) still largely dominant.

This is scientifically unsatisfactory. When broader cross-linguistic studies have been carried out, they have proved enlightening and, sometimes, even revolutionary. Research into Arabic, the most widely spoken Semitic language, has led to major developments in linguistic theory, as attested by the groundbreaking work of John McCarthy on morphophonology (McCarthy, 1981). Ongoing psycholinguistic research into Arabic is also helping to constrain our knowledge of and theorizing about how different linguistic components, such as morphology, phonology, orthography, and semantics, are processed by and represented in the human brain/mind (Boudelaa & Gaskell, 2002; Boudelaa & Marslen-Wilson, 2005; Boudelaa, Pulvermüller, Hauk, Shtyrov, & Marslen-Wilson, 2010; Idrissi, Prunet, & Béland, 2008; Plunkett & Nakisa, 1997; Prunet, Béland, & Idrissi, 2000).

The importance of Arabic as a subject of investigation stems from its distinctive structural properties that have made it an important test case for competing theories in many fields. Among these properties is its nonconcatenative morphology, whereby surface word forms (e.g., [katab] *write*) arguably result from the interleaving of a consonantal root {ktb}, which conveys semantics, and a vocalic word pattern {faʕal},[1] which conveys morphosyntactic and phonological information. A second property relates to its right-to-left consonantal cursive orthography, where the same letter has different shapes depending on where it occurs in the word. A third specificity of this language relates to its set of pharyngeal consonants, which are associated with two places of articulation. A fourth property of this language is its inextricably diglossic situation where Modern Standard Arabic (MSA) coexists with several regional dialects.

Research into Arabic has been severely hampered, however, by the lack of an adequate lexical database that would provide information about its distributional and structural characteristics. The absence of information on the frequencies of Arabic morphological constituents, for example, puts research into Arabic at a major disadvantage, as compared with other languages for which such resources have long been available. This also makes it hard to compare the results of psycholinguistic experiments on Arabic with the results of equivalent experiments on other languages. Given the special properties of the Arabic language, such comparisons are critical for adjudicating between competing linguistic and cognitive theories and for advancing our overall understanding of human language functions.

S. Boudelaa, sami.boudelaa@mrc-cbu.cam.ac.uk

The need for an Arabic lexical database, providing a variety of distributional and structural information about Arabic words and morphemes, has led to the creation of Aralex. In what follows, we describe how we created Aralex, how it is structured, and what kind of information it provides. Before doing this, we give a brief account of Arabic, characterizing the special properties and challenges associated with this language in psycholinguistic, educational, and computational domains.

A Basic Description of Arabic

The label *Arabic* is a general term referring to the different dialectal Arabic (DA) varieties spoken in different regions of the Arab world and to MSA. MSA is the pan-Arabic variety of Arabic shared by educated speakers throughout the Arab world. It is the language used for written and formal oral communication, such as broadcast news, courtroom language, and university lectures, and is generally the language of the mass media (radio, television, newspapers). Everyday communication, however, is more likely to be carried out in one of the various regional dialects. The coexistence of MSA and DA defines a typical situation of diglossia where MSA is the "high" variety and the regional local dialect is the "low" variety. MSA and DA usually present with similar (although not identical) phonological, syntactic, and lexical systems but fulfill distinct sociolinguistic functions, as indicated above (Ferguson, 1959; Holes, 1995; Versteegh, 1997).

Aralex is a database of MSA only and does not offer information about any DA varieties (see note 1). Despite the sociolinguistic differences between MSA and DA, there seems to be no reason in principle why the architecture of Aralex as it stands could not be used to cover DA corpora in due course. Current psycholinguistic evidence indicates that MSA and DA share the same underlying decompositional mechanisms for word formation and lexical processing (Boudelaa & Marslen-Wilson, 2005).[2]

We focus here on two specific features of Arabic that are of particular psycholinguistic interest: its rich and highly inflected nonconcatenative morphology and its writing system. In Arabic, every content word and many function words can be analyzed into a root and a word pattern. The root is made up exclusively of consonants, typically three, and conveys general semantic information that will be expressed to different extents in the various surface forms featuring that root. By contrast, the word pattern consists primarily of vowels (although it can include some consonants as well) and conveys morphosyntactic and phonological information (Holes, 1995; Versteegh, 1997). These two minimal units of form and meaning are not appended together one after the other, like stems and affixes in Indo-European languages, but are interleaved within each other. For example, when the root {ktm}, with the general meaning of *hiding*, is interleaved with the pattern {faʕal},[3] with the morphosyntactic meaning of *active*, *perfective*, it yields the verbal surface form [katam] *hide*. When the same root is combined with the *agentive* pattern {faaʕil}, the resulting form is [kaatim] *someone who hides or conceals*. The meaning of any surface form is by no means always a composition of the meaning of the root and the pattern, but there is a reasonable amount of predictability (McCarthy, 1981). To these surface forms, inflectional affixes and enclitics are added. For example, the complex surface form [wabitʕalaaqatihi] *and with his eloquence* is made up of the proclitics [wa] *and*, [bi] *with*, the stem or surface form [tʕalaaqat] *eloquence*, the enclitic [i] as a *genitive marker*, and the third person masculine possessive pronoun enclitic [hi] *his*. The surface form [tʕalaaqat][4] is further analyzed into the root {tʕlq} *being loose* and the pattern {faʕaalah} *deverbal noun*, *feminine* (Holes, 1995; Versteegh, 1997).

The complexity of Arabic word structure is compounded, for the reader, by its cursive writing system where (1) short vowels and other diacritics are not written, except in religious texts or in reading materials for children; (2) letters are joined to each other even in typescript; and (3) the shape of the letter changes depending on where it occurs in the word, creating allographic variation. The absence of short vowels and other diacritic marks makes any stretch of MSA text potentially ambiguous at morphological, phonological, syntactic, and lexical levels. For instance, the written form علم as typically experienced by MSA readers can be read as [ʕalima] *know*, [ʕulima] *be known*, [ʕilm] *science*, [ʕalam] *flag*, or [ʕallam] *teach*. The cursive nature of the script adds another layer of ambiguity, since the clitics and the affixes, as the above phrase [wabitʕalaaqatihi] demonstrates, are directly attached to the surface form. Finally, the allographic variation of MSA graphemes generates a nontrivial many-to-one mapping between letter forms and their internal orthographic representations.

The rich morphological system and the special orthographic characteristics of Arabic make it an interesting and challenging subject of research for fields such as theoretical linguistics, psycholinguistics, and natural language processing. Its morphological system raises the problem, for example, of how the component morphemes of a word (the root and the word pattern) can be recovered from a written text where the word pattern morpheme is systematically not supplied and the root morpheme is intermingled with and flanked by other clitics and inflectional material that make it difficult to extract the correct morpheme. Similarly, the orthographic allography that pervades the writing system makes it a complex task to process Arabic text automatically.

Building Up Aralex

The importance of morphology as a domain of knowledge in Arabic cannot be overstated. Our own experimental research and similar research into other Semitic languages, particularly Hebrew, underlines the primacy of morphemes, showing how roots and word patterns govern lexical organization and lexical processing in this language family (Boudelaa & Marslen-Wilson, 2005; Boudelaa et al., 2010; Frost, Forster, & Deutsch, 1997). This means that any lexical resource that does not provide reliable statistics about roots and word patterns will be of limited use to the experimental psychologist interested in the study of Arabic and, indeed, to anyone interested in the statistical structure of Arabic, whether a language

learner or a language practitioner. Accordingly, we set out to build Aralex as a resource that not only provides the classical distributional information about word token frequency and bigram and trigram frequency found in typical databases like CELEX (Baayen et al., 1993), but also was designed to give information about type and token frequencies of roots and word patterns.

This dual aspect of Aralex led us to collect information from two sources: (1) a reliable and widely used dictionary—namely, the Hans Wehr *Dictionary of Modern Written Arabic* (Wehr, 1994)—and (2) a representative corpus of 40 million words derived from various Arabic newspapers covering a wide spectrum of subjects, such as politics, sport, literature, and so forth. We did not seek to cross-check the corpus against the dictionary or vice versa, although this has been a common practice among database developers in other languages, such as English (Davis, 2005), Spanish (Perea et al., 2006), and Greek (Ktori, Van Heuven, & Pitchford, 2008). The advantage of such cross-checking is that it eliminates entries that do not occur in the dictionary, improving the accuracy of the corpus, as well as cleaning up the database by removing misspellings, nonlexical abbreviations, and nonalphabetic characters. In the case of MSA, however, the disadvantages of this procedure substantially outweigh its advantages. In particular, it results in a reference vocabulary comprising only the citation forms of words. For instance, all the imperfective forms of a verb like [yaktub] would be deleted from the corpus because once the {ya~} is stripped off, the residual {ktub} is not listed in a dictionary. This means that cross-matching would remove the majority of Arabic verb forms. Similarly, regularly inflected plural forms like [najjaaruun] *carpenters* and [mufakkiruun] *thinkers*, which make up about half of the MSA inflected nouns (Boudelaa & Gaskell, 2002) and are very common in any corpus, would suffer the same fate, since they are not listed in a dictionary. For these reasons, the dictionary is made up of words in their citation form, whereas the corpus includes not only all the citation forms of the dictionary, but also all the noncitation forms.

The Aralex dictionary. The source dictionary we worked with was the *dictionary of stems* used by the Arabic Morphological Analyzer (hereafter, AraMorph; Buckwalter, 2002). Buckwalter cross-checked his dictionary of stems with the Hans Wehr dictionary and the Larousse Arabic–French dictionary (Reig, 1999). The dictionary of stems provides an exhaustive coverage of Arabic stems and roots, omitting lexical items that have fallen into disuse (e.g., [kitAbxAnah] *library*). The version of the stem dictionary as we used it, after exhaustive checking, consists of 37,494 different stems. These include native Arabic words, assimilated and unassimilated foreign words, and proper Arabic and foreign nouns. For each stem, we determined by hand the appropriate root and word pattern. The absolute number of root and word pattern types is 6,804 and 2,329, respectively. Excluding Arabic and foreign proper nouns brings the number of roots down to 5,336 and the number of word patterns to 2,324, which is, in effect, the total number of roots and word patterns currently used in MSA.[5] This dictionary is used as a lookup table to provide a deterministic parse into a root and a word pattern of every stem in the corpus. Thus, for a given stem, we can have the root and the word pattern, along with their respective type frequencies (or family sizes)—that is, the number of forms that feature that particular root or word pattern.

The Aralex corpus. The corpus consists of 40 million written MSA words drawn from various Arabic newspapers available online. The most challenging aspect of this corpus was the absence of diacritics from the script. This makes any stretch of text fraught with ambiguities at all levels of linguistic description.[6] First, the corpus was stripped of its html tags, converted into manageable text files, and then submitted to AraMorph (Buckwalter, 2002), which takes a text in Arabic Windows encoding and outputs a file with a full morphological analysis and part-of-speech (POS) tags. For each input word, defined as a string of letters with white space on either side of it, AraMorph provides (1) a fully "vowelled" solution of all the possible alternative readings of the word at hand, with the appropriate short vowels reinstated to give the full phonological surface form; (2) a breakdown of the constituent morphemes of the word, including affixes, clitics, and stems, but not roots and word patterns; and (3) their POS and corresponding English glosses.

To choose the correct vowelled solution from among the several alternatives provided by AraMorph for each orthographic form in the corpus, we developed a novel automated technique based on the use of support vector machines (SVM) (Wilding, 2006). This extended the standard unigram approach to classification and combined the output of AraMorph with a set of concatenation and disambiguation rules. The output of this technique is a probability score, reflecting the accuracy of the automatic vowelling, and an entropy score that measures the amount of uncertainty in the probability score.[7] In initial testing on 792,000 words from the Arabic Treebank, the accuracy of this automatic vowel-restoring program was over 93% when case endings were stripped off and over 85% with case endings included. When applied to the 40-million-word corpus, the accuracy of the program remains high, averaging 80% for fully diacritized forms and 90% for forms without case endings. These figures were further cross-validated against a randomly selected 500,000-word sample of automatically vowelled words that were also hand-annotated by a team of native Arabic speakers in Egypt.[8] The validation yielded an overall accuracy of 77.9%, meaning that the solutions chosen by the annotators were also likely to be chosen by the automatic diacritizer.

Integrating the corpus and the dictionary. To be able to provide both type frequency counts and token frequency counts, we combined the dictionary and the corpus into an integrated database.[9] For every item in the corpus that had a stem in the dictionary, we determined the root and the word pattern, using the dictionary as a deterministic lookup table. Around 0.44% of the corpus stems are not listed in the dictionary. These are mainly either compound proper nouns like [ʕabd-assayyid] or CCVC-stems like [ʒlis], which AraMorph systematically outputs as stems for imperfective verb forms with a pre-

fix (e.g., na+ʒlis] *we sit*). What this means in practice is that we do not provide a type frequency count for roots and patterns in compound proper nouns and imperfective verbs, but we do provide token frequency counts for such forms. Compound proper nouns are not counted as part of the root or pattern morphological families simply because the Hans Wehr dictionary on which we based our lookup table does not systematically list this type of noun. Consequently, the algorithm that we use to parse the surface forms into a root and a pattern cannot currently handle this type of noun. The simple proper nouns (e.g., [ħusny mubaarak]), however, are treated like ordinary lexical items and are integrated with the rest of the corpus. As regards imperfective stems, these are inflectional variants of a given word, and if they need to be included in the family size count of roots and patterns, they can easily be added as a constant based on the inflectional paradigm of the verb at hand.

For those forms that are attested in both the corpus and the dictionary, constituting 99.56% of the data, we provide frequency counts for the orthographic form, the unpointed stem (i.e., the stem without vowels), the pointed stem (i.e., the stem with the vowels), the root, the word pattern, and the bigram and trigram frequencies of the orthographic form, the root, and the word pattern.

We defined the orthographic form as the graphic entity that occurs with white space on either side of it. Given the nature of the Arabic script, an orthographic form can be a whole phrase (e.g., ويستقبله [wystqblh] *and he welcomes him*) or a noun without any enclitics or affixes (e.g., مكتب [mktb] *office*). The unpointed stem is the stem as output by AraMorph once the clitics and the affixes have been stripped off. For the form [wystqblh], for example, the unpointed stem is [stqbl]. The pointed stem of this form, output by the SVM diacritizer, is [staqbil]. For each pointed stem, we provide the root and the word pattern, on the basis of the dictionary. For example, the pointed stem [maktab] will have the root {ktb} and the word pattern {mafʕal}. Finally, for each orthographic form, root, and word pattern, a bigram and a trigram frequency count is provided. We do not provide a bigram or a trigram frequency count for pointed or unpointed stems, because the same frequency counts for the unpointed stem are covered by those provided for the orthographic form, whereas the bigram and trigram frequencies for pointed stems would be unrealistic, since typical Arabic script does not feature any vowels.

The token frequency statistics are computed from occurrence counts in the 40-million-word corpus as the rate of occurrence per 1 million words of text, given by

$$\text{Freq}(w) = \frac{\text{occ}(w)}{T/k},$$

where occ(w) is the number of occurrences of word w in the corpus, T is the total number of words in the corpus, and $k = 1{,}000{,}000$. The generation of token frequencies for orthographic forms consists simply in counting and normalizing the number of times each distinct orthographic form occurs in the corpus. Where the token frequencies of stems, roots, and word patterns are concerned, the following procedure was followed. For each record in the corpus, the pointed and unpointed stems were extracted; then their corresponding root and word pattern were located in the dictionary, and the occurrence of each of these four units (i.e., the pointed stem, the unpointed stem, the root, and the word pattern) was recorded. If a pointed stem was not found in the dictionary, the unpointed stem was used to match with dictionary entries without diacritics to get a set of pointed stem candidates. Then all the corresponding roots and patterns for that set of stems were located and recorded, thus increasing recall at the cost of potentially decreasing precision.

Turning to type frequencies of roots and patterns, these are simply raw counts and are extracted from the dictionary. Specifically, the type frequency (or the morphological family size) for a particular root or word pattern is the number of stems containing that particular root or pattern. Finally, the character n-gram frequencies (bigrams and trigrams) are computed from the 40-million-word corpus for orthographic forms, roots, and word patterns as follows:

$$\text{Freq}(g) = \frac{\text{occ}(g)}{T/k},$$

where occ(g) is the number of occurrences of n-gram g in the corpus, T is the total number of n-grams in the corpus, and $k = 1{,}000{,}000$.

The Web Interface

Once Aralex was completed, we wanted to maximize its usefulness for different types of researchers, with different needs and skills. We therefore developed two interfaces for searching it: a JSP/Java-based Web interface and a Java-based command line interface (CLI). Both interfaces are based on the *Apache Lucene* index tool (http://lucene.apache.org/java) and provide advanced query functionality with rapid response times.

The Web interface is aimed at the majority of users, whose needs can be met by a set of predefined queries. It allows the user to query the database using either Buckwalter's (2002) transliteration scheme or Arabic script. Users can request the surface frequency for an orthographic form, a pointed stem, an unpointed stem, a root, and a word pattern. They can also request the type frequencies for roots and patterns and the bigram and trigram frequencies for orthographic forms, roots, and patterns. The output can be sorted by a search unit (e.g., orthographic form frequency or root type frequency) ordered in ascending or descending order. All the user needs to do is to enter a search term in the appropriate window, tick the appropriate boxes or, indeed, check all the boxes to have exhaustive information about the search string, and hit the search button at the bottom right-hand corner. Figure 1 illustrates the output of Aralex using the orthographic form [wystqbl] as a search string.

To get the bigram and trigram frequency counts, the user simply needs to check the *show n-gram* stats box at the right-hand corner. Figure 2 shows the bigram and tri-

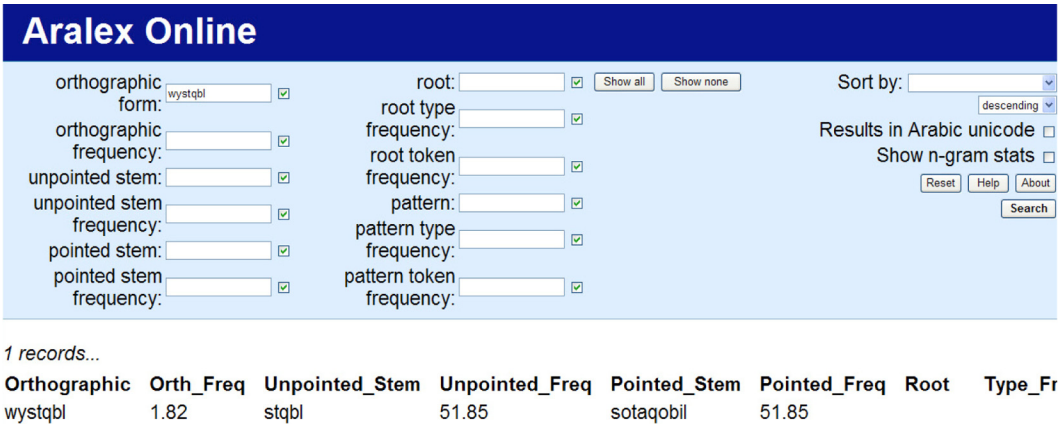


Figure 1. Example of a query using the orthographic string "wystqblh."

gram frequencies for the search string [mstqbl] *future*. These counts are ordered by overall frequency based on combining the bigram or trigram frequency for the orthographic form, the root, and the word pattern.

Researchers interested in compiling a list of items for an experiment in which the words have, for example, a type root frequency and a type word pattern frequency above a certain threshold can do that as well. For example, to obtain a full listing of all the stems that have a root type frequency of more than 15, a word pattern type frequency of more than 300, and an orthographic frequency of more than 10, one needs to enter >15 in the *root type frequency* window, >300 in the word *pattern type frequency* window, and >10 in the *orthographic frequency* box and tick the appropriate boxes. Alternatively, select the *show all* button, and then press *search*. Figure 3 illustrates this query.

The Command-Line Interface

Since the Web interface cannot cover every possible query that a potential user might want to carry out, a CLI is also provided. The CLI offers a powerful, customizable method for the interrogation of the Aralex database. The input to the CLI can be a single word or a text file, allowing batch processing, and the output can be written into a file or displayed on the screen.

To use the Aralex CLI, the user needs to ensure that Java JDK 5.0 or later and Lucene 2.3.2 or later have been installed. An Aralex CLI Java class file and an Aralex Lucene database index are also required and can be downloaded from the Aralex Web site. Once these components are available and the Lucene core JAR is on the system classpath for the Aralex CLI, the interface can be invoked by the command *java SearchDB*. If successful, this should display the input argument format, options, and field

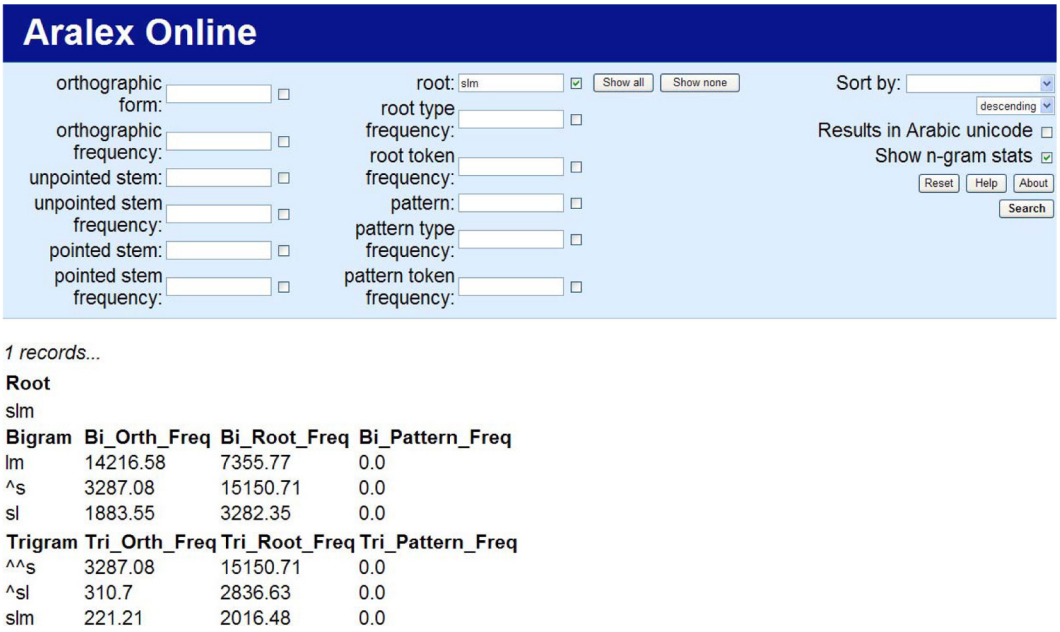
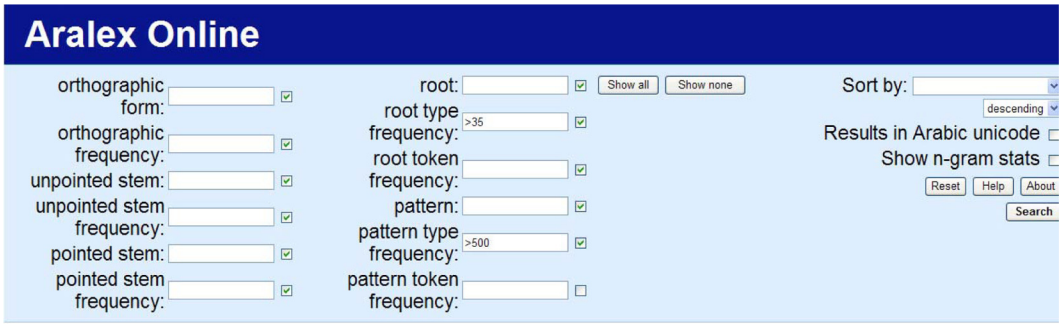


Figure 2. Example of a query using the root {slm} to request its *n*-gram frequencies.



5057 records...

Orthographic	Orth_Freq	Unpointed_Stem	Unpointed_Freq	Pointed_Stem	Pointed_Freq	Root	Type_F
fslmtk	0.03	slm	253.64	sal~am	130.78	slm	42
wAlqA}mp	2.44	qA}m	472.04	qA}im	472.04	qwm	43
fAljmyE	2.29	jmyE	1129.52	jamiyE	1129.52	jmE	41
bAlqAdmyn	0.18	qAdm	322.05	qAdim	322.05	qdm	43
tmAvlhm	0.05	mAvl	19.63	mAvil	19.63	mvl	40
ltHwylhmA	0.03	tHwyl	156.56	taHowiyl	156.56	Hwl	54
vAnytA	0.03	vAny	1386.12	vAniy	1386.12	vny	51
Drbh	1.87	Drb	273.96	Darob	166.78	Drb	42
ytErDwn	6.37	tErD	366.8	taEar~aD	327.02	ErD	47
tqArb	9.86	qArb	42.21	qArib	38.04	qrb	38
AlbvAnw	3.38	bvAnw	5.98	bivAnuw	5.98	bvn	43

Figure 3. Example of a query using numerical input to request a list of items meeting a set of predefined criteria.

names. At this stage, the program requires the directory containing the Aralex index files to be specified. Invoking the command *java SearchDB index_dir*, where *index_dir* is the location of the database index, yields the prompt *Enter query*. From now on, any valid Lucene query can be entered.[10] Note, however, that in the Aralex CLI, the wildcard operator is "%," and not the "*," since the asterisk is part of the Buckwalter transliteration scheme.

To take an example of the batch-processing facility offered by the Aralex CLI, suppose that we have an input file called *test.in* with the following lines:

orthographic: brys

orthographic: Alrys

orthographic: lryAsp.

Should we enter the command *java SearchDB index_dir –b test.in*, the queries will be processed sequentially, and all the frequency information concerning the words in the input file will be returned. For information about *n*-gram frequencies in orthographic forms, roots, and word patterns, the user can use the *n*-gram CLI, which functions in the same way as the main database CLI and takes as input either single-word queries or a whole file in Arabic script or in Buckwalter's (2002) transliteration.

Summary and Future Development

Aralex, a lexical database consisting of 40 million MSA words, is a unique resource that includes information about orthographic forms, pointed stems, unpointed stems, roots, and word patterns. It also provides information about bigram and trigram frequencies in orthographic forms, roots, and word patterns. As such, it offers a unique combination of statistical information about words and morphemes. Other databases for Arabic use untagged corpora and, hence, give approximate frequency figures (e.g., Parkinson, 2006) or simply compile a dictionary of Arabic into computerized format and are thus unable to offer frequency information (e.g., Dichy & Hassoun, 1998).

Aralex should allow, for the first time, a stringent control over the selection of word stimuli and the creation of nonword stimuli for experimental research into Arabic. It can also serve as a source of information for natural language processing development, and it can fulfill an educational purpose by providing basic vocabulary lists and helping second-language learners to automatically find the component morphemes of a given orthographic form.

The flexible architecture of Aralex makes it easy to envisage future development of the database by enlarging the corpus and updating the dictionary. One important development of the database should involve the inclusion of the number of meanings of roots and word patterns and information about the argument structure of verbs. This information will allow us to gain further insight into the structure of the language and the processing mechanisms subserving it.

AUTHOR NOTE

This work was financed by British Academy Large Research Grant LRG 42466 and by the U.K. Medical Research Council U.1055.04.002.00001.01. The authors are most grateful to Sameh Al-

Ansary, Ted Briscoe, Tim Buckwalter, Hubert Jin, Mohamed Maamouri, Fermin Moscoso del Prado Martin, Dilworth B. Parkinson, and Mark Wilding for their help at different stages of the project (as detailed in the text). Correspondence concerning this article should be addressed to S. Boudelaa, MRC Cognition and Brain Sciences Unit, 15 Chaucer Road, Cambridge CB2 2EF, England (e-mail: sami.boudelaa@mrc-cbu.cam.ac.uk).

REFERENCES

BAAYEN, R. H., PIEPENBROCK, R., & VAN RIJN, H. (1993). *The CELEX lexical database* (Release 1) [CD-ROM]. Philadelphia: University of Pennsylvania, Linguistic Data Consortium.

BOUDELAA, S., & GASKELL, M. G. (2002). A re-examination of the default system for Arabic plurals. *Language & Cognitive Processes*, **17**, 321-343.

BOUDELAA, S., & MARSLEN-WILSON, W. D. (2005). Discontinuous morphology in time: Incremental masked priming in Arabic. *Language & Cognitive Processes*, **20**, 207-260.

BOUDELAA, S., PULVERMÜLLER, F., HAUK, O., SHTYROV, Y., & MARSLEN-WILSON, W. D. (2010). Arabic morphology in the neural language system. *Journal of Cognitive Neuroscience*, **22**, 998-1010.

BUCKWALTER, T. (2002). *Buckwalter Arabic Morphological Analyzer, Version 1.0*. Philadelphia: Linguistic Data Consortium, Catalog No. LDC2002L49, ISBN 1-58563-257-0.

CONTENT, A., MOUSTY, P., & RADEAU, M. (1990). Brulex: Une base de données lexicales informatisée pour le francais écrit et parlé. *L'année Psychologique*, **90**, 551-556.

DAVIS, C. J. (2005). N-Watch: A program for deriving neighborhood size and other psycholinguistic statistics. *Behavior Research Methods*, **37**, 65-70.

DICHY, J., & HASSOUN, M. O. (1998). Some aspects of the DIINAR-MRC research programme. In *Proceedings of the 6th International Conference and Exhibition on Multilingual Computing* (pp. 1-6). University of Cambridge, Centre Middle Eastern and Islamic Studies.

FERGUSON, C. A. (1959). Diglossia. *Word*, **15**, 325-340.

FROST, R., FORSTER, K. I., & DEUTSCH, A. (1997). What can we learn from the morphology of Hebrew? A masked priming investigation of morphological representation. *Journal of Experimental Psychology: Learning, Memory, & Cognition*, **23**, 829-856.

HOLES, C. (1995). *Modern Arabic: Structures, functions, and varieties*. London: Longman.

IDRISSI, A., PRUNET, J. F., & BÉLAND, R. (2008). On the mental representation of Arabic roots. *Linguistic Inquiry*, **39**, 221-239.

KTORI, M., VAN HEUVEN, W. J. B., & PITCHFORD, N. J. (2008). GreekLex: A lexical database of Modern Greek. *Behavior Research Methods*, **40**, 773-783.

LÉTÉ, B., SPRENGER-CHAROLLES, L., & COLÉ, P. (2004). MANULEX: A grade-level lexical database from French elementary school readers. *Behavior Research Methods, Instruments, & Computers*, **36**, 156-166.

MCCARTHY, J. J. (1981). A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, **12**, 373-418.

NEW, B., PALLIER, C., BRYSBAERT, M., & FERRAND, L. (2004). *Lexique 2*: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, **36**, 516-524.

PARKINSON, D. B. (2006). *ArabiCorpus*. Available at http://arabicorpus.byu.edu/.

PEREA, M., URKIA, M., DAVIS, C. J., AGIRRE, A., LASEKA, E., & CARREIRAS, M. (2006). E-Hitz: A word frequency list and a program for deriving psycholinguistic statistics in an agglutinative language (Basque). *Behavior Research Methods*, **38**, 610-615.

PLUNKETT, K., & NAKISA, R. C. (1997). A connectionist model of the Arabic plural system. *Language & Cognitive Processes*, **12**, 807-836.

PRUNET, J. F., BÉLAND, R., & IDRISSI, A. (2000). The mental representation of Semitic words. *Linguistic Inquiry*, **31**, 609-648.

REIG, D. (1999). *Dictionnaire arabe-français, français-arabe: As-Sabil al-Wasit*. Paris: Larousse.

SEBASTIÁN-GALLÉS, N., MARTI, M. A., CUETOS, F., & CARREIRAS, M. (2000). *LEXESP: Léxico informarizado del español*. Barcelona: Edicions de la Universitat de Barcelona.

STADTHAGEN-GONZALEZ, H., & DAVIS, C. J. (2006). *The Bristol norms for age of acquisition, imageability, and familiarity*. Available at http://language.psy.bris.ac.uk/bristol_norms.html.

VERSTEEGH, K. (1997). *The Arabic language*. Edinburgh: Edinburgh University Press.

WEHR, H. (1994). *Arabic–English dictionary* (J. M. Cowan, Ed.). Ithaca, NY: Spoken Language Services.

WILDING, M. (2006). *Bootstrapping Arabic pointing and morphological structure*. Unpublished master's thesis, University of Cambridge, St Catherine's College.

NOTES

1. The corpus on which Aralex is built is primarily newspaper text. This contained very few DA words. Where these occurred, we retained them in the corpus but flagged them as dialectal.

2. The fact that DA is generally not written means that relevant corpora are currently less readily available. However, as written versions of DA become more widespread—as, for example, in online women's forums—it will become possible to build appropriate lexical data sets.

3. We are adhering to the traditional notation of the word pattern, where the letters "f, ʕ, l" are used as placeholders to refer to the first, second, and third letters of the root, respectively. Note that we use the IPA symbols to represent Arabic phonemes.

4. This is the citation form of [tˤalaaqah] *eloquence*; however, in connected speech, the final [h] is deleted and replaced by a [t], thus marking the word as a feminine noun and avoiding hiatus.

5. Note that the figure for the word patterns includes many analogical patterns like {fiʕluuqraatˤy}, the pattern for the assimilated foreign word [diimuuqraatˤy] *democratic*.

6. This part of the project was carried out in collaboration with Mark Wilding and Fermin Moscoso del Prado Martin.

7. For a full presentation of the diacritization procedures applied to our corpus, see Wilding (2006).

8. This cross-validation was conducted under the leadership of Sameh Al-Ansary of the University of Alexandria.

9. The integration of the corpus and the dictionary and the development of the front-end interface for Aralex were developed in collaboration with Ted Briscoe and Ben Medlock, in a contract with the iLexIR company.

10. See http://lucene.apache.org/java/2-3-2/queryparsersyntax.html for an overview of the basic Lucene query syntax.

(Manuscript received August 10, 2009;
revision accepted for publication January 12, 2010.)