Developing LMF-XML Bilingual Dictionaries for Colloquial Arabic Dialects

David Graff, Mohamed Maamouri
Linguistic Data Consortium
University of Pennsylvania
E-mail: graff@ldc.upenn.edu, maamouri@ldc.upenn.edu

Abstract

The Linguistic Data Consortium and Georgetown University Press are collaborating to create updated editions of bilingual dictionaries that had originally been published in the 1960's for English-speaking learners of Moroccan, Syrian and Iraqi Arabic. In their first editions, these dictionaries used *ad hoc* Latin-alphabet orthography for each colloquial Arabic dialect, but adopted some properties of Arabic-based writing (collation order of Arabic headwords, clitic attachment to word forms in example phrases); despite their common features, there are notable differences among the three books that impede comparisons across the dialects, as well as comparisons of each dialect to Modern Standard Arabic. In updating these volumes, we use both Arabic script and International Phonetic Alphabet orthographies; the former provides a common basis for word recognition across dialects, while the latter provides dialect-specific pronunciations. Our goal is to preserve the full content of the original publications, supplement the Arabic headword inventory with new usages, and produce a uniform lexicon structure expressible via the Lexical Markup Framework (LMF, ISO 24613). To this end, we developed a relational database schema that applies consistently to each dialect, and HTTP-based tools for searching, editing, workflow, review and inventory management.

Keywords: Arabic dialect, bilingual dictionary, lexical markup

1. Introduction

Beginning in 2008, the Linguistic Data Consortium (LDC) and Georgetown University Press (GUP) have been collaborating on a project to enhance and update three Arabic dialectal dictionaries published by GUP. The source dictionaries are:

(1) *A Dictionary of Moroccan Arabic*, *Moroccan-English, English-Moroccan* (Harrell and Sobelman 2004) [1966]) (hereafter "GUP Moroccan");

(2) *A Dictionary of Syrian Arabic: English-Arabic* (Stowasser and Ani 2004 [1964]) (hereafter "GUP Syrian"); and

(3) *A Dictionary of Iraqi Arabic, English-Arabic, Arabic-English* (Clarity, et al. 2003 [1965]) (hereafter "GUP Iraqi").

These dictionaries have some basic features in common: (a) Arabic words are rendered in Latin letters, except that Arabic letter glyphs for 'ʕayn' ع and 'ħaa' ح are used as supplements to the Latin alphabet to represent these phonemes; (b) Arabic text is distinguished from English text by means of italic vs. normal font; (c) the collation of Arabic headwords uses an ordering of the Latin letters that mimics the canonical Arabic alphabetic sequence; (d) when example Arabic phrases are given, word spellings vary to reflect clitic attachment and contextualized pronunciation (e.g. elisions and consonant assimilations). But differences among the dictionaries are significant:

(a) Arabic headword entries in GUP Iraqi and Syrian are organized according to "consonant skeleton" classes, but GUP Moroccan is not. While the skeleton classes are analogous to the organization by consonantal roots that is typical in dictionaries of Modern Standard Arabic (MSA), GUP classes are based on current dialectal pronunciations; as a result, cognate terms in the two dialects would sometimes fall into distinct skeleton groups, due to the divergent sound changes that distinguish these dialects relative to their common ancestor.
(b) As a consequence of (a), GUP Iraqi and Syrian also follow MSA standards by placing entries for various derived and inflected Arabic forms under their root heading, e.g. the form "maktab" *(office, bureau)* is only found under "k-t-b" (chapter "k", not chapter "m"). In GUP Moroccan, all forms are ordered in a flat sequence according to their pronunciation in isolation ("maktab" is in chapter "m"), and many entries are defined by merely referring to their citation forms (e.g. "passive of …", "plural of …", "active participle of …", etc).
(c) Each dictionary uses a slightly distinct level of phonetic detail, and sometimes uses an idiosyncratic symbol for a common phonetic value.

The published volumes had been keyboarded into word processor files prior to the start of the project, in a manner that preserved the typesetting details (page and paragraph breaks, indentations, and alternations of normal, bold and italic fonts); these files were the primary source for populating database tables that would drive annotation for each dialect.

The Arabic-English portion of GUP Syrian was never published, but the Georgetown University library still retained the full archive of original index cards, as well as a partial set of galley proof sheets. The index cards

were keyboarded and stored in a distinct database table as part of the current project, so that data from the cards could be used as primary source material; copies of the galley proofs were supplied as an aid to annotators.

Our goals for each dialect were as follows:

- Preserve all the content of the original dictionary (except where it is determined to be in error)
- Add new Arabic look-up entries (headwords) with English definitions based on available corpora of current usage in the given dialect
- Transliterate the original Latin-based spellings of Arabic headwords, definitions and phrases into both Arabic script letters for a common native orthography, and International Phonetic Alphabet (IPA) characters for pronunciation
- Decode grammatical details (usually expressed as abbreviations or ordered text fragments) into explicit features (e.g. irregular plural forms, verbal noun forms, etc.)
- Organize and maintain all content in a consistent relational database structure for each dialect
- Extract the database content into XML using the Lexical Markup Framework (LMF, ISO 24613)
- Release materials both in printed book form (for use by linguists and language learners) and as electronically accessible corpora (for use in online or localized search tools and in various natural language processing applications).

The remaining sections of this paper will go into detail regarding the specification of a uniform orthographic strategy across Arabic dialects, design of the database schema and annotation tools, inclusion of novel Arabic headwords from more recent data sources, application of LMF XML structure to these lexicons, and issues raised by the juxtaposition of Arabic script and IPA spellings.

2. Establishing orthographic conventions

Modern Standard Arabic (MSA) is the only form of Arabic for which an orthographic standard exists. The discrepancies between MSA and the various colloquial dialects such as Iraqi, Moroccan and Syrian are widely recognized to be significant, but the divergence among the dialects themselves can often be much more significant, and more disorienting to native speakers of the respective dialects. From the perspective of both language learning and NLP, our expectation is that there is much to be gained by developing a pan-dialectal orthographic convention that makes use, as much as possible, of the common-core etymological bases that all dialects share with MSA.

We therefore adopted the view that, when it comes to representing colloquial dialects in Arabic script for general use, it is better for the "meanings" of the Arabic letters to be more ideographic than phonetic, encoding historical relationships with MSA rather than current pronunciations. This in turn places greater emphasis on the need for an IPA representation as an essential supplement to specify pronunciation and clarify the phonetic divergence among dialects (Maamouri et.al. 2004 a, b).
In order to establish a firm and consistent foundation for using the Arabic alphabet in a pan-dialectal orthographic convention, we devoted significant amounts of time in training, discussions and annotation effort to the assessment of consonantal root classes in each dialect. In all cases, Wehr (1994) was used as a primary reference to identify the root classes for MSA. Whenever dialectal word forms preserved the consonant structure of a cognate root class in MSA - allowing for regular rules of sound change typical to the given dialect - the MSA consonant sequence was adopted in the spelling of the dialectal forms.

In general, the sound change rules tend to involve shifts and mergers among apical consonants, or among dorsal consonants, such that place and/or manner features (e.g. alveolar vs. dental/palatal, stop vs. affricate/fricative) are modified or neutralized. Table 1 provides some prominent examples of correlations among consonants in MSA and the various dialects. Given the regularity of these rules, and the fact that native speakers are generally familiar with at least some of the cognate relations between MSA and their own dialect, we believe that the use of "MSA semantics" (rather than "phonetic semantics") for Arabic letters will not seriously impede, and will likely enhance, the essential function of promoting word recognition when the dialect is presented in written form. For those learning a given dialect as a second language, the highlighting of cognate relations with MSA and other dialects is bound to be a significant boon.

MSA	Iraqi	Syrian	Moroc.
ق	q, g (k, ǰ)	ʔ (q, g)	q (g)
ك	k, č (g)	k	k
ث	θ (f, t)	s, t	t

Table 1: Prominent sound change relations among MSA and colloquial dialects.

As of this writing, our work on the Iraqi dictionary is complete, the Moroccan dictionary is in its final stages of quality checking, and work on the Syrian Arabic headword inventory is still in progress. We can therefore summarize the overall statistics for root-class overlap among the dialects, though the results for Syrian are incomplete. Table 2 shows the total number of distinct "consonant skeleton classes" in each dialect, along with the number of classes identified as shared MSA roots; below that we count up the shared MSA roots that appear in pairs of dialects, and among all three. Note that these are counting only distinct consonantal skeletons, not the number of distinct Arabic headwords (many roots contain several headwords each). The somewhat larger proportion of non-MSA classes in Iraqi is due in part to having a larger number of headword entries incorporated

from recent speech corpora. If we take the union of distinct root classes that are shared with MSA, rather than the intersections tallied in Table 2, we find that over 2400 MSA roots are represented in one or more of the three dialects, which is roughly two-thirds of the root/skeleton inventory in Wehr (1994).

	Iraqi	Syrian	Moroc.
Total classes	4368	3323	3014
Shared w/MSA	1993	2030	1590
Shared w/others	I/S: 1676	S/M: 1157	I/M: 1433
	I/S/M: 1116		

Table 2: Tally of consonantal skeleton classes by dialect (classes shared with MSA are Semitic roots).

It's likely that a number of non-MSA skeleton patterns are also cognates across the dialects, owing to common borrowings, but we haven't attempted to confirm these as such.

3. Database design

As we addressed the Iraqi Arabic-to-English dictionary in the first phase of the project, we began with a relational database schema that was fairly isomorphic with the structure of the original volume: there was a table of consonant skeletons; each row in this table had one or more related rows in a headword table; each headword entry had one or more related rows in a "sense" (English definition) table, and whenever a given definition included an example phrase, this was stored as a related row in a phrase table. In the headword and phrase tables, we stored the original GUP Latin-based orthography, and provided separate fields for the Arabic script and IPA spellings; the headword table also had additional fields for storing the spellings of additional inflected forms (plural, feminine, etc). The English-to-Arabic portion was likewise converted directly to a relational table structure, which was simpler: a table of English headword entries, a table of Arabic definition terms and phrases related to each headword, and a table of example phrases related to the Arabic definition terms.

In moving on to the Moroccan and Syrian dictionaries, the English-to-Arabic portions lent themselves well to equivalent treatment, but the Arabic-to-English content posed some distinct problems. In the Moroccan source material, headword entries were not organized by root/skeleton, and indeed contained no indication of what the root/skeleton was for a given entry; also, many items appearing as "headword entries" were in fact derivative or inflected forms that, in a more conventional Arabic dictionary, would have been placed within or adjacent to the related citation form, rather than being headwords in their own right. In the case of Syrian, the inventories of skeleton patterns, headwords, definitions and example phrases were rather speculative, being based on over 70,000 index cards, many of which were fragmentary, redundant, or of no practical use.

In order to provide a stable and consistent approach for handling Moroccan and Syrian, we chose to decouple the notions of "word form" and "lemma", which had been conflated in the notion of the "headword" element of the Iraqi dictionary. In effect, we replaced the headword table of the Iraqi database structure with two tables, "lemma" and "wordform". In terms of relations to other tables, the lemma table functions just like the earlier headword table: one or more rows of lemmas are related to a given consonant skeleton entry, and one or more rows of senses are related to a given lemma. But in addition to that, each lemma would have one or more related rows in the wordform table, containing Arabic and IPA spellings for each form. At most one wordform row would be labelled as the "citation" form for the given lemma, and other rows would be given suitable labels such as "plural", "feminine", etc.

Although we concluded our work on Iraqi without converting its data to this new table structure, such a conversion does not pose any serious difficulty, and will be applied as we move forward with future uses of the data, so that further tool development can employ a single data model in addressing all dialects. In particular, the tools currently in use for Moroccan and Syrian (which consist of a single software code base applied uniformly to both dialects) can be used for Iraqi as well, once we reorganize its data into the newer table structure. This also supports queries and tool development for cross-dialect efforts: all dialects are contained in a single database, and each dialect is instantiated as a consistent set of tables whose names include the dialect's ISO 693-3 abbreviation (i.e. "ary_lemma", "ary_wordform", etc. for Moroccan, "apc_lemma", "apc_wordform", etc. for Syrian, "acm_lemma", "acm_wordform", etc. for Iraqi).

In loading the tables from primary source data to prepare for annotation, the bulk of the effort was focused on determining how to "decode" the typesetting mark-up into functional information structure. The previously published volumes had been keyboarded into Microsoft Word document files; we used that application to export the data as HTML, and then used an open-source HTML parsing library to extract the text content with its associated typesetting features. Where necessary, the extracted HTML data was edited manually to rectify errors or anomalies in the original keyboarding (which were sometimes due to a faithful rendering of typesetting errors in the original publication); for example, italic font (which was supposed to represent Arabic text) was sometimes used on English content (and normal font was sometimes used on Arabic), or bold font (intended for definition numbering and for English headwords) was mistakenly present or absent. While page breaks in the original mark-up had to be carefully excised (because they could occur between or within headword entries), they needed to be kept track of (because annotators, having the paper

publication at hand, would want to know the page numbers for particular entries). The majority of headword entries tended to employ a consistent range of structural templates, so the mapping of typesetting transitions to structural boundaries was fairly clear, but there was a long tail of variations that would ultimately require manual interpretation as part of the annotation process.

4. Annotation tools

Given the need to handle varied character sets, bidirectional text, transactions with a database, and a varied computing environment, our tool development was essentially focused on HTTP-based user interfaces to support the various stages of directed workflow and the *ad hoc* searches for general review and quality control.

We implemented workflow management by defining annotation tasks with respect to a specific table (the Iraqi headwords or Moroccan/Syrian lemmas, the English headwords, or the example phrases), and then providing tools to present all the information needed entry (some of which might come from related tables) to perform the task for a given table. The workflow was then a matter of marking table rows as needing a particular phase or pass of annotation, assigning the next available entry to an annotator engaged in that task, applying the updates submitted by the annotator, and iterating until all the rows in question had been addressed.

Two predominant search methods are supported: (a) a flexible parameterized search page to list Arabic or English lemma entries that match user-specified criteria (Arabic or IPA orthography, consonantal skeleton, definition content, part-of-speech, etc); and (b) a two-stage browsing interface for Arabic lemmas, in which the Arabic alphabet is presented as a row of links across the top of the page, clicking any letter presents the set of consonant skeletons starting with that letter in a narrow left-hand column of the page, and clicking any skeleton presents the set of lemma entries associated with that skeleton in the main body of the page.

In both search methods, the resulting list of hits is presented as a table containing a variety of information about each entry, and one of the columns in the table provides a link with the appropriate url to go directly to the appropriate editing function with the given entry loaded for use.

5. Introduction of novel Arabic headwords

For Iraqi, we had a fairly large collection of recorded conversational speech that had already been transcribed using orthographic guidelines consistent with the principles described in section 2 above. In particular, Appen (2006) provides a total of nearly 25 hours of recorded and transcribed telephone conversations among native speakers of Iraqi Arabic; the corpus contains nearly 120,000 word tokens, comprising about 18,000 distinct word forms. We were also able to use a much larger collection of transcribed speech that had been created for the DARPA TRANSTAC program, where the focus was on developing speech-to-speech machine translation to support dialogs between English-speaking "subject matter experts" (especially military, medical or administrative personnel) and monolingual Arabic speakers. LDC had participated in TRANSTAC corpus development by distilling a lexicon of Arabic word forms from the transcripts, to provide part-of-speech tagging, clitic segmentation and English glosses (Graff et al., 2006).

The basic procedure was to pre-filter the existing transcript lexicon to extract citation forms that were not already present in the GUP Iraqi dictionary. This was somewhat speculative, owing to differences in orthography, so native-speaker annotators were responsible for determining what the "correct" citation form spelling should be for a given word candidate, and checking whether that spelling was already present in the database.

Nearly 4100 Iraqi terms were added as headwords to the Arabic-to-English dictionary using these sources. At present, we are unable to enhance the inventories for the other dialects on a similar scale, because we lack an adequate corpus of transcribed speech from the specific regions represented by those GUP dictionaries. However, in the course of conducting annotations and searches on the contents of those published volumes, and as a result of web searches by annotators to investigate current usage, we have added several hundred new lemma entries for each dialect.

In all cases, we have left the original inventory of the English-to-Arabic dictionaries unmodified; it was decided early on that the addition of new English look-up terms, even to correlate with newly added Arabic terms, would fall beyond the scope of the current project.

6. Application of LMF XML

Our initial and most essential goal in extracting the lexical database contents into a portable format was to ensure that GUP would be able to render the data efficiently and accurately when typesetting a new edition of the book publication. In this regard, the Lexical Markup Framework XML standard (ISO 24613) seemed to provide a significant advantage, because vendors for book printing services were both willing and readily able to accommodate data in that format. Given the unavoidable complexity of typesetting any bilingual dictionary, let alone one so densely packed with bi-directional text, there was significant value for all parties concerned in having a stable data structure whose design was reasonably well documented and widely recognized.

Available descriptions and examples of LMF and its use cases are not particularly detailed on the matter of applying its XML structures to an Arabic-English bilingual

dictionary. Nevertheless, the standard appears to provide enough flexibility and robustness to support a range of reasonable implementations without creating hardships for creators or consumers of the data.

In order to ensure adequate trace-back to our database from the delivered XML structure, we adjusted the LMF DTD to accommodate "id" attributes on some elements where this had not been part of the published ISO specification. These adjustments have not had any noticeable impact on downstream users of our data.

The process of extracting from our database into LMF XML is controlled at three levels: manipulating the contents of certain status fields in the tables, crafting the conditions and sequencing of queries for data extraction, and post-processing the query results during the construction of the XML tree for the dictionary as a whole. In combination, these steps ensure that the data release contains only the intended inventory (e.g. excluding items that annotators have flagged as "not needed"), that the released items are fully intact and coherent (checking that all essential features are present for each element, etc), and that the ordering of elements in the XML stream is appropriate.

With regard to arrangement and ordering of LexicalEntry elements for the print release of the dictionary, we elected to use this element to instantiate the consonantal root/skeleton headings that dominate one or more lemma/headword entries, as well as using it for the lemmas themselves. The root/skeleton entries are sorted in Arabic alphabetic collation order, and within the more complex and productive root classes, the lemmas are further sorted to place verbs first, then nouns, adjectives, adverbs and other parts of speech; for classes with multiple verb lemmas, these in turn are sorted according to the canonical ordering of verbal form classes as established in traditional Arabic grammar.

The ordering of multiple Sense elements within a LexicalEntry was likewise used to establish the sequence of definitions for a given headword. Apart from these ordering relations, the sequencing of other tags within the XML stream, though implemented in a consistent manner, was not significant, and could readily be manipulated via XSLT transformations.

7. Juxtaposition of Arabic and IPA

There were special challenges posed by the requirement that annotators correct and confirm orthographic forms in both Arabic script and IPA. We considered it important to include diacritics for all short vowels (except for a small set of easily definable environments where the presence and quality of the vowel were fully predictable from the immediate context), and we used Buckwalter transliteration for keyboarding Arabic characters. The partial overlap of synonymous letters between Buckwalter and IPA, combined with the unusual level of attention needed to compare the two spellings for consistency, made this aspect of the annotation a significant load (on top of the sometimes difficult decisions about how the Arabic form should be spelled, which consonant root it should be assigned to, how it should be glossed, etc). The problem was exacerbated by the fact that the correlation between an Arabic string and the corresponding IPA string was actually somewhat irregular, especially with regard to the example phrases, where an Arabic spelling with a sequence of distinct consonants would be rendered in IPA as single long consonant, due to phonetic assimilations at morpheme boundaries. This made it difficult to establish search techniques that would locate potential spelling discrepancies between Arabic and IPA with reasonable accuracy. We are able to accumulate and apply some rules that properly account for many of the acceptable alternations that relate Arabic to IPA, but it remains a problem for quality assessment of the annotation.

8. Conclusions

We have made substantial progress in the direction of establishing a reasonable and pedagogically sound set of conventions for normalizing the application of Arabic script orthography to multiple colloquial dialects, and are applying this approach in creating newly revised editions of three seminal works in colloquial Arabic lexicography. The database structure we've created to support this work promises to be an important resource for pursuing research on cross-dialect issues. We have also created a portable electronic version of these resources, relying on the established standard of LMF XML, which we hope will prove to be valuable for future NLP research involving these dialects, in addition to supporting broader publication of the materials.

9. Acknowledgements

We gratefully acknowledge the sponsorship of the U.S. Department of Education, whose International Research Study (IRS) Grant No. P017A050040-07-05 supported our work on this project. The views, opinions and/or findings contained in this article are those of the authors and should not be interpreted as representing the official views or policies, either expressed or implied, of the U.S. Department of Education's International Research Study program.

10. References

Clarity, B., Stowasser, K., Wolfe, G. eds. (2003 [1965]) *A Dictionary of Iraqi Arabic.* Georgetown University Press. Washington, D.C.

Appen Pty. Ltd. (2006) *Iraqi Arabic Conversational Telephone Speech and Transcripts*. LDC Catalog Nos.: LDC2006S45, LDC2006T16. Linguistic Data Consortium. Philadelphia, PA.

Graff, D., Buckwalter, T., Jin, H., Maamouri, M. (2006) 'Lexicon Development for Varieties of Spoken Col-

loquial Arabic.' In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)* May 22-28, 2006. Pages 999-1004.

Harrel, R., Sobleman, H, eds. (2008 [1996]) *A Dictionary of Moroccan Arabic*. Georgetown University Press. Washington, D.C.

Maamouri, M, Graff, D., Jin, H., Cieri, C. (2004a) 'Dialectal Arabic Orthography-based Transcription & CTS Levantine Arabic Collection.' EARS PI Meeting and RT-04 Workshop, IBM Executive Conference Center, Palisades, NY, USA; November 7-11, 2004. http://www.sainc.com/richtrans2004/

Maamouri, M., Buckwalter, T., Cieri, C. (2004b) 'Dialectal Arabic Telephone Speech Corpus: Principles, Tool design and Transcription Convention.' In *Proceedings of the NEMLAR Arabic Language Resources and Tools Conference* Sept. 22-23, 2004. Pages 55-60. Cairo, Egypt.

Stowasser, K., Ani, M. (2004 [1964]) *A Dictionary of Syrian Arabic.* Georgetown University Press. Washington, D.C.

Wehr, H. (1994) Arabic English Dictionary. The Hans Wehr Dictionary of Modern Written Arabic. Editor: J M.Cowan. (Fourth Edition). Spoken Language Services, Inc. Urbana, IL.