Lexicon Development for Varieties of Spoken Colloquial Arabic

David Graff, Tim Buckwalter, Hubert Jin, Mohamed Maamouri

Linguistic Data Consortium (LDC), University of Pennsylvania
3600 Market St. Suite 810, Philadelphia PA 19104
{graff, timbuck2, hubertj, maamouri}@ldc.upenn.edu

Abstract

In Arabic speech communities, there is a diglossic gap between written/formal Modern Standard Arabic (MSA) and spoken/casual colloquial dialectal Arabic (DA): the common spoken language has no standard representation in written form, while the language observed in texts has limited occurrence in speech. Hence the task of developing language resources to describe and model DA speech involves extra work to establish conventions for orthography and grammatical analysis. We describe work being done at the LDC to develop lexicons for DA, comprising pronunciation, morphology and part-of-speech labeling for word forms in recorded speech. Components of the approach are: (a) a two-layer transcription, providing a consonant-skeleton form and a pronunciation form; (b) manual annotation of morphology, part-of-speech and English gloss, followed by development of automatic word parsers modeled on the Buckwalter Morphological Analyzer for MSA; (c) customized user interfaces and supporting tools for all stages of annotation; and (d) a relational database for storing, emending and publishing the transcription corpus as well as the lexicon.

1. The status of Dialectal Arabic

Since the mid-20th century, the term *diglossia* has been used to describe the sociolinguistic situation in Arabic speech communities (Ferguson, 1959): in each Arabic-speaking region, there are two distinct language systems in use.

Modern Standard Arabic (MSA) is the sole system for official communication in government, news reporting and academia. While MSA is used orally in speeches, broadcast news and formal settings, only a small minority of the population has practical experience or facility in speaking it; for most users of MSA, it is like a second language, somewhat related to their primary spoken language, and its use consists mainly if not exclusively in reading, writing and listening.

Spoken Arabic dialects are the primary languages in these communities, but their usage is almost exclusively oral. All formal instruction in reading and writing is conducted in and for MSA: being literate means reading and writing MSA. Differences between MSA and DA involve a variety of diachronic sound changes affecting both manner and place of articulation for several consonants, as well as alterations in derivational and inflectional morphology that may reflect a restructuring of some underlying paradigms. Efforts to establish an explicit standardization for DA orthography and grammar have arisen only recently and are very rare; for the most part, such standardization does not exist.

So, even though DA speakers may be familiar with a writing system and a wide range of textual resources, this is only marginally related to their daily usage of speech. In creating corpora and linguistic annotations for DA – to address the spoken language for purposes of human language technologies – we lack some of the basic underlying resources that are typically available in other literate languages.

We must first establish an orthography that will serve as an adequate "canonical" written form, but still preserve evidence of significant pronunciation variants where possible, because in the absence of an established orthographic standard, and with only limited samples of speech to work from, observed variations could have equal standing as the basis for canonical spellings and further analysis of the language. Next, given the complex morphological structure of DA, we need to develop a lexicon that will support the automation of morphological analysis, by creating a sufficient body of manual annotations. In the process, we need a means to assure that all annotations can be revisited, amended and refined in an efficient and reliable manner while both transcription and manual analysis are in progress, with suitable feedback to annotators as further transcription and manual analysis are done.

1.1. Issues for MSA-based annotation of DA

The differences between MSA and DA created by diachronic sound changes are significant enough that MSA is unsuitable as a standard orthography for DA. Still, the Arabic script-based writing system is familiar to all literate speakers of DA, and there is a fairly large base of common cognate vocabulary between MSA and DA, making the use of Arabic script, and of orthographic practices that closely resemble those of MSA, an effective means for transcription of DA speech (Maamouri et al., 2004a,b). The keyboard layout for these characters can be learned fairly quickly, and transcribers find it easier to read and verify their typing when it is presented in Arabic script, rather than Latin/ASCII transliteration. It is also useful to distinguish two forms for each word: a "consonant skeleton" form, consistent with standard orthographic practice in MSA, and a "diacritized" form, using the common Arabic diacritic marks (for short vowels, consonant gemination, etc), to represent pronunciation.

For morphological analysis, the situation is more difficult. Given that DA is directly related to MSA, and we have very good tools for analyzing the morphology of MSA, we first made an attempt to adapt the Buckwalter Morphological Analyzer (Buckwalter, 2004) so that it could provide candidate analyses of DA word forms. The task for annotators, we hoped, would then be simply identifying which of several possible analyses was the appropriate one for a given word. When we tried this approach on a set of transcripts drawn from a corpus of Levantine Arabic telephone conversations (Maamouri et al., 2006), we found that the differences between MSA

and Levantine Arabic (LA), in terms of both phonological and morphological composition, were much greater than anticipated. A majority of LA forms needed manual editing of analyzer output, to come up with an acceptable segmentation of the word into morphemes, and to provide correct part-of-speech (POS) labels for the morphemes.

The difficulty was compounded by other design features of the annotation process, which had worked quite well for MSA text drawn from newswire sources in the creation of various Arabic Treebank corpora. In essence, annotators would work linearly through the corpus text, because to identify the correct morphological segmentation, POS tag and English gloss for a word form (MPG annotation), words must be assessed in their phrasal context. If viewed in isolation, a consonant-skeleton form (the only form available in MSA data) may be ambiguous, because different patterns of short vowels, which are left out of this spelling, would signal different morphological structures and related meanings. On the other hand, the pronunciation form of a word (available in LA transcripts) could be misleading when viewed in isolation, because this spelling is intended to preserve pronunciation variants.

Working linearly through a set of texts means that words that occur frequently need to be annotated repeatedly. The success of this annotation on MSA data was due in large part to the fact that the morphological analyzer was already well trained on this language, and the choice of possible analyses that it provided for each word included one that was fully correct for the given context in the vast majority of cases – in its current release, the morphological analyzer provides a correct parse for over 99% of words in MSA text.

When applied to LA, however, the extensive mismatch between the analyzer's models and the forms to be analyzed yielded a much longer annotation process, because so much manual revision was required; but even worse, it also yielded a much less consistent set of annotations, because any frequently occurring form, whether word or morpheme, tended to receive different treatments over the course of the project. The natural tendency for repetitive manual tasks to induce an unavoidable error rate was compounded by the difficulty and indeterminacy that native LA speakers encountered when attempting a formal text-based analysis of their native language for the first time.

2. Basic concepts for DA annotation

Following our experiences with the annotation of LA, we needed to develop a better overall process for the next dialect, which in this case was Iraqi Arabic (IA), based in part on a corpus of recorded telephone conversations (created in 2004 by Appen Pty. Ltd. of Sydney, Australia, soon to be released as a corpus by LDC). The two-layer transcription process was considered to be stable, reliable and valuable, and did not change in any significant way relative to the LA project. However, the MPG annotation process was clearly in need of revision.

An essential requirement was to get away from the linear progression through the text corpus: to have a means for manually analyzing each distinct word form only once, and distributing this single annotation over the text corpus with a minimum of manual effort, regardless of the frequency of the form. Another requirement was to have the annotations stored in such a way that a dialect-specific morphological analyzer could easily be built on the basis of manual annotations, and iteratively refined as improvements were made to existing annotations and more MPG analysis was done.

The remainder of this section goes into detail about how the stages of annotation are structured, and section 3 describes how this structure was implemented as a set of annotation tools and a relational database.

2.1. Two-layered transcription

As described briefly in section 1, native speakers of DA who are literate in MSA can adapt fairly quickly to using a standard computer keyboard in order to transcribe speech in their dialect. But literacy in MSA conveys a unique property: the reader is more accustomed to seeing words in terms of their "consonant skeletons", minus the diacritic marks that identify short vowels. (For most readers of Arabic, the presence of vowel diacritics is associated with reading materials used during the early years of schooling as an aid to acquiring skills of word recognition.) Once a student becomes a competent reader, the absence of explicit vowel marks becomes more of a shortcut than an impediment to understanding the text.

For transcribing DA, we use the distinction of "skeletal" vs. "diacritized" spellings to differentiate between two distinct but equally important types of transcription: one that uses a normative spelling for words, and one that provides accurate phonological details of how the words were actually pronounced in a speech corpus. These two types of transcription, treated as separate layers of annotation for speech data, are summarized briefly below; a more detailed discussion of the principles and rational are provided in Maamouri et al. 2004a,b.

2.1.1. Consonant skeleton

For the transcription of DA, the first stage of effort is to create the non-diacritized spellings for spoken words. This form would be comparable to the standard orthographic practice in MSA text, and part of the effort to establish orthographic conventions for a given dialect consists in striking an appropriate balance between the use of MSA spellings to reinforce word recognition, and the use of novel spellings to reflect more accurately the current phonological structure of the dialect.

By seeing the transcribed text in Arabic script without diacritics during the transcription process, annotators are able to converge more quickly on "consensus" spellings for both stems and affixes. Even though the phonological structure of a given morpheme may vary noticeably across numerous occurrences in recorded speech (due to inherent variability within the dialect, and/or morphophonemic processes that alter pronunciation in particular contexts), it is relatively easy to establish a sense of stem or affix identity and to use that sense to arrive at a normative spelling for each morpheme. In this way, the potential ambiguity of non-diacritized spelling works in our favor: it simply omits much of the variability that would impede uniform word identity, and it provides a textual display that is more intuitively legible to native speakers.

2.1.2. Pronunciation

Once the consonant skeleton transcription is complete for a given recording, a second pass of transcription is done to provide a spelling for each word that is segmentally complete (all consonants and vowels are represented) and qualitatively accurate (in terms of the broad phonetic or phonologically relevant distinctions among both consonants and vowels). The results of this stage are stored as a separate annotation layer, so that both forms of transcription are available for display and processing, either separately or in parallel.

In addition to providing full specification of all vowels, this layer also marks changes in consonantal structure due to morphophonemic rules or inherent variability, including substitution or gemination of consonants. This level of detail provides useful information for the next layer of annotation

2.2. Morphology / part-of-speech / gloss (MPG) annotation

Once we have a sufficient body of speech data that have received both layers of transcription, as described in the previous section, the next task is to extract a structured word list, in which we count all token occurrences according to their "skeletal" spelling while also maintaining token counts for each of the distinct pronunciation spellings associated with a given skeletal form.

An additional feature of the structured word list is that it provides, for each skeleton/pronunciation pairing, a complete index for locating all occurrences of this pairing in the transcript corpus. The index includes the filename, time-stamps and channel information, which link each transcribed utterance to the segment of audio recording that it represents.

We then order this listing according to the overall frequency of each skeletal form, placing the most frequent forms at the top and all singleton forms at the bottom; this sets the order in which MPG annotation will be carried out.

The MPG annotator addresses one skeletal form at a time, and is presented with all the pronunciation forms associated with it in the transcripts, together with their respective frequencies of occurrence. The full index of token occurrences for this skeletal form is also provided, ordered initially according to the pronunciation assigned to each token (the least frequent pronunciation forms are listed first). The token occurrences are presented in a concordance format, showing the preceding and following phrasal context from the utterances where they occurred.

2.2.1. Segmenting / labeling of morphemes

The first task in MPG annotation is to divide the word form into morpheme segments (stem and affixes), unless the word happens to be monomorphemic. This is done by inserting "+" (the ASCII "plus" symbol) to mark morpheme boundaries, as shown in Figure 1.

Next, the annotator needs to add short vowels to each of the morphemes to create a fully diacritized "canonical" (normative) spelling. In general, the selection of short vowels and other diacritics should be based on a pronunciation form already provided in the transcription, but this is left to the MPG annotator's discretion: it may be that a given transcript form reflects a variant pronunciation deemed unsuitable as a canonical spelling, or that among a set of existing pronunciations, the most desirable normative spelling involves a combination of elements from two or more distinct pronunciations. Figure 2 shows an example of this step.

bhAlslfwn → b + h + Al + slfwn

Figure 1: An example of morpheme segmentation

b + h + Al + slfwn → bi + ha + Al + salfuwn

Figure 2: An example of diactritizing morphemes

For each morpheme segment, the annotator must assign a POS label and English gloss. At the beginning of this process for a given dialect, the set of available POS labels, and their possible combinations into poly-morphemic structures, are derived from prior annotations on other corpora. In the annotation of LA (the first dialect to receive MPG annotation), the initial inventory of POS labels came from MSA annotations; for the case of IA, the initial POS label inventory was derived from the existing LA annotations. Figure 3 shows an example of adding POS labels and glosses.

bi	+	ha	+ Al	+ salfuwn
PREP	+	DEM_PRON	+ DET	+ NOUN
with	+	this	+ the	+ cell phone

Figure 3: Assigning POS labels and English glosses

Although we begin with an established inventory of POS labels, we must also allow for the possibility that new labels may be needed for a given dialect, either because it shows a morphological distinction not observed in other varieties of Arabic, or because it merges categories that other varieties have kept distinct. But this becomes problematic if too much allowance is given to the discretion of annotators, especially at the beginning of annotation for a given dialect, because their notions about the grammatical structure of the dialect are still predominantly intuitive and potentially fluid, rather than being based on an externalized formal analysis. Also, as with any other manual annotation, the more novel POS labels are accepted from keyboard input, the more we should expect to find typographic errors and variant renderings of these labels.

2.2.2. Assigning analyses to transcript tokens

Since an intrinsic property of our skeletal orthographic forms is their potential ambiguity, an essential goal for MPG analysis in our current approach is to identify all possible morphological parses of a given skeletal form, or at least to identify enough distinct parses to account for all token occurrences in the given transcript corpus. Upon completing a first iteration of the steps outlined in the previous section to create an initial MPG record for the word, the annotator reviews the listing of token occurrences, and identifies all instances to which that specific analysis is applicable. If additional token occurrences still remain, the steps for creating a new MPG record are repeated, with appropriate differences in the placement of morpheme boundaries, insertion of short vowel segments and other diacritics, and assignment of POS labels and glosses. Token occurrences are then identified to go with this new analysis, and the entire process is repeated until all token occurrences have been accounted for.

2.3. Compiling the lexicon

A variety of important results can be drawn from the accumulated output of MPG annotators:

- The inventory of distinct skeleton/ pronunciation/MPG tuples;
- The quantity and locations of tokens assigned to each such tuple;
- The distinct set of conjoined POS label strings applied to words of two or more morphemes, and their frequencies of occurrence;
- The final inventory of distinct, single-morpheme POS labels (including those created during annotation), along with the frequency of occurrence for each;
- The inventory of distinct pairings of POS labels with canonical morpheme orthography, and the frequency of each pairing.

Each of these summaries can be used to identify various types of "outliers" – potential inconsistencies or mistakes in the annotation. In that regard, it's important to begin using summaries of this sort as soon as possible in the annotation process, especially since the most frequent forms are annotated first. Not only do these initial terms cover the broadest range of tokens in the corpus, but the elements that are established at the beginning will tend to be re-used as the annotation progresses to less frequent forms, so that the overall impact of initial annotation problems is doubly amplified if not addressed early on.

Another complicating factor, which affects both MPG annotation and subsequent lexicon creation, is the likelihood of encountering errors or inconsistencies in the original transcripts, involving one or both transcription layers. Completion of the lexicon logically depends on getting these problems fixed at the source, in such a way that subsequent annotations can be updated accordingly without fear of confusing or corrupting the overall project. To address these concerns, we chose a relational database design as the infrastructure for the project, as described in the next section.

3. Implementation: tools and database

As indicated earlier, we consider the two-layer transcription process to be relatively stable and effective for DA, though we are still actively pursuing further enhancements to the tool and procedures. The source code for this tool (written in Python and using the Qt graphical interface library) is available from the authors on request.

Based on our experience with the tool described in section 3.2 for creating MPG annotation, we believe that this is also very close to being an ideal annotation engine for bootstrapping morphological analysis and lexicon creation for additional regional varieties of DA, though again we know there are several enhancements that would improve its effectiveness. This tool is also written in Python using the Qt library, and although it is somewhat less mature than the transcription tool, we are happy to make it available on request as well.

The database design and the project-specific tools for loading, querying and updating the database contents have reached a state of completion sufficient to successfully create and deliver a pilot lexicon of IA, based on a 20-hour corpus of recorded conversations, comprising over 118,000 IA word tokens, from which we derived 13,000 distinct skeletal forms, 17,600 distinct pairings of skeletal and pronunciation forms, and 18,000 distinct combinations of skeleton, pronunciation and MPG annotations. The pilot delivery included not only the fully detailed lexicon itself, but also the complete 20-hour set of transcripts rendered in a multi-linear form that provides, for each utterance: (a) the speaker/channel/time-stamp identification; (b) the skeletal transcription layer; (c) the pronunciation transcription layer; (d) the morphologically segmented orthographic forms with canonical diacritized spelling; and (e) the corresponding concatenation of morphemic POS labels.

The entire pilot lexicon project (not counting the creation of the two-layer transcriptions, which were completed before the lexicon effort began) spanned a period of about six calendar months of active involvement by the authors and an annotation crew of four, with no one working full time on the project during this period. This included design and creation of the database tables and procedures, creation of the MPG annotation tool, manual MPG annotation on all 13,000 distinct word forms, and a significant amount of review, validation and correction of MPG entries by the authors.

3.1. Transcription (AMADAT)

The user interface for DA transcription takes as its initial input a transcript file that already contains time-stamps that correctly delineate the utterances in a corresponding audio file of recorded speech. The person transcribing is presented with a numbered list of time-stamped utterance segments, and control methods, both at the keyboard and on the screen (for selection/activation via the mouse), for scrolling up and down over the utterance list, playing the audio for the current utterance, and inserting a variety of standard annotation tags, to mark acoustic events such as laughter, cough, hesitation sound, external noise, and so on.

Keyboard input of characters for transcription is mapped using the Buckwalter transliteration scheme

(BWT), which assigns a character on the standard ASCII keyboard for each of the Arabic characters needed in transcription. During the initial (skeletal) phase of transcription, the keyboard input is displayed in a text pane at the bottom of the tool, and as soon as the user types the "return/enter" key, the current text is copied to the corresponding utterance line in the scrolling text pane at the top of the tool, where the full set of utterances can be reviewed. In the pronunciation phase, the existing skeletal transcription is provided again at the bottom, and the user has the option of making corrections to that layer of annotation. Completion of the pronunciation layer involves using two central text panes, where both Arabic script and BWT strings are displayed; because of the difficulty of text cursor navigation in diacritized Arabic script, only the BWT display is used for editing and adding diacritics in this layer of annotation.

The final output of the tool is a plain-text transcript file (written in BWT) with one line per time-stamped utterance, containing delimited fields for the time stamps, speaker and channel ID's, the skeletal transcription, and the pronunciation transcription.

3.2. MPG creation (ABUMORPH)

This tool takes as input a plain-text file containing one or more data structures, where each structure contains the following information for a given skeletal word form:

1. The skeletal orthographic form itself
2. One or more pronunciation forms associated with this skeleton;
3. One or more existing MPG analyses applied to this skeleton (initially, there are none of these in the file – they are created by the annotator);
4. A complete list of token occurrences for this form, indexed by the pronunciation applied to each token.

In addition to being indexed by pronunciation, the token list also includes all the information needed to assess the token and assign it a specific MPG analysis. This includes the file name, channel and time-stamps for the utterance containing the token, along with the preceding and following context, if any.

As the user creates MPG analyses and assigns tokens to each one, the analysis strings are added both at the top of the structure and at the end of each token-occurrence record. Once all tokens in a given data structure have been assigned to analyses, the user moves on to the data structure for the next skeletal word form.

3.3. Relational table schema

In order to implement the transduction from the original transcript files into the data structure files for MPG annotation, we defined a relational table schema that would store both the transcripts and the structured word list. Also, in order to manage the annotations of individual morphemes, the schema would need to include a separate table for these units.

3.3.1. Core data tables: files, turns, lex, morph

We began by defining the primary types of external data in the schema. To handle the transcripts, there is a table listing the transcript files, along with essential metadata about each file: the file name is used as the primary key field, and metadata includes who did the transcription and when the file was imported into the database.

A separate table is used to maintain the list of utterances, referred to as "turns" for convenience; each turn entry is given a unique numeric ID, and cites its file ID as a foreign key relation; additional information about the turn includes its time stamps, speaker-ID and channel.

The lexicon table ("lex") is initially used as a "token-type" table: reading through a given transcript file turn by turn, each time a new orthographic type is encountered – whether an Arabic word or any sort of non-lexical annotation – it is added to the lex table and given a unique numeric ID, while previously seen types are simply indexed to existing lex records, using the "tokes" mapping table described below. If the token is a punctuation mark, word fragment or non-lexical annotation (e.g. "noise", "laugh", etc), this is reflected in the lex entry to distinguish it from DA word forms and exclude it from later annotation. For the DA word forms, a new entry is created if the current token represents a novel combination of skeletal and pronunciation spellings. This leads to having multiple lex entries with the same skeletal form, because different pronunciations were posited for it, and can also yield multiple entries with the same pronunciation, if this has been posited for more than one skeletal form.

The "morph" table does not come into play until some amount of MPG annotation has been done and the results are imported into the database, as described in section 3.4 below. Information associated with each morph entry includes a numeric ID, the canonical spelling, the POS label, and the gloss. Different entries are created for each distinct combination of spelling and POS label; if two different MPG annotations (i.e. drawn from two distinct skeletal forms) show the same spelling and POS label, their respective glosses are combined (if different) or collapsed (if identical), and a single morph entry is maintained.

3.3.2. Mapping / relation tables: tokes, segs

As tokens are read from the transcript and distinct types are entered into the lex table, the "tokes" table is populated to maintain the relationship between turns and types: for each token drawn sequentially from a turn, a tokes table entry stores the numeric turn ID, the numeric lex ID for tokens of this type, the sequence number of the token within the turn, and any additional annotation that may be specific to this one token occurrence (e.g. if it was marked as mispronounced or not clearly audible).

In a similar manner, as MPG annotations are read in and each analysis is split into its component morphemes, the "segs" table keeps track of the relation between these components and the lex entries that are built from them. For each morpheme component of an analyzed skeletal pronunciation/MPG form, the segs table stores the numeric morph ID for the morpheme, the numeric lex ID for the word form, and the sequence number of the morpheme within the word.

3.4. Querying / editing lexicon and transcripts

Once the initial transcripts have been loaded into the database, the information stored there makes it possible to recreate the transcript files in their full original detail using a sequence of relatively simple queries. It is also fairly trivial to produce a variety of summaries for things like n-gram distributions, pronunciation variants, speaking rate, etc. An important type of query for us is the creation of the data structure files used as input to MPG annotation, as described in section 3.2.

But the most important factor in the design is the range of operations that can be implemented as updates to the various tables. Primary among these is the process of importing the MPG analyses. Currently, this process operates as follows:

1. For each distinct skeletal/pronunciation/MPG annotation, a new entry is created in the lex table.
2. All token occurrences associated with this annotation (i.e. entries in the "tokes" table) are updated so that they cite the new lex entry instead of the old one that was cited in the initial loading of transcripts into the database.
3. Entries in the morph table are located or inserted as needed to identify each of the morpheme components in the MPG analysis, and rows are inserted into the "segs" table to relate these components to the new lex entry.

Assuming that the MPG annotation has exhaustively assigned all token occurrences to analyzed forms, the end result of the import process is that the various lex entries created by the initial loading of transcripts will now show up as "unattested forms" – that is, all the word tokens are now associated with the newer, analyzed lex entries instead. If at a later time additional transcript files are loaded into the database, the unanalyzed forms may become "attested" again, and at that point, the MPG annotation becomes a matter of determining which if any of the existing analyzed forms is appropriate for the newly added tokens.

Other operations include (but are not limited to):

- Spelling normalization: merging distinct skeletal forms and/or pronunciations by remapping tokens from one lex entry to another
- Error correction: updating a spelling or label field to fix typographic mistakes
- Revision of MPG analyses: updating relations in the "segs" table to associate lex entries with a new set of morphological components

In all cases, the updates have an immediate and consistent effect on subsequent queries to reconstruct the transcript corpus.

3.5. DA-specific morphological analyzers

Having done manual MPG analysis on 13,000 distinct word forms for IA, we are confident that we can achieve a significant improvement in efficiency by building a process to extrapolate analyses automatically from existing morphological annotations. Since manual annotation covers the most frequent forms first, we quickly acquire a body of annotations that will cover a large percentage of tokens in unseen data. In addition to word form coverage, we can use the morpheme component and sequence data from the morph and segs tables to construct a dialect-specific morphological analyzer, on the model already established for MSA (Buckwalter, 2004). This consists of simplified look-up tables that list possible (combinations of) prefixes and suffixes, as well as known word stems, and an appropriate set of string matching algorithms that will propose possible or "allowable" segmentations for previously unseen word forms. This should make it possible to expand the lexicon and the coverage of MPG annotation for a given form of DA at a significantly accelerated rate.

Other avenues for research along these lines include investigations of the relatedness / distance between different regional forms of DA, in terms of shared morphological and phonological structures, and the use of analyses from one dialect to seed the development of resources in another.

4. Conclusions and future work

The approach described above has been used successfully to create a moderately sized lexicon of Iraqi Dialectal Arabic, providing a complete set of pronunciation, morphology, part-of-speech and English gloss annotations for a transcript corpus of nearly 120,000 word tokens. Although the breadth of vocabulary in this pilot corpus is relatively small, we believe the procedures and data structures created in this effort hold substantial promise for future annotation efforts in this domain, not only in terms of rapid expansion for the colloquial Arabic lexicons already in development, but also in terms of creating equivalent corpora and lexicons for other dialects where these resources do not yet exist.

5. References

Buckwalter, T. (2004): 'Buckwalter Arabic Morphological Analyzer Version 2.0', *LDC Corpus Catalog No.LDC2004L02*.

Ferguson, Charles (1959b): 'Diglossia', *Word*, 15, pp. 325-340.

Maamouri, M., T. Buckwalter, C. Cieri (2004a): "Dialectal Arabic Telephone Speech Corpus: Principles, Tool Design, and Transcription Conventions", NEMLAR International Conference on Arabic Language Resources and Tools, Cairo; http://papers.ldc.upenn.edu/NEMLAR2004/Dialectal-Arabic-telephone-speech-corpus.pdf

Maamouri, M., D. Graff, H. Jin, C. Cieri, T. Buckwalter (2004b): 'Dialectal Arabic Orthography-based Transcriptions and CTS Levantine Arabic Collection', EARS RT-04 Workshop; http://www.ldc.upenn.edu/Projects/EARS/Arabic/EARSMtgFINAL.12142004.doc

Maamouri, Mohamed, Tim Buckwater, Hubert Jin, David Graff (2006): Levantine Arabic Transcripts, *LDC Corpus Catalog No.LDC2006T07*.