Finding Entries in an On-line Arabic Dictionary

Sarah C. Wayland, C. Anton Rytting, David Zajic, Timothy Buckwalter, Jason White, Corey Miller, Jeffrey Carnes, Nathanael Lynn, Paul Rodrigues, Michael Maxwell, Evelyn Browne
Center for Advanced Study of Language
Contact: {swayland, crytting, dzajic, tbuckwalter, jwhite, cmiller, jcarnes, nlynn, prodrigues, mmaxwell, ebrowne}@casl.umd.edu
http://www.casl.umd.edu/node/525

In this paper, we describe a new Arabic spelling correction tool called Arabic Did You Mean….? designed specifically to help non-native learners look up words in Arabic electronic dictionaries.

Mistakes that hinder successful dictionary lookup can stem from a few underlying causes:

1. Difficulty discerning phonemic contrasts (such as /q/ and /k/),
2. Difficulty discriminating visually similar letters (such as غـ and فـ),
3. Incorrectly reconstructing a citation form of an inflected word (such as trying to find the singular of مباريات by looking up the non-word مبارية).
4. Words read on Arabic-language opinion forums and blogs, or heard in audio or video webcasts are usually in dialect, and as such have different orthography, morphology, and lexical content from the Modern Standard Arabic commonly taught in foreign language classrooms.
5. If the language learner doesn't have access to Arabic keyboards, they sometimes make errors based on an unfamiliar or unintuitive Romanization scheme.
6. Users sometimes make simple typographical errors, such as hitting a key adjacent to the one intended.

Because Arabic language learners can generate a wide range of alternate inputs, a tool that suggests spelling corrections has the potential to make it easier for the learner to find unfamiliar words in existing lexicons.

We have created a spelling corrector for Arabic dictionary lookup that accepts input in a Romanization system known as the Standard Arabic Technical Transliteration System, or SATTS, verifies whether or not the query matches a citation form[1] in a bilingual Iraqi Arabic-to-English dictionary (Woodhead & Beene, 2003), and suggests similar citation forms the user may have intended.

Following Beesley (Beesley, 1998; Beesley & Karttunen, 2003) the underlying spelling correction algorithm relies on a weighted finite state transducer (FST) that assigns weights based on confusion matrices representing the various error types, including sound confusions, visual (Arabic letter) confusions, and keyboard proximity errors. These confusion matrices can be modified in order to adapt the spell-checker to the error types associated with different Arabic dialects, keyboard layouts, or even the listening errors of students who speak a native language other than English.

As we know of no existing spelling error corpus for Arabic, we simulated error data by generating a corpus using a Noisy-Channel model for error production. For the noise model, we learned the kinds of errors a nonnative speaker could make from a corpus of transcribed Arabic speech elicited from learners of Arabic during an imitation task (Sethy, et al., 2005). The elicitations were common MSA greetings and simple conversations uttered by native speakers of Levantine Arabic and Iraqi Arabic. We evaluated our system by comparing it with a baseline based on a standard spell-checking method, known as Levenshtein (1966) distance, which assumes no language-specific knowledge. Our system got a significantly higher Mean Reciprocal Rank (MRR) score on a test corpus of Arabic query strings and intended words than the baseline Levenshtein version ($\Delta t = 10.95$, $p=0.0001$). Both sound-based confusions and variant spellings contributed to the improved MRR score (Rytting et al., in press.)

Once we determined that our approach was reasonable, we needed to create a user interface that would display the results in a way that was meaningful to English speakers looking up words in Arabic (see Figure 1, below.) After interviewing a number of Arabic language learners at a variety of skill levels, we determined that users should be able to enter their queries using either native Modern Standard Arabic (MSA) or SATTS.

[1] A *citation form* is the form of a word that heads a lexical entry and is alphabetized in a dictionary.

These potential users told us that the output of our spelling correction algorithm should list not only the **citation form** and the type of **inflection** of the word (e.g., plural, citation, etc.), but also the **root form**[2] of the word, as well as its **part-of-speech** (POS). In addition, they requested that the system list the **definitions** associated with the root form.

The citation forms returned by the spelling correction algorithm have a rating associated with them that reflects the weights returned by the FST. We translate these weights into an integer rating system that ranges from one to five, with one being the least likely match, and five being the most likely match. These **ratings are reflected as the number of stars** in the user interface. Perfect matches are highlighted in yellow. Because rankings are based on the citation forms and not the root forms, a single root form may have more than one citation form associated with it; each of these citation forms may have a different ranking. Thus, entries are grouped by root, with the order of the roots determined by the rating of its highest-ranking citation form.

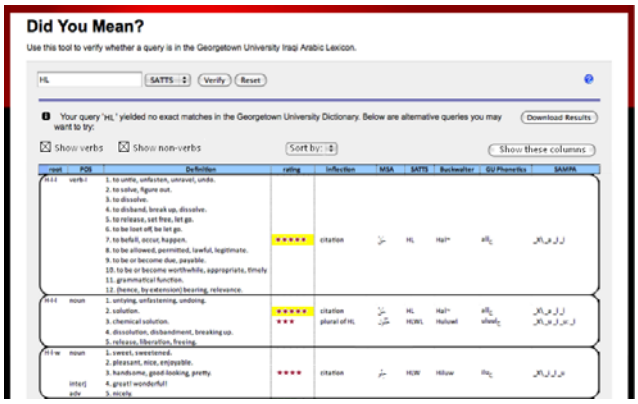
Other columns show the inflected form of the root as returned by the FST in a variety of output formats, including Modern Standard Arabic (**MSA**), **SATTS**, **Buckwalter** (another Romanized form of Arabic), **GU Phonetics** (the sound-based form listed in the Iraqi Arabic dictionary published by Georgetown University), and **SAMPA** (a computer-readable phonetic script based on 7-bit printable ASCII characters known as the Speech Assessment Methods Phonetic Alphabet). A dialog box allows the user to **show or hide any of the columns**.

To limit the output, users can ask the system to show only **verbs** or **non-verbs** (a distinction that is more meaningful to Arabic speakers than to non-Arabic speakers). In addition, although the default output is ordered by rank, we allow users to **sort** by whatever column makes sense to them.

Lastly, our users wanted to be able to download the results into a text file for further processing in other program. To facilitate that, we allow them to **download the results** into a comma separated value (CSV) text file.

We are in the process of finalizing our design for this tool; the next step is to collect feedback and usage information comparing Arabic dictionary lookup when users have access to the "Did You Mean...?" tool, and when they do not.

We have plans to extend our system to allow lookup of non-citation word forms, such as inflected and morphologically complex forms, as well as named entities.



REFERENCES

1. Beesley, K. R. (1998). Arabic morphology using only finite-state operations. In *Proceedings of the Workshop on Computational Approaches to Semitic Languages*, 50-57, Montreal, Quebec.
2. Beesley, K., & Karttunen, L. (2003). *Finite state morphology*. Stanford, CA: CSLI Publications.
3. Center for Advanced Study of Language, University of Maryland. (2008). *Iraqi Arabic corpus*. [Data file].
4. Erwin, W. M. (2004). *A short reference grammar of Iraqi Arabic*. Washington, DC: Georgetown University Press.
5. Levenshtein, V. I. (1966). "Binary codes capable of correcting deletions, insertions, and reversals." *Soviet Physics Doklady*, 10, 707–710.
6. Rytting, C.A., Buckwalter, T., Wayland, S.C., Hettick, C., Rodrigues, P., Zajic, D., (in press). Spelling Correction for Dialectal Arabic Dictionary Lookup. ACM Transactions on Asian Language Information Processing.
7. Sethy, A. Mote, N., Narayanan, S. and Johnson, L. (2005). "Modeling and automating detection of errors in Arabic language learner speech," in *Proc. of Eurospeech, Interspeech*, Lisbon.
8. Standard Arabic Technical Transliteration System. (2010). Retrieved February 03, 2010, from *Wikipedia.org*: http://en.wikipedia.org/wiki/SATTS
9. Woodhead, D. R., & Beene, W. (2003). *A Dictionary of Iraqi Arabic*. Washington, DC: Georgetown University Press.

[2] The *root form* of an Arabic word is the base template form (three to five consonants) that, by the application of strict morphonemic rules, decline into noun and verb forms reflecting gender, plurality, voice, and other aspects of the word's meaning.