Making Vocabulary Corporeal: Arabic Learners, Vocabulary Development, & arabiCorpus[1]

Amy Johnson (Massachusetts Institute of Technology)
Mike Raish (Georgetown University)

Abstract

Recent years have seen increasing integration of language corpora into second language instruction. Arabic's rich polysemy and multiglossic spectrum can complicate vocabulary development for learners. Corpora thus offer great potential as complementary tools for Arabic learners. This study examines two variants of a corpus-based exercise used with fourth-semester adult Arabic learners. Learners analyzed data for select Arabic vocabulary items in the online arabiCorpus. Students in both variant groups proved themselves adept at using the corpus, interpreting data, and articulating metalinguistic hypotheses about nuances of meaning.

Keywords: Corpus-based language learning, vocabulary building, Arabic, arabiCorpus.

One of the most challenging tasks for learners of a foreign language is to understand *which* vocabulary item to use *when* and *why*. This problem can be conceptualized in a number of ways. Within Peirce's framework of object-sign-interpretant and icon-index-symbol (1955[1940]), learners must move beyond the object a vocabulary item signifies to understand what social or linguistic elements it indexes. In the context of Bakhtin's theory of speech genres (1986), learners must recognize the correct genre and then deploy a genre-appropriate individual style within it. And in Austin's schema of

[1] Our thanks to the students who participated in this study, as well as to Hanaa Kilany, Dilworth Parkinson, Karin Ryding, and the article's anonymous reviewers for encouragement, assistance, and thoughtful critique.

speech acts (1975[1962]), learners must understand how vocabulary items interact with felicity conditions to acquire different kinds of performativity. What all of these theories and their later refinements and developments share is an awareness that vocabulary items are used not in isolation, but in context—context that is in turn shaped by many other contexts. Use is sociolinguistic and strategic.

Although language learners may be exposed to vocabulary items through nontextbook materials such as media in the target language or direct communication with a teacher or language partner, the structure of evaluation in most classes is such that the primary exposure points through which students acquire new vocabulary items are lists, exercises, and perhaps dictionaries or glossaries at the back of a textbook. And although these are, themselves, contexts of a sort, they aren't contexts that offer students much information with regard to even basic larger contextual features such as collocations or frequency of use, much less information about the registers in which they are used or the emotional resonance (Kramsch, 2011) they may have acquired for different groups.

Building a vocabulary is critical to all language learning, but Arabic offers a number of particular vocabulary-building challenges for the language learner. Its history and breadth of use has caused Arabic to develop into a richly polysemic language, with a standard written form as well as multiple significantly diverse dialects. Thus, not only may vocabulary items have numerous distinct meanings, Arabic is also replete with synonyms and lexical items strongly associated with different registers. Further, when Arabic is written without full diacritical marks—as it is in most printed forms other than religious texts and texts for children—vocabulary items that represent different *ʔawzaːn* or voweling patterns can appear to be homographs. Thus, *qabila* [to receive], *qabbala* [to kiss], *qablu* [before], *qubl* [fore part] and *qibal* [power] all typically appear as قبل

However, as Parkinson notes in his 1985 analysis of *Elementary Modern Standard Arabic*, in the context of proficiency-based language teaching, it is critical to develop vocabulary in the early stages of learning Arabic. Exposure to only a modest range of vocabulary items in these stages can lead to advanced Arabic learners who feel they don't have the words to express themselves. How,

then, can we help learners to navigate these various challenges in order to build the nuanced vocabularies they need?

Linguistic corpora offer a solution to some, though not all, of these problems. (At the moment, with a few exceptions, most Arabic corpora consist primarily of texts that were originally composed for a written context; this necessarily limits their relevance with regard to the vocabulary of spoken Arabic, which can have significant differences in terms of both form and meaning. Further, even in those corpora that include transcriptions of spoken Arabic, transcription can be problematic and/or inconsistent due to diglossia-linked language ideologies.) Because corpora collect large datasets of authentic material, they can assist students to situate vocabulary items in context, offer direct feedback on frequency of use and common collocations, and provide a foundation for the construction of laminated, sociolinguistic meaning.

Here we describe a vocabulary-building exercise using arabiCorpus with intermediate students of Arabic, its results, and future directions for the use of corpora in the Arabic classroom. First, however, we will briefly examine the use of corpora in language learning more broadly.

Corpora in Language Learning

Corpora can influence language learning in a number of ways, such as through the development of textbooks, reference grammars, and dictionaries (e.g., the *Touchstone* series (McCarthy et al., 2005); *Longman Grammar of Spoken and Written English* (Biber et al., 1999), *A Frequency Dictionary of Arabic: Core Vocabulary for Learners* (Buckwalter & Parkinson, 2011); corpus-based learning activities; and studies of learner corpora. Further, corpus studies not targeted specifically to language learning can nonetheless yield important results for language teachers and learners. Römer (2011) divides applications of corpora in language instruction into two categories: corpus tools, or "the actual text collections and software packages for corpus access"; and corpus methods, or "the analytic techniques that are used when we work with corpus data" (206). These two may often be integrated in practice. Thus, in our study, students were introduced to analytic techniques (corpus methods) while using arabiCorpus (corpus tool) as part of a vocabulary-building exercise.

As many note, the descriptivist nature of corpora—its collection of authentic materials—aligns it with language learning that emphasizes Hymes's notion of communicative competence (1972). Within this context, selection and design in corpora figure considerably into its value as a learning tool. Thus, although the web itself serves as the most comprehensive corpus that is also easily accessible, its lack of explicit categorization and annotation can make utilizing its search results challenging. Davies (2010) suggests that the benefits of using a large designed corpus, such as the Corpus of Contemporary American English (COCA) or the British National Corpus (BNC), include the organization of a wide range of material into separate, distinct genres. Language learners engaging with such corpora thus benefit not only from exposure to authentic materials, but also from the metalinguistic system inherent in the separation of texts into genres.

Research suggests that corpora can be useful in helping learners to identify, among others, formulaic expressions that native speakers are sensitized to (Ellis & Simpson-Vlach, 2009), discourse functions of structures like the existential *there* (Palacios-Martínez & Martínez-Insua, 2006), the rhetorical and interpersonal functions that items like contrastive adverbs perform (Charles, 2011), metaphorical usage of vocabulary items (Chambers, 2011), and the lexical patterns that recur in translated texts as opposed to nontranslated texts (Dayrell, 2008). Further, researchers suggest corpus-based exercises can be designed to teach phraseology (Philip, 2011), as well as improve reading comprehension and explore discourse functions of fillers (Reppen, 2010). Corpora have also been used to examine vocabulary introduced by textbooks in aggregate. A survey of popular Spanish textbooks found that only 10–50 percent of textbooks' vocabulary items corresponded with lemmas identified as most frequent in a frequency dictionary based on the Corpus del Español (Davies & Face, 2006).

A number of studies have underlined the need to engage with language teachers in order to help them adopt and adapt corpora for classrooms (e.g., Al-Sulaiti & Atwell, 2006; Römer, 2011). We hope that by describing the process and results of our corpus-based vocabulary-building exercise we can offer insight into additional ways of utilizing corpora in the language classroom.

arabiCorpus

arabiCorpus (arabicorpus.byu.edu) is a medium-sized, plain text corpus of nearly 200 million words, created by Dilworth Parkinson and hosted by Brigham Young University. Though divided into five genres (Newspapers, Modern Literature, Nonfiction, Egyptian Colloquial, and Premodern), the core of the corpus is the Newspapers category—and thus Modern Standard Arabic with an emphasis on media Arabic. This category consists of full-year datasets of newspapers from different regions of the Arab world—including Egypt, Morocco, and Syria—as well as pan-Arab newspapers such as *Al-Hayat*. The other categories draw from a variety of sources, including modern literature, classical literature, scientific texts, the Qur'an, chats, and plays. One of the current goals of the project is to add texts to the Modern Literature and Nonfiction genres, as well as develop the corpus's historical depth (Parkinson, 2011).

As its focus on written texts demonstrates, arabiCorpus was not designed to present a balanced sample of Arabic genres. Indeed, Parkinson describes its initial selection principle as that of ""take whatever you can get your hands on" and try to fill in the holes later" (2011). It is neither part-of-speech annotated nor lemmatized, but it applies parameters to determine word boundaries and search for common morphological variations depending on what part of speech the user has selected. E.g., if the user identifies her search term as a noun it returns not only the bare form, but also examples with the definite article, pronominal suffixes, and connectors such as *wa-*, *fa-*, *ka-*, etc.; for verbs, in addition to the bare form, the corpus returns conjugated forms and examples with pronominal suffixes, connectors, etc. In contrast with research-focused corpora, which often restrict access and require technical skills to navigate, arabiCorpus was specifically designed to be broadly accessible and straightforward to search, in order that it might be useful to language learners and teachers (2011).

arabiCorpus users can search for a particular lexical item or string of lexical items within the corpus as a whole, within the different genres, or within specific genre subsections. A search in arabiCorpus returns a number of different statistics, including: the overall frequency of the term per 100,000 words in the selected

search category; the various word forms found, as well as their context and frequency; the most common words found before and after the item; the corpus subsections in which the item is found; the immediate context in which the item is embedded; and often large sections of the original article or passage in which the item appears. Searching within a genre subsection, such as the 2010 dataset of *Al-Masri al-Yawm* newspaper, additionally returns results organized by the categories contained within the paper itself—economy, sports, national news, cinema, etc.

In addition to its genre divisions, the geographical breadth of the corpus allows users to compare variations in occurrences of particular lexical items across different regions. Thus, for example, Parkinson has used corpus data to elucidate regional differences in journalistic MSA, including synonym choice (e.g., *hātif* vs. *tilifūn*) and variation of the future particle (2010).

The arabiCorpus Exercise

We assigned a vocabulary exercise utilizing arabiCorpus to two groups of intermediate Arabic learners in the second half of the fourth semester of university Arabic instruction. Our goals were twofold: 1) to observe how learners adapted to using the corpus and its analytic functions as a learning tool; and 2) to evaluate how to successfully incorporate corpus-based exercises into structured study of Arabic. The assignment was used in conjunction with the introduction of new vocabulary in chapter 9 of part two of the *Al-Kitaab* series (2006).

A brief description of vocabulary presentation in *Al-Kitaab* is in order. The first and second editions of the series typically introduce new vocabulary items in the form of an alphabetical list included at the start of each chapter. Each Arabic item is paired with its English gloss; in parts two and three of the series, items related by word root to previous vocabulary items are also paired with the more familiar cognate. Not only does such an approach (by no means limited to this series) strip vocabulary items of useful context, such vocabulary lists often oversimplify meaning and suggest inaccurate usage patterns. Three areas of simplification that we sought to address with the arabiCorpus exercise include the presentation of synonyms, the presentation of prepositions associated with verbs,

and the limited definitions provided for words with multiple meanings. Thus, for example, the vocabulary list for chapter 9 introduces the words *ʔadilla* and *dalāʔil* [evidence] as synonyms, with no indication that their usage or meaning can differ; similarly, the item *naḥwa* is defined as "toward," although when students encounter this item outside of the textbook, it is more likely to mean "approximately."

We used two variants of the same exercise; although both variants explored our two goals, the first variant was weighted more toward observation of how learners adapted to using the corpus, while the second focused more on the integration of a corpus-based exercise into a structured learning environment. Each exercise asked students to research several preselected target items in the corpus and analyze their associated frequency data to hypothesize about nuances of usage.

The seven students of Group A were each given three vocabulary items from the chapter list to investigate through the lens of arabiCorpus. Two of these three items were simply individual vocabulary words/phrases; the third was a *challenge item*, a pair of related vocabulary items to compare (e.g., the alternate plural forms *ʔadilla* and *dalāʔil* [evidence], presented as semantic equivalents in the chapter). For each item or comparison, students were instructed to select three examples that illustrated their analysis of the item's usage. After gathering the data, students presented their findings to their peers via class blogs and discussed each other's results in class.

The fourteen students of Group B were each given four vocabulary items to explore with arabiCorpus in three stages. Similar to the exercise given to Group A, three of the vocabulary items were individual words/phrases and one was a challenge item. For the first stage (analysis), students were asked to answer nine questions regarding frequency and usage for every item. (E.g., "Is the word more likely to occur in novels or newspapers?" "What kind of word(s) commonly follow the target word? Be as specific as you can." And, "From the context, what does the target word mean in each of these sentences? Try to be as nuanced as possible."). Further, they were instructed to analyze and articulate the usage differences of the challenge item. As with Group A, they presented their findings to their peers via class blogs and discussed results in class. For the

second stage (production), students were instructed to read four of their classmates' blog entries, and then add an exemplary sentence of their own in the comment section of the blog. For the third and final stage (evaluation), students were asked to respond to these comments, assessing their usage in comparison with the data collected through arabiCorpus.

Findings

The blog posts for both groups demonstrated that students were universally comfortable using the corpus as a tool for analysis and gathering metalinguistic data. In their posts students noted, among other things: 1) usage differences with regard to prepositional verb phrases; for example, noting that the verb *taja>waba* [to respond positively] is most frequently used with the preposition *maʕ* [with] although it can also occur with *min* [from] (the textbook does not indicate *taja>waba min* as a permissible combination); 2) contextual variation of alternative plural forms of nouns, as in the previously mentioned *ʔadilla/dalāʔil* example; 3) register usage distinctions; for example, identifying that *taʕbīr* [expression] does not occur in the Classical Arabic genre but occurs frequently in the Egyptian Colloquial genre; and 4) metaphorical nuance and meanings that significantly differ from those listed in the textbook; for example discovering that the verbal noun form *taṭʕīm bi-*, though defined in the textbook as part of the verb that means "to infuse with" and "to pollinate," is much more frequently used with the meaning of "to vaccinate."

The exercise was not without its hiccups. Several students misspelled search terms and thus returned anomalous results. Further, it was clear that the distinctions linked to choice of part of speech for the search term had not been adequately explained to students. arabiCorpus offers a wealth of search functionalities; for those unfamiliar with corpora, the differences among some of these may initially be unclear; continued use of the corpus will likely help students to hone their search results more effectively. Finally, at least one student also ran into a bug with the corpus for which we were able to find no explanation.

Overall, however, students in both groups indicated that they had enjoyed the arabiCorpus exercise; they described using arabiCorpus as straightforward but time-consuming. During the class following the exercise, students asked noticeably more questions about nuances of meaning, particularly with regard to items presented in the textbook as synonyms, and which prepositions to use with vocabulary items when one or more was included on the chapter list.

Each group also offered specific insights of its own.

Group A

Although a few students in Group A engaged extensively with arabiCorpus through this exercise, most chose to use the corpus only cursorily. From this perspective, a more structured approach seems to offer greater probability for consistent, deeper engagement with the tool and consequent development of metalinguistic hypotheses. Further, although students demonstrated comfort with using arabiCorpus, it is not clear if this single interaction provides enough experience with the corpus to encourage them to embrace it as a tool for their independent use.

One benefit of the more open-ended structure of this variant, however, was that a number of students completed the assignment in Arabic (the language of completion wasn't specified in the assignment description), demonstrating that intermediate Arabic students have the ability to engage in metalinguistic discourse in the target language. Importantly, students' hypotheses about the nuanced use of the vocabulary items developed as a result of their own investigation of empirical data, rather than as a result of explicit instruction from the instructor.

Group B

The structure of the arabiCorpus exercise used with Group B appeared to encourage students to engage at greater depth with the corpus. Students were able to sustain engagement with the corpus without being overwhelmed by the specificity or size of the datasets that corpora searches can return. Further, the three-stage process facilitated not only interaction, but also additional use of arabiCorpus: a number of students mentioned revisiting it during their evaluation

of their classmates' exemplary sentences. It seems likely that repeated use of arabiCorpus in structured and semi-structured frameworks will encourage students to utilize it independently.

Although one student responded to this more structured variant of the exercise in a mixture of Arabic and English, the remainder of the students in Group B answered the exercise's various questions using only English. It may be that asking students such specific questions in English primed them to reply in English; posing similar questions in Arabic should encourage the development of metalinguistic discourse in Arabic. The responses of students in Group A indicate that students at this level can use the target language to articulate the hypotheses they have formed through examining the corpus.

Two unexpected findings became visible through the three-stage process of the variant used with Group B: the exercise both encouraged students to articulate psycholinguistic conceptualizations and revealed social dynamics that could affect classroom learning. Thus, for example, one student, while responding to another's exemplary sentences, drew a distinction between two different vocabulary possibilities for "answer" (*jawāb* and *ʔijāba*) based on the degree of specificity of the question-answer process described in the sentence provided. Knowing how students are conceptualizing different linguistic features can help teachers identify when and how to offer additional explanation. Also, one of the students engaged in significant policing behavior—on peers' blogs as well as her own—informing other students of mistakes not only with regard to the meaning of the target vocabulary item, but also in terms of other spelling and word usage. For students who are struggling with challenges erected by affective filters (Krashen & Terrell, 1983; Krashen, 1985; Krashen, 1988), this may be intimidating; this kind of behavior may also be present in other group activities.

Future Directions

The current study is exploratory and thus limited in both scope and sample size. Studies that draw on larger participant samples as well as repeated use of corpus-based exercises are needed. Also useful would be research that explores methods for integrating corpus-based tools with other language-learning tools for Arabic learners of different

levels. One particularly provocative question we were left with is how learners' use of a tool like arabiCorpus may affect long-term retention of vocabulary items. Exercises such as the one we employed draw students' attention to indexical features of language. Will this engender a robust cognitive network of interconnected vocabulary items? Further research is required.

Within the context of specific corpus-based exercises, design elements that offer intriguing possibilities include utilizing string-based searches to tap into chunking benefits more directly, deploying a peer evaluation system to enhance metalinguistic awareness through active dialogue, and integrating corpus-based analysis with more extensive production tasks so that learners have numerous opportunities to rehearse and solidify their new knowledge.

Finally, as valuable as corpora are in detailing usage patterns, some types of indexical features that native speakers are sensitized to remain uncaptured by traditional corpora. Mollin (2009) found that elicitation tests yielded different responses than might be predicted through corpus searches (e.g., though word association tests pair "millionaire" most frequently with "money," this was not a top collocate combination listed in the BNC). Some of these indexical features may overlap with the emotional resonance that lexical items acquire (Kramsch, 2011). Another fascinating future project would thus be to explore ways to identify and introduce these features to learners, whether through exercises, instruction, or some other learning activity.

Conclusion

Using corpora in language learning can enhance students' understanding of specific vocabulary items through examining the relationship between context and nuance of meaning. Corpus-based exercises can also help students to focus on lexical bundles and communicating in chunks, and can provide students with a foundation for thinking about how language indexes different social features. Further, corpora can facilitate instruction by furnishing both learners and instructors with a large number of examples from authentic texts.

Students in both groups of our study proved themselves adept at analyzing corpus data to form hypotheses about the usage of

vocabulary items. Differences between the results of the two variants suggest students should be provided with enough structure to enable them to engage deeply with the corpus exercise and to prevent them from being overwhelmed by the size and specificity of the results a corpus search can yield. However, it may be worthwhile to explore posing some or all of the exercise questions in Arabic, so as to encourage students to develop their metalinguistic abilities in the target language. Multistage exercises can integrate the use of arabiCorpus with intended learning outcomes of varying complexity and sociality; such exercises also seem to encourage more independent use of arabiCorpus, although further research would be necessary to determine long-term use patterns.

Given how easily students adapted to using arabiCorpus, we believe it would be possible to implement a corpus exercise even earlier than the fourth semester of university instruction—certainly during the third semester of instruction, possibly even toward the end of the second semester—provided that students are afforded a structured framework within which to research vocabulary items. As several students noted, using arabiCorpus can be time consuming, so it is best used in a complementary role, targeted to specific vocabulary items rather than entire lists. arabiCorpus is, nonetheless, a powerful tool for extending vocabulary and exploring nuance.

References

Austin, J. L. (1975 [1962]). *How to do things with words* (2nd ed.). J. O. Urmson & M. Sbisà (Ed.). Cambridge, MA: Harvard University Press.

Bakhtin, M. M. (1986). *Speech genres and other late essays*. (V. W. McGee, Trans.). C. Emerson & M. Holquist (Ed.). Austin, TX: University of Texas Press.

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Harlow, UK: Pearson Education Limited.

Buckwalter, T., & Parkinson, D. (2011). *A frequency dictionary of Arabic: Core vocabulary for learners*. London, UK: Routledge.

Chambers, A. (2011). Language learning as discourse analysis: Playing games in a corpus of French journalistic discourse. In N. Kübler (Ed.), *Corpora, language, teaching, and resources: From theory to practice* (pp. 97–112). Bern: Peter Lang.

Charles, M. (2011). Corpus evidence for teaching adverbial connectors of contrast: *however, yet, rather, instead* and *in contrast*. In N. Kübler (Ed.), *Corpora, language, teaching, and resources: From theory to practice* (pp. 113–132). Bern: Peter Lang.

Davies, M., & Face, T. L.. (2006). Vocabulary coverage in Spanish textbooks: How representative is it? In N. Sagarra & A. J. Toribio (Ed.), *Selected proceedings of the 9th Hispanic linguistics symposium* (132–143). Somerville, MA: Cascadilla Proceedings Project.

Davies, M. (2010). More than a peephole: Using large and diverse online corpora. *International Journal of Corpus Linguistics*, *15(3)*, 412–418.

Dayrell, C. (2008). Investigating the preference of translators for recurrent lexical patterns: A corpus-based study. *trans-kom*,

1(1), 36–57.

Ellis, N. C., & Simpson-Vlach, R. (2009). Formulaic language in native speakers: Triangulating psycholinguistics, corpus linguistics, and education. *Corpus Linguistics and Linguistic Theory*, *5(1)*, 61–78.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Ed.) *Sociolinguistics: Selected readings* (269–293). Harmondsworth, UK: Penguin.

Kramsch, C. (2011, March). New perspectives on culture and foreign language study. Paper presented at Georgetown University, Washington, DC.

Krashen, S. D. (1985). *The input hypothesis: Issues and implications*. London: Longman.

Krashen, S. D. (1988). *Second language acquisition and second language learning*. New York, NY: Prentice Hall.

Krashen, S. D., & Terrell, T. D. (1983). *The natural approach*. Oxford, UK: Pergamon Press.

McCarthy, M., McCarten, J., & Sandiford, H. (2005). *Touchstone level 1: Student's book A*. Cambridge, UK: Cambridge University Press.

Mollin, S. (2009). Combining corpus linguistic and psychological data on word co-occurrences: Corpus collocates versus word associations. *Corpus Linguistics and Linguistic Theory*, *5(2)*, 175–200.

Palacios-Martínez, I., & Martínez-Insua, A. (2006). Connecting linguistic description and language teaching: Native and learner use of existential *there*. *International Journal of Applied Linguistics*, *16(2)*, 213–231.

Parkinson, D. B. (1985). Proficiency to do what? Developing oral proficiency in students of Modern Standard Arabic. *Al-Arabiyya*, *18(1–2)*, 11–43.

Parkinson, D. (2010). Communities of use in Arabic newspaper language: The meaning of the country effect. In R. Bassiouney (Ed.) *Arabic and the media: Linguistic analyses and applications* (pp. 47–60). Leiden: Brill.

Parkinson, D. (2011, April). *Under the hood of arabiCorpus.byu.edu*. Paper presented at the Arabic Corpus Linguistics Workshop, Brigham Young University, Provo, UT.

Peirce, C. (1955 [1940]). *Philosophical writings of Peirce*. J. Buchler (Ed.). New York, NY: Dover.

Philip, G. (2011). "…and I dropped my jaw with fear": The role of corpora in teaching phraseology. In N. Kübler (Ed.), *Corpora, language, teaching, and resources: From theory to practice* (pp. 49–68), Bern: Peter Lang.

Reppen, R. (2010). *Using corpora in the language classroom*. Cambridge, UK: Cambridge University Press.

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics*, *31*, 205–225.

Al-Sulaiti, L., & Atwell, E. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, *11(2)*, 135–171.