# ArCADE: An Arabic Corpus of Auditory Dictation Errors

**C. Anton Rytting**
**Paul Rodrigues**
**Tim Buckwalter**
**Valerie Novak**
**Aric Bills**
University of Maryland
7005 52nd Avenue
College Park, MD 20742
{crytting,prr,tbuckwal,
vnovak,abills}@umd.edu

**Noah H. Silbert**
Communication
Sciences & Disorders
University of Cincinnati
2600 Clifton Avenue
Cincinnati, Ohio
silbernh
@ucmail.uc.edu

**Mohini Madgavkar**
Independent Researcher
6120 Dhaka Pl. 20189-6120
Dhaka, Bangladesh
mohini.madgavkar
@gmail.com

## Abstract

We present a new corpus of word-level listening errors collected from 62 native English speakers learning Arabic designed to inform models of spell checking for this learner population. While we use the corpus to assist in automated detection and correction of auditory errors in electronic dictionary lookup, the corpus can also be used as a phonological error layer, to be combined with a composition error layer in a more complex spell-checking system for non-native speakers. The corpus may be useful to instructors of Arabic as a second language, and researchers who study second language phonology and listening perception.

## 1 Introduction

Learner corpora have received attention as an important resource both for guiding teachers in curriculum development (Nesselhauf, 2004) and for providing training and evaluation material the development of tools for computer-assisted language learning (CALL). One of the most commonly used technologies in CALL is spell correction. Spell correction is used for providing automated feedback to language learners (cf. Warschauer and Ware, 2006), automatic assessment (Bestgen and Granger, 2011), and in providing cleaner input to downstream natural language processing (NLP) tools, thereby improving their performance (e.g. Nagata et al., 2011). However, off-the-shelf spell correctors developed for native speakers of the target language are of only limited use for repairing language learners' spelling errors, since their error patterns are different (e.g. Hovermale, 2011; Mitton and Okada, 2007; Okada, 2005).

Most learner corpora (and spell correctors) are understandably focused on learner-written texts. Thus, they allow a greater understanding (and improvement) of learners' writing skills. However, another important aspect of language learning is listening comprehension (cf. Field, 2008; Prince, 2012). A better understanding of listening errors can guide teachers and curriculum development just as written production errors do. Listening error data may also be helpful for improving technologies for listening training tools, by helping prioritize the most critical pairs of phonemes for discrimination, and pointing out the most troublesome contexts for phoneme discrimination.

Finally, spell correction specifically designed to correct listening errors may aid listening comprehension and vocabulary acquisition. If learners are unable to hear, recall and record accurately what they heard, they will be less able to search dictionaries or the Web for more information on new vocabulary items they otherwise could have learned from listening exercises. While data-driven spelling correction on popular search engines may catch some non-native errors, native errors are likely to 'drown out' any non-native errors they conflict with due to larger numbers of native users of these search engines. On the other hand, if the most common listening and transcription errors are automatically corrected within a search tool, learners will have greater success in finding the new vocabulary items they may have misheard in speech.

Learner corpora focused on written production may not have enough samples of phonologically-based errors to aid in developing such tools, and



*Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications* , pages 109–115,
Baltimore, Maryland USA, June 26, 2014. 

even in a large corpus, word avoidance strategies and other biases would make the source unreliable for estimating relative magnitudes of listening problems accurately. It may be more effective to target listening errors directly, through other tasks such as listening dictation.

## 2 Related Work

Tools for language learning and maintenance, and learner corpora from which to build them, typically focus on language pairs for which there is a large market. Learner corpora for native English learners of low resource languages such as Arabic have been until recently comparatively rare, and often too small to be of practical use for the development of educational technology. In the past few years, however, a number of learner corpora for Arabic have become available, including a corpus of 19 non-native (mostly Malaysian) students at Al Al-Bayt University (Abu al-Rub, 2007); the *Arabic Interlanguage Database* (ARIDA; Abuhakema et al., 2008, 2009); the *Arabic Learners Written Corpus* from the University of Arizona Center for Educational Resources in Culture, Language, and Literacy (CERCLL; Farwaneh and Tamimi, 2012);[1] and the *Arabic Learner Corpus v1* (Alfaifi and Atwell, 2013).[2]

These corpora are all derived from learner writing samples, such as essays, and as such they contain many different types of errors, including errors in morphology, syntax, and word choice. Spelling errors are also observed, but relatively rarely, and the relevance of these spelling errors to listening competence is unclear. Hence, while they are likely to be useful for many applications in teaching Arabic writing, their usefulness for other purposes, such as examining listening skills and the effects of learner phonology on spelling, is limited.

Corpora or datasets focused on speaking and listening skills in Arabic are rarer. One such corpus, the West Point Arabic Speech Corpus, available from the LDC, contains one hour of non-native (learner) speech (LaRocca and Chouairi, 2002) Sethy et al. (2005) describe a corpus of elicited Arabic speech, but because none of the participants had prior exposure to Arabic, its use for understanding learner Arabic is limited. While there have been a few studies of Arabic listening skills (e.g. Huthaily, 2008; Faircloth, 2013), their coverage was not sufficiently broad to make reuse of their data likely to inform such purposes as the development of phoneme discrimination training or other CALL technology.

## 3 Motivation

We present here the *Arabic Corpus of Auditory Dictation Errors* (ArCADE) version 1, a corpus of Arabic words as transcribed by 62 native English speakers learning Arabic. This corpus fills the current gap in non-native spelling error corpora, and particularly for spelling errors due to listening difficulties. Unlike error corpora collected from non-native Arabic writing samples, it is designed to elicit spelling errors arising from perceptual errors; it provides more naturalistic data than is typical in phoneme identification or confusion studies.

A principal purpose for creating the corpus was to aid in the development and evaluation of tools for detecting and correcting listening errors to aid in dictionary lookup of words learners encountered in spoken language (cf. Rytting et al., 2010). As such, it serves as a complementary dataset for the dictionary search engine's query logs, since in this case the intended target of each transcription is known (rather than having to be inferred, in the case of query logs). We list three other potential uses for this corpus in Section 5.

## 4 Corpus Design and Creation

The ArCADE corpus was created through an elicitation experiment, similar in structure to an American-style spelling test. The principal difference (other than the language) is that in this case, the participants are expected to be unfamiliar with the words, and thus forced to rely on what they hear in the moment, rather than their lexical knowledge. We selected words from a commonly-used dictionary of Modern Standard Arabic such that the set of words would contain a complete set of non-glide consonants in various phonetic contexts.

### 4.1 Selection of Stimulus Words

Since the corpus was originally collected for a study focused on the perception of consonants within the context of real Arabic words, the stimulus set was designed with three purposes in

[1] Available from `http://l2arabiccorpus.cercll.arizona.edu/?q=homepage`.

[2] As of February 2014, a second version, with about 130K words from non-native speakers, is available from `http://www.arabiclearnercorpus.com/`. It also has a small (three hour) speech component.

mind: coverage of target sounds, exclusion of basic words, and brevity (so that participants could complete the task in one sitting).

In order to differentiate consonants that are relatively unpredictable (and thus test listening ability) from consonants whose value could be predicted from non-acoustic cues (such as prior knowledge of morphological structure), the corpus is annotated for *target* consonants vs. non-target consonants. A target consonant is defined as a consonant that should not be predictable (assuming the word is unknown to the listener) except by the acoustic cues alone. Glides */w/* and */j/* were not targeted in the study because orthographic ambiguities between glides and vowels would complicate the error analysis.

Each Arabic consonant other than the glides occurs as a target consonant in the stimulus set in six consonant/vowel/word-boundary contexts: C_V, V_C, V_V, #_V, V_#, and C_#.[3] (The contexts #_C and C_C are phonotactically illegal in Modern Standard Arabic.)

Consonants that were judged morphologically predictable within a word were considered non-target consonants. These included: (1) non-root consonants, when Semitic roots were known to the researchers; (2) consonants participating in a reduplicative pattern such as */tamtam/* and */zalzala/*; and (3) Consonants found in doubled (R2=R3) roots if the two consonants surfaced separately (e.g., in broken plurals such as */ʔasnan/*).

We excluded words from our stimulus set if we anticipated that an intermediate Arabic student would already be familiar with them or would easily be able to guess their spellings. Items found in vocabulary lists associated with two commonly-used introductory textbooks (*Al-Kitaab* and *Alif-Baa*) were excluded (Brustad et al., 2004a,b). Loanwords from Western languages were also excluded, as were well-known place names (e.g., */ʔiskotlanda/* = "Scotland"). Words found only in colloquial dialects and terms that might be offensive or otherwise distracting (as judged by native speaker of Arabic) were removed, as well.

In order to keep the stimulus set as short as possible while maintaining coverage of the full set of target stimuli consonants in each targeted context, we chose words with multiple target consonants whenever possible. The final set of 261 words contained 649 instances of target consonants: one instance of each geminate consonant and between 17 and 50 instances of each singleton consonant (at least two instances for each of the six contexts), with a few exceptions.[4] Although glides and vowels were not specifically targeted, 6 instances of */w/*, 10 instances of */j/*, and at least 12 instances of each of the monophthong vowels (*/a/, /i/, /u/, /a:/, /i:/, /u:/*) occur in the stimulus set.

[3] C = consonant, V = vowel, # = word boundary, and '_' (underscore) = location of target consonant.

### 4.2 Recording of the Stimuli

The audio data used in the dictation was recorded in a sound-proof booth with a unidirectional microphone (Earthworks SR30/HC) equipped with a pop filter, and saved as WAV files (stereo, 44.1kHz, 32-bit) with Adobe Audition. The stimuli were spoken at a medium-fast rate. The audio files were segmented and normalized with respect to peak amplitude with Matlab.

The native Arabic speaker in the audio recording is of Egyptian and Levantine background, but was instructed to speak with a neutral ("BBC Arabic") accent.

### 4.3 Participants and Methodology

Seventy-five participants were recruited from six universities. To be eligible, participants had to be 18 years of age or older, native speakers of English, and have no known history of speech language pathology or hearing loss. Participants were required to have completed at least two semesters of university level Arabic courses in order to ensure that they were able to correctly write the Arabic characters and to transcribe Arabic speech. Heritage speakers of Arabic and non-English dominant bilinguals were excluded from the study. The corpus contains responses from 62 participants. The mean duration of Arabic study completed was 5.6 semesters (median 4).

Before beginning the experiment, participants were asked to fill out a biographical questionnaire. This included questions about language exposure during childhood and languages studied in a classroom setting. There were additional questions about time spent outside of the United States to ascertain possible exposure to languages not addressed in previous questions.

[4] These exceptions include only one instance of a phone rather than two for the following contexts: (1) */h/* in the context C_#, (2) */f/* in the context V_#, and (3) */z/* in the context #_V. One geminate consonant, */x:/*, was inadvertently omitted from the stimulus set.

Participants wrote their responses to the 261 stimulus words on a response sheet that contained numbered boxes. They were asked to use Arabic orthography with full diacritics and short vowels (*fatha, damma, kasra, shadda* and *sukun*). The *shadda* (gemination) mark was required in order to analyze the participants' perception of geminate consonants; the other diacritics were included so as to not single out *shadda* for special attention (since participants were naïve to the purpose of the study) and also to increase the value of the resulting error corpus for later analysis of short vowels.

### 4.4 Presentation of the Stumuli

The proctors who ran the experiment supplied an iPod Touch tablet to each participant, pre-loaded with a custom stimuli presentation application.

In this custom iPod application, 261 Arabic words were randomized into 9 stimulus sets. Each stimulus set was preceded by four practice items which were not scored; thus each participant saw 265 items. Each touch screen tablet was initialized by the testers to deliver a specific stimulus set. A button on the touch screen allowed the participants to begin the experiment. After a few seconds' delay, the first word was played. A stimulus number identifying the word appeared in a large font to aid the participants in recording the word on paper. Participants were given 15 seconds to write their response, before the tablet automatically advanced to the next word. Participants were not able to replay a word.

The participants used noise-canceling headphones (Audio-Technica ATH-ANC7 or ATH-ANC7B) for listening to the audio stimuli. The experiment was performed in a quiet classroom.

### 4.5 Data Coding

The participants' handwritten responses were typed in as they were written, using Arabic Unicode characters. Any diacritics (short vowels or gemination) written by the participants were preserved. An automatic post-process was used to ensure that the gemination mark was ordered properly with respect to an adjacent short vowel mark.

The corpus consists of two main sections: orthographic and phonemic. The orthographic section is very simple: each stimulus word is given in its target orthography (with diacritics) and in each participant's corresponding orthographic transcription (including diacritics if the participant provided them as instructed). The phonemic section is more elaborate, containing additional fields designed for a phone level analysis of target consonants. Its construction is described in further detail below.

Both the orthographic response and the canonical (reference) spelling were automatically converted to a phonemic representation. This conversion normalizes certain orthographic distinctions, such as various spellings for word-final vowels. This phonemic representation of the response for each stimulus item was then compared with the phonemic representation of the item's canonical pronunciation, and each phoneme of the response was aligned automatically with the most probable phoneme (or set of equally plausible phonemes) in the canonical phonemic representation of the auditory stimulus. This alignment was done via dynamic programming with a weighted Levenshtein edit distance metric. Specifically, weights were used to favor the alignment of vowels and glides with each other rather than with non-glide consonants (since the scope of our original study was non-glide consonants). Thus substitutions between short vowels, long vowels, and glides are given preference over other confusions. This is intended to reduce the ambiguity of the alignments and to ensure that non-glide consonants are aligned with non-glide consonants when possible, without introducing any bias in the non-glide consonants alignments. When one unique alignment had the lowest cost, it was used as the alignment for that item. In some cases, multiple alignments were tied for minimal cost. In this case, all alignments were used and assigned equal probability.

Once the least-cost alignment(s) were found between a response string and the reference string for an item, the target consonants within the reference string were then each paired with the corresponding phonemes in the response, and an error category (<substitution>, <deletion>, or <match> for no error) was assigned. In the case of geminate phonemes, two subtypes of <substitution> were introduced: <gemination> and <degemination>.

Where an entire word had no response, 'NA' was used to indicate that no edit operation can be assigned. (A total of 112 items were missing).

Note that insertions were not marked, because only the 649 instances of target consonants were analyzed for the phonemic portion of the corpus, and no other material in each stimulus word (including any possible insertion points for additional material) were annotated for errors. Insertions can

be recovered from the orthographic portion of the corpus.

The coding method described above yielded a set of 41,121 target consonant records of participants' responses to target consonants (not counting the 112 non-response items), including 29,634 matches (72.1%) and 11,487 errors (27.9%). At the word level, there are 16,217 words, of which 8321 (48.2%) contain at least one error in a targeted consonant, and 5969 (37.1%) are spelled perfectly (excluding diacritics).

## 5 Potential Uses of the Corpus

In addition to the uses described in Section 3, we believe the data could be used for several other uses, such as examining linguistic correlates of proficiency, developing phonemic training, and investigating non-native Arabic handwriting.

One potential use of the corpus is to analyze the errors by individual learners to determine which sounds are confused only by relatively beginning learners (after two semesters) and which are confused by beginning and experienced learners alike. While hard measures of proficiency are not available for the participants, the language questionnaire includes time of study and self-report measures of proficiency. To the extent to which these proxies are reliable, the corpus may lead to the development of hypotheses which can be tested in more targeted studies.

Since the corpus allows quantitative evidence for the relative difficulty of particular sound pairs in particular contexts, it may guide the prioritization of foci for phonemic discrimination training and other listening exercises. At the most basic level, a teacher can take our original audio stimuli and use them as dictation exercises for beginning students (who may not be ready for sentence or paragraph level dictation). It may also form the basis for automated phonemic discrimination training, such as Michael et al. (2013). Cf. Bradlow (2008) for a review.

Since the participants handwrote their responses, the corpus contains, as a byproduct, a set of 16,329 words in non-native handwriting and their digital transcriptions. As Alfaifi and Atwell (2013) note, this could be used as a corpus of non-native handwriting for training or evaluating OCR on L2 Arabic script. If corresponding native transcriptions of the same (or similar) strings were obtained, the corpus could also be used to differentiate native from non-native handwriting (cf. Farooq et al., 2006; Ramaiah et al., 2013).

## 6 Limitations and future work

The corpus as it currently stands has some limitations worth noting. First, there is no control set of native Arabic listeners to provide a comparison point for distinguishing non-native perceptual errors from acoustic errors that even native speakers are subject to. Second, the survey does not contain proficiency ratings (except self-report) for the participants, making direct correlation of particular confusion patterns with proficiency level more difficult.

Statistical analysis of the participants' accuracy at distinguishing Arabic consonants is currently underway (Silbert et al., in preparation). An investigation of the utility of the corpus for training and evaluating spelling correction for L1 English late learners of Arabic, including the effects of training corpus size on accuracy, is also in progress.

## 7 Conclusion

The Arabic Corpus of Auditory Dictation Errors (ArCADE) version 1 provides a corpus of word-level transcriptions of Arabic speech by native English speakers learning Arabic, ideal for the analysis of within-word listening errors, as well as the development and evaluation of NLP tools that seek to aid either in developing listening skill or in compensating for typical non-native deficits in listening. Since most learner corpora only include written composition or spoken production from students, this corpus fills a gap in the resources available for the study of Arabic as a second language.

The corpus, along with the original audio stimuli and participants' handwriting samples, is available at `http://www.casl.umd.edu/datasets/cade/arcade/index.html`.

## Acknowledgments

This material is based on work supported, in whole or in part, with funding from the United States Government. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the University of Maryland, College Park and/or any agency or entity of the United States Government.

## References


Muhammad Abu al-Rub. 2007. تحليل الأخطاء الكتابية على مستوى الإملاء لدى متعلمي اللغة العربية الناطقين بغيرها. Taḥlīl al-akhṭā' al-kitābīyah 'ala mustawá al-imlā' ladá muta'allimī al-lughah al-'arabīyah al-nāṭiqīna bi-ghayrihā [Analysis of written spelling errors among non-native speaking learners of Arabic]. دراسات، العلوم الإنسانية والاجتماعية. *Dirāsāt, al-'Ulūm al-Insānīyah wa-al-Ijtimā'īyah [Humanities and Social Sciences]*, 34(2). http://journals.ju.edu.jo/DirasatHum/article/view/1911/1898.

Ghazi Abuhakema, Anna Feldman, and Eileen Fitzpatrick. 2008. Annotating an Arabic learner corpus for error. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2008)*. Marrakech, Morocco.

Ghazi Abuhakema, Anna Feldman, and Eileen Fitzpatrick. 2009. ARIDA: An Arabic interlanguage database and its applications: A pilot study. *Journal of the National Council of Less Commonly Taught Languages (NCOLCTL)*, 7:161–184.

Abdullah Alfaifi and Eric Atwell. 2013. Potential uses of the Arabic Learner Corpus. In *Leeds Language, Linguistics, and Translation PGR Conference 2013*. University of Leeds, Leeds, UK.

Yves Bestgen and Sylvaine Granger. 2011. Categorising spelling errors to assess L2 writing. *International Journal of Continuing Engineering Education and Life-Long Learning*, 21(2/3):235–252.

Ann Bradlow. 2008. Training non-native language sound patterns. In *Phonology and Second Language Acquisition*, Benjamins, Amsterdam and Philadelphia, pages 287–308.

Kristin Brustad, Mahmoud Al-Batal, and Abbas Al-Tonsi. 2004a. *Al-Kitaab fii Ta'allum al-'Arabiyya*, volume 1. Georgetown University Press, Washington, DC, 1st edition.

Kristin Brustad, Mahmoud Al-Batal, and Abbas Al-Tonsi. 2004b. *Alif Baa: Introduction to Arabic Letters and Sounds*. Georgetown University Press, Washington, DC, 2nd edition.

Laura Rose Faircloth. 2013. *The L2 Perception of Phonemic Distinctions in Arabic by English Speakers*. BA Thesis, The College of William and Mary. https://digitalarchive.wm.edu/bitstream/handle/10288/18160/FairclothLauraRose2013Thesis.pdf?sequence=1.

Faisal Farooq, Liana Lorigo, and Venu Govindaraju. 2006. On the accent in handwriting of individuals. In *Tenth International Workshop on Frontiers in Handwriting Recognition*. La Baule, France. http://hal.inria.fr/docs/00/11/26/30/PDF/cr103741695994.pdf.

Samira Farwaneh and Mohammed Tamimi. 2012. Arabic learners written corpus: A resource for research and learning. Available from the University of Arizona Center for Educational Resources in Culture, Language, and Literacy web site. http://l2arabiccorpus.cercll.arizona.edu/?q=homepage.

John Field. 2008. *Listening in the Language Classroom*. Cambridge University Press, Cambridge, UK.

DJ Hovermale. 2011. *Erron: A Phrase-Based Machine Traslation Approach to Customized Spelling Correction*. Ph.D. thesis, The Ohio State University.

Khaled Yahya Huthaily. 2008. *Second Language Instruction with Phonological Knowledge: Teaching Arabic to Speakers of English*. Ph.D. thesis, The University of Montana.

Col. Stephen A. LaRocca and Rajaa Chouairi. 2002. West Point Arabic speech corpus. Technical report, LDC, Philadelphia.

Erica B. Michael, Greg Colflesh, Valerie Karuzis, Michael Key, Svetlana Cook, Noah H. Silbert, Christopher Green, Evelyn Browne, C. Anton Rytting, Eric Pelzl, and Michael Bunting. 2013. Perceptual training for second language speech perception: Validation study to assess the efficacy of a new training regimen (TTO 2013). Technical report, University of Maryland Center for Advanced Study of Language, College Park, MD.

Roger Mitton and Takeshi Okada. 2007. The adaptation of an English spellchecker for Japanese writers. Birbeck ePrints, London. http://eprints.bbk.ac.uk/archive/00000592.

Ryo Nagata, Edward Whittaker, and Vera Sheinman. 2011. Creating a manually error-tagged and shallow-parsed learner corpus. In *Proceedings of the 49th Annual Meeting of the Asso-*

*ciation for Computational Linguistics*. Association for Computational Linguistics, Portland, OR, pages 1210–1219.

Nadja Nesselhauf. 2004. Learner corpora and their potential in language teaching. In *How to Use Corpora in Language Teaching*, Benjamins, Amsterdam and Philadelphia, pages 125–152.

Takeshi Okada. 2005. Spelling errors made by Japanese EFL writers: with reference to errors occurring at the word-initial and word-final positions. In Vivian Cook and Benedetta Bassetti, editors, *Second language writing systems*, Multilingual Matters, Clevedon, UK, pages 164–183.

Peter Prince. 2012. Writing it down: Issues relating to the use of restitution tasks in listening comprehension. *TESOL Journal*, 3(1):65–86.

Chetan Ramaiah, Arti Shivram, and Venu Govindaraju. 2013. A Baysian framework for modeling accents in handwriting. In *12th International Conference on Document Analysis and Recognition (ICDAR)*. `http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6628752`.

C. Anton Rytting, Paul Rodrigues, Tim Buckwalter, David M. Zajic, Bridget Hirsch, Jeff Carnes, Nathanael Lynn, Sarah Wayland, Chris Taylor, Jason White, Charles Blake, Evelyn Browne, Corey Miller, and Tristan Purvis. 2010. Error correction for Arabic dictionary lookup. In *Seventh International Conference on Language Resources and Evaluation (LREC 2010)*. Valletta, Malta.

Abhinav Sethy, Shrikanth Narayanan, Nicolaus Mote, and W. Lewis Johnson. 2005. Modeling and automating detection of errors in Arabic language learner speech. In *INTERSPEECH-2005*. pages 177–180.

Noah H. Silbert, C. Anton Rytting, Paul Rodrigues, Tim Buckwalter, Valerie Novak, Mohini Madgavkar, Katharine Burk, and Aric Bills. in preparation. Similarity and bias in non-native Arabic consonant perception.

Mark Warschauer and Paige Ware. 2006. Automated writing evaluation: Defining the classroom research agenda. *Language Teaching Research*, 10(2):157–180.