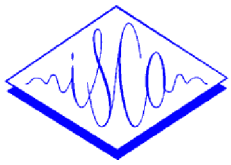

Speech and Language Technology in Education (SLaTE2009)
Wroxall Abbey Estate, Warwickshire, UK
September 3-5, 2009

Evaluating Diglossic Aspects of an Automated Test of Spoken Modern Standard Arabic

Jared Bernstein, Masanori Suzuki, Jian Cheng, & Ulrike Pado.

Pearson Knowledge Technologies
299 S. California Ave., Palo Alto, California, 94306 U.S.A.
Jared.Bernstein at Pearson.com

Abstract

A fully automatic test of facility with spoken Modern Standard Arabic (MSA) was developed and evaluated. The paper notes the diglossic situation of MSA (where colloquial and formal languages are quite distinct), and presents the structure and scoring of the test. Evaluation of the reliability and validity of the test is described, with added analyses that compare not just learners and native speakers, but also educated and uneducated speakers of the formal dialect. Results suggest scores from this commercial test are suitable in selecting MSA speakers.

1. Introduction

In this paper, we describe the design and validation of the Versant Arabic Test (VAT), a fully automated test of facility with spoken Modern Standard Arabic (MSA). The automated test can be administered over the telephone or on a computer in approximately 17 minutes. Despite its short format, test scores on the VAT closely correspond to scores from a 40-minute ILR Oral Proficiency Interview. In the evaluation, we illustrate the validation of scoring in a diglossic situation.

The paper is structured as follows: we describe Modern Standard Arabic and introduce the test construct, then describe the structure of the VAT in Section 3 and present evidence for its reliability and validity in Section 4.

2. Facility in Modern Standard Arabic

We describe an operational test of facility with spoken MSA that closely follows the tests described in Balogh and Bernstein (2007) in structure and method. To understand the test, one should know what MSA is and what facility is.

Modern Standard Arabic is a non-colloquial language used throughout the Arabic-speaking world for writing and in spoken communication within public, literary, and educational settings. It differs from the colloquial dialects of Arabic that are spoken in the countries of North Africa and the Middle East in lexicon and in syntax, for example in the use of explicit case and mood marking.

Written MSA can be identified by its specific syntactic style and lexical forms. However, since all short vowels are omitted in normal printed material, the word-final short vowels indicating case and mood are provided by the speaker, even when reading MSA aloud. This means that a text that is syntactically and lexically MSA can be read in a way that exhibits features of the regional dialect of the speaker if case and mood vowels are omitted or phonemes are realized in regional pronunciations. Also, a speaker's dialectal and educational background may influence the choice of lexical items and syntactic structures in spontaneous speech. The MSA spoken on radio and television in the Arab world therefore shows a significant variation of syntax, phonology, and lexicon.

We define *facility* in spoken MSA as the ability to understand and speak contemporary MSA as it is used in international communication for broadcast, for commerce, and for professional collaboration. Listening and speaking skills are assessed by observing test-taker performance on spoken tasks that demand understanding a spoken prompt, and formulating and articulating a response in real time.

Success on the real-time language tasks depends on whether the test-taker can process spoken material efficiently. Automaticity is an important underlying factor in such efficient language processing (Cutler, 2003). If processing is automatic, the listener/speaker can focus on the communicative content rather than on how the language code is structured. Latency and pace of the spoken response can be seen as partial manifestation of the test-taker's automaticity.

3. Versant Arabic Test

The VAT consists of five tasks with a total of 69 items. Four diagnostic subscores as well as an overall score are returned. Test administration and scoring is fully automated and utilizes speech processing technology to estimate features of the speech signal and extract response content. The VAT items were designed to represent core syntactic constructions of MSA and probe a wide range of ability levels. To make sure that the VAT items used realistic language structures, texts were adapted from spontaneous spoken utterances found in international televised broadcasts with the vocabulary altered to contain common words that a learner of Arabic may have encountered.

3.1. Test Tasks and Structure

The VAT has five task types that are arranged in six sections (Parts A through F): Readings, Repeats (presented in two sections), Short Answer Questions, Sentence Builds, and Passage Retellings. These item types provide multiple, fully independent measures that underlie facility with spoken MSA, including phonological fluency, sentence construction and comprehension, passive and active vocabulary use, and pronunciation of rhythmic and segmental units.

Part A: Reading (6 items) In this task, test-takers read six (out of eight) printed sentences, one at a time, in the order requested by the examiner voice. Reading items are printed in Arabic script with short vowels indicated as they would be in a basal school reader.

Parts B and E: Repeats (2x15 items) Test-takers hear sentences and are asked to repeat them verbatim. The sentences were recorded by native speakers of Arabic at a

conversational pace. Sentences range in length from three words to at most twelve words, although few items are longer than nine words. The ability to repeat longer items indicates more automaticity with phrase and clause structures.

Part C: Short Answer Questions (20 items) Test-takers listen to spoken questions in MSA and answer each question with a single word or short phrase. Each question asks for basic information or requires simple inferences based on time, sequence, number, lexical content, or logic. The questions are designed not to presume any specialist knowledge of specific facts of Arabic culture or other subject matter.

Part D: Sentence Building (10 items) Test-takers are presented with three short phrases. The phrases are presented in a random order (excluding the original, naturally occurring phrase order), and the test-taker is asked to respond with a reasonable sentence that comprises exactly the three given phrases.

Part F: Passage Retelling (3 items) In the final task, test-takers listen to a spoken passage (19 to 50 words long) and then are asked to retell the passage in their own words. Currently, this task is not automatically scored in this test.

3.2. Test Administration

Administration of the test takes about 17 minutes and the test can be taken over the phone or via a computer. A single examiner voice presents all the spoken instructions in either English or Arabic and all the spoken instructions are also printed verbatim on a test paper or displayed on the computer screen. Test items are presented in Arabic by native speaker voices that are distinct from the examiner voice. Each test administration contains 69 items selected by a stratified random draw from a large item pool. Scores are available online within a few minutes after the test is completed.

3.3. Scoring Dimensions

The VAT provides four diagnostic subscores that indicate the test-taker's ability profile over various dimensions of facility with spoken MSA. The subscores are

- *Sentence Mastery*: Understanding, recalling, and producing MSA phrases and clauses in sentences.
- *Vocabulary*: Understanding and producing common words spoken in continuous sentence context.
- *Fluency*: Appropriate rhythm, phrasing and timing when constructing, reading and repeating sentences.
- *Pronunciation*: Producing consonants, vowels, and lexical stress in a native-like manner in sentence context.

The VAT also reports an Overall score, which is a weighted average of the four subscores (Sentence Mastery 30%, Vocabulary 20%, Fluency 30%, and Pronunciation 20%).

3.4. Automated Scoring

The VAT's automated scoring system was trained on native and non-native responses to the test items as well as human ability judgments.

Data Resources. For the development of the VAT, a total of 246 hours of speech in response to the test items was collected from natives and learners and was transcribed by educated native speakers of Arabic. Subsets of the response data were also rated for proficiency. Three trained native speakers produced about 7,500 judgments for each of the Fluency and the Pronunciation subscores. The raters agreed well with one another (r=0.79 for Pronunciation, r=0.83 for Fluency). All test administrations included in the concurrent validation study were excluded from the training data.

Automatic Speech Recognition. Recognition is performed by an HMM-based recognizer built using the HTK toolkit (Young et al. 2000). Three-state triphone acoustic models were trained on 130 hours of non-native and 116 hours of native MSA speech. The expected-response networks for each item were induced from the transcriptions of native and non-native responses.

Since standard written Arabic does not mark short vowels, the pronunciation and meaning of written words is often ambiguous and words do not show case and mood markings. Words were represented with their fully voweled pronunciation. The orthographic transcript of a test-taker utterance in standard, unvoweled form is still ambiguous with regard to the actual words uttered, since the same consonant string can have different meanings depending on the vowels that are inserted. Moreover, the different words written in this way are usually semantically related, making them potentially confusable for language learners. This partial voweling procedure deviates from the standard way of writing, but it facilitated system-internal comparison of target answers with observed test-taker utterances since the target pronunciation was made explicit.

Scoring Methods The Sentence Mastery and Vocabulary scores are derived from the accuracy of the test-taker's response, and the presence or absence of expected words in correct sequences, respectively. The Fluency and Pronunciation subscores are calculated by measuring the latency of the response, the rate of speaking, the position and length of pauses, the stress and segmental forms of the words, and the pronunciation of the segments in the words within their lexical and phrasal context. The subscores are based on a non-linear combination of these features. The non-linear model is trained on normalized feature values and human judgments for native and non-native speech.

4. Evaluation

Two properties of a test are crucial: *reliability* and *validity*. Reliability represents how consistent and replicable the test scores are. Validity represents the extent to which one can justify making certain inferences on the basis of test scores. Reliability is a necessary condition for validity. To investigate the reliability and the validity of the VAT, a concurrent validation study was conducted in which a group of test-takers took both the VAT and the ILR OPI. If the VAT scores are comparable to scores from a reliable traditional measure of oral proficiency in MSA, this suggests that the VAT captures important aspects of test-takers' abilities in using spoken MSA. As additional evidence to establish the validity of the VAT, we examined the performance of various speaker groups

4.1. Concurrent Validation Study

ILR OPIs. The ILR Oral Proficiency Interview is a well-established test of spoken language performance, and serves as the standard evaluation tool used by United States government agencies (see www.govtilr.org). The test is a structured interview that elicits spoken performances that are graded according to the ILR skill levels. These levels describe

the test-taker's ability in terms of communicative functioning in the target language. The OPI test construct is therefore different from that of the VAT, which measures facility with spoken Arabic, and not communicative ability, as such.

Concurrent testing. A total of 118 test-takers (112 non-natives and six Arabic natives) took two VATs and two ILR OPIs. Each test-taker completed all four tests within a 15 day window. The mean age of the test-takers was 27 years old (SD = 7) and the male-to-female split was 60-to-58. Seven active government-certified oral proficiency interviewers conducted the ILR OPIs over the telephone. The average inter-rater correlation between one rater and the average score given by the other two raters administering the same test-taker's other interview was 0.90.

4.2. Reliability

Since each test-taker took the VAT twice, we can estimate the VAT's reliability using the test-retest method. The correlation between the scores from the first administration and the scores from the second administration was found to be at r=0.97, indicating high reliability of the VAT test. The scores from one test administration explain 0.97^2=94% of the score variance in another test administration to the same group of test-takers.

We also compute the reliability of the ILR OPI scores for each test taker by correlating the averages of the ratings for each of the two test administrations. The OPI scores are reliable at r=0.91 (thus 83% of the variance in the test scores are shared by the scores of another administration). This indicates that the OPI procedure implemented in the validation study was relatively consistent.

4.3. Validity

Evidence here for VAT score validity comes from two sources: the prediction of ILR OPI scores (assumed for now to be valid) and the performance distributions of different groups of test takers.

Prediction of ILR OPI Test Scores. For the comparison of the VAT to the ILR OPI, a scaled average OPI score was computed for each test-taker from all the available ILR OPI ratings. Figure 1 is a scatterplot of the ILR OPI scores and VAT scores for the concurrent validation sample (N=118). IRT scaling of the ILR scores allows a mapping of the scaled OPI scores and the VAT scores onto the original OPI levels, which are given on the inside of the plot axes. The correlation coefficient of the two test scores is r=0.87. This is roughly in the same range as both the ILR OPI reliability and the average ILR OPI inter-rater correlation. The test scores on the VAT account for 76% of the variation in the ILR OPI scores (in contrast to 83% accounted for by another ILR OPI test administration and 81% accounted for by one other ILR OPI interviewer).

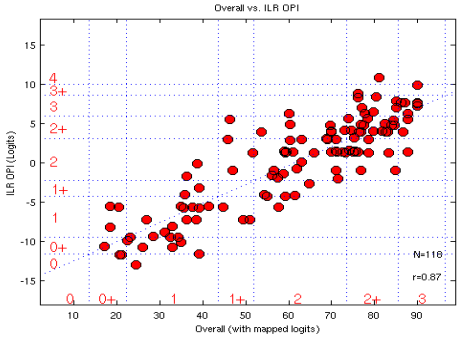

Figure 1: Test-takers' ILR OPI scores as a function of VAT scores (r=0.87; N=118).

4.4. Group Performance

Since the test claims to measure facility in understanding and speaking MSA, most educated native speakers should do quite well on the test, whereas the scores of the non-native test-takers should spread out according to their ability level. Furthermore, one would also expect that educated native speakers would perform equally well regardless of specific national dialect backgrounds and no important score differences among different national groups of educated native speakers should be observed.

Finally, we examine the score distributions for different groups of test-takers to investigate whether three basic expectations are met:

- Native speakers all perform well; while non-natives show a range of ability levels (the test can distinguish well between a range of non-native ability levels)
- Native speakers from different countries perform similarly (national origin does not predict performance) \
- Uneducated natives do not perform as well as educated.

We compare the score distributions of test-taker groups in the training data set, which contains 1309 native and 1337 non-native tests. For each test in the data set, an Overall score is computed. Figure 2 presents cumulative distribution functions of the VAT overall scores, showing for each score which percentage of test-takers performs at or below that level. This figure compares two speaker groups: Educated native speakers of Arabic and learners of Arabic. The score distributions of the native speakers and the learner sample are clearly different. For example, fewer than 5% of the native speakers score below 70, while fewer than 10% of the learners score above 70. Further, the shape of the learner curve indicates a wide distribution of scores, suggesting that the VAT discriminates well in the range of abilities of learners of Arabic as a foreign language.

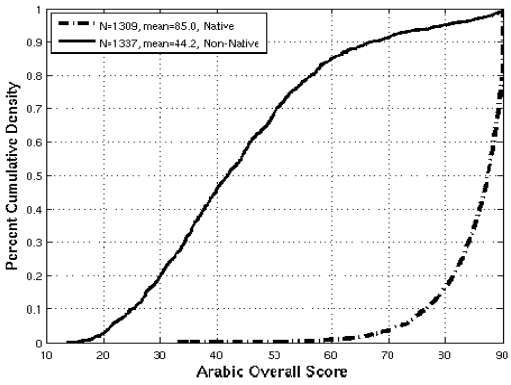

Figure 2: Cumulative score distributions for native and non-native speakers.

Figure 3 compares the cumulative distribution of educated speakers' scores with a sample of Egyptian service workers employed at the American Univesity of Cairo. Although many of the uneducated group perform well, as a group they do not perform in MSA like the educated sample. This confirms both that the VAT is a test of the educated form and that Egyptian dialect speakers are still generally better at MSA than most non-Arab learners.

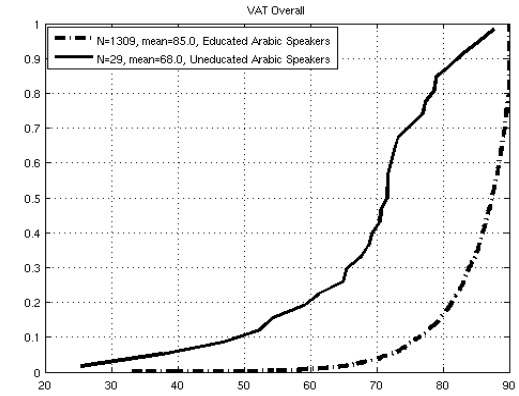

Figure 3. Cumumative distributions of educated and uneducated samples of native speakers.

Figure 4 is also a cumulative distribution functions, but it shows score distributions for native speakers by country of origin (showing only countries with at least 40 test-takers). The curves for Egyptian, Syrian, Iraqi, Palestinian, Saudi and Yemeni speakers are indistinguishable. The Moroccan speakers are slightly separate from the other native speakers, but only a negligible number of them scores lower than 70, a score that less than 10% of learners achieve. This finding supports the notion that the VAT scores reflect a speaker's facility in spoken MSA, irrespective of the speaker's country of origin.

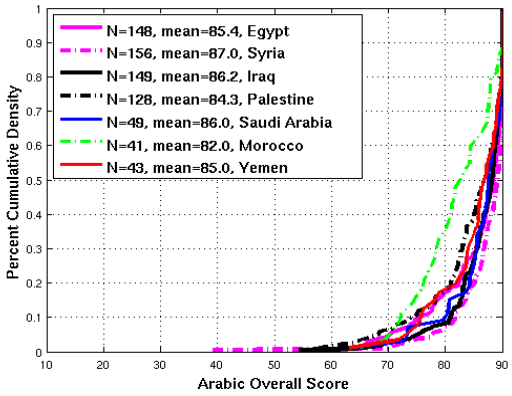

Figure 4: Cumulative score distributions for native speakers of different countries of origin.

5. Conclusion

We have presented an automatically scored test of facility with spoken Modern Standard Arabic (MSA). The test yields an ability profile over four subscores, Fluency and Pronunciation (manner-of-speaking) as well as Sentence Mastery and Vocabulary (content), and generates a single Overall score as the weighted average of the subscores. We have presented data from a validation study with native and non-native test-takers that shows the VAT to be highly reliable (test-retest r=0.97). We also have presented validity evidence for justifying the use of VAT scores as a measure of oral proficiency in MSA. Educated native speakers of Arabic can score high on the test regardless of their country of origin because they all possess high facility in spoken MSA. Uneducated Egyptians perform respectably, but nearly half of them score below 70. Non-native learners of Arabic are spread across the score scale according to their ability levels. Furthermore, the VAT test scores account for most of the variance in the interview-based ILR OPI for MSA, indicating that the VAT captures a major feature of oral proficiency.

In summary, the empirical validation data suggests that the VAT can be an efficient, practical alternative to interview-based proficiency testing in many settings, and that VAT scores can be used to inform decisions in which a person's listening and speaking ability in Modern Standard Arabic should play a part.

6. Acknowledgments

Work was conducted under contract W912SU-06-P-0041 from the U.S. Dept. of the Army. The authors thank Andy Freeman, Waheed Samy, Naima Bousofara Omar, Eli Andrews, Mohamed Al-Saffar, Nazir Kikhia, Rula Kikhia, and Linda Istanbulli for development support.

7. References

J. Balogh and J. Bernstein. Workable models of standard performance in English and Spanish. In Y. Matsumoto, D. Oshima, O. Robinson, and P. Sells, editors, *Diversity in Language: Perspectives and Implications* (CSLI Lecture Notes, 176), pages 271-292. CSLI, Stanford, CA, 2007.

J. Bernstein, M. Cohen, H. Murveit, D. Rtischev, and M. Weintraub. Automatic evaluation and training in English pronunciation. *In Proceedings of ICSLP*, 1990.

A. Cutler. Lexical access. In L. Nadel, editor, *Encyclopedia of Cognitive Science*, volume 2, pages 858-864. Nature Publishing Group, 2003.

M. Eskenazi. Detection of foreign speakers' pronunciation errors for second language training – preliminary results. In *Proceedings of ICSLP*, 1996.

H. Franco, V. Abrash, K. Precoda, H. Bratt, R.Rao, J.Butzberger, R. Rossier, and F. Cesari, The SRI EduSpeak system: Recognition and pronunciation scoring for language learning. In *Proceedings of InSTiLL*, 2000.

B. Townshend, Jared Bernstein, Ognjen Todic & Eryk Warren (1998): "Estimation of Spoken Language Proficiency," in *STiLL: Speech Technology in Language Learning*, pp. 177-180, http://www.speech.kth.se/still/still_proceed.html.

S. Young, D. Kershaw, J. Odell, D. Ollason, V.Valtchev, and P. Woodland. *The HTK Book Version 3.0*. Cambridge University, Cambridge, England, 2000.