Speech Processing Technology in Second Language Testing

Marina Dodigovic
Weil Cornell Medical College in Qatar
mad2142@qatar-med.cornell.edu

Abstract

The purpose of the study described in this article was to investigate the effectiveness of one application of speech recognition technology in the assessment of spoken English and overall English proficiency. The application referred to is called Versant and is a fully automated test of English as a second language, a test which utilizes speech recognition technology rather than a human rater. The use of this technology could lead to considerably reduced cost of testing as well as to the reduction of test anxiety. Especially the developing world might benefit from these two improvements. This paper uses the correlation between Versant and another established English proficiency test (TOEFL) to assess the effectiveness of speech recognition technology in large-scale second language testing. The results of the study suggest that Versant scores compare well with TOEFL scores, even in different educational contexts.

1. Introduction

Success at high-stakes, high-anxiety and high-cost standardized language proficiency tests such as TOEFL (Test of English as a Foreign Language), IELTS (International English Testing System) or TOEIC (Test of English for International Communication) opens the door to prestigious international schools, jobs and awards. However, not everyone can afford to take such a test, and even if he or she can, sometimes test anxiety can ruin the outcome. This becomes even more of a concern as such tests take longer and longer to complete [5]. However, there is a way to reduce the time, cost and anxiety associated with English proficiency testing. Speech recognition technology could be a solution. In fact, an application of this technology (Versant, formerly known as Ordinate SET or Phonepass test) is currently being used to test English proficiency. This test only takes about fifteen to twenty minutes to complete, is fully automated and costs only US$ 40 per session. Compared to the four hours that the latest version of TOEFL requires to complete and the price of several hundred dollars per session, the time and cost of Versant are really a significant improvement. Added to that is the fact is that the TOEFL exams take time to convey the results to the test taker, while Versant results are available on-line within only a few minutes of test completion.

With the use of Versant as a predictor of TOEFL success, students would not unnecessarily sit costly and labor intensive exams, such as TOEFL, before they were actually ready to pass. Such savings would first and foremost benefit individual students, but would also be particularly useful to English educators in developing countries. This paper discusses a research study in which the TOEFL scores of a student population are compared with their scores on Versant, a fully automated, speech recognition based test of English.

Demonstrating a good correlation of Versant with a significant international English proficiency test, such as TOEFL, thus becomes a goal worthy of pursuing, especially since this technique has had an established history in probing test validity and has been recommended as a way to obtain evidence of criterion related predictive validity [25]. Whereas test validity is the degree to which a test measures what it claims to measure [2] (Bachman & Palmer, 1996), predictive validity "is achieved by establishing how well the performance on the test of interest predicts performance on some other test", [25, 22]. While TOEFL is deemed to measure the ability in academic English language, Versant is claimed to be a predictor of general language ability [23], [24], so a perfect correlation cannot be expected.

This study mirrors one part of the validation study for the Versant test performed by the Ordinate Corporation[1] in the United States [23], [24], in which

[1] As of July 2008, Ordinate Corporation has become Pearson Knowledge Technologies.

the overall Versant score was correlated with the overall TOEFL score. The outcome of the Ordinate validation study suggests that Versant compares well with TOEFL and is likely to be a good predictor of success on the TOEFL. However, Ordinate's validation study was conducted in the US, an English speaking country, which may raise doubts regarding Versant's predictive value in a non-English speaking context [4]. Thus the research hypothesis in this study was that Versant would correlate well with TOEFL, even if taken in a country in which English is not the first language of communication.

To bring the issue closer to the audience, this paper will first address the theoretical underpinnings of the use of speech recognition technology in second language assessment. It will then focus on the validation study conducted in the United Arab Emirates by correlating the Versant and TOEFL scores of a number of university students of non-English speaking background. In doing so, the paper is seeking to establish at least one meaningful use of speech recognition technology in second language testing. On the recognition that language assessment is a vast area, and issues such as test construct, test validity or reliability could easily fill up a separate article, the focus will here remain on the technology. More detail on assessment issues related to Versant can be found in articles by Chun [4] and Downey et al [7].

2. Speech recognition technology and second language learners

Voice-interactive systems, based on speech recognition technology, have only recently started being used in Computer Assisted Language Learning (CALL) and Testing (CALT) [26], [14], [9], [19], [8], [10], to teach and test pronunciation, reading and conversation skills [23], [17], [27], [8]. L'Haire & Faltin [20: 481] make the following observation about contemporary CALL: "Voice recognition software and speech synthesizers are certainly the most prominent sellable features of current commercial CALL software." Spoken English Test (SET), developed by Ordinate and currently known as Versant is one such system. The developer's claim regarding this system is that it can assess facility in spoken English based on the speech performance of the test taker in four areas: vocabulary, syntax, pronunciation and fluency [23], [24].

Recognizing and understanding human speech requires a considerable amount of linguistic knowledge at the phonological, lexical, semantic, grammatical and pragmatic level [28]. Thus, such a system can, to some extent, assess the range of the learner's grammar and vocabulary in English as a second language. Its particular strength is probably the ability to determine the level to which a learner's speech sounds non-native-like.

More recent research [3] emphasizes the role of digitized and processed speech in second language learning. In particular, computer-assisted speech instruction and assessment seem to be gaining in importance [26]. Within this context, speech processing technologies are seen as leading edge technologies, capable of improving the efficiency of the second language speaking instruction and testing [14], [9].

However, as Handley and Hamel [14] point out, systematic evaluation of such tools has so far been sparse. To counter this trend, these two authors have developed a benchmarking procedure, which yields a comparison score between the output of a speech processing system and another standardized procedure. The trend in technology evaluation set by Handley and Hamel [14] is partly based on assessing its adequacy, viability and potential benefits. This is complemented very well by the tendency in language assessment to gauge test usefulness [25]. The study described here is based on the same principles.

Of particular concern is the extent to which computerized oral proficiency assessment would be able to contribute to the assessment of productive language performance [21]. So far, computers have only been extensively used to assess reading or listening comprehension as well as the understanding of vocabulary or grammar [6], [13], [15], [16]. These skills are known as receptive skills and do not require highly sophisticated computer programs to test. Simple gap filling, multiple choice or string matching are more than adequate to test such skills.

More recently, though, speech processing technologies have made it possible for computers to perform a very complex task of automatically scoring features such as speech fluency, pronunciation of individual or connected speech sounds, speech and intonation [21]. The Ordinate Corporation has been mentioned in the literature [21] as one of the pioneers in this field. Developing software of such capability requires collecting a large amount of non-native speech data [9], investigating its features and determining which measurable acoustic properties are sufficiently discriminating for description and grading purposes.

Speaking is a language skill that has particularly gained in significance within the communicative language learning framework [8]. According to Ehsani & Knodt [10] "foreign language curricula focus on

productive skills with special emphasis on communicative competence." This is why Eskenazi [12: 62] makes the following statement: "Below a certain level, even if grammar and vocabulary are completely correct, effective communication cannot take place without correct pronunciation (Celce Murcia & Goodwin, 1991 in Eskenazi, 1999) because poor phonetics and prosody can distract the listener and impede comprehension of the message."

However, attaining native-like pronunciation as an adult second language learner is not an easy task [10]. Looking at the sheer physiology of speech and hearing, some claim that auditory nerves specialize for the auditory tasks early on in a person's life, thus restricting the range of sounds heard and interpreted [11]. This makes the task of recognizing and subsequently repeating speech sounds of another language correctly more difficult. Therefore, an adult second language learner must take a series of time-consuming steps in order to improve his or her pronunciation. These steps include producing a large number of sentences, receiving pertinent corrective feedback, hearing many different native models, emphasizing prosody (amplitude, duration, and pitch) and feeling at ease in the language learning situation [12].

Thus, the adult second language learner needs to perform a difficult task with limited learning capacity, without feeling ill at ease or lacking confidence, which is a sort of contradiction in terms. Ideally, the sheer amount of output needed for this endeavor, would be attained in one-on-one interactive language situations [12], which are often both impractical and too costly. The situation lends itself perfectly to computer assisted language learning and testing. For this particular purpose the upcoming speech processing technology offers a promise [27], [8]. Thus, it could be said that speech recognition technology has the potential to reduce test anxiety.

Feedback, correction and grading of speech errors seem to be a very sensitive area with adult learners, as they appear to lose confidence if criticized in front of their peers [12], [27]. One-on-one instructional situation seems to work best in this case, with emphasis on the amount of interruption a learner can tolerate, avoidance of negative feedback and focus on positive reinforcement. The profile of the instructor and rater that ideally matches the requirements outlined above can be found in the latest speech processing systems as nowadays slightly more daringly applied in CALL/CALT [3], [14], [6], [16]. In order to understand the unique advantages and some of the disadvantages of such systems, one needs to understand the basics of the underlying technology, which will be reviewed in the following.

3. System design

The linguistic discipline involved in the design of speech processing technology is the subdiscipline of computational linguistics known as computational phonetics and phonology. While phonetics is concerned with speech sounds in general, phonology or phonemics is concerned with phonemes or ideal sounds of a natural language. Computational phonetics and phonology are applied in two distinct approaches to speech: speech synthesis and speech analysis. The latter has a much longer tradition that was initiated well before the advent of the computer [22] and is the one that is more relevant to language assessment. At the background of this or any subsequent technology is the fact that each sound can be broken down into its fundamental wave forms. This procedure is known as spectrographic analysis and the graphic representation of sound waves is called spectrogram. A spectrogram is a diagram representing the duration of an utterance on the horizontal axis and the different wave frequencies on the vertical axis. The main frequencies or formants are marked for vowels because they have more intensity than other frequencies. Such acoustic analysis is used to isolate and represent typical speech sounds. This is not an overly easy task since visually similar waveforms do not necessarily indicate similar sounds [18]. Some of the reasons are indicated below.

The task of speech recognition is to take speech sound waves and decode them [22]. This task is much easier for human beings than it is for computers [19]. Broken down into steps, a human being has no problems coping with fast, informal and muffled speech, including faulty utterances in a continuous stream of sound, even under exacerbating conditions such as background noise. For a computer, all these things create problems [27], which is why some systems require slow input with pauses between words, a limited vocabulary, speaker dependency (only recognizing one speaker) and the exclusion of outside noise by using special microphones.

When dealing with the speaking skill within the second language learning paradigm, especially as related to CALL/CALT, and in particular if thinking of error diagnosis, correction and grading, speech recognition and its quality become the critical issue.

Recognizing and understanding human speech requires a considerable amount of linguistic knowledge at the phonological, lexical, semantic, grammatical and pragmatic level. While the linguistic competence of an

adult native speaker covers a broad range of recognition tasks and communicative activities, computer programs perform best when designed to operate in clearly outlined sub-domains of linguistics. Ehsani & Knodt [10] identify four different speech recognition tasks: that of a court reporter transcribing a court session, a voice activated dictation system, a computerized reading tutor highlighting difficult words and providing reading assistance and finally that of a toddler being asked to fetch mum's slippers and getting a different type of shoe. The argument is that a human being, e.g., the court reporter, would perform all four tasks with similar competence whereas the computer is best used for a task for which it has been programmed or specialized.

It needs to be pointed out that speech recognition systems in general vary in type. They can be suited for the recognition of isolated words or for continuous speech recognition [27]. To use the former, one has to pause after each word, whereas with the latter one speaks normally [1]. Isolated word recognition is older and has found application in issuing voice commands to computerized systems and in vocabulary focused CALL [27].

Another distinguishing feature of speech recognition systems is vocabulary size. Low end recognizers are often limited to not more than 30 words, while large-vocabulary systems can contain tens of thousands of words. Systems also vary from speaker-dependent (only recognizing one speaker) to speaker-independent (recognizing a wide range of speakers). Some speaker-independent systems can be additionally trained to suit one person for more efficiency. Training involves speech sampling at a certain rate and sound modeling. It necessitates a large amount of speech data representative of the type the system is expected to recognize.

Generally speaking, an automatic speech recognizer is not capable of processing speech that differs significantly from the speech it has been trained on [10]. Thus, speaker-independent continuous speech systems with large vocabulary are normally trained on tens of thousands of utterances read by a variety of speakers, including different dialects and age-groups.

When it comes to systems designed to teach or test a second language, "the underlying speech processing technology tends to be complex since it must be customized to recognize and evaluate the disfluent speech of language learners" [10: 51]. For the reasons stated above, eliciting speech data from non-native speakers is a very important task when it comes to training large vocabulary speaker-independent continuous speech recognizers [19]. There are, however, other observable differences between native and non-native speech. While with native speakers' spontaneous speech contains disfluencies, filler words, conversational devices and a choice of syntactic devices which are often characteristic of a particular speech style, non-native speakers may exhibit an extreme measure of disfluencies, pauses between words and errors without any signs of a developed conversational style. Read speech, on the other hand, contains reading errors and stumbling that may not occur in spontaneous speech. Non-native speakers may exhibit a larger number of reading errors, especially with unfamiliar vocabulary.

Automatic speech recognition begins with the analysis of the incoming speech signal. When a person speaks into a microphone, the computer samples the input and creates a precise description of the speech signal. Next, a number of acoustic parameters such as the information on energy, spectral features, and pitch are derived from the speech signal. This information is used differently, depending on whether the system is in the training phase or the recognition phase. In the training phase, it is used for the purpose of modeling the speech signal. In the recognition phase, the speech signal is matched against the already existing model.

4. Versant

Ordinate's Versant is a high-end, continuous speech, speaker-independent system, using both read and conversational speech, which is additionally robust enough to cope with the disfluencies of non-native speech, especially in reading. The Versant speech recognizer was trained on a large speech sample of native speakers from a variety of backgrounds, including British, North American and Australian accents. It was also trained on a wide range of non-native speakers, of different ethnic backgrounds, at different levels of proficiency [23], [24]. The scores produced by its automatic scoring system are reported to be consistent with those of human raters [23], [24].

The Versant's underlying system, which uses the same framework for each of the several target languages it tests, includes an acoustic model of speech sounds, a dictionary and a language model, based on syntactic probability. Statistical models are used to identify the test taker's utterance. Finally, the system utilizes a pronunciation and fluency model, trained on the judgment of human raters, to evaluate the test taker's speech.

Testing with Versant is easy. There are two ways of taking this test: firstly over the phone and secondly by using a computer with Internet connection. The latter is

more cost effective in terms of communications. The cost-free software, called CDT (Computer-Delivered Test), enables one to run a test on a computer with an Internet connection, which can, but need not be active at the time of test taking. One further needs a headset with a microphone and a prepaid and downloaded test. This test takes about 15-20 minutes to complete. SET 10, the most complex in the Versant series, which was utilized in this project, has five parts: a) reading sentences from screen, b) repeating sentences heard over the headsets, c) replying to a vocabulary testing question with one word, d) repeating jumbled sentences in the syntactically correct order and e) expressing opinion on a given issue. The latter is not assessed, but can be used by human raters in case there are any doubts. A few minutes after test completion, both the test administrator and the test taker can access the test taker's score via the Internet.

The test report contains an overall score between 20 and 80, as well as its four major components (also ranging 20 – 80): sentence mastery, vocabulary, fluency and pronunciation. Sentence mastery reflects the test taker's ability to understand, repeat and produce syntactic structures. Vocabulary measures what seems to be a combination of receptive and productive vocabulary command (Nation, 1990). While the pronunciation reflects the test taker's ability to deal effectively with the segmental level, i.e. the speech sounds, pronunciation focuses on the suprasegmental level, i.e. rhythm, stress and intonation [10]. The score for each component, as well as the overall score, places the test taker in one of the six competency bands, each of which is accompanied with a descriptor that serves as a diagnostic. On request, the test taker can see all of the ranges and thus rate herself against other test takers.

According to the developer [23: 5], Versant "probes the psycholinguistic elements of spoken language performance rather than the social and rhetorical elements of communication". It "measures facility in spoken language" [23: 4], or in other words, "the ease and immediacy in understanding and producing appropriate conversational English". It was designed to be a predictor of language ability [23]. This answers questions that are sometimes asked about the test task nature [4] to the extent that is necessary for this study. More information on this and other assessment related issues can be found in Downey et al. [7].

5. Study goals and methodology

With such a potential in language assessment and a price which is exceedingly competitive, this test seems to present an affordable alternative to high-cost international English proficiency tests such as TOEFL or IELTS. Although it is not being suggested here that Versant should replace TOEFL or IELTS, since the strength of their validation research, secure and reliable administration and international acclaim is too great to discard, Versant might prove to be a good indication of the level a student could achieve on the TOEFL. Especially in the developing world and in particular at its English medium institutions, where a certain level of attainment on the TOEFL is expected of all applicants, Versant, if proven to correlate well with TOEFL, or other important tests, in the given context, would be of much benefit and would most likely be instrumental to considerable savings. The students could avoid numerous and discouraging TOEFL re-takes because Versant would provide them with reliable information about their chances of attaining the required score on TOEFL.

Therefore it was the investigator's task in this study to find out whether and how this test correlates with the results of the TOEFL exam, when taken in a country in which English is not the first language. Although a significant level of correlation between Versant and TOEFL has already been established with test takers residing in the USA under immersion conditions [23], [24], it needed to be investigated whether the same level could be achieved in the Middle East, where English is not the first medium of communication.

The procedure, funded through a Faculty Research Grant at the American University of Sahrjah, and conducted under the strict guidelines of its Institutional Board Review to guarantee the rights of human subjects, involved 104 study participants. Many of them had completed an intensive English course shortly prior to taking the Versant test and were at different levels of proficiency. These were applicants for entry in the university's first year. However, the sample also contained a fair amount of university students at various stages of progress through their undergraduate or graduate studies. Most of these were however first year university students, as they were the ones most likely to have a recent TOEFL score. The same is true of new graduate students. After an initial training session, used to familiarize the subjects with Versant, the study subjects took the Versant test within a short period of time (48 hours from the training session). The recent TOEFL scores, ideally obtained not longer than a month before or after taking the Versant, were noted for each participant. Correlations between the two sets of scores, Versant results and TOEFL scores were calculated using the Pearson Product-moment correlation coefficient, which is generally used to

obtain criterion related evidence of concurrent validity [25].

It was deemed that a good correlation between TOEFL and Versant scores would indicate that Versant is a reliable performance predictor for the TOEFL. It has to be said that students may often sit the TOEFL test before they are actually ready to perform at the required level. If Versant could be relied on as an effective predictor of potential success on TOEFL, costly and discouraging repeated failure on a high-stakes exam such as TOEFL could be avoided. This would not only bring considerable savings to the test takers and their sponsors, but would also counter the trend of giving up further attempts at TOEFL due to discouragement. Thus, all of the involved parties might benefit from the use of Versant. In addition, taken as a placement test, Versant might be utilized as an additional measurement to improve the placement of students in intensive English programs, thus contributing to overall effectiveness of such programs.

6. Results

The main step in this study was to compare scores of the two tests (Versant and TOEFL) for criterion-related evidence of predictive validity [25]. This means comparing the performance on assessment measurement instrument in question (Versant) with an established, recognized measurement instrument (TOEFL) and trying to predict a test taker's performance on the latter by his or her performance on the former, which is only likely to occur in case of high correlation between the two sets of scores.

The TOEFL scores of over 100 students were collected by self reporting and, with the test taker's permission, confirmed by the university's admissions office, whenever possible. Eventually, 104 sets of TOEFL and Versant scores were deemed viable and admitted into the pool. One clear outlier could be identified, with the Versant highest score 80 and only 530 on the TOEFL scale. Since the TOEFL score and date for this study participant could not be verified, it was removed from the database.

The Versant test scores, currently available through Ordinate on-line system, and TOEFL scores, partly available in hard copy format, were entered into an Excel spread sheet and then processed, using the Pearson function, to calculate Pearson product-moment correlation coefficient. The table below shows the obtained correlation coefficient in comparison with the one obtained by Ordinate [23] in the USA, while the scatter plot in Figure 1 is a graphic representation of the score correlations in our study, with TOEFL score on the x axis and Versant score on the y axis. The latter also includes the regression line.

Table 1

Study / Value	r	n
Ordinate USA study	0.75	392
UAE study	0.80	104

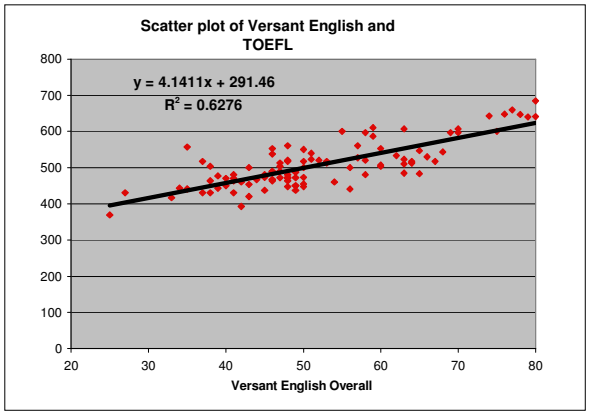

Figure1: Scatter plot and regression line

7. Discussion

It is interesting that r2, the coefficient of determination, is as high as 0.6275. Thus approximately 63% of the variance in the TOEFL scores is accounted for by the Versant scores. This means that the strength of the correlation between the two examined variables, TOEFL and Versant, is high. According to the table of critical values for the Pearson product moment correlation coefficient, the correlation level obtained for the 104 participants in the UAE study is statistically significant, and is slightly higher than the correlation level obtained by Ordinate [23].

This difference may be due to a number of variables that could have affected the data in our UAE based study. These variables include, but may not be restricted to the self reporting of TOEFL scores by some study subjects, for whom it was difficult to obtain the official results; test conditions, including external noise and technical problems associated with the new technology; as well as some test-taker related variables, such as the mental and physical state of the study subjects. However, these variables should have adversely affected the correlation coefficient, causing it to be less rather than greater than the one found in Ordinate's study.

The background of the study subjects constituted another variable that could have influenced the correlation coefficient. Most test takers were native speakers of Arabic. This seemingly unifying factor may have been responsible for the fact that no further outliers could be identified. Filtering of candidates through the university system with its bottom cut-off points for all levels, including the pre-university intensive English program, may have been another factor contributing to the even distribution of the test scores.

As hypothesized, by correlating well with TOEFL in a country where English is not the first language, Versant proved itself of value not only as a predictor of success at a major gate-keeping English proficiency test. In the developing world, where, just like in this study, English is prevalently not spoken as the first language, this finding is likely to have a socio-economic impact by allowing prospective TOEFL takers to gauge their level of TOEFL readiness in a cost-effective and yet reliable way.

8. Conclusion

This study investigates the compatibility of Versant with TOEFL by comparing TOEFL scores of non-English speaking background students with their success on Versant, a fully automated, speech recognition based test of English. So far, 104 students have successfully completed the Versant test. The participants' TOEFL scores were registered as well. The Pearson correlation product-moment coefficient calculation was used to establish the evidence of criterion related predictive test validity. Although this study resulted in a higher correlation level than Ordinate's own 2005 study, conducted in the USA, the correlation coefficient can still be interpreted as significantly close to the one in Ordinate's study. More importantly, it is a statistically significant correlation. Thus one could assume that Versant correlates well with TOEFL and for this reason can be used to gauge a language student's likelihood of success at TOEFL. By being able to test TOEFL readiness level through an inexpensive, low stakes and low anxiety and yet reliable instrument like Versant would bring considerable savings to English language students across the globe and would be seen as particularly beneficial to the developing world.

9. References

[1] J. Allen. Natural language understanding. Redwood City: Benjamin/Cummings. 1995.

[2] L. F Bachman,., and Palmer, A. S. Language Testing in Practice. Oxford: Oxford University Press, 1996.

[3] G. M.Chinnery. Going to the MALL: Mobile assisted language learning. Language Learning and Technology, 2006, 10 (1), pp 9 – 16.

[4] C. W. Chun, C. W. Commentary: An Analysis of a Language Test for Employment: The Authenticity of the PhonePass Test. Language Assessment Quarterly, 2006 , 3 (3), pp 295 – 306.

[5] M. Corrado, M. Interacting with the next generation TEOFL. TESOL Arabia Annual Conference, March 2007.

[6] M. Dodigovic. M. Raising error awareness in second language learning: An artificial intelligence perspective. Clevedon: Multilingual Matters, 2005.

[7] R Downey, H. Farhady, , R Present-Thomas,., M. Suzuki, and V. Moere, A. Evaluation of the usefulness of the Versant for English test: A response. *Language Assessment Quarterly*, 2008, 5 (2), pp 160 – 167.

[8] K. B. Egan. Speaking: A critical skill and challenge, CALICO Journal 1999. 16 (3), pp 277 – 294.

[9] R. Godwin-Jones. Language in action: From Webquests to virtual realities. Language Learning and Technology,, 2004, 8(3), – 14.
(http://llt.msu.edu/vol8num3/pdf/emerging.pdf,12 November 2006)

[10] F.Ehsani, and E. Knodt. Speech technology in Computer-Aided Language Learning: Strengths and limitations of a New CALL Paradigm. Language Learning & Technology, 1998, 2 (1), pp 45 – 60.
(http://llt.msu.edu/vol2num1/article3/, 18 October 2003)

[11] N. Ellis. Memory for language. In P. Robinson (ed.) Cognition and Second Language Instruction (pp. 33 – 68). Cambridge: Cambridge University Press, 2001.

[12] M. Eskenazi,. Using automatic speech processing for foreign language pronunciation tutoring: Some issues and a prototype. Language Learning & Technology, 1999, 2 (2), pp 62 – 76.
(http://llt.msu.edu/vol2num2/article3/, 18 October 2003)

[13] S. Granger. Error-tagged learner corpora and CALL: A promising synergy. CALICO Journal, 2003, 20 (3), pp 465 – 480.

[14] Z. Handley, & M.-J. Hamel. Establishing a methodology for benchmarking speech synthesis for Computer-Assisted Language Learning (CALL). Language Learning and Technology, 2005, 9 (3), pp 99 – 120.

[15] T. Heift and M. Schulze. Error diagnosis and error correction in CALL. CALICO Journal, 2003, 20 (3), pp 437 – 450.

[16] T. Heift and M. Schulze. Errors and intelligence in computer-assisted language learning: Parsers and pedagogues. New York: Routledge, 2007.

[17] M. V. Holl and. Tutors that listen. CALICO Journal 16 (3), 1999, pp245 – 250.

[18] F. Jelinek. Statistical methods for speech recognition. Cambridge: MIT Press, 1997.

[19] G.A. Levow and M. B Olsen. Modeling the language assessment process and result: Proposed architecture for an automatic oral proficiency assessment, ACL-IALL symposium. 1999.
(http://www.ets.org/research/dload/acl99rev.pdf, 14 October 2003)

[20] S. L'Haire, and A. V Faltin. Error diagnosis in the FreeText project. CALICO Journal, 2003, 20 (3), pp 481 – 495.

[21] J. M. Norris. Concerns with computerized adaptive oral proficiency assessment. Language Learning and Technology, 2001, 5 (2), pp 99 – 205.

[22] W. O'Grady, M Dobrovolsky and M. Arnoff. Contemporary linguistics. Boston: Bedford/St. Martin's, 1997.

[23] Ordinate. SET-10: Test description – Validation summary. Menlo Park: Harcourt. 2005.

[24] Ordinate. *Versant for English – Technical manual*. Menlo Park: Harcourt. 2007.

[25] S. Stoynoff and C. Chapelle. ESOL tests and testing. Alexandria: Teachers of English to Speakers of Other Languages. 2005.

[26] L. M. Volle. Analyzing oral skills in voice e-mail and oral interviews. Language Learning and Technology, 2005, 9 (3), pp 146 – 163.

[27] K. A. Wachowicz and B.Scott. Software that listens: It's not a question of whether, it's a question of how. CALICO Journal, 1999, 16 (3), pp 253 – 276.

[28] D. Jurafsky and J. H. Martin. Speech and language processing. Prentice Hall: Upper Saddle River. 2000.