CHAPTER 18

THE RELATIONSHIP BETWEEN THE AUTOMATIC ASSESSMENT OF ORAL PROFICIENCY AND OTHER INDICATORS OF FIRST YEAR STUDENTS' LINGUISTIC ABILITIES

Febe de Wet[1], Thomas Niesler[2], Christa van der Walt[3]

[1]*Meraka HLT Group, Council for Scientific and Industrial Research (CSIR)*
[1]*Stellenbosch University Centre for Speech and Language Technology*
fdwet@csir.co.za
[2]*Department of Electrical & Electronic Engineering, Stellenbosch University*
trn@sun.ac.za
[3]*Department of Curriculum Studies, Stellenbosch University*
cvdwalt@sun.ac.za

1. INTRODUCTION

Academic literacy proficiency is key to the success of a student at university. Currently, the large-scale assessment of language proficiency, particularly at higher education levels, is dominated by reading and writing tests because listening and speaking skills are thought to be too difficult to evaluate. The assessment of oral and aural skills is particularly challenging when large groups of students need to be considered simultaneously and when the results have to be available within a short space of time. However, to make a meaningful assessment, a balanced picture of a student's language proficiency is required, and this must include information on oral proficiency as well as listening comprehension. The application of automatic speech recognition (ASR) techniques in the automatic assessment of these skills is one of the ways in which the logistical challenges associated with testing listening and oral proficiency can be addressed.

The design and development of an automatic test for oral proficiency and listening comprehension poses a number of challenges. Firstly, the assessment of fluency above word and sentence level has to take the subjective judgements of assessors into account. In the case of extended writing assessment, attempts to obtain more objectivity include the use of rubrics or assessment schedules and multiple assessments of the same material by different assessors. The same methods can be employed for oral proficiency assessment, although it must be borne in mind that speech carries (among other attributes) the accent and gender of the testee, both of which can increase the possibility of bias. For speech fluency assessment above sentence level, which often takes the form of oral proficiency interviews, the interaction between assessor and testee may influence performance on the test,

especially when the two parties are acquainted. Hence, the possibility that subjective elements may influence the final score is much greater in oral proficiency assessment than in written tests.

Although multiple assessments of the same oral products can be introduced to obviate problems with subjectivity, they are logistically more problematic, particularly when the results must be available in a short space of time. The main difference between oral and written proficiency assessment is that, whereas we can read faster than we can write, and the assessment of writing is therefore faster, listening to speech takes as much time as the speaking itself. Even if it were feasible to record and subsequently assess a large group of students by using rubrics and an extensive group of markers, it would take much longer to conclude the assessment. Furthermore, testees may feel that listening and responding to prompts or doing an oral interview with a lecturer does not, for example, reflect their ability to teach Biology to secondary-school learners. These considerations mean that a test of oral proficiency must meet all the requirements of good assessment procedures: it must be feasible (i.e. to test large groups), reliable (to obviate problems of subjectivity) and valid (particularly in terms of face validity).

For the purposes of the project discussed in this chapter, the issue of feasibility was the main consideration for developing a test that can be administered to a large group of students in such a way that reliable results are available within a day or two. The particular setting was one in which students must be streamed into appropriate language support modules at the start of their university education. As such the test was not a final, high-stakes assessment, but the starting point for a series of more open-ended oral proficiency assessments.

Nevertheless, an automatic test can only be used as an objective measure of proficiency if the test results show a high degree of similarity with multiple human assessments of the same data. For these human assessments, the rating scales used by the evaluators must be designed very carefully to ensure reliable results. Previous attempts to design such scales showed promising results with advanced, postgraduate students (Van der Walt *et al*. 2008). In this case, however, the instrument was used with first-year students. Since students come from a variety of language backgrounds it was decided to include a wider variety of language samples, varying in linguistic and syntactic complexity. Although the students were registered for different degree programmes, they all followed a common first-year module in English Studies, which includes aspects of literature (e.g. short stories, drama and film studies) as well as aspects of language-in-use, such as the analysis of advertisements and the development of academic language proficiency. Being proficient in academic English is extremely important because even if they did not continue their study of English after the first year, the students would still

use the language in their academic subjects. Using and understanding English for cognitive, academic purposes is very important for their academic development.

In this study, all correlation values are expressed in terms of Spearman's rank correlation coefficients, which are used as a means to quantify the degree of similarity between automatic and human assessments. We report on the correlation between human and automatic ratings for a test population of first year students, as described in the preceding paragraphs. We also consider the correlation between these assessments and the performance of the same students in two written tests as well as a test of academic listening skills. In addition, we comment on the feasibility of using readability index, quantified in terms of the Flesch Reading Ease (FRE) scale, as a design criterion for test items.

The next section describes the design and implementation of the automatic test that was used in this study. Section 3 reports on the human assessment of the data and Section 4 describes the ASR system that was used to evaluate the data automatically. Section 5 introduces the other indicators of linguistic ability that our measurements will be compared to. Results are presented in Section 6. Concluding remarks and future directions are discussed in Section 7.

2. AUTOMATIC ORAL PROFICIENCY TESTING

The test that was used to assess students automatically is the result of a number of piloting experiments and subsequent adjustments. The students' language background and the lecturers' expectations of their language proficiency were important considerations in the design of the test and the associated rating criteria. All the students are active bi- and multilinguals that need to use English as an academic and professional language, rather than just for everyday interpersonal communication. This orientation meant that the test needed to include context-sensitive content (i.e. educational and academic context). The rating criteria did not assess the students' proficiency in home-language speaker terms but rather in terms of intelligibility and comprehensibility. Although accent and grammatical correctness played a role in the assessment criteria, they were only taken into account when comprehensibility was affected.

Our initial aim was to develop a test of oral and listening proficiency in English for postgraduate students doing a certification course to become secondary school teachers. The first version of the test included seven sections and was implemented as a telephone-based spoken dialogue system (SDS) (Van der Walt *et al*. 2008). However, the students took much longer to complete the test than was anticipated. Subsequent versions of the test were therefore limited to a reading and a repeating (elicited imitation) task. The purpose of the reading task is to test comprehension of the instructions and the ability to read fluently using appropriate intonation and

pronunciation, while the elicited imitation task aims to determine the extent to which students can grasp the meaning of what they hear and repeat and/or rephrase what they heard.

Experiments involving different groups of post-graduate students consistently showed that the students did not find the reading task challenging, and obtained very high scores. This left little room for discrimination between different levels of proficiency. The elicited imitation task, on the other hand, resulted in a wider range of scores and showed more potential as an indicator of oral proficiency (De Wet *et al*. 2009, De Wet *et al*. 2010).

This kind of task is controversial and seems to have originated in the field of language therapy with a view to predicting spontaneous speech production (Fujiki and Brinton 1987). The prompts consist of sentences that are just longer than can be accommodated with ease by the students' working memory. Graham *et al.* (2008:1604) explain the process as follows:

> Since short-term or working memory is limited, the retention of a representation there is, by most accounts, dependent upon the number of units being processed. As the length of utterances becomes greater, it necessitates the chunking of information ... It is believed that language competence is what facilitates this chunking process.

Since pilot versions of the test showed that this kind of test item discriminated more consistently among advanced users of English, it seemed a promising direction on which to focus in subsequent versions of our automatic assessment system. Moreover, it is an important skill for higher education students to listen and make notes in lectures and therefore the ability to retain and process complex sentences, albeit in this case in writing. This is an important motivation for assessment by elicited imitation. The current version of the test consists of a SDS, running on a desktop PC and administered in a multi-media computer laboratory using headsets with directional, noise cancelling microphones. During the test, students were prompted to read sentences from a test sheet as well as to repeat utterances produced by the SDS. On average, the students took around seven minutes to complete the test.

2.1. Prompt Design

In an effort to test the effect of varying sentence difficulty on reading ability and repetition accuracy, it was necessary to find a consistent means of calculating both the complexity and length of sentences. The level of sentence difficulty for the test was therefore linked to the FRE scale. The FRE scale is a standard means to quantify textual sentence complexity and provides an indication of how easy or difficult a given text is to read: higher FRE scores are associated with easier texts

and lower scores with more difficult texts. We chose to use the FRE scale because it is readily available and easy to use, e.g. it is included as a standard tool in most word processing packages. Such readability scores are often criticised, because they are generated at sentence level and do not take the overall structure of a text into account (Shehadeh and Strother 1994). However, since the test was designed with a focus on the sentence level, the FRE scale was regarded as suitable for our purposes.

For the reading task, the readability scores ranged from 28.5 to 83.0 (average = 52.3). The prompts for the elicited imitation task were divided into two sets: a fairly easy (*Repeat A*) and more challenging (*Repeat B*) set. The readability scores ranged from 65.7 to 85.2 (average = 70.5) and from 46.6 to 57.7 (average = 50.8) in *Repeat A* and *Repeat B*, respectively. Since previous attempts to use elicited imitation indicated the importance of context when using this technique (Fujiki and Brinton 1987:302) and since the purpose here was to assess the oral proficiency of multilingual, higher education students who use English mainly for educational purposes, the vocabulary of both tasks was controlled by focusing on educational settings with which participants would be familiar, for example:

- Read: *First year students find that they lack academic skills*. (FRE = 71.7)
- Repeat A: *I don't see useful teaching techniques in the schools.* (FRE = 85.2)
- Repeat B: *Teachers often resist change and don't want to see new methods, unfortunately.* (FRE = 54.2)

In total, there were fifteen possible sentences in the reading task, seven sentences in *Repeat A* and eight in *Repeat B*. During the test, the SDS randomly selected six sentences to be read: three easier and three more challenging repeat prompts from *Repeat A* and *Repeat B* respectively. Each student was therefore prompted to read and repeat six utterances.

2.2. Test Population

The automatic test was taken by 58 first-year undergraduate students at Stellenbosch University. The students were divided into three groups in terms of their overall performance in their first year English Language and Literature module. The first group contained those students achieving an average mark between 40 and 49%, the second between 50 and 59%, and the third between 60 and 75%.

A mark of 40% is the lowest possible, while 75% represents the top mark among the 58 students. From these three groups, random selections were made to obtain a spread of participants from low to high scoring individuals. In this way, the number of participants per language proficiency group could be kept equal across

groups to facilitate comparison between oral proficiency scores and written test scores achieved in their English module.

2.3. Test Administration

Oral instructions were given to the students before the test. They also completed an audio test to verify playback and recording volume. We included the audio test because, in previous studies, we found that some students spoke too loudly (causing clipping in the audio file) while others spoke too softly to perform meaningful ASR.

In addition to the instructions given by the SDS, a printed copy of the test instructions was provided. The students were allowed to read through the instructions before taking the test. No staff was present while the students were taking the test.

3. HUMAN ASSESSMENT

Six teachers of English as a second or foreign language were asked to rate the students' responses. Each teacher also rated at least three students twice. The intra-rater reliability for the majority of the teachers was above 0.9. Inter-rater agreement was determined in terms of two-way, intra-class correlation coefficients (ICCs). The ICC values indicated that the evaluations of two of the teachers differed substantially from those of the other four (ICC < 0.2). The ratings of these two teachers were therefore not taken into account during the rest of the study. The inter-rater agreement for the four remaining teachers varied between 0.87 and 0.88.

The rating scales for the reading and elicited imitation tasks are illustrated in Figures 1 and 2 respectively. These scales were developed by taking the results of a number of previous rating experiments into consideration (De Wet *et al.* 2009). For the reading task, the aim was to assess the students' ability to read without hesitation and with pronunciation and phrasing that closely approximates what would be regarded as educated South African English in which an L2 accent may be discernible. This scale allows for active bi- and multilinguals who are not necessarily attempting to sound like home language speakers of English but who are competent users of the language for academic purposes.

The scale for the elicited imitation task was designed to assess sentences in terms of the degree of hesitation as well as the accuracy of *repetition* or *interpretation,* which meant that students could still be given the highest score even if they did not use the exact words or the exact word order of the prompt. In language processing terms an accurate interpretation would mean that the working memory is able to make meaning of the incoming message by repeating its essence, a phenomenon

that was observed in earlier versions of the test. For example the sentence, *"Teachers often resist change and don't want to see new methods, unfortunately"* could be reinterpreted as "*Unfortunately teachers don't want to change or use new methods*" and still be given the highest score on the human ratings. One could even argue that an accurate interpretation shows language proficiency at a more advanced level than mere repetition of the original prompt.

Figure 1: *Scale used by humans to rate read prompts*.

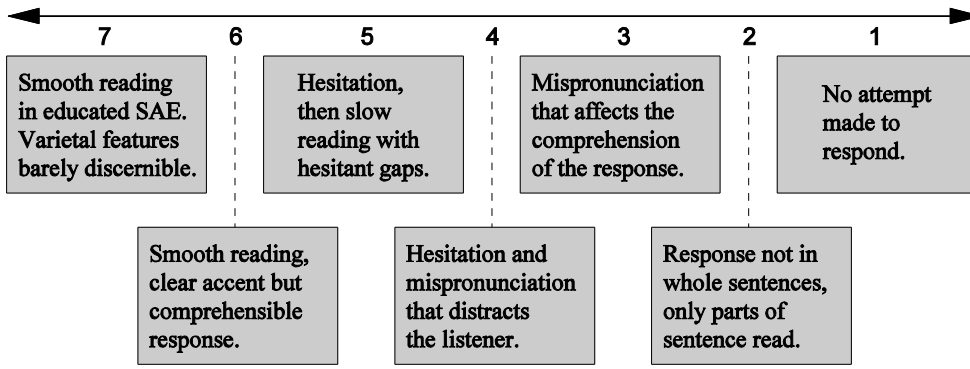
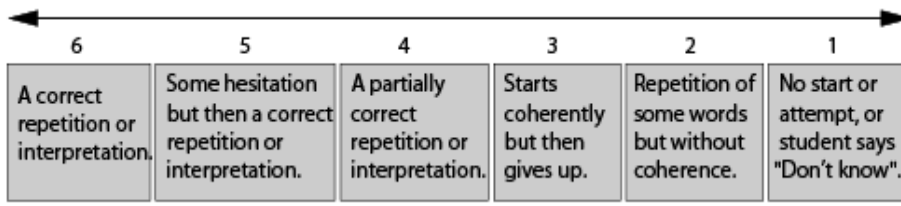


Figure 2: *Scale used by humans to rate repeated prompts*.



The raters were not provided with the numerical values indicated in the figures. These were used only to quantify the ratings for subsequent correlation with machine scores.

4. AUTOMATIC ASSESSMENT

The output of an ASR system can be used in various ways to extract quantitative features from speech signals. Various techniques to derive these so-called *machine scores* from speech data have been reported on in the literature (Speech Communication 2000, Speech Communication 2009). In previous studies, we consistently found low correlation between human ratings and what are known as posterior scores (De Wet *et al.* 2009). This investigation will therefore be restricted to scores derived from segmentation information and from repeat accuracy, because, to date, these have shown the highest correlation with human ratings (De Wet 2010).

4.1. The ASR System

The Hidden Markov Model Toolkit (HTK) version 3.4 was used for ASR (Young *et al.* 2006). The hidden Markov models (HMMs) used by the speech recogniser were trained on approximately 6 hours of telephone quality speech from English mother-tongue speakers. An equal number of male and female speakers were included in the speaker population. This data is part of the African Speech Technology corpus, and consists of phonetically and orthographically annotated speech gathered over South African fixed as well as mobile telephone networks (Roux *et al.* 2004).

The test data was bandlimited during feature extraction to match the frequency range of the training data. Features were extracted from 25 ms overlapping frames of acoustic data and subsequent frames were extracted every 10 ms. Each acoustic data frame was encoded as 12 Mel-Frequency Cepstral Coefficients (MFCCs) and an energy feature (C0). Cepstral mean normalisation was applied at utterance level. The first and second order derivatives were extracted from the static coefficients and appended to the feature vector.

Triphone HMMs were obtained by means of decision-tree state clustering and embedded Baum-Welsh re-estimation. The final set of triphone HMMs consisted of 4797 tied states based on a set of 52 phones, and a maximum of 8 Gaussian mixtures per HMM state.

Finite State Grammars (FSGs) were used for the automatic recognition of the reading task. It is expected that the students, who generally have good English reading skills, would make very few errors while reading prompts from a test sheet. Hence the use of a strict finite state grammar (FSG) is an appropriate recognition method for this task. For each prompt in the reading task an FSG was created allowing the desired utterance, as well as “I don't know” or simply “don't know”. The branch allowing the desired utterance expects all words to be present

in the correct order, but allows inserted silence, noise and filled pauses. These prompt-specific grammars were defined using extended Backus-Naur form (EBNF) notation and were parsed to lattice files that were used during recognition.

Unigram language models (LMs) were used for the automatic recognition of the repeated prompts. For this task, provision must be made for missing words and changes in word order. A separate unigram language model (LM) was therefore created for each prompt of the elicited imitation task. Each LM consisted of an unweighted word loop, with word-to-word transitions between the words in the prompt all having an equal probability. Silence, noise and filled pauses were allowed between words.

Since the word insertion penalty and language model scale factors were optimised on a development set in previous experiments (De Wet *et al.* 2009), all data collected in this study was available for use as a test set.

4.2. Segmentation-Based Scores

The scores based on segmentation information focus on the *temporal* features of speech, rather than on its acoustic characteristics, and are calculated from phone level alignments. A distinction is made between the *speech phones* (those forming part of words) and *non-speech phones* (those forming part of silence or noise) in each utterance. In a previous study, four segmentation-based scores were investigated: rate of speech, articulation rate, phonation/time ratio and segment duration scores. The highest correlation between a segmentation based score and the human ratings of the same data was observed for rate of speech (De Wet 2010), and the scope of this study will therefore be restricted to this measure.

4.2.1 Rate of Speech

The *Rate of Speech* (*ROS*) of an utterance is defined in Cucchiarini *et al.* (2000) as the number of speech phones per second, calculated using the number of speech phones in the utterance M_{Speech}, and the total duration of the utterance T_{Total}, in seconds:

$$ROS = \frac{M_{Speech}}{T_{Total}}$$

Any silences leading or trailing the utterance are ignored when determining the total duration.

4.3. Scores Derived from Repeat Accuracy

Speech recognition accuracy can also be used as a score for automatic assessment. Previous experiments have shown that most university level students are able to achieve a perfect reading accuracy for the majority of the prompts in the exercise. Reading accuracy is therefore not useful in scoring proficiency automatically. Repeat accuracy, on the other hand, is more variable and two closely-related alternatives were considered in this study: ASR Accuracy and ASR Correct.

4.3.1 ASR Accuracy

The score *ASR Accuracy* (Acc_{ASR}) is calculated using the HTK tool *HResults*, which uses a dynamic programming-based string alignment procedure to align the recogniser output with the reference transcription (Young *et al.* 2006). It counts the number of correctly aligned words (H), the number of insertions (I), and the number of words in the reference transcription (W).

The score is then calculated as:

$$Acc_{ASR} = \frac{H - I}{W} \times 100\%$$

Note that this score is penalised by insertions. When the number of insertions exceeds the number of correctly recognised words, the score is negative.

4.3.2 ASR Correct

The score *ASR Correct* (Cor_{ASR}) indicates the percentage of reference transcription words present in the recogniser output (Young *et al*. 2006).

$$Cor_{ASR} = \frac{H}{W} \times 100\%$$

In contrast to Acc_{ASR}, this score does not take insertions into account, but simply reflects the percentage of correctly-aligned words.

5. INDICATORS OF LINGUISTIC ABILITY

The two written tests that were chosen as indicators of linguistic ability are the so-called *Test of Academic Literacy Levels* and the *Early Assessment* test. Students' performance on the *Academic Listening Test* was also taken into consideration.

5.1. Test of Academic Literacy Levels

The Test of Academic Literacy Levels (TALL) is a multiple choice test of comprehension, academic vocabulary, inference, coherence and register (Van Dyk and Weideman 2004). Its purpose is to assess the existing academic literacy of incoming students with a view to streaming them into appropriate language support modules. All first-year students are expected to complete this test.

5.2. Early Assessment Test

The Early Assessment test (EA) is a university-wide measure of how first-year students perform in their various modules after six weeks in the first semester of the first year of undergraduate study. The purpose is to identify students who are at risk of not passing and to provide appropriate academic support. The EA score can include a number of assessments, depending on the structure of particular modules. In the case of the students who participated in this study, the EA consisted of an academic essay.

5.3. Academic Listening Test

The Academic Listening Test (ALT) was developed at Stellenbosch University with the express purpose of appropriately assessing the listening proficiency of first year students, using academic material and an academic context (Marais and van Dyk 2010). Students completed the computer-based test by answering multiple-choice questions that elicited responses in four tasks:

- Students were required to structure information;
- Students were required to watch a video-recorded lecture and subsequently answer questions on, for example, its main and supporting idea;
- Students were required to watch a video-recorded discussion by two students and subsequently answer questions on, for example, the represented attitudes;
- Students were required to listen to a video-recorded lecture and subsequently fill in words that had been omitted from a transcript.

The ALT consisted of multiple-choice questions throughout.

6. RESULTS

The previous sections have described how language proficiency can be assessed by means of written tests, by means of oral assessments with human evaluators, and by means of automatically-derived oral proficiency indicators. We will now investigate how well these various approaches relate to one another.

6.1. Relationship between Human and Automatic Oral Proficiency Assessments

The correlations between the machine scores defined in Section 4 and the ratings given by the English teachers are shown in Table 1. All the correlations in the table are statistically significant (p-values < 0.05). The results for the elicited imitation task are shown separately for *Repeat A* and *Repeat B*, as well as for the exercise as a whole (*Repeat*). For each utterance, the human rating was taken to be the average of the individual scores given by the four judges, as described in Section 3.

Table 1: *Correlation between human ratings and automatically-derived scores*

	ROS	Acc_{ASR}	Cor_{ASR}
Read	0.40	-	-
Repeat A	0.47	0.42	0.42
Repeat B	0.65	0.49	0.84
Repeat	0.55	0.36	0.67

From Table 1 we see that the correlation between *ROS* and the human ratings for the reading task is low. Other researchers have shown *ROS* for read speech correlates very strongly with human ratings of fluency (Cucchiarini *et al.* 2000). However, a number of our experiments have shown that *ROS* no longer correlates well with human ratings when the test subjects are very proficient speakers (De Wet *et al*. 2009).

The results in Table 1 indicate that the correlation between Cor_{ASR} and the human ratings of repeat accuracy (as defined in Figure 2) are much higher than the corresponding values for Acc_{ASR}. *ROS* is also better-correlated with the human ratings of repeat accuracy than Acc_{ASR}.

In general, the correlations associated with *Repeat B* are higher than those measured for *Repeat A*. This observation is attributed to the much wider range of human ratings and corresponding machine scores in *Repeat B* than in *Repeat A*. Higher

correlations are usually observed for scores spanning wider ranges of the assessment scale than for those limited to a small interval. The correlations between the human and automatic scores are also in the same range as in previous studies, even those where a bigger group of human raters were involved in the assessment. These trends and results are in good agreement with those from previous studies involving post-graduate students, which indicates some consistency of the test over different test populations (De Wet *et al.* 2009, De Wet *et al.* 2010).

6.2. Relationship between Written and Oral Language Proficiency Assessments

When considering the language proficiency assessments described in Section 5 in isolation, we find that they are poor indicators of one another. For example, the correlation between the results of the TALL and the ALT is 0.52, while it is just 0.33 for the TALL and the EA ($p > 0.05$ in both cases). These low correlations indicate that the three tests probably assess different aspects of linguistic ability.

When we consider how well the written tests mirror the results of human or automatic oral assessments, the picture described by Table 2 emerges. Only correlations that are statistically significant are shown ($p < 0.05$).

Table 2: *Correlations between oral proficiency assessments (human and machine) and other indicators of linguistic ability*

	TALL	EA	ALT
Human raters (Read)	0.48	0.42	0.47
ROS (Read)	-	-	0.44
Human raters (Repeat B)	0.55	-	0.49
ROS (Repeat B)	0.39	-	0.41
Cor_{ASR} (Repeat B)	0.35	-	0.37

The correlations between *ROS* and Cor_{ASR} for *Repeat B* and the TALL results are lower than the correlation between the human oral proficiency assessments for *Repeat B* and TALL (0.39 and 0.35 as opposed to 0.55). This indicates that the two automatically derived measures are poor indicators of written proficiency.

From Table 1 we recall that the correlation between *ROS* and the human assessments for *Repeat B* was 0.65 ($p < 0.05$). The corresponding value for Cor_{ASR} was 0.84 ($p < 0.05$). These figures are considerably higher than the correlation between the TALL or ALT results and the human assessments for *Repeat B* (0.55

and 0.49 respectively). This indicates that the automatically-derived measures are better indicators of the human assessments of oral proficiency than either of the written tests.

For the EA test, the correlation with the human assessments for read speech are low and significant (0.42). This indicates that the human ratings cannot be used to predict performance on the early assessment tests.

Overall, the results in Table 2 seem to indicate that the oral assessment is not a good indicator of performance in the EA tests. However, neither is the TALL, despite also being a written test. It should be borne in mind, though, that the TALL is in multiple choice format while the EA test is in the form of an essay. Although a positive correlation was found between the TALL scores and a short, open-ended writing piece that was included in early versions of TALL (Van Dyk and Weideman 2004), this is not the case in our study.

6.3. Relationship between Human Assessments and the FRE Scale

Table 3 shows the correlation between the per-utterance average of the human scores and the difficulty of the corresponding utterance on the FRE scale. As in Table 1, the results for the elicited imitation task are shown for *Repeat A* and *Repeat B* separately as well as for the whole exercise (*Repeat*). The p-values associated with the correlations are given in the third column of the table.

Table 3: *Correlation between human ratings (judge average) and utterance difficulty according to the FRE scale*

	Correlation	p-value
Read	0.37	0.17
Repeat A	0.49	0.26
Repeat B	0.57	0.14
Repeat	0.86	<0.00

The results in the first row of Table 3 reject the hypothesis that there is a correlation between the FRE scale level and the human ratings of the read prompts. The same trend is observed if *Repeat A* and *Repeat B* are considered separately. However, for a combination of the easy and more challenging tasks, there is a high and significant correlation between the FRE scale levels and the human ratings of repeat accuracy. This seems to indicate that FRE scale values can be used as a design criterion, provided that the exercises include utterances with varying levels of difficulty - as

is, indeed, required by test designs that attempt to assess overall speaking proficiency (Graham *et al.* 2008).

7. DISCUSSION AND CONCLUSIONS

Our journey through the various versions of the ASR assessment has led to questions on a number of levels and in a variety of fields.

Firstly, the use of ASR to assess oral proficiency shows the complex interplay between psycholinguistic processing, like the role of working memory, and sociolinguistic factors, like the role of context, in the oral production of language. Secondly, and at the same time, it requires of researchers from very different disciplinary backgrounds to collaborate on an assessment that has consequences for students in terms of the level and kind of academic support that they need. Thirdly, the original intention of this project - to investigate the possibility of ASR as a valid, reliable and feasible instrument - led to comparisons with other kinds of assessment. Although the results discussed above provide information about the use of ASR as a measure of oral proficiency, they also say much about the kinds of assessment conducted with first year students and the lack of correlation among these tests.

In earlier versions of the test it was not possible to find a positive correlation between ASR scores and ratings of short, open-ended oral proficiency tasks, which may be an indication of the degree to which these tasks differ in context and content (Müller 2010). Although one could argue that these assessments provide an overall picture of proficiency in different areas, it remains the case that there is no easy way for lecturers in language support to assess and predict performance for an overarching construct such as `academic language proficiency'.

The ASR test shows positive correlations with human ratings of the speech sample, with the highest rating for Cor_{ASR} on the elicited imitation tasks. The human rating scales on these tasks required that they award highest score for a correct repetition or interpretation, which meant that students could still be given the highest score even if they did not use the exact words in the exact order. The human ratings therefore allowed for a meaning-making process that would repeat the essence of the original prompt. For an ASR system this is quite a challenge and yet Cor_{ASR} appears to measure this adequately and produced the highest correlation between the human ratings and machine scores. The fact that this measure does not penalise for insertions or require a 100% accurate repetition of words in the same order as that of the original (as in the Acc_{ASR} measure) means that an appropriate degree of flexibility is built into the system, resulting in a high correlation with human ratings.

The second highest correlation with human ratings was with the *ROS* measure. At first glance, and based on other studies, the benefit of *ROS* as a measure of oral fluency is that it can be measured by simply providing students with texts to read (Cucchiarini *et al.* 2000). However, in this study, the *ROS* scores on the elicited imitation task correlated far better with human ratings. This result also confirms the value of using readability scores to develop test items for elicited imitation tasks.

Finally, our experiments confirm the widely-held belief that written tests are a poor indicator of oral proficiency. Furthermore, we demonstrate that measures derived automatically from a recorded speech signal are better indicators of oral proficiency than the written tests, because they show substantially higher agreement with the opinions of human judges. This indicates that an automatic oral proficiency assessment system has a clear additional role to play in the evaluation of language skills.

Future research will focus on a number of issues. Firstly, we will consider the implementation of an ASR system trained on a representative variety of South African English accents for the calculation of the automatic proficiency indicators. By experimental evaluation, we will determine whether more advanced ASR benefits the accuracy of the automatic assessment system. Secondly, we would like to automatically include synonyms in our unigram language models as used to determine the repeat accuracy. This would allow the system to be more flexible and not to penalise responses that are semantically equivalent to the prompt. Finally, we would like to expand the number of test items while maintaining a wide variety of difficulty levels. This we believe can be achieved in a semi-automatic manner with the help of the FRE scale.

ACKNOWLEDGEMENTS

This research was supported by an NRF Focus Area Grant for research on *English Language Teaching in Multilingual Settings* as well as NRF grants TTK2007041000010 and GUN2072874 and the *Development of Resources for Intelligent Computer-Assisted Language Learning* project sponsored by the NHN.

REFERENCES

Cucchiarini, C., H. Strik & L. Boves. 2000. Different aspects of expert pronunciation quality ratings and their relation to scores produced by speech recognition algorithms. *Speech Communication* 30:109-119.

De Wet, F., C. van der Walt & T.R. Niesler. 2009. Automatic assessment of oral language proficiency and listening comprehension. *Speech Communication* 51(10):864-874.

De Wet, F., P.F. de V. Müller, C. van der Walt & T.R. Niesler. 2010. *Using segmentation and accuracy-based scores to automatically assess the oral proficiency of proficient L2 speakers*. Proceedings of the 21st Annual Symposium of the Pattern Recognition Association of South Africa. Stellenbosch, South Africa, 2010.

Fujiki, M. & B. Brinton. 1987. Elicited imitation revisited: A comparison with spontaneous language production. *Language, speech and hearing services in schools* 18:310-311.

Graham, C.R., D. Lonsdale, C. Kennington, A. Johnson & J. McGhee. 2008. *Elicited imitation as an oral proficiency measure with ASR scoring*. Proceedings of the 6th Conference on Language Resources and Evaluation (LREC). Marrakech, Morocco.1604-1610.

Marais, F.C. & T.J. van Dyk. 2010. Putting listening to the test: An aid to decision-making in language placement. *Per Linguam* 26(2):34-51.

Müller, P. F. de V. 2010. *Automatic Oral Proficiency Assessment of Second Language Speakers of South African English*. Master's Thesis. Stellenbosch: Stellenbosch University.

Roux, J.C., P.H. Louw & T.R. Niesler. 2004. The African Speech Technology Project: An Assessment. *Proceedings of LREC*. Lisbon, Portugal. Vol.1:93-96.

Shehadeh, C.M.H. & J.B. Strother. 1994. *The use of computerized readability scores: Bane or blessing?* Proceedings of the Annual Conference of the Society for Technical Communication.41:225.

Speech Communication. Various Authors. 2000. *Special Issue on Language Learning. Speech Communication* 30(2-3).

Speech Communication. Various Authors. 2009. *Special Issue on Spoken Language Technology for Education. Speech Communication* 51(10):831-1038.

Van der Walt, C., F. de Wet & T.R. Niesler. 2008. Oral proficiency assessment: the use of automatic speech recognition systems. *Southern African Linguistics and Applied Language Studies* 26:135-146.

Van Dyk, T.J. & A.J. Weideman. 2004. Switching constructs: on the selection of an appropriate blueprint for academic literacy assessment. *Journal for language teaching* 38(1):1-13.

Young, S., G. Evermann, M. Gales, T. Hain, D. Kershaw, X. Liu, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev & P. Woodland. 2006. *The HTK book, version 3.4*. Cambridge: Cambridge University Engineering Department.