2011-06-27

Oral Proficiency Assessment of French Using an Elicited Imitation Test and Automatic Speech Recognition

Benjamin J. Millard
Brigham Young University - Provo

Oral Proficiency Assessment of French Using an Elicited Imitation Test and Automatic Speech Recognition

Benjamin Millard

A thesis submitted to the faculty of
Brigham Young University
in partial fulfillment of the requirements for the degree of

Master of Arts

Deryle Lonsdale, Chair
Dan P. Dewey
Robert Erickson

Department of Linguistics and English Language

Brigham Young University

June 2011

ABSTRACT

Oral Proficiency Assessment of French Using an Elicited Imitation Test
and Automatic Speech Recognition

Benjamin Millard
Department of Linguistics and English Language
Master of Arts

Testing oral proficiency is an important, but often neglected part of the foreign language classroom. Currently accepted methods in testing oral proficiency are timely and expensive. Some work has been done to test and implement new assessment methods, but have focused primarily on English or Spanish (Graham et al. 2008). In this thesis, I demonstrate that the processes established for English and Spanish elicited imitation (EI) testing are relevant to French EI testing. First, I document the development, implementation and evaluation of an EI test to assess French oral proficiency. I also detail the incorporation of the use of automatic speech recognition to score French EI items. Last, I substantiate with statistical analyses that carefully engineered, automatically scored French EI items correlate to a high degree with French OPI scores.

Keywords: oral proficiency assessment, automatic speech recognition, ASR, elicited imitation, EI, sentence repetition, SR, SRT, French, French testing, global oral proficiency, elicited response.

ACKNOWLEDGEMENTS

My sincere thanks to:

- My wife and children. Thank you for borrowing your husband and father to this thesis and enduring the long days and nights of reading, working and writing over the last couple of years. It has been a long process and I thank you for your undying support and understanding.
- My thesis chair, Deryle Lonsdale. He taught me the techniques and skills necessary to complete this thesis. His patience, mentoring and diligence in this process have been essential to its success.
- Drs. Dan Dewey and Robert Erickson for taking the time and effort necessary to mentor and lead me through the thesis process.
- The PSST research group, for building a foundation upon which I could stand. I especially need to thank Nate Glenn for his many hours of technical support with regards to speech recognition.
- The Center for Language Studies staff and administration, for helping to ensure the success of this thesis. Not only because of the funding that was provided through their office, but because of their unwavering dedication to its success; thank you Dr. Ray Clifford, Agnes Welch and Lindsey Tipps.
- The French and Italian department, for allowing me access to French classrooms and for encouraging their students to participate.

Table of Contents

List of Figures

List of Tables

Chapter 1: Introduction

Assessing oral proficiency in language learners and second language speakers is an important but difficult task. Foreign language classes have traditionally focused on testing skills that are more easily measured, such as knowledge of grammar principles. Many language classes simply do not test speaking ability because of the time it takes for testing, scoring and receiving results and feedback. Luoma confirms this by stating: "Assessing speaking is challenging . . . because there are so many factors that influence our impression of how well someone can speak a language . . ." (Luoma 2004). There is a need for a simple, effective way to assess oral proficiency.

In partial attempt to meet this need for French, this thesis evaluates the validity and reliability of elicited imitation (EI) and automatic speech recognition (ASR) to estimate French oral proficiency. First, it demonstrates that the processes established for English and Spanish EI test development, administration and evaluation are relevant to French testing (Erlam 2006; Graham et al. forthcoming; Graham et al. 2008). It also documents my specific contributions in the development, implementation and evaluation of an EI test to assess French oral proficiency. Furthermore, it details the incorporation of automatic speech recognition to score French EI items. Last, it substantiates with statistical analyses that carefully engineered and automatically scored French EI items correlate with a high degree to French OPI scores and are usable in automatic and adaptive EI tests.

1.1 Thesis structure

Chapter 2 looks at different methods for evaluating oral proficiency and their respective strong and weak points. It also details past and present uses of elicited imitation and some arguments for and against its use. It describes different methods set up by several organizations that use EI for a variety of tasks and the lessons they have learned. Finally, it looks into some issues in using automatic speech recognition to assess oral proficiency.

In chapter 3, I delineate the implementation of lessons learned from previous test development in English and Spanish testing. I list the steps necessary for developing EI tests using natural language processing techniques and resources and in accordance with criteria from OPI training guidelines. In this chapter, I also specify the design and delivery of the test and the

inclusion of student participants. I detail the preparations necessary for web-delivery to a group of participants and the methods used to solicit their participation. Last, I discuss some of the problems that I encountered and how they were resolved.

The next chapter explains the methods and tools I used to rate and score the participant responses by the human raters and the ASR. It shows the implementation of best practices learned from research by various groups. It discusses the ASR configuration for scoring and the different configurable parts. It also briefly shows some of the algorithms set up through previous work and how they work to score the EI items.

In the results and analysis chapter I demonstrate the ability of elicited imitation and automatic speech recognition to estimate oral proficiency in a variety of French speakers. To do so, I show the different scoring methods with ASR and how they correlate to human ratings and OPI scores of the participants. I also evaluate the test items using Item Response Theory (IRT) and then show how associating each stimulus sentence to a proficiency level allows for a more in granular analysis of each participant's score at each level and lends to the possibility of automatic and adaptive testing.

The last chapter specifies the conclusions that were reached and their impact on the field of oral proficiency assessment. It also provides best practices and lessons learned as well as future directions for testers wishing to use elicited imitation to assess oral proficiency.

Chapter 2: Survey of Literature

I begin by reviewing four main topics in the survey of literature. First, I present an analysis of oral proficiency testing in its current state. Second, I thoroughly review elicited imitation, its past, its present usage and differing opinions about its use. Third, I look into the topic of automatic speech recognition. I show how it has been used in similar projects to work side by side with different methods to determine oral proficiency. Last, I assess different methods of statistical analysis typically used to evaluate proficiency tests. Along with the typical methods, I also investigate the usage of criterion-referenced analysis in language testing.

2.1 Oral Proficiency Assessment

Several methods are presently used to assess proficiency in language learners and L2 speakers. A few of these methods include oral interviews - specifically the oral proficiency interview (OPI), semi-direct methods and automated tests (Bernstein et al. 2010). Each of these methods has its weak and strong points and these will be discussed in the sections below.

The OPI is an interactive test in which the interviewer assesses the oral proficiency of the examinee. This interview technique has its strong and weak points. First, as a personal interview, it is able to more realistically duplicate a communicative event so the actual communicative proficiency is gauged. This is one reason why OPI is the accepted best measure of oral proficiency. Part of this stems from the fact that the interviewer is supposed to be able to, through probing, determine the linguistic ceiling of the examinee (Liskin-Gasparro 1982). Validation studies of most other testing methods will attempt to show a high correlation between their test results and the results of the same speakers on the OPI or another accepted oral proficiency measure (Bernstein et al. 2010; Radloff 1991).

Another type of testing that has come about in recent years is the semi-direct proficiency testing method. This type of test has similar grading mechanisms, but uses computers or audio recorders to administer tests instead of trained professionals. Such a method allows for a more standardized and presumably less-expensive version of the OPI. The U.S. Government has even funded projects such as the Computer Assisted Screening Tool (CAST) through Fullbright Grants (Malone 2007). The idea is that with a less expensive test, a greater frequency of proficiency testing can be conducted, allowing for better-directed classrooms.

Semi-direct tests have some of the same pitfalls as the OPI in that they have kept a similar test while cutting out the test administrator, using a computer to administer it instead of a human. There is, however, still a need for highly trained personnel for grading and upkeep throughout the entire life of the project. Furthermore, it loses some of the authenticity and uniqueness that the OPI offers with highly trained and specialized OPI administrators. Overall they are a viable replacement for low-stakes testing.

The third type of test is the automated proficiency test. These tests are typically administered via a computer program and speech is either elicited via questions and tasks, or they are asked to repeat sentences. One major problem with automated testing is that it relies on technology, which is not always 100% reliable. It is also not as accurate as the other testing methods, partly because proficiency is inferred by correlation rather than directly measured. This method is thus normally only used as a screening method and/or as an estimate of proficiency. The positives of this method may outweigh the problems in certain situations. The ability to instantly test a high quantity of speakers and provide quick results is a very attractive option. Often the questions that must be answered when attempting an automated test are (1) which types of tests to use and (2) what is to be measured. Some focus on measuring fluency through repetition and reading tasks (Müller et al. 2010). Others measure mostly accuracy similarly through reading, repetition or elicited speech tasks (Bernstein et al. 2010). These tests often include at least one elicited imitation task, though they may call it sentence repetition or a repeating task.

2.2 Elicited Imitation: Past and Present

I now turn to the history of EI, some of the theoretical assumptions behind using EI to measure oral proficiency and some of the downfalls and benefits of using EI to do so. I also show present studies, organizations and other testing formats that use EI as part of their testing.

Elicited Imitation (EI) has been around as a tool to approximate communicative capacity since the 1960's when Menyuk (1964) and Slobin & Welsh (1971) first began using it. Since then its usage has increased or decreased as different organizations or researchers advocated its use or argued against it. EI was at the peak of use in the 1970's while being used to estimate communicative competence in L2 learners of French (Naiman 1974) and normal and abnormal

development in L1 speakers of English (Carrow 1974). It lost popularity and validity in the eyes of many when a 1978 study conducted by Hood & Lightbrown found that EI did not, at that time, fit very well into their requirements for a research method. These requirements were that it be valid, reliable, quick and easy to administer. They focused in on the contradiction in studies conducted by EI researchers as some said that EI could measure comprehension and others production. They listed the following points to discourage its use: "The relationship of elicited imitation to the normal functions of language is not at all clear, nor is its validity and reliability for measuring linguistic knowledge" (Hood & Lightbrown 1978). This study pointed out some serious flaws in EI that needed to be worked out at the time, but which helped shape the future of EI usage.

As EI developed and evolved over the years and into the 80's, researchers found practical uses for the practice that moved away from attempting to discover communicative competence to approximating proficiency or getting at a learner's implicit knowledge of a language. By 1995 Chaudron, from the University of Hawaii, began using EI to estimate global proficiency. He found, in his 1994 study with Bley-Vroman, that "EI is a reasonable measure global proficiency" (1994).

Another organization that used EI as an estimate of oral proficiency was also present in the early 1990's. Researchers from the Summer Institute of Linguistics (SIL) used EI as a quick and effective estimate of oral proficiency in bilingual communities such as Pakistan. They found EI to be a great tool in an environment where there is not a lot of time to conduct lengthy interviews and the local norms did not permit long interactions with a foreigner. Their validity and reliability results have been beneficial to those looking to produce similar results (Radloff 1991).

Measuring the implicit knowledge of language learners is a central research topic of Erlam from the University of Auckland, New Zealand. She takes a different approach for using EI, but has had success in validating the use of EI for such a purpose. Her studies have also been instrumental in determining some of the guidelines for the development of EI instruments (Erlam 2006; Erlam 2009).

Much of current EI theory associated with oral proficiency measures is still derived from the systems reported in Bley-Vroman & Chaudron (1994). Though they are not the first to report these systems, they articulate the process the mostly clearly. The basic underlying assumption of EI they present is that when a sentence is elicited from learners, several systems are involved. They list these systems as (1) the speech comprehension system, (2) the representation, (3) memory and (4) the speech production system. The learner must first process the sentence they heard through their speech comprehension system, and then form a representation. This representation is stored in what they called at the time short term memory (STM). The representation is then passed through the speech production system and the learner reproduces the sentence. The idea is that once a certain length threshold is reached, thought to be 7 chunks at the time, the learner would no longer be able to repeat the sentence by pure rote imitation, but would have to pass the sentence through the above-mentioned systems during the imitation process.

Sentence length and working memory has always been an important part in the assumptions of EI theory (Slobin & Welsh 1971; Spitze & Fischer 1981). Though some of the ideas have changed over the years about short-term memory and working memory, the fundamental theory for EI is still the same. Once the EI item reaches a certain length threshold (varying in syllable length according to the proficiency of the learner) the learner must process the item through their comprehension system to produce a representation which is, in turn, passed through the production system. The production system is seen as containing the current linguistic implicit knowledge of the language learner (Bley-Vroman & Chaudron 1994; Spitze & Fischer 1981). The length threshold is measured in syllables and is currently thought to be around 4-7 chunks, or groups of syllables or words in this case. For more on the relationship between working memory and EI see Okura (2011).

2.2.1 *Current Uses of EI*

The history section above mentioned the study by Hood & Lightbrown (1978), which pointed out some serious concerns that have not been entirely overcome. To portray EI as a perfect tool for understanding a learner's communicative competence or production capability would be misleading. EI is admittedly imperfect. It has shown to be valid and reliable tool for certain tasks and not so valid for others. The use of EI for too lofty goals may produce results

less beneficial than expected. Researchers seem to have realized this as there has been a decline in EI studies trying to provide a measurement of linguistic competence or production in children and learners. Now studies focus on either discovering parts of implicit linguistic knowledge in language learners or estimating global oral proficiency (Bley-Vroman & Chaudron 1994; Erlam 2006; Radloff 1991).

Many current studies have had great success with the use of EI. The benefits of using EI are seen when fitting them into the criteria mentioned by one of its greatest opponents. Hood & Lightbrown (1978) demand that a method for linguistic research be valid, reliable, quick and easy to administer. EI was widely used at that time because it was quickly and easily graded. It is also very easy to administer, and is becoming even easier to administer with the advances of computer and internet technologies, especially automatic speech recognition (ASR). EI has lacked in the past in the areas of validity and reliability. This has recently changed, however, as EI has been shown to reliably measure implicit knowledge and oral proficiency in language learners (Bernstein et al. 2010; Bley-Vroman & Chaudron 1994; Erlam 2009; Müller et al. 2010; Radloff 1991).

Currently, EI is being incorporated into some of the current automatic tests. One of the most prominent of these examples is the Versant Speaking test provided by Pearson Education, Inc. The test they use for English includes what they call a Sentence Repeating task where the L2 speaker is instructed to repeat sentences out of context as closely as possible to the original. They use this method to inform the fluency, pronunciation and sentence mastery portions of their speaking test. They correlate their scores to the New Test of English as a Foreign Language (TOEFL) Speaking, TOEFL iBT Speaking, Common European Framework, Interagency Language Roundtable (ILR) Speaking and Internation English Language Testing System (IELTS) Speaking tests. They show correlations between .75 and .94 between their speaking tests and those listed.[1]

Another recent test that included an EI task is a study conducted at Stellenbosch University in South Africa. Though the study focused heavily on fluency measures, one of the

[1] https://www.ordinate.com/technology/VersantEnglishTestValidation.pdf.

main tests used to gather data was an elicited imitation task, called a repeating task (Müller et al. 2010).

2.3 EI Test Design

Much work has been done in the area of test development for EI. Whether it be geared to measure implicit knowledge (Tomita et al. 2009), bilingual oral proficiency (Radloff 1991) or global oral proficiency (Chaudron et al. 2005; Graham et al. 2008), much work has ensured that the tests are measuring what they are expected to measure. These are some of the guidelines that have been consolidated by Tomita et al. (2009):

1. Stimulus sentences are sufficiently long to surpass the working memory of the participants

2. If specific features are being evaluated in the stimulus sentences, they should be placed as close to the middle of the stimulus as possible because of the saliency of the beginning and end of the stimuli

3. If measuring implicit knowledge, participants must be able to attend to the meaning and not the form

4. Stimulus sentences should not be too easy or too hard to avoid floor and ceiling effects

5. Instructions to participants should be clear and succinct to ensure the participants know exactly what they need to do

Though not all of these guidelines apply to every type of testing that EI can be used for, they are nonetheless a good start for any test developers.

2.3.1 *Stimulus Development*

Developing stimulus sentences for an EI test can be done in several ways. The first is somewhat rudimentary, but is the least time and work intensive. The researcher need only decide which features are to be investigated and create sentences with those features included. The second way is to administer another proficiency measure, like the OPI, and use utterances from this test as EI stimuli. The last way is to try to extract naturally occurring sentences from a

corpus of natural language. Each of these methods and how they have been used in the past studies, as well as their effectiveness as reported by the investigator, will be reviewed.

The first method has been used in many studies dating back from Slobin & Welsh (1971) and Carrow (1974) to more recent studies (Graham et al. 2008; Graham et al. 2010). Most of the early studies do not actually report on the effectiveness of their stimulus items in relation to other methods. Their results, however, are presented as providing relatively good correlations to external proficiency measurements when that is the goal. Later studies have found that sentences devised from the mind of the test developer can often end up being unnatural or awkward, affecting the effectiveness of the stimulus (Christensen et al. 2010).

The second method, on the other hand, is reported as providing valid and reliable assessments of bilingual proficiency (Radloff 1991). Radloff found that stimulus sentences taken from the OPI recordings of native speakers were effective at predicting proficiency in speakers of all levels except between levels 2 and 3 or higher on the ILR proficiency scale (which scale was also used to initially assess their proficiency). This method is an effective way to find naturally occurring sentences as long as the recordings are accessible to the researchers and they used a variety of linguistic features – grammatical, syntactic, etc.

Progress in the field of natural language processing (NLP) has made possible the last method of stimulus development. Now that large corpora of natural language for multiple languages exist along with processing tools to evaluate these corpora, the development of EI test stimuli can be more systematic. A recent study on this topic found that extracting EI stimuli from natural language corpora allowed for rapid test development and higher correlations with other proficiency measures – Second Language Acquisition and Teaching certification in this case (Christensen et al. 2010). Another recent study found that by including natural sentences with certain syntactic features extracted from a corpus of naturally occurring Japanese actually increased an EI test's ability to distinguish between advanced and near-native speakers (Matsushita et al. 2010).

Another important aspect of the third method is that certain target linguistics features can be annotated using NLP techniques. Graham et al. (forthcoming) found that using OPI guidelines to construct EI stimuli rendered better correlations. Though automatic processes were not used in

this particular study, the studies mentioned in the previous paragraph have shown that annotating features and using them in the test development process is beneficial.

The research on EI test development suggests that doing work on the front end of test development can help make the test better in several ways. First, the test includes sentences that have actually been spoken by a native speaker of the language being studied. Second, the stimulus sentences include features which are to be investigated, or are listed in a proficiency testing guideline, while maintaining naturalness. Last, the many factors that play a role in the repeatability of a stimulus sentence, such as sentence length, can be controlled.

2.4 EI Test Administration

The method of administration of an EI test is a very important part of the success of the test. Many factors influence the successful accomplishment of the test, especially when using technology. Some of these include simplicity of instruction, method of administration (web vs. computer vs. hand-delivered) and participant recruiting. In addition to the listed factors, variations in stimuli delivery have also been investigated in EI studies. This section will focus mostly on the successful administration of an EI test, but will also briefly discuss some of the variations in stimuli delivery.

Variations in stimuli delivery methods have been investigated recently by Erlam as she has been attempting to measure implicit linguistic knowledge. She maintains that if the stimulus sentence and the participant response are divided by a task of some sort, the EI test is better able to measure the implicit knowledge of the participants. This is because when the participant is required to answer an agreement question in between the stimulus and the response, they are less likely to rely on a simple memory repetition and that they will have less access to surface representations. This forces the participant to use only features that have been acquired and a "deeper" representation of what they understood, focusing the student more on meaning and less on form (Erlam 2006; Erlam 2009).

Like stimuli delivery variations, several different methods have been used and advocated for the actual administration of the test. At first, sentences were delivered directly from the mouth of the researcher to the participants. Responses were recorded either by hand or recorded

using whatever audio recording technologies were available at the time – usually a tape recorder even up through 2006 (Connell & Myles-Zitzer 1982; Erlam 2006; Gallimore & Tharp 1981).

Currently, the state of the art is to deliver the test online and record responses using computer software. The responses are then saved as WAV files to a server and the location and participant ID are saved to a database. Human raters then grade the audio via an online grading tool and an application using automatic speech recognition and a scoring algorithm.

2.5 Rating and Scoring EI Tests

There is widespread variety in scoring algorithms and guidelines. Some advocate scoring on a scale, providing a score of 3, 4 or 5 for each stimulus sentence and calculating a cumulative overall score from the sum of the individual scores. Within these methods there are two options. First, as in Ortega's dissertation study, a subjective score is given according to the imitation of each stimulus meeting certain criteria (Ortega 2000). Most others use a more objective method and subtract one point from the best possible score (3 for some, 4 for most others) off for inaccurate words or syllables within the utterance (Chaudron et al. 2005; Graham et al. 2008; Radloff 1992).

Another method, when looking for implicit grammar knowledge, is to score only the one feature that is under investigation. The correct imitation of just that one portion of the stimulus renders a binary score of 1 or 0. This method seems to have worked well when investigating implicit knowledge of certain features (Erlam 2006). Similar to Erlam's method, some research has weighted the features that are believed to be the most semantically important when scoring EI (Müller et al. 2010).

The last method of scoring EI tests is to give a simple percentage score of words or syllables correct by stimulus sentence. As with other methods, all individual stimulus scores are summed to provide an overall percentage score. Percentage scores, however, have not been found to be significantly better at correlating to oral proficiency tests than the 3, 4 or 5 score methods (Lonsdale et al. 2009).

All of the scoring methods mentioned above can be achieved by human raters or ASR. Human rating is, for the most part, relatively straight forward in that relatively few factors can

influence the human rating other than lack of training. ASR scoring, on the other hand, includes quite a bit of variability. Whichever method is chosen, most of the scoring calculations are done as post-processing once the stimuli have been scored on the word or syllable level.

2.6 EI and Automatic Speech Recognition

Automatic Speech Recognition (ASR) has made great improvements in the last ten years. Every person with a cell phone or a computer has access to speech recognition technology. This has allowed for great advances in linguistic research and the facilitation of automatic testing, especially oral proficiency testing. One program that has been extremely beneficial to the linguistic community is Sphinx. Developed at Carnegie Mellon University (CMU) and offered as open-source products, Sphinx has become widely used. This program is the basis upon which the recognition software for this thesis is built.

When incorporating ASR and EI testing, the Pedagogical Software and Speech Technology (PSST) research group at BYU has been involved in this type of research for several years[2]. The existing PSST framework for English and Spanish EI tests using ASR will be modified to accommodate French. Resources developed by university research programs, like the one at the University of LeMans, France, produce free language and acoustic models as well as dictionaries for Sphinx. These models will be integrated into the existing framework to create a working ASR program for recognizing EI input in French.

EI and ASR work well together as we know exactly what is expected to be spoken by the participant being tested. We are able to create a finite state grammar (FSG) for each stimulus sentence. This limits the recognizer to only look for those specific words that are in the grammar. For this thesis, a few features of the ASR heavily influence its ability to effectively recognize the utterances of those taking the test, especially since EI tests are geared to test all proficiency levels. Finite state grammars, or just grammars for short, are used to tell the recognizer exactly which words to look for and which order. Language and acoustic models, on the other hand, include information about how the sound waves match up to the individual phonemes and common trigrams, or sets of 3 contiguous words, of the language, respectively. Only these main

[2] http://psst.byu.edu/wiki/index.php/Main_Page

parts will be discussed in this review. For a more detailed summary of how ASR has traditionally been incorporated into EI tests and the state of the art in the area, see Graham et al. (2008).

Two main problems with EI testing have yet to be solved by the current ASR scenario. First, since an EI test is geared to test a wide variety of proficiency levels, it must recognize not only native and near native speakers of the target language, but novices whose speech might be heavily accented. Furthermore, as the lower-level learners have not acquired all the aspects of any given stimulus, they are likely to insert words that were not in the stimulus sentence or transpose words or syllable. The current configuration handles such errors by either recognizing none of the utterance or trying to squeeze what the participant said into the list of words provided in the grammar. To this point, it has shown relatively good results and correlations, but one study has shown that there may be ways to mitigate such errors. Matsushita et al. (2011) have shown that by making a language model for each individual stimulus, the recognizer is able to catch what the participant actually said and grade against the expected sentence to give better correlations to the external tests.

A few projects have attempted to solve the problem of improving word recognition rates of non-native speech. The most recent is a study conducted by Educational Testing Services (ETS). They were attempting to improve the recognition of non-native spontaneous speech. One of the improvements they made was to train their recognizer with non-native speech. They still admit, however, that despite this their word recognition accuracy is not as good as it should be, mostly because they are attempting to recognize spontaneous speech (Zechner et al. 2009). There have yet to be any studies done on improving recognition of non-native speech in an elicited imitation test by training the acoustic model on non-native data.

The last method used to improve recognition is to use an FSG, often simply called a grammar. As mentioned above, an FSG is a method that tells the recognizer to look for certain representations in a certain order. This method is especially beneficial for tasks such as elicited imitation tests where a specific utterance is expected. To ensure that multiple possibilities are captured, including repeats and corrections, logical symbols like the Kleene star (*) and the plus sign (+) can be used. They are able to capture zero or more instances of a given state for the kleene star and one or more instances of a given state for the plus sign. In addition to this, the grammar can include the actual words of an expected utterance or a syllable by syllable

representation. The effectiveness of each of these methods as well as an analysis on their effect on an EI test's ability to estimate proficiency are detailed by Graham et al. (2008).

2.7 Statistical Methods of Evaluation

Oral proficiency assessment tools have been traditionally evaluated as a whole using simple correlations to a gold standard such as the OPI. One other measure, however, has not been used to this point but will be investigated as part of this thesis. This measure is called criterion-referenced analysis. When measuring the effectiveness of individual items within the test, item response theory (IRT) analysis has been used to determine the effectiveness of individual items to distinguish between learner levels (Graham et al. 2008). Some of these types of analyses and how they have been used in proficiency testing will be reviewed as part of this survey of literature.

Alternate proficiency tests have long been compared to other tests such as the OPI, TOEFL speaking test or other test deemed to be a good standard. When using EI as an oral proficiency measure, this method has been used extensively. In fact, almost every EI test ever given (SRT, SR, EI, etc.) has been correlated to some external measure (Bley-Vroman & Chaudron 1994; Radloff 1991). To start, testing conducted by researchers in Pakistan used correlations between OPIs and their sentence repetition test (SRT) to calibrate their test after initial testing and use those correlations to determine test scores of future subjects (Grimes 1992).

More recent tests, such as the repetition task, which is included as part of proficiency evaluation provided through Pearson testing, correlate to the TOEFL speaking test, ILR speaking test and several others (Bernstein et al. 2010). Bernstein et al. found that their test correlates to these tests usually between .75 and .85. This correlation is one of the main arguments that their test is able to correctly and consistently estimate proficiency.

A similar study by Christensen et al. (2010) showed correlations of an EI test to the OPI of .71. Though their focus was not specifically on the correlations, they used the correlation as a means to show improvement between their test and previous tests and to justify the usage of certain principles. Overall, correlations play an important role in validating tests.

Item response theory (IRT) has been used in previous EI studies to determine the ability of each item to distinguish between the learner levels. Once the best-discriminating items are found, the test is shortened and re-tested. This, or similar methods has been used by Grimes (1992) and Graham (2008) in testing, calibrating and re-testing EI items.

One method that is emerging in the field of language assessment is criterion-referenced analysis (Brown & Hudson 2002). This type of assessment has not traditionally been used to analyze EI tests because of the lack of incorporating criteria into proficiency testing. Partly because of this lack, norm-referenced interpretations have been more pervasive. The aid of NLP techniques, however, is changing the ability of researchers to include criteria in test development and allow for criterion-referenced interpretations.

One part of criterion-referenced analysis that is important to this study is the evaluation of participant consistency in each portion of a test. For this test in particular, participants with OPI scores can be evaluated on their performance at each proficiency level and a threshold of consistency can be established between the participant scores at the two levels. This type of analysis is useful for the calibration of an EI test and for adaptive computerized testing.

Chapter 3: System Development

This thesis is an extension of the existing English oral proficiency assessment project underway at Brigham Young University under the direction of the Pedagogical Software and Speech Technology (PSST) research group. The idea for this project originated in research on using EI to measure global oral proficiency (Chaudron et al. 2005). As the research continued, it evolved to using speech technologies to aid the process of assessing oral proficiency with elicited imitation. Since acceptable results were accomplished for English, the next logical step was to take the lessons learned to other languages. French was a logical choice as several in the group were familiar with French, there are many of the same resources for French as there are for English, and many students at BYU and throughout North America are studying French without much chance for oral proficiency evaluation. Much of the information in this section is from a presentation given at the Linguistic Symposium on Romance Languages (Millard & Lonsdale 2011).

With three main sections, this chapter depicts the systems that were either in place or that I developed as part of this thesis. First, the overall system is shown and described. The second section is test development. In this section, I elaborate upon the processes and resources I used to develop the EI French test. Lastly, I list the methods used to administer the test in the test administration section.

3.1 System Design

Much of the framework necessary for this thesis was designed prior to this project by members of the research group. My work has been in developing a French-specific module to be placed into this framework. There are several main parts to the system used for this thesis. These parts are: (1) test design, (2) speech recognizer and ASR scoring, (3) human rating, and (4) statistical analyses and correlation. Figure 1: Overall French EI System Design shows a graph of how they work together. The illustrated pipeline includes two parts: one for the system during the development phase – using human raters – and another for the ideal system once fully calibrated and completed.

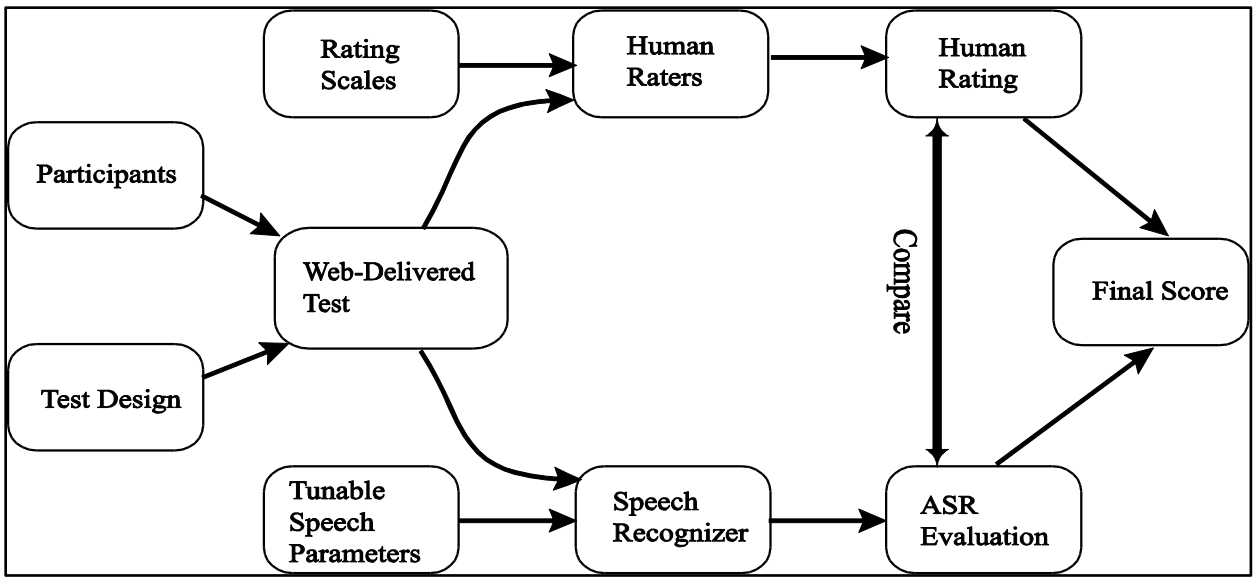


Figure 1: Overall French EI System Design

3.2 Test Development and Design

The purpose of this section is to show the development and design of the elicited imitation test for French. It is broken into three main sections, each discussing an important sub-topic for EI test design. The resources section discusses resources necessary for the ideal test development. The automatic retrieval of natural sentences section demonstrates the steps taken to extract sentences from a corpus of natural French, put them into a database, and annotate them with necessary features. Last, the item selection section elaborates on the process of taking the large database of sentences and narrowing them down to a sufficiently small bank of items for the test.

For this thesis I utilized several natural language resources in the development of the EI test. See Table 1 below for a comprehensive list of resources and their purpose. The foundation for this test development is a large corpus of naturally occurring French, the GigaWord[3] corpus. This corpus includes French newswire from two sources: Agence France-Presse (afp_fre) May 1994 - Dec 2008 & Associated Press Worldstream, French (apw_fre) Nov 1994 - Dec 2008. In these newswire is an assortment of written and spoken French. I used this corpus as it is a very large body of naturally occurring text.

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2009T28

Table 1: Resources Necessary for French Feature Tagging

Resource	Purpose
French GigaWord[4]	large corpus of naturally occurring sentences
Perl[5]	sentence extraction from the GigaWord corpus
Lexique[6]	resource for French lexical information
BDLex[7]	syllable counts
Small table of irregular French verbs[8]	used to annotate certain word-level features
Bonsai[9]	platform for French parsing
Berkeley parser[10]	incorporates into the Bonsai platform
TreeTagger[11]	part of speech (PoS) tagger for French
The Frequency Dictionary of French (Lonsdale & LeBras 2009)	lexical frequency information
TRegex[12]	analyze parsed sentences
MySQL[13]	storage and retrieval of sentences

3.2.1 *Automatic Retrieval of Natural Sentences*

Since the GigaWord corpus is not broken up in a way that is ready for analysis, parsing, tagging or database entry, I extracted sentences using text processing technologies, specifically the Perl scripting language. They were extracted according to simple length criteria since there are floor and ceiling effects in EI. If stimulus sentences are too short, they are easily imitated by rote repetition and if they are too long, they exceed the short-term memory capacity of even the most educated and proficient native speaker (Erlam 2006; Hamayan et al. 1977). For this thesis, I chose sentences between 5 and 20 words and excluded any sentences that did not at least meet that simple criterion.

As mentioned earlier, certain features taken from the ILR Handbook on Oral Interview Testing are critical to evaluating performance in an oral proficiency testing (Lowe 1982). Some, but not all, of these features can be tagged at the word and sentence level. Appendix 1 shows the word- and sentence-level features that were tagged for this study.

---

4 http://www.ldc.upenn.edu/Catalog/catalogEntry.jsp?catalogId=LDC2006T17
5 http://www.perl.org/
6 http://www.lexique.org/
7 http://catalog.elra.info/product_info.php?products_id=34
8 http://www.orbilat.com/Languages/French/Grammar/Verbs/index.html
9 http://alpage.inria.fr/statgram/frdep/fr_stat_dep_parsing.html
10 http://alpage.inria.fr/statgram/frdep/fr_stat_dep_bky.html
11 http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/
12 http://nlp.stanford.edu/software/tregex.shtml
13 http://www.mysql.com/

In order to reconcile the difference between the ACTFL and ILR OPI scales – as the ILR guidelines were used for test design and the ACTFL OPIs were administered to certain participants – I used a 4-level approach for stimulus level associations. This is the only way to match up the two methods. Table 2 below shows how the two scales line up and how associations are made during test design.

Table 2: Proficiency Level Correspondences

ACTFL	ILR
Novice	0
Intermediate	1
Advanced	2
Superior	3

One way to get at many of the features listed above is to extract them from a lexical database that contains word level information. For this study, two such databases were used for word level information extraction. *Lexique* is an open-source database of lexical information from spoken frequency to morphological count (New et al. 2004). It contains 30+ columns of information for around 140k words. This database is the main source of much of the word-level analysis. The other database, BDLex, contains less information about each entry, but has around 450k words (De Calmès & Pérennou 1998). BDLex is used primarily for retrieval of the syllable count.

In order for all of this information to be used in a manner in which each source can provide information to another, it must be entered into a searchable database. MySQL was chosen for this study as it is open-source, fast and flexible. All of the sentences, words from each sentence, and the lexical databases were each entered into their own MySQL table. The sentence table is the main table and contains all of the features listed in Appendix 1. The words table contains word-level information about each word from every sentence and is linked to the sentence table by an index of the sentence identification number (senID).

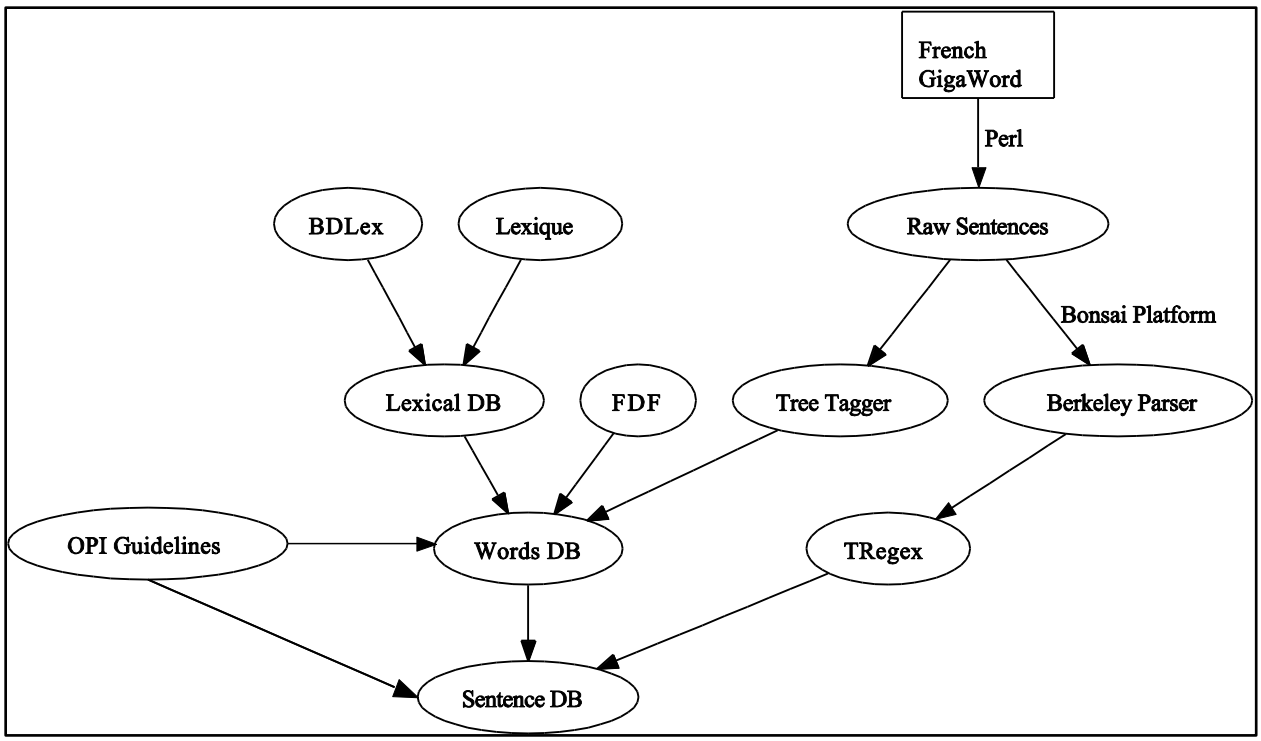


Figure 2: Item Design Resource Allocation

Many of the features listed as important features for measuring oral proficiency can be drawn from a word-level analysis. For instance, whether or not the sentence contains a frequent verb in the future tense can be annotated from a word level analysis. To accommodate this process, the sentences were first run through a part of speech (PoS) tagger. TreeTagger was used for this and it provided the PoS and the lemma for each word in each sentence. This information was gathered into the database where the word-level features could be annotated from the combined lexical database (Lexique and BDLex), the Frequency Dictionary of French (FDF) (Lonsdale & LeBras 2009) and a small table of irregular verbs (see Table 1 for links to these resources).

Annotating each sentence with its features is the culmination of all the other efforts; doing so came from two lines of effort. Figure 2 shows how the different resources all feed into feature annotation for each sentence. The first set of features is from the word-level feature annotations. They were grouped together, rendering a Boolean value of 1 or 0 showing whether or not the sentence had that particular feature. All of the word-level features were done in this manner. An example of how this works can be seen in the tag IMPF. This tag points to the existence of a verb in the imperfect tense (imparfait). This feature was tagged at the word level

by the part of speech tagger and entered into the database. This particular feature was queried in MySQL and grouped by sentence ID. This was then fed to the sentence table where each sentence containing a 1 for the IMPF feature in the word table was updated to the sentences table.

After entering all the data, I updated sentence level features by running the sentences through a syntactic parser for French. I downloaded and used Bonsai, a platform developed for running French parsers, to run the Berkeley parser for French. Once all the sentences were parsed – keeping their sentence IDs for efficient entry into the database – they were analyzed using TRegex. See (1) below for an example of the output from the Berkeley parser in Treebank format and Figure 3 below for a sample TRegex query for finding the idiomatic use of *'il y a'* with a time clause which roughly translates to 'ago,' e.g. *il y a longtemps* (a long time ago), *il y a un an* (a year ago). The number at the beginning of the sentence is the database key SenID, which helps maintain the identity of the sentence.

```
1 ( (SENT 17247 (NP (PRO Tout)) (VN (V a) (VPP débuté)) (PP (P
il_y_a) (NP (ADV plus) (P d') (DET un) (NC an))) (PONCT .)))
```

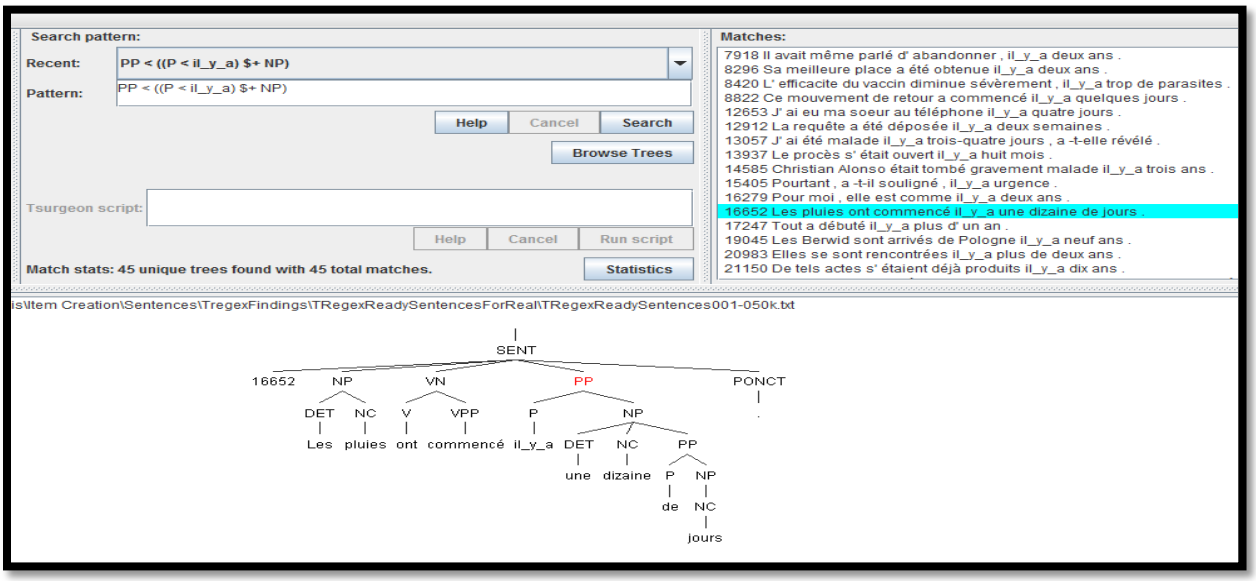


Figure 3 : Sample TRegex Query of Parsed Sentences

Though many of these queries and results are less than perfect they give the developer a great advantage since the database provides a large amount of sentences to choose from, despite the fact that there may be the occasional sentence in the returned data that does not have the desired feature. It is very easy, once the desired features have been determined, to sort through

the sentences provided by the database and choose the ones that best suit the need for that particular test or test item.

The end product of gathering and analyzing all of this data brings the project to a single point of data – the proficiency level association for each sentence. Once each sentence has an associated proficiency level, sentences can be quickly and efficiently selected by a test developer, and the sentences can be tested in an elicited imitation task.

To determine which items to select for actual testing, I simply queried the database that was created to store the French sentences. First, I queried for sentences from each proficiency level. I then looked at the guidelines and narrowed the search down by certain features. I included sentences for each feature and with a variety of sentences ranging from 7 to 25 syllables for each level.

Searching and narrowing the sentences down by features provided about 150 good sentences that had a variety of features and lengths in syllables. A native speaker and several other L2 French speakers then evaluated the sentences. This was necessary because sentences taken out of context can lose meaning or become awkward. I asked the evaluators to read over the sentences I chose and mark any that did not make sense out of context. Any sentences marked by the evaluators were dropped from the test. The completed test had a pool of 82 stimulus sentences.

Once the items were selected and had been vetted, they were read aloud and recorded by two native speakers in a high quality sound booth. Each native speaker recorded every sentence while the other listened for any irregularities or mistakes. The sentences were recorded as 16 bit WAV files at 44100 kHz. After the audio was cleaned up and normalized using Audacity[14], it was converted to MP3 format using FFMPEG[15] for inclusion into the Flex-based[16] web-delivery system and saved to a folder on our testing server. Lastly, the audio path on the server was entered into our MySQL database which is tied into the tester using PHP scripts.

[14] http://audacity.sourceforge.net/
[15] http://www.ffmpeg.org/
[16] http://www.adobe.com/devnet/flex.html

3.3 Test Administration

For this thesis, 94 participants from the French and Italian Department at Brigham Young University (BYU) were tested. Participants come from a variety of classes and proficiency levels, from French 101 – the entry level course – to graduate students and native speakers. Three participants, acting as a control group, were actually recruited because of their lack of any French. These three participants show the ability of a participant with no French background and are listed below as absolute beginners in Table 4 and as the 3 novice lows in Table 3. They are the only participants listed in Table 3 who were not actually given official ACTFL OPIs. All other participants were recruited in cooperation with the French and Italian department and the Center for Language Studies (CLS) at BYU. Each participant was briefed individually or as a group as to their rights as a research participant and asked to sign a consent form (see Appendix 2).

Table 3: Participants with OPI Scores

ACTFL OPI Score	Participants
Novice Low (assumed)	3
Novice Mid	0
Novice High	0
Intermediate Low	3
Intermediate Mid	1
Intermediate High	4
Advanced Low	3
Advanced Mid	4
Advanced High	2
Superior	5
Total	25

Table 4: All Participants and Class / Level

Class or Level	Participants
Absolute beginner	3
101 (first semester)	20
102 (second semester)	26
200 (second year)	21
300 (third year)	11
400 (fourth year)	8
500 (graduate)	2
Native	3
Total	94

All participants were tested in a computer lab at BYU that is available for research or other collaborative activities at BYU. It has 12 Apple computers that can be used on the Windows or Mac operating systems (OS). Most tests were delivered using the Mac operating system, but some were done on Windows. For the browser, most of the testing was done using the Chrome browser. At the beginning of testing, however, and partially dependent on the test proctor,

Firefox was also used. All testing was proctored by the author or one other designated test proctor.

The test is delivered using an online Flex application. Each participant's ID is entered into our database before testing and they log into the application using this ID. Upon entrance into the program, an audio check is conducted to ensure good audio quality through the current microphone and audio settings. See Figure 4 for logging in and audio check.

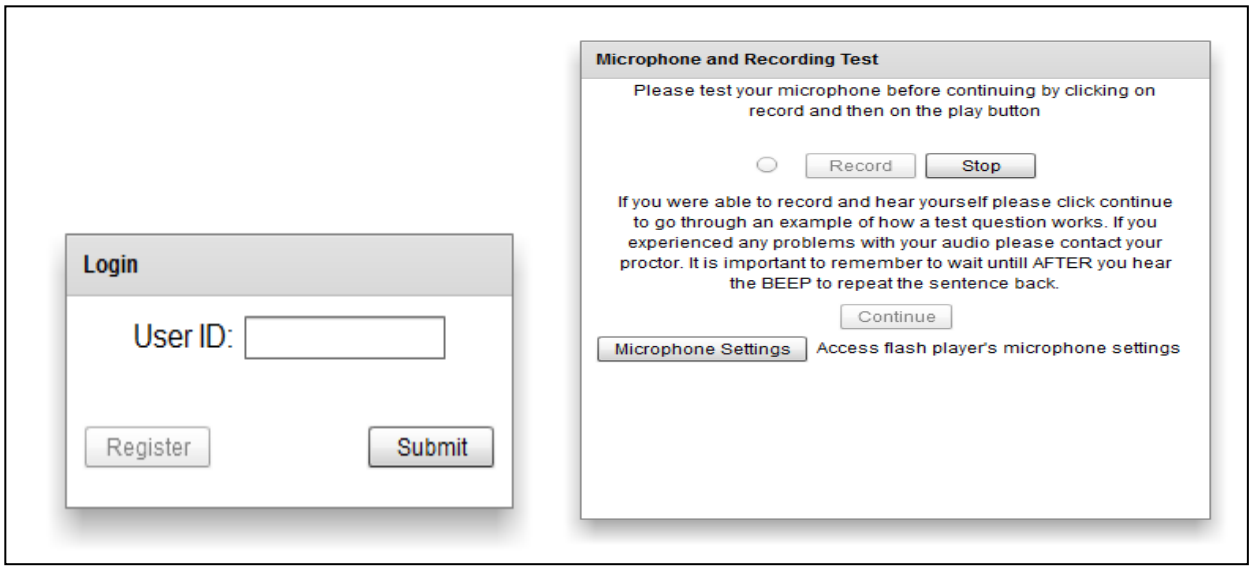


Figure 4: Log in and Audio Check Screenshots

If the audio settings are correct and the program finds the test recording to be adequate, the participant proceeds to a test item – see Figure 5.

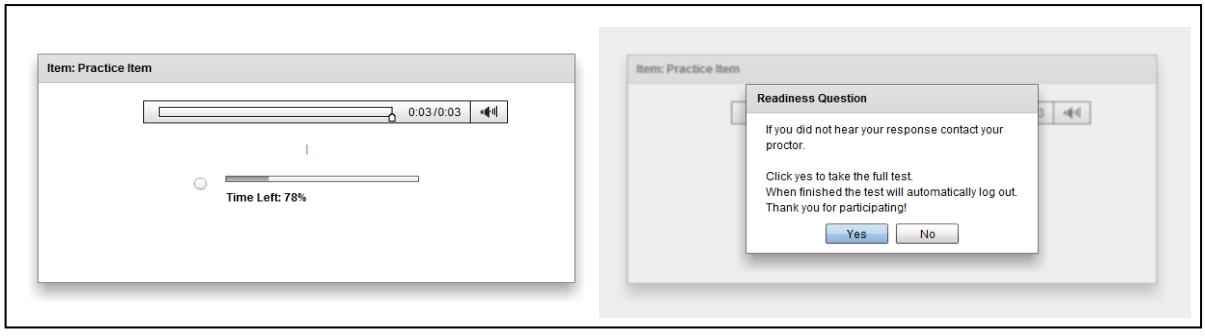


Figure 5: Practice Item Screenshots

At this point, if the participant has no questions or concerns, the test begins. After each item a beep indicates the point at which the participant should begin their response. The responses from each participant are saved to the server by the interface and the location is stored in the database.

After the test has been administered, the next step in the process is for each response to be rated and scored. The next section details the processes of human rating as well as how the responses are scored using automatic speech recognition (ASR). I also provide an overview of each of the 3 different scoring methods which can be used for the human rating and the ASR scoring.

Chapter 4: Rating and Scoring

In the review of literature, different scoring methods were discussed. For this thesis, three different scoring rubrics are all utilized. The first is the scalar method, often called the 4-score method (Chaudron et al. 2005; Ortega 2000). It subtracts a point off for each incorrect syllable in the response, rendering a score from 0 to 4. Second, a 1 or 0 can be given to each response according to whether or not the participant correctly repeated the stimulus in its entirety. Last, a simple percentage score is given according to the number of correct syllables repeated. As each of these methods is used to analyze human ratings and provide an ASR score, they are organized in the following sections by scoring method, following by how they are implemented by the human raters and ASR scoring.

One caveat is that some of the response audio recordings, upon initial examination and scoring, showed irregularities. For example, one of the native speakers that was tested showed abnormally low ASR scores compared to the other native speakers and even the advanced speakers. The audio for those who had abnormal scores was first evaluated to check for anomalies. Each of them had a section of silence followed by some loud filler noises, such as microphone breathing, background noise or other. This last section of filler noise was then cleaned up using a Praat script that recursively evaluated each WAV file for sound and silence sections. When there was a silence of 1 or more seconds, it was annotated. The script then selected only the first section that was completely sound and cut off the rest. This scrubbed audio was then rerun through the ASR scoring to verify its effectiveness. Cleaning the affected audio files in this manner significantly increased the speech recognizer's ability to recognize this audio. The results from this work are discussed in more detail in Results and Analysis.

To determine the scalar score, the responses are simply graded on a syllable basis. The score for each syllable for each response of every participant, along with its associated identification key, are stored in a centralized database. A scalar score can then be extracted from the database through a query that subtracts the number of correct syllables from the total number of syllables from 4 (4 – (total_syllables – total correct)). These scores can then be analyzed by participant or averaged for a total score.

The binary score is the simplest scoring method as it gives a 1 if the response was exactly the same as the prompt, and a 0 if there are deviations that fall outside of those allowed. This score can be derived by setting a binary_score field to 1 where the number of correct syllables is the same as the total number of syllables.

The last type of score under evaluation is the percentage score. The percentage score, like the other two scoring methods, is simple. A simple percentage correct is calculated for each response by each participant. This is done by dividing the number of syllables marked correct by the number of total syllables. The percentage scores are averaged across the total number of responses for each participant, rendering an overall percentage score.

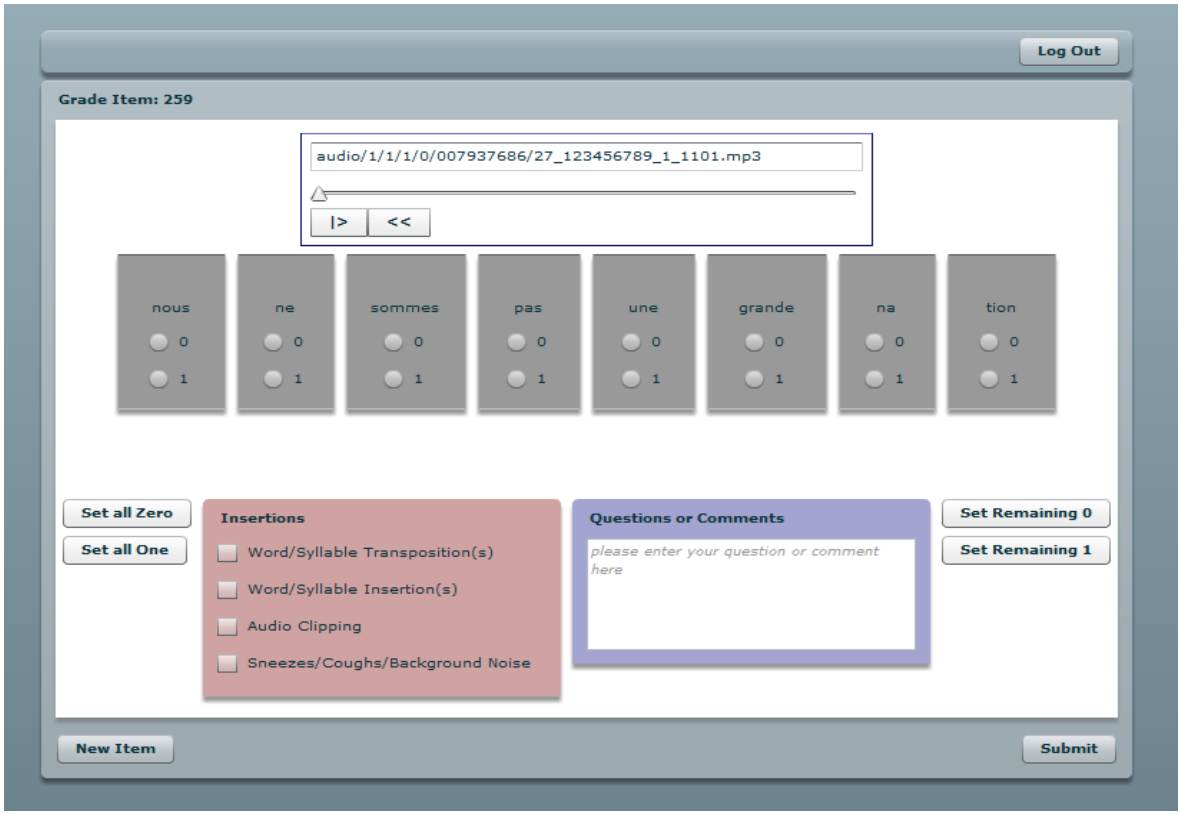


Figure 6: Rater Interface

4.1 Human Raters

For rating the EI responses, 4 raters, of varying proficiency worked on the project. There were one native and 3 non-native French speakers. Each was trained using grading criteria

established for the PSST research group and in accordance with prior research[17]. The raters where also trained on how to use the rater interface that was developed originally to grade English responses. Figure 6 shows the layout for that interface. According to the criteria they were given, raters are able to decide on which syllables to mark correct or incorrect. These scores are stored into the central database on the group server and used during response analysis.

4.2 ASR Scoring

ASR scoring is done in a similar manner to human rating. The speech recognizer reads into a directory of participant responses and grades each WAV file, storing the participant ID, response ID (item_number) and the score for each response into a local server. One of the three scoring methods can then be utilized to score the items depending on the configuration, which I will discuss in the next chapter. Simply put, most of the scores given by the ASR can be considered to be binary scores.

The engine used to do automatic speech recognition (ASR) scoring is the Sphinx4 engine developed by Lamere et al. (2003). This engine has been incorporated into the PSST pipeline and is modular enough to work with any language that has language and acoustic models appropriate for Sphinx4. The French language and acoustic models are those developed by the University of LeMans, France (Deléglise et al. 2009), which can be found on the SourceForge site for Sphinx4[18].

Using these French acoustic and language models in Sphinx4 provides flexibility in the configuration. For this test, several configurations were tested initially to determine which worked best with the right amount of speed and accuracy. It was determined that the flatLinguist, a simple configuration that uses only the acoustic model and a grammar, would be sufficient for EI testing. The flatLinguist is very fast, but normally less accurate. It works relatively well with EI, however, because the exact sequence of expected words is known and can is provided to the recognizer. There are other configurations available, but the inclusion of a general language model slows down the process and significantly reduces recognition as it must guess from all possible words.

[17] http://psst.byu.edu/wiki/index.php/Updated_Grading_FAQS
[18] http://sourceforge.net/projects/cmusphinx/files/Acoustic%20and%20Language%20Models/

4.2.1 *ASR Configuration*

Two different configuration files marshaled the two main types of grammars used in the ASR scoring. One was for word-based analysis and the other for syllable-based analysis. Since the pipeline developed provides command line options that allow for the specification of the configuration file, this was a feasible option. Each of these two grammars is discussed in the following sections.

Within the flatLinguist, which falls into the category of "Linguist" in Figure 7, there are several tunable ASR parameters. These parameters include beam widths and insertion probabilities.

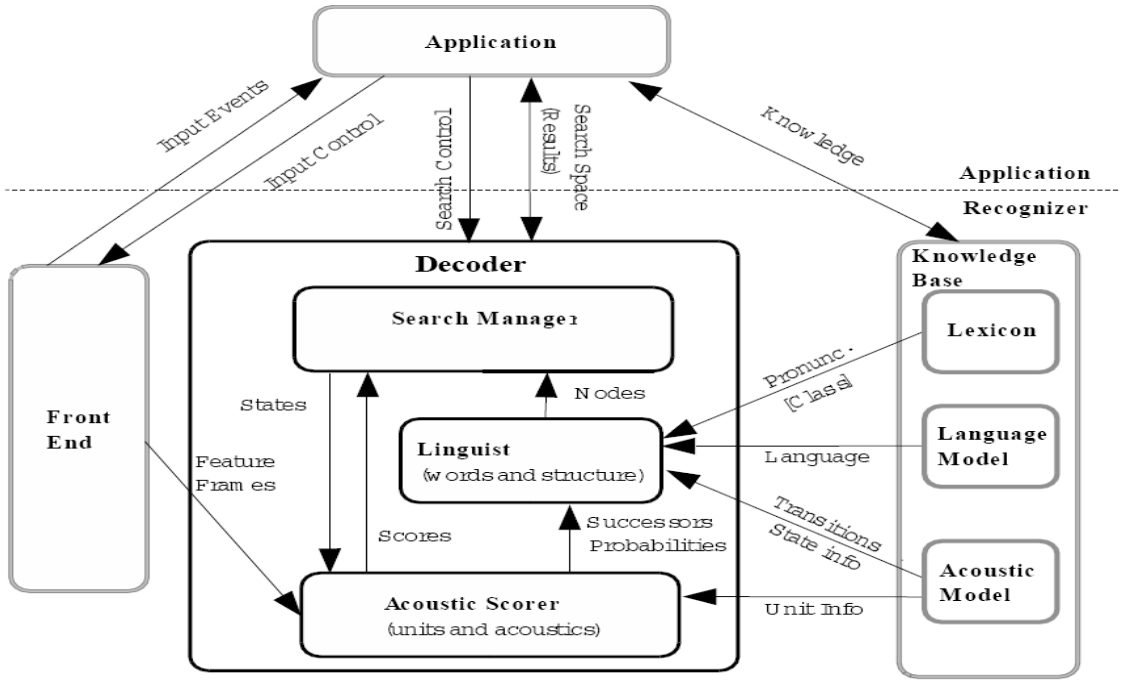


Figure 7: Sphinx4 Structure (Lamere et al. 2003)

Each tells the recognizer how much time to spend on making a correct guess or how probable it is that certain extralinguistic sounds, like silence or fillers, might be inserted. Below is a comparison between the English and the French configurations of the flatLinguist.

English Configuration

```
<property name="absoluteBeamWidth"          value="-1"/>
<property name="relativeBeamWidth"          value="1E-80"/>
<property name="wordInsertionProbability"   value=".1"/>
<property name="languageWeight"             value="8"/>
```

```
<property name="silenceInsertionProbability" value="1"/>
<property name="fillerInsertionProbability"  value="1E-10"/>
```

French Configuration

```
<property name="absoluteBeamWidth"           value="200"/>
<property name="relativeBeamWidth"           value="1E-800"/>
<property name="absoluteWordBeamWidth"       value="200"/>
<property name="relativeWordBeamWidth"       value="1E-90"/>
<property name="wordInsertionProbability"    value="1"/>
<property name="fillerInsertionProbability"  value="1E-2"/>
<property name="languageWeight"              value="8"/>
<property name="silenceInsertionProbability" value="1E-10"/>
<property name="skip"                        value="1"/>
```

Figure 8: English to French ASR Configuration Comparison

The ASR is, as of yet, still separate from the server where testing and human rating is done. For complete integration, the participant responses (as WAV files) must be downloaded to a local machine, scored with the ASR and saved to the local database. The database information must then be moved from the local database to the server database for complete analysis.

4.2.2 *Word-based Grammars*

The above paragraph mentions that different grammars, short for finite state grammars (FSGs), can be incorporated into the flatLinguist to tell the recognizer exactly which set of phonemes to expect in which order. In EI testing, two main categories of grammars can be used. They are the traditional word-based grammar and the syllable-based grammar.

Up to this point, word-based grammars have been the norm for scoring EI items with ASR. This is the easiest approach since a large of list of words is already defined as part of the acoustic model and dictionary. To make a word-based grammar, one simply has to enter the expected response and point the recognizer to that sentence. For example, if the prompt was something like: *nous avons beaucoup travaillé* (we worked a lot), then a simple grammar with that exact same sentence could then be used. In a finite state grammar (FSG), there is a certain node marked as the start point and another marked as the end point. These are designated with the double lines. In a normal grammar, there is only one start point and end point and any deviation from the expected order will cause the system to crash.

nous avons beaucoup travaillé

The ASR will look for that exact sequence of words in that exact order. Any deviation, in general, will cause the recognition to fail. This is the main reason why ASR can be perceived as a sort of binary score for most grammars – for normal grammars it is all or nothing. Some exceptions to deviations, however, will not cause failure. Restarts, for example, are acceptable in this grammar as long as the sentence is given in its entirety at some point. For the example sentence above one may see: *nous . . . nous avons . . . nous avons beaucoup travaillé.* Corrections, repairs and other speech phenomena are not handled as well with this specific grammar.

Another word-based grammar often used in ASR is called the Kleene star grammar. This grammar has more freedom in recognizing normal discourse occurrences like corrections, restarts, stutters, etc. and is much more forgiving than the normal word-based grammar. It matches 0 or more occurrences of any of the listed words. Sometimes this flexibility causes the recognizer to be less accurate, but by adding the Kleene stars, it allows each node to be a start and end point. This allows the system to start on any node, skip any node, end on any node, and count any node as many times as necessary. The above example sentence for a Kleene star grammar would look like this:

nous avons* beaucoup* travaillé**

Now deviations like "*nous avons . . . avons . . . beaucoup . . . beaucoup travaillé"* become acceptable, even recognizable, to the ASR. It does, however, often overgeneralize and overcompensate, making it, at times, inaccurate. The above response may be recognized as:

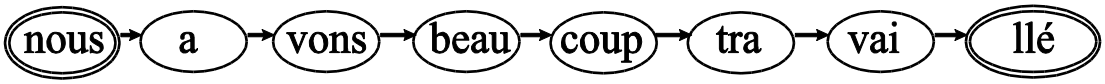
nous avons avons avons avons travaillé

Despite its imperfections, the Kleene star grammar is the only current grammar that can provide results closer to a percentage score or a 4-score.

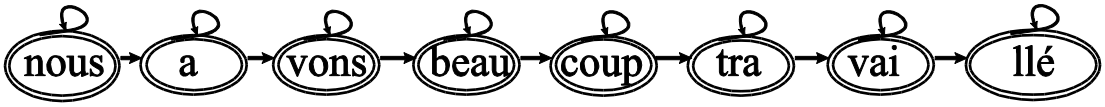
4.2.3 *Syllable-based Grammars*

The syllable-based grammars are very similar to the word-based grammars except that they are broken into their constituent syllables. In previous EI testing, we have only approximated syllable scoring by breaking up the words into other words that were similar to the desired syllables. To make these syllable grammars for French the words were broken up and the syllables that were not actual words were added to the dictionary file. There were about two hundred syllables that had to be added by hand to the dictionary. If the example sentence from above is used again, a syllable grammar would look like this for the normal syllable grammar and the syllable Kleene star grammar:

nous a vons beau coup tra vai llé



nous* a* vons* beau* coup* tra* vai* llé*



Making this change allowed for a closer similarity to the human ratings as they are done on a syllable basis.

Chapter 5: Results and Analysis

5.1 Results in View of Traditional Evaluation Methods

5.1.1 *Correlations*

For a general overview of test effectiveness, I correlated test scores from each scoring method against the OPI. This analysis allows for a general overview of the effectiveness of the test and of each scoring method. Though it is not a perfect indication of the test's effectiveness in comparison to the OPI, it is a point of departure. To begin, I calculated the correlation between OPI results and the human ratings only for those who have OPI scores. Table 5 shows these results.

Table 5: Pearson Correlations between OPI Results and Human Ratings

	Binary Scores	Percentage Scores	4-Scores
ACTFL OPI Scores	. 872	. 918	. 908
ILR OPI Grouped Equivalents	.825	.787	.818

Previous correlations between EI human ratings and OPI results have shown similar results. Graham et al. (Graham et al. forthcoming; Graham et al. 2008) presented correlations for English EI of .750. Recent work in the area of EI test has increased the correlations for English human ratings with OPIs to .890 and to .925 for Spanish EI human ratings with OPIs (Graham et al. forthcoming).

The next set of analyses compares participant ASR scores to their OPI results. Table 6 illustrates correlations between the different scoring methods used for ASR and human ratings. These correlations are slightly lower than the human rating correlations, but they are nonetheless high correlations. Note the high correlation of the syllable binary scores to OPI results.

Table 6: Pearson Correlations between OPI Results and ASR Scores

	word binary	word 4-score	word percent	syllable binary	syllable 4-score	syllable percent
ACTFL OPI Results	0.877	0.670	0.814	0.883	0.669	0.822

Correlating the ASR scores to the human rating is also another important step. Table 7 shows each of the ASR scoring methods and how they correlate to the human rating methods. This is important because we already know now which of the human rating methods correlates best and it will be important to replicate that with the ASR scoring. I highlight the correlation between the human 4-score results and the syllable binary results as they have the highest correlation.

Table 7: Peasrson Correlations between ASR Scoring and Human Rating Methods

Correlations - OPI Data	word binary	word 4-score	word percent	syllable binary	syllable 4-score	syllable percent	human 4-score	human binary	human percent
word binary	1.000	0.843	0.964	0.991	0.835	0.974	0.859	0.832	0.843
word 4-score		1.000	0.937	0.787	0.980	0.908	0.606	0.595	0.591
word percent			1.000	0.939	0.927	0.992	0.767	0.743	0.750
syllable binary				1.000	0.791	0.955	0.883	0.848	0.869
syllable 4-score					1.000	0.911	0.617	0.602	0.609
syllable percent						1.000	0.790	0.758	0.778
human 4-score							1.000	0.971	0.979
human binary								1.000	0.911
human percent									1.000

5.1.2 *Item Analysis*

For this thesis, I use Item Response Theory (IRT), calculated using Winsteps[19], to gauge item (stimulus sentence) difficulty. This method can provide the investigator with a look into the each item's ability to discriminate between participants at different proficiency levels. Figure 9, specifically, shows the relative difficulty of the given items for this group of participants. The participants used in this analysis are those with OPI scores that were human-rated using the 4-score method listed in Table 3.

According to the results of the IRT analysis in Figure 9, it confirms that the most difficult items are primarily the items associated with level 3 (superior) proficiencies. Figure 9 shows that many of the items do discriminate well since the participants on the left-hand side are nicely spread out across the levels. Only one group clusters near the upper levels of proficiency. This is similar to what I found in the analyses listed in section 5.2 where the advanced speakers were

[19] http://www.winsteps.com/winsteps.htm

clustered together with the non-native superiors. Examples 2 and 3 below show two stimulus sentences that were found to be the most difficult and least difficult, respectively.

(2) 1203 - d'importantes mesures de sécurité avaient été prises autour du tribunal (21 syllables)
(3) 1131 - il fait les choses bien mieux que moi (8 syllables)

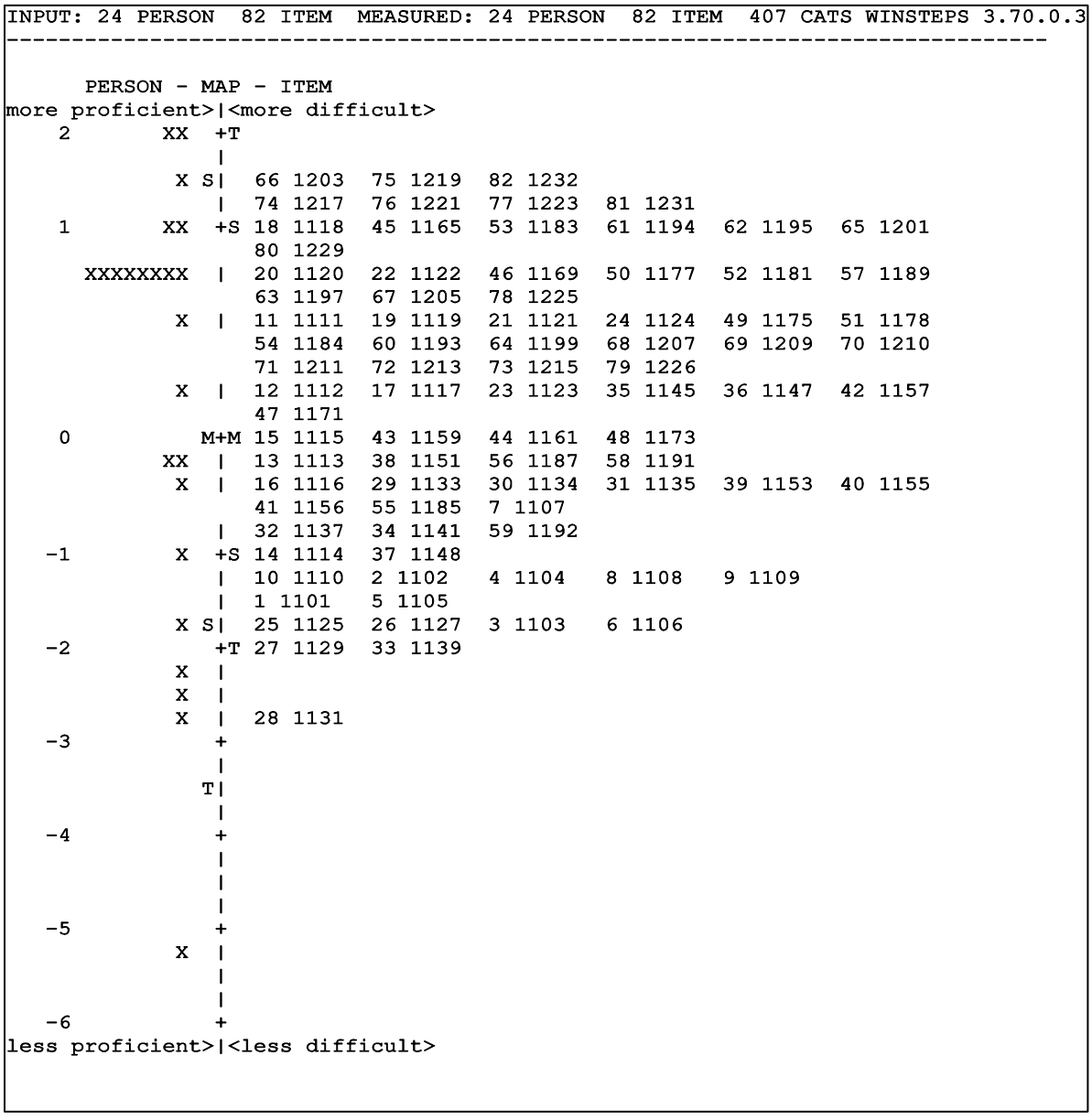


Figure 9: Human Rated 4-Score IRT Item Map

There is a normal distribution on the item (right hand) side of the map, showing that there is a good distribution of item difficulty; it is just skewed to the top. This means that many of the items are considered difficult, even to this group of participants despite the fact that 14 out of the

24 used for this analysis are advanced or superior speakers. I confirmed these results by taking the top 61 items according to the IRT and rerunning the correlation measures. Table 8 shows these results. There is very little variation in the correlations when the top 61 items are used instead of all 82. The ASR correlations increased slightly by .3 and the others stayed close to the same.

Table 8: Pearson Correlations between OPI Results and Human & ASR Scores after IRT

	ASR Syllable Binary	Human 4-score	Human Percentage
ACTFL OPI Results	0.886	0.902	0.919
ASR Syllable Binary	1	0.893	0.874
Human 4-score		1	0.973
Human Binary			0.886
Human Percentage			1

The analysis in Figure 10 shows a different perspective. Using all 92 of the participants (with two excluded due to bad audio) and the syllable binary ASR scores, I show that the picture changes slightly. This analysis shows many of the same items are listed as the most difficult, but it spreads them out across a greater distribution and, in turn, spreads out the distribution of proficiency levels on the left-hand side of the figure. Table 4 illustrates that 67 out of 92 of the participants who took this test were actually from the 100 and 200 level classes at BYU. The results are now skewed in the opposite direction of the previous analysis.

With such a large amount of what are likely intermediate speakers, the IRT now shows a clustering of participants at what is likely the intermediate level (near -2). I can now show with this analysis that the items developed for this test distinguish between the participants to a high degree. The large spread on the person side of the map confirms this, by placing some participants and others at the highest level on the person-to-item map in Figure 10.

```
INPUT: 90 PERSON  82 ITEM  MEASURED: 89 PERSON  82 ITEM  164 CATS WINSTEPS 3.70.0.3
-------------------------------------------------------------------------------

        PERSON - MAP - ITEM
 <more proficient>|<more difficult>
   4            X  +  75 1219
                X  |  46 1169
                   |  61 1194
   3            X  +  49 1175  50 1177  62 1195  76 1221  77 1223
                X  |
                X T|  66 1203
   2           XX  +S 51 1178  68 1207
                X  |  11 1111  37 1148  72 1213
               XX  |  19 1119  24 1124  47 1171  53 1183  57 1189  63 1197
                      65 1201  67 1205  71 1211  80 1229
   1          XXX  +  18 1118  21 1121  29 1133  45 1165  64 1199  69 1209
                      70 1210  79 1226
               XX  |  1 1101   13 1113  22 1122  44 1161  78 1225  81 1231
                X S|  12 1112  15 1115  16 1116  23 1123  27 1129  52 1181
                      58 1191  60 1193  7 1107
   0            X  +M 20 1120  41 1156  43 1159  48 1173  74 1217  82 1232
               XX  |  36 1147  73 1215
               XX  |  32 1137  38 1151
  -1         XXXXX  +  14 1114  25 1125  26 1127  35 1145  40 1155  59 1192
              XXX  |  42 1157
           XXXXXX M|  10 1110  28 1131
  -2     XXXXXXXX  +S 31 1135  39 1153
     XXXXXXXXXXX  |  33 1139  34 1141  5 1105
          XXXXXX  |  2 1102   30 1134  55 1185  8 1108
  -3     XXXXXXX  +  9 1109
         XXXXXXX  |
           XXXXX S|  54 1184  6 1106
  -4          XX  +  56 1187
              XX  |T
                  |
  -5          XX  +  17 1117
                  |
                 T|
  -6              +
                  |  3 1103
                  |
  -7              +  4 1104
                  |
                  |
  -8         XXXX  +
 <less proficient>|<less difficult>
```

Figure 10: ASR Scored Syllable Binary IRT Item Map

To support the item analysis, I ran every item and each of its features through the Tilburg Memory-Based Learner (TiMBL) machine learning software. I did this to show that a machine learning program, given the features of each sentence, could predict the proficiency level outcome for any given sentence. I used 95% of the sentences (over 600k) as training data and

tested 5%. TiMBL was able to predict the correct outcome 99.5% of the time. The most important measure I got from TiMBL is a feature ranking. TiMBL ranks each feature in its importance to the prediction. It found that these features were the most significant:

39 – Pluperfect
10 – Conditional
43 – Existence of the verb "rester"
34 – Inversion
07 – Complex subjunctive (other than with falloir and vouloir)
42 – Existence of the verb "manquer"
05 – Subjunctive
41 – Future past (passé intérieur)
40 – Reflexive in the past tense

All of these features are closely associated to ILR level 3 sentences that were annotated during test development. Part of the reason for the inclusion of such a large amount of level 3 features is that a large amount of the data consists of level 2 sentences.

Table 9: Best Items from Human Rated IRT and Best Features from TiMBL

item number	Prof Level	syll Count	French 5kavg	French 1kavg	Complex Subj	Plupfct	Manquer	Rester
1118	1	13	1	1				
1183	3	10	0.42	0.42		1		
1194	3	13	1	0.9	1			
1195	3	14	0.75	0.66	1			
1201	3	15	1	0.8	1			
1203	3	15	0.9	0.81		1		
1217	3	16	0.92	0.92	1			
1219	3	17	0.81	0.45				1
1221	3	19	0.63	0.45				1
1223	3	19	0.92	0.84			1	
1231	3	22	0.8	0.6		1		
1232	3	23	0.86	0.73		1		

Upon a close assessment of these two evaluations, there is an interesting overlap. Each of the 12 most discriminating items listed, except the level 1 item, contains one of the features that TiMBL found to be significant. Table 9 shows these relationships and lists the most effective items from the IRT human rated analysis along with the significant features listed by TiMBL. See Appendix 1 for a more detailed description of the sentence features.

The most significant results in Table 9 are the first two listed in the table. The first item is supposed to be a relatively simple item with only 13 syllables and all of its vocabulary coming from the top 1000 most frequent words in French. It also contains no significantly difficult features. According to the IRT, however, it considered a more difficult item because, upon close inspection of the results, it includes an irregular adverb placement between the auxiliary verb and past participle. Many participants, including some of the more advanced speakers, incorrectly reconstructed the stimulus in their response, placing the adverb after the auxiliary verb and past participle. The second item is significant for two reasons. First, it contains one of the nine most significant features as listed by TiMBL. Second, its vocabulary is quite infrequent. Only 42% of the words from this sentence even appear in the top 1000 most frequent words of French. The two features listed as French 1k and French 5k are the percentage of words in each stimulus sentence that are on the top 1000 and 5000 words of French according to Lonsdale & Lebras (2009). These frequency features were not listed in the guidelines, but were kept for future analyses.

5.2 Results in View of Criterion-referenced Analysis

During the test design, each stimulus sentence was associated to an ILR proficiency level between 1 and 3. The test was designed in this manner under the hypothesis that doing so would allow for more in depth analyses and better methods for level distinction. I hypothesized that stimuli created in this manner would improve the ability of an EI test to differentiate participant proficiency levels. I show that this type of analysis can be utilized to prepare the test for computerized adaptive testing.

As ACTFL OPI ratings were assigned to participants, EI scores from both human raters and the ASR could be correlated. The test can then be calibrated on the provided data from the OPIs that were given to a variety of participants at different proficiency levels. The results are calculated by plotting all the OPI results against the human ratings and ASR scores. Linear regression is used to determine the best fit line. A boundary is set where the separation between the proficiency levels is most distinct – represented in the figures by a dotted line. This boundary is the cutoff score for that proficiency level. Any scores falling in the top left or bottom right quadrant after the lines are drawn are deemed to be inaccuracies or outliers. Each cut off threshold is then tested against the entire data set and any that fall below each threshold are deemed to be part of the lower group and vice versa. Doing that for each of the three groups

provides a good indication of which participants should fall into which proficiency level. In the following figures, any score that falls to the right of the line at that level is determined as a passing score for that level. The next sections show these results.

5.2.1 *Human Ratings for Level 0 Proficiency*

For this thesis, no sentences were associated to level 0 since participants can be assigned to level 0 on the sole basis of their inability to perform consistently at level 1. Figure 11 shows that participants who fall below the dotted threshold line during the evaluation of level 1 sentences would be considered level 0. For the scoring methods used, there is almost always a separation between the Intermediate Low participants (level 4 on the ACTFL scale) and the absolute beginners who were tested and listed as level 1.

5.2.2 *Human Ratings for Level 1 Proficiency*

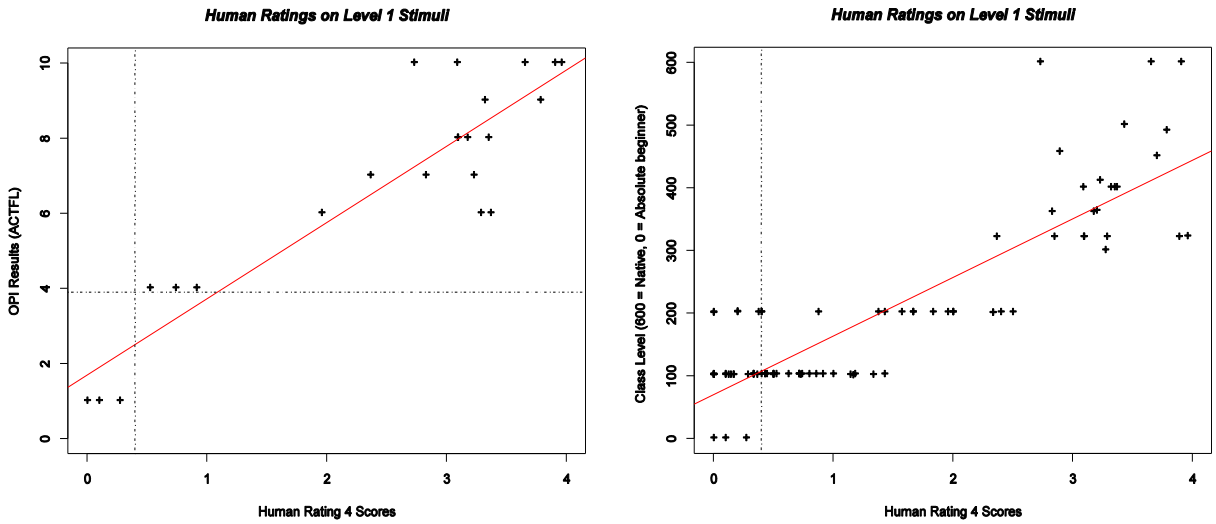


Figure 11: Human 4-Score Ratings and OPI Results on Level 1 Sentences

The most important part of the level 1 analysis is the ability of the different scoring methods to distinguish between intermediate (level 1) and novice (level 0) participants. Only 3 novices took the test – all of whom are absolute beginners with no French. The human 4-score method sets the clearest boundary between level 0 and level 1participants. This analysis evaluates very few students in the 200 level (second year) classes or above as novice. This seems to fall in line with expected results and with the few students from that level that were tested.

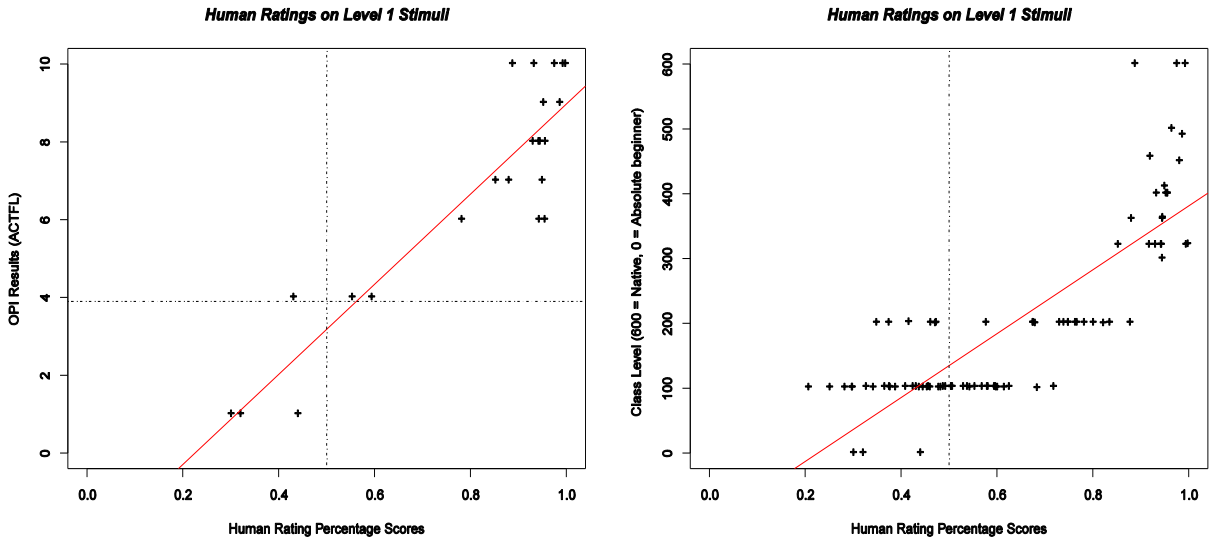


Figure 12: Human Percentage Score Ratings and OPI Results on Level 1 Sentences

The human percentage score for level 1 sentences, shown in Figure 12, has a reasonably clean separation between the level 1 and level 0 participants. When this same threshold is applied to all participants, it appears to distinguish well between the two levels. The only participants who fall below the threshold are the absolute beginners, a good amount of French 101 and 102 students and a few students in the 200 level classes. With only one participant outside the threshold, it gives this method a 96% accuracy rate.

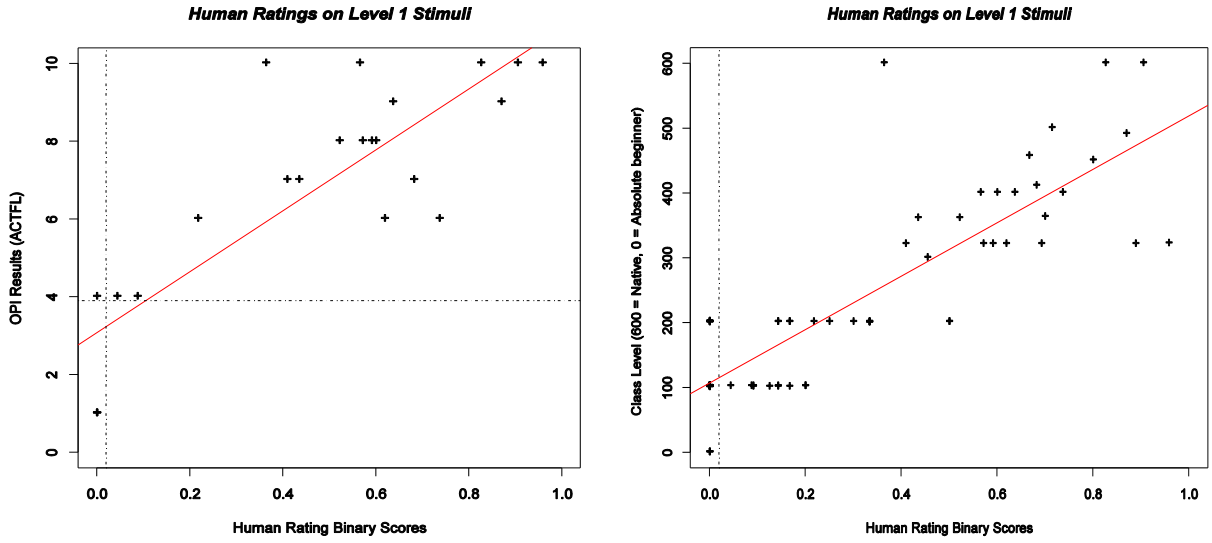


Figure 13: Human Binary Score Ratings and OPI Results on Level 1 Sentences

The binary scores performed worst in the correlations and appear to perform the worst at each of the proficiency levels. Since this scoring method has very little tolerance for error, it does not provide for a strong separation between the levels. This, along with the other results, may actually indicate a lack of easy items since there are floor effects apparent in the results for each method.

5.2.3 *Human Ratings for Level 2 Proficiency*

The level 2 thresholds are the most problematic for each of the scoring methods. There is a serious discrepancy between the intermediate high (6) participant score at level 2. This discrepancy reduces the accuracy of discrimination.

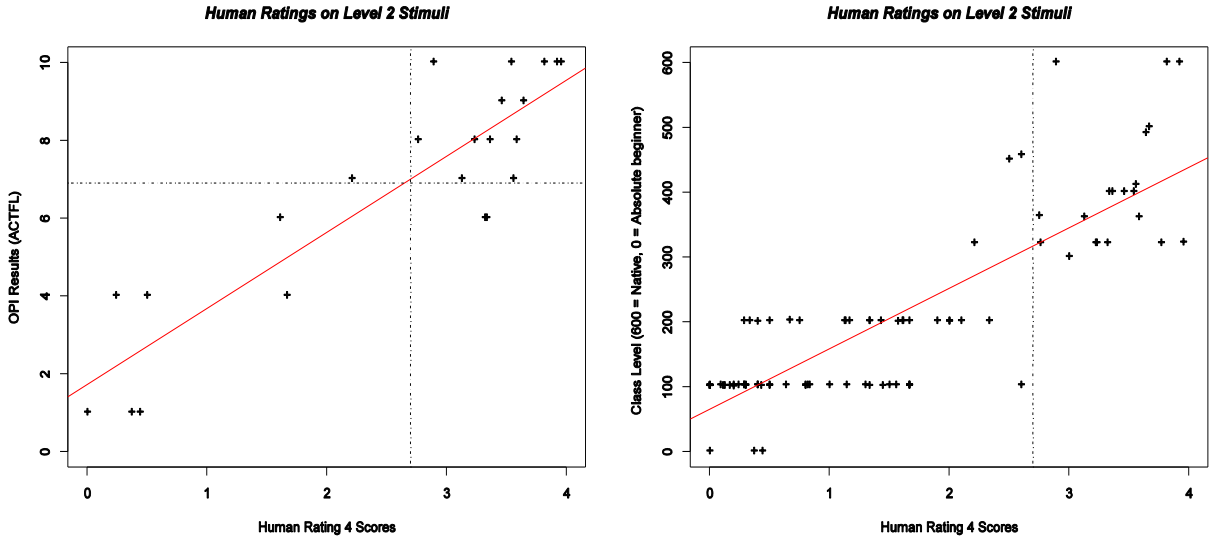


Figure 14: Human 4-Score Ratings and OPI Results on Level 2 Sentences

Human 4-scores, however, are the most accurate out of the three methods, the threshold allows 2 intermediate highs to slip past and does not allow one of the advanced lows. This means that if these three were taking the test, the two intermediate high speakers would be predicted as advanced and the advanced low would be predicted as intermediate. It is possible that there are some items in the level 2 set that do not discriminate well enough. This may be solved as the bad items are culled out. See the item analysis section above for more on this topic.

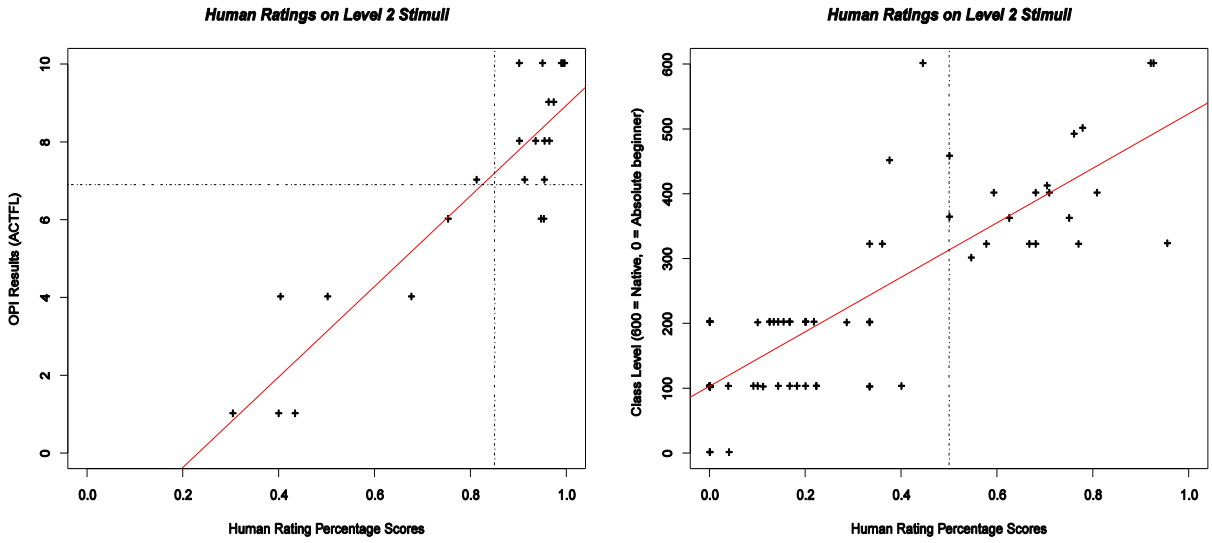


Figure 15: Human Percentage Score Ratings and OPI Results on Level 2 Sentences

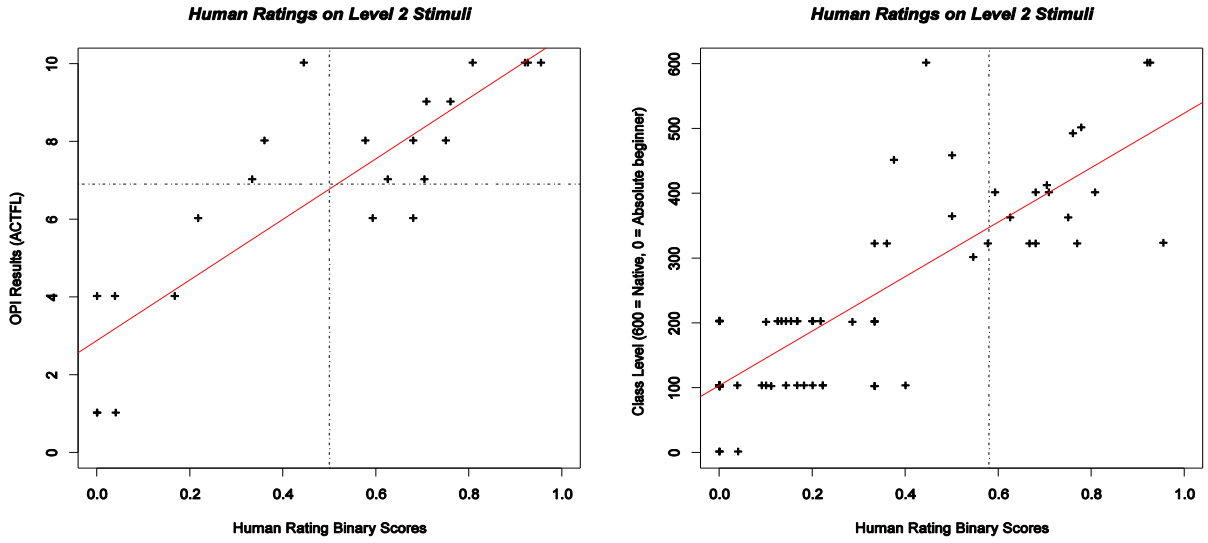


Figure 16: Human Binary Score Ratings and OPI Results on Level 2 Sentences

If level 2 sentences are unable to effectively separate intermediate and advanced speakers, it may be possible to use level 3 sentences to do the same job. Though the level 3 sentences were not originally established for such distinction, they may end up doing a better job at distinguishing not only between advanced and superior speakers, but also between intermediate and advanced speakers.

5.2.4 *Human Ratings for Level 3 Proficiency*

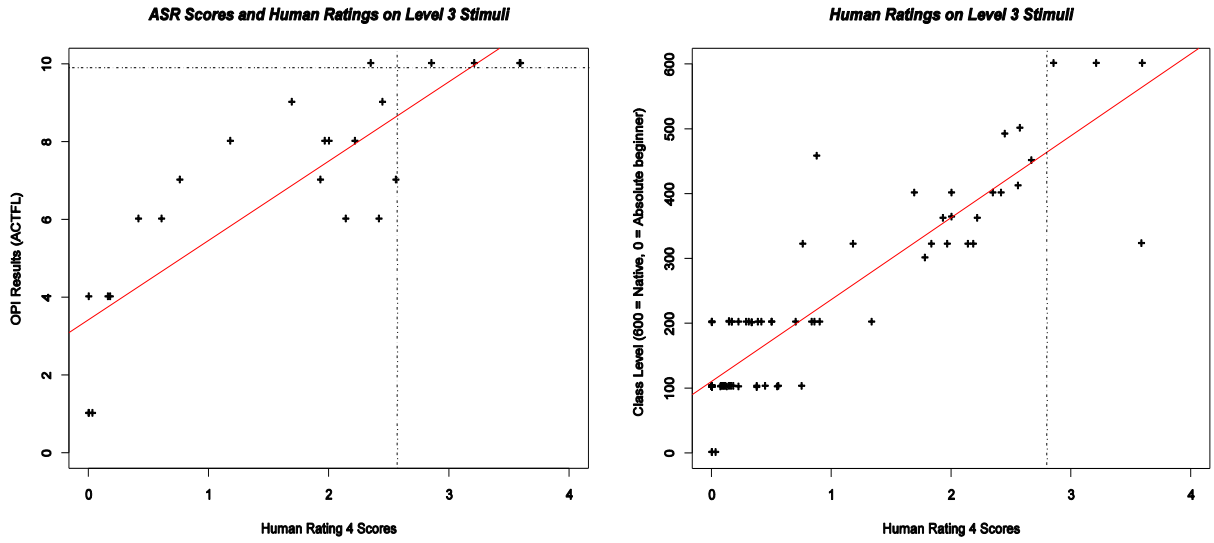


Figure 17: Human 4-Score Ratings and OPI Results on Level 3 Sentences

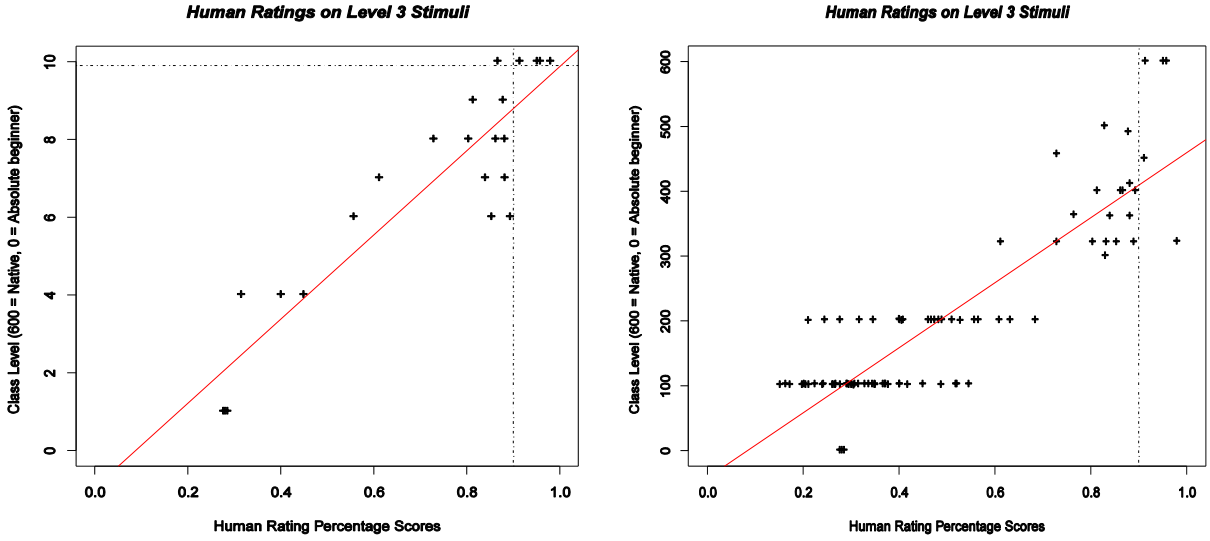


Figure 18: Human Percentage Score Ratings and OPI Results on Level 1 Sentences

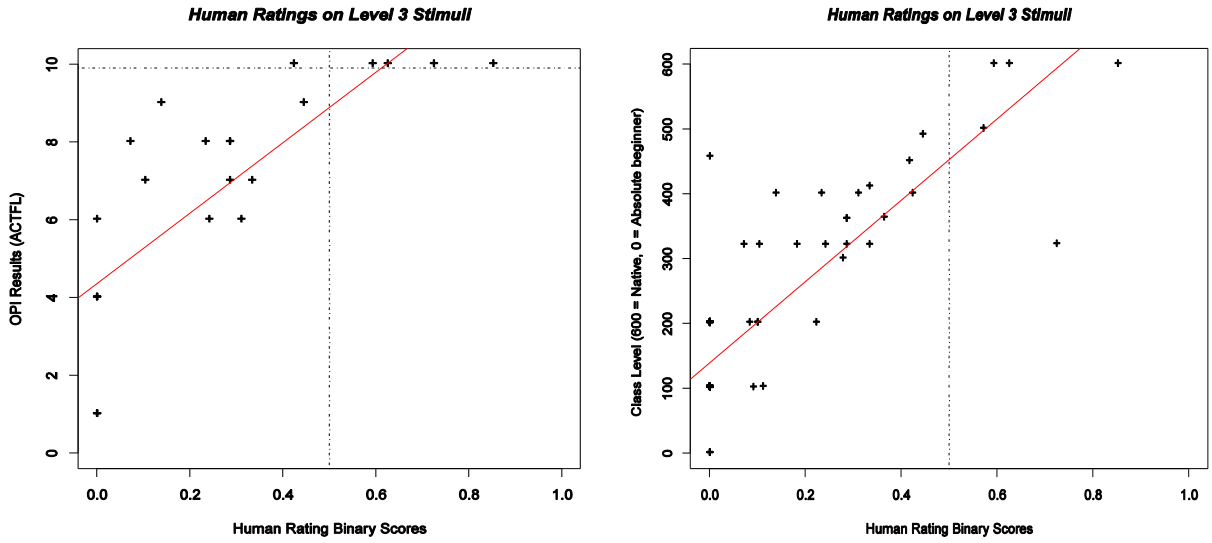


Figure 19: Human Percentage Score Ratings and OPI Results on Level 1 Sentences

5.2.5 *ASR Scores for Level 1 Proficiency*

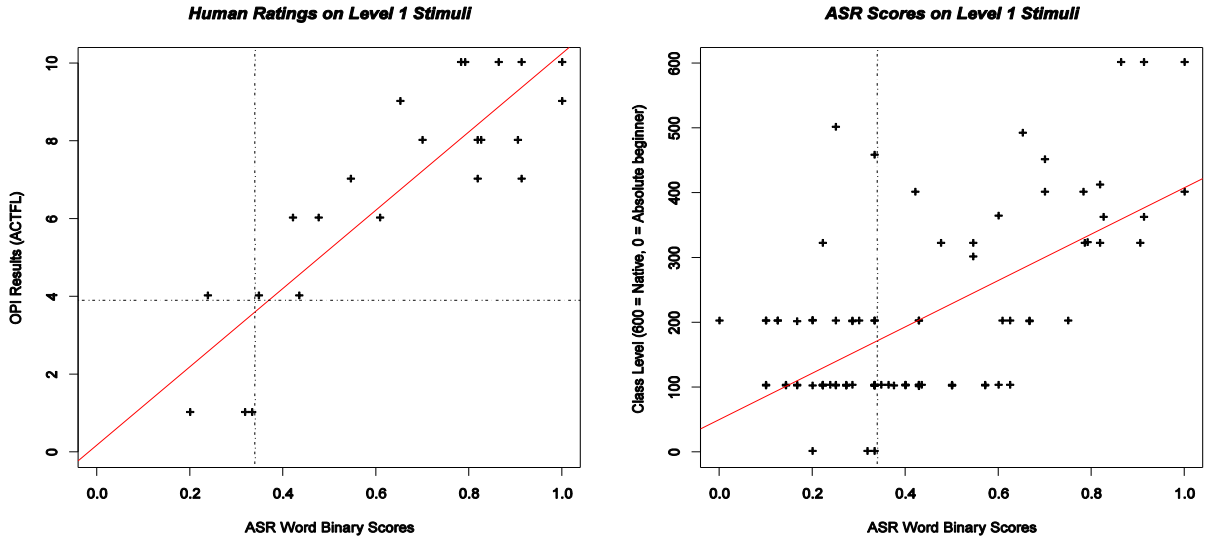


Figure 20: ASR Word Binary Scores and Level 1 Analysis

Level 1 scores are the basis for evaluating novice and intermediate speakers. As level 0 and l speakers, the level 1 sentences should provide a way to discriminate the two levels. Stimulus sentences associated to level 1 (ACTFL Intermediate 4-6) are correlated to actual proficiency level using the ACTFL 9 point scale (1-10). Each participant with an OPI score is

graphed for two different scoring methods given by the ASR scorer – binary and percentage score – with both syllable and word-based grammars.

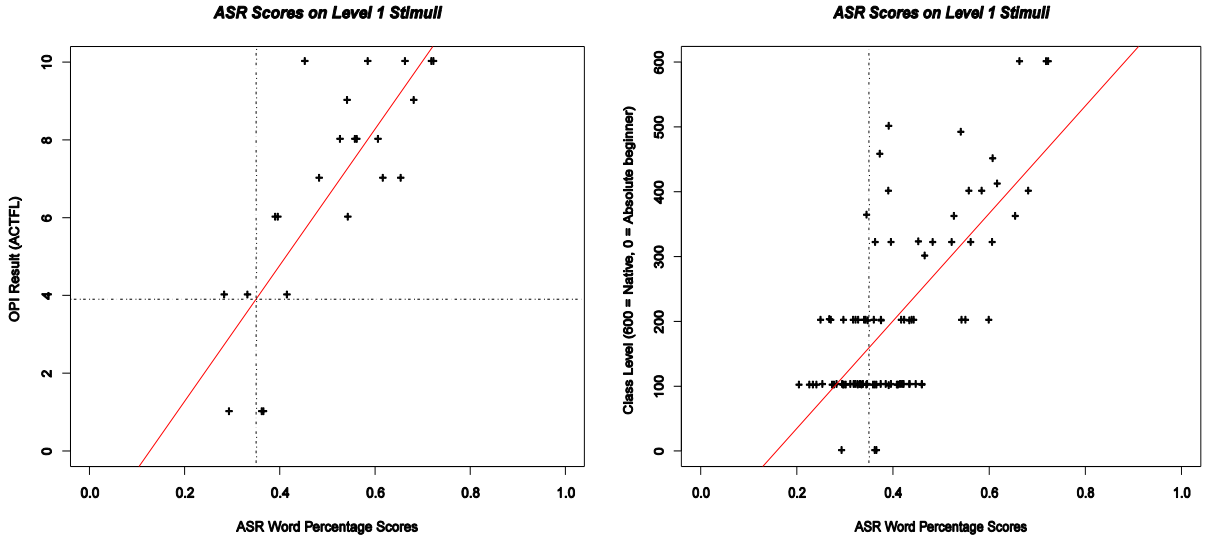


Figure 21: ASR Word Percentage Score and Level 1 Analysis

For level 1 stimuli evaluated by the word grammars, there is a considerable gap between the number of first year and second year students who would be considered below the threshold for level 1. Figure 20 and Figure 21 show the two methods and the differences between the two. It would appear that the percentage score given by the Kleene grammar is much more generous at the bottom of the scale, but blurs the lines between proficiency levels. With most scores so tightly clustered in the middle, it would appear initially that the kleene grammar with a percentage score may not be the best method for evaluating proficiency level.

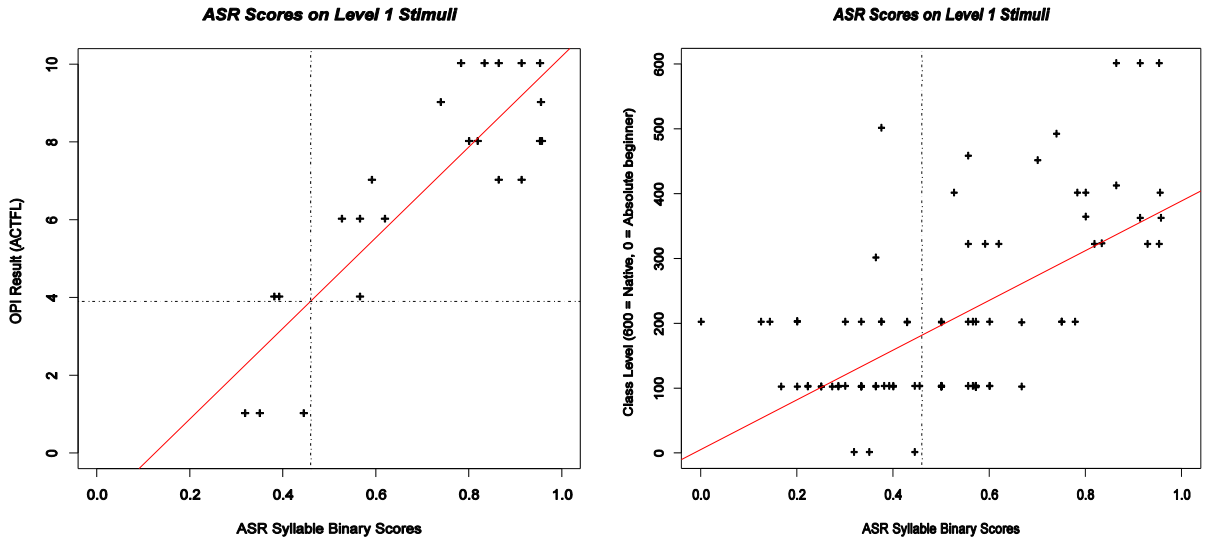


Figure 22: ASR Syllable Binary Scores and Level 1 Analysis

At this first level of analysis, one could infer that the evaluation of all level 1 (intermediate) or higher participants would pass this portion of the test. One noticeable difficulty is the lack of data for novice speakers. Though there were several first semester students who took the EI test, none were willing to take an OPI. Their data, however, is still useful as it can be we can now at least see how many of them would, using the current metrics, be considered novice speakers.

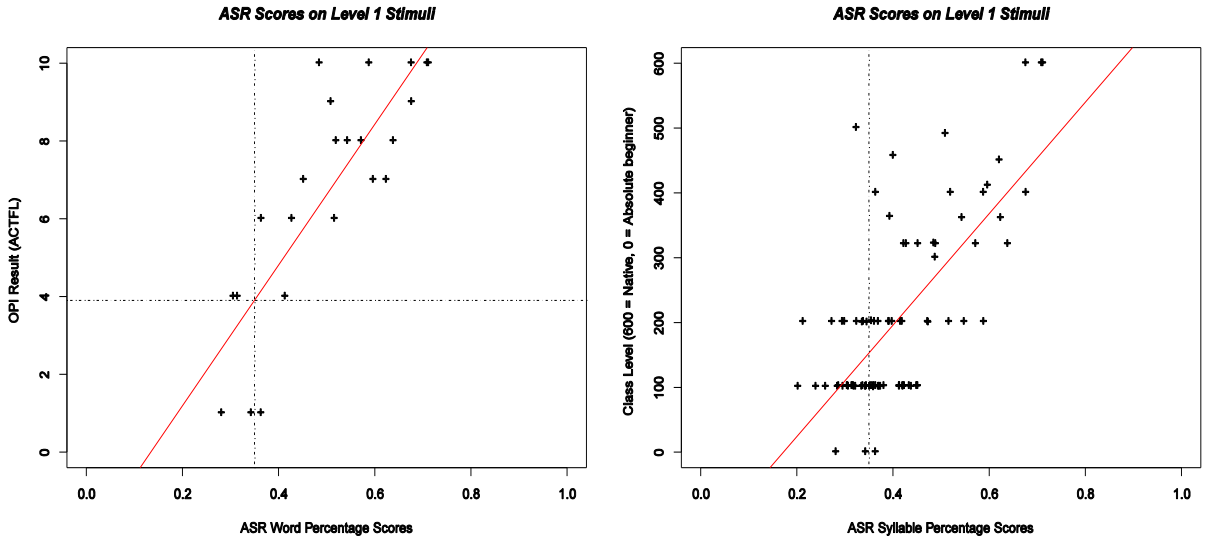


Figure 23: ASR Syllable 4-Scores and Level 1 Analysis

5.2.6 *ASR Scores for Level 2 Proficiency*

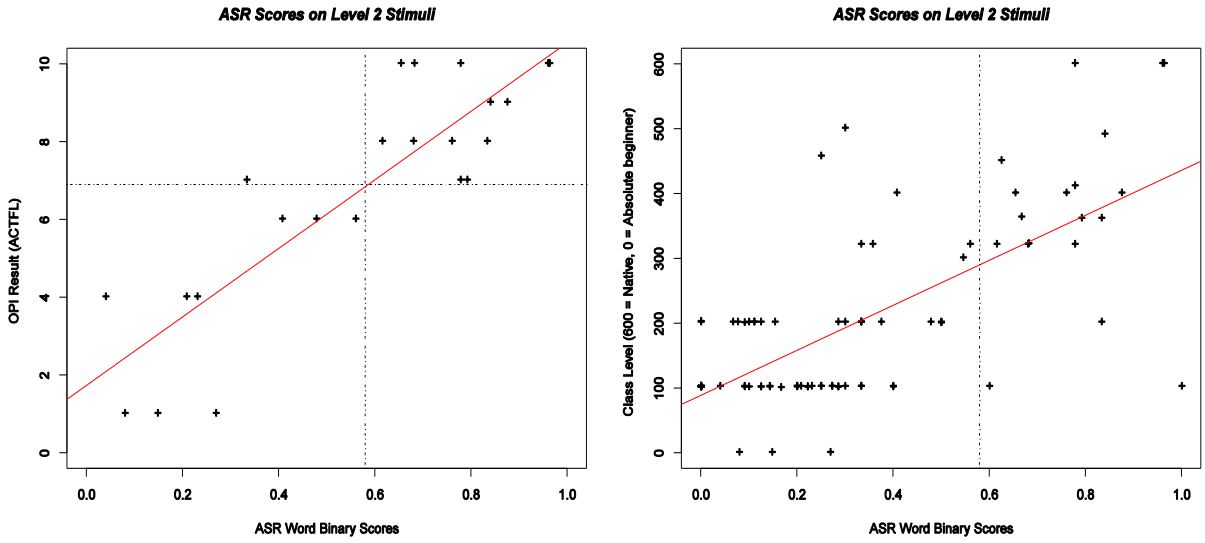


Figure 24: ASR Word Binary Scores and Level 2 Analysis

The level 2 analyses are probably the most informative. They show, in almost every method, that there would not be a single first year and very few second year students who fall into the level 2 (ACTFL advanced) category. This would seem to be a safe assumption. And in the binary score evaluations (Figure 24 and Figure 26) only one participant score falls outside of the expected thresholds. This would give the binary scoring method a 95% accuracy in distinguishing between participants at a level lower than level 2 and those at level 2 or higher on the given data. This would decrease the accuracy of the binary method to 97.5% for participants up to level 2.

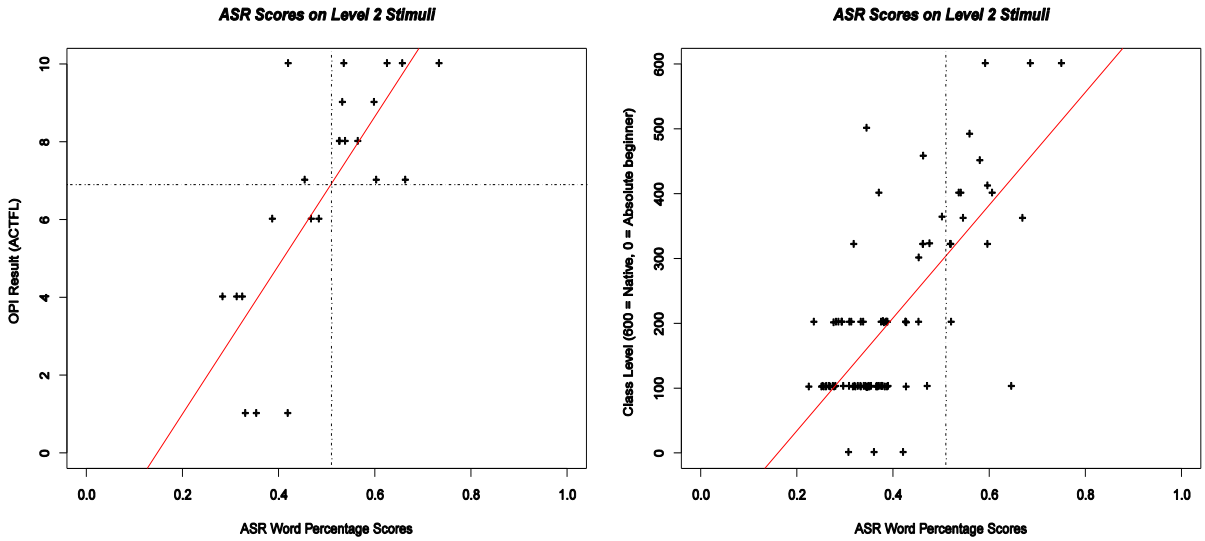


Figure 25: ASR Word Percentage Scores and Level 2 Analysis

Each ASR scoring methods appears to perform similarly to the human scoring methods shown in previous sections. Though the ASR scores are admittedly imperfect, they are able to accurately separate levels. Each of these methods, like the human rating methods, cut off almost all students in 100 or 200 level classes. This seems to fall in line with expectations as most students, even with 4 semesters of formal French, would not be expected to speak at an advanced level.

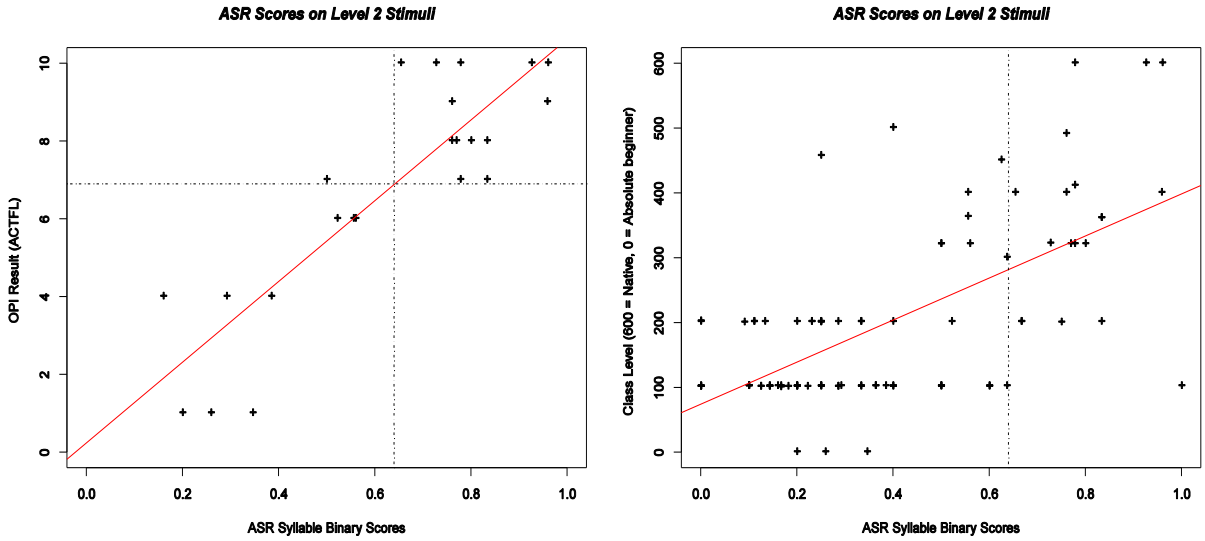


Figure 26: ASR Syllable Binary Scores and Level 2 Analysis

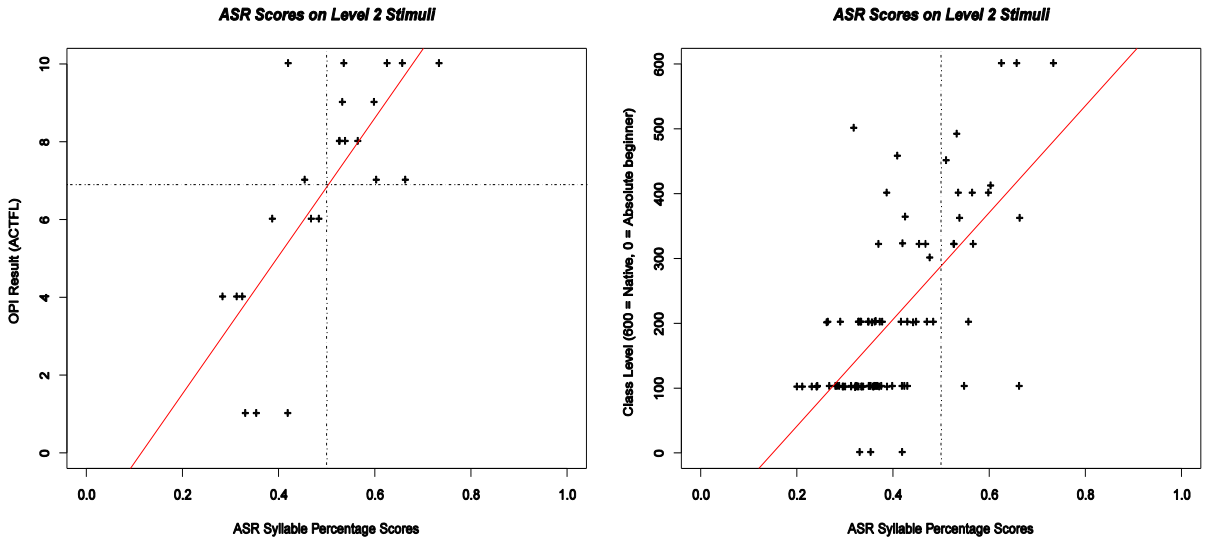


Figure 27: ASR Syllable Percentage Scores and Level 2 Analysis

One interesting result seen in the ASR graphics is a greater ability to distinguish between the intermediate and advanced speakers. The only real explanation for this phenomenon is that there was possibly some audio that was cut off and the ASR was either more generous to the advanced speakers or less generous to the intermediates. The only real way to know is to conduct more testing to verify the thresholds.

5.2.7 *ASR Scores for Level 3 Proficiency*

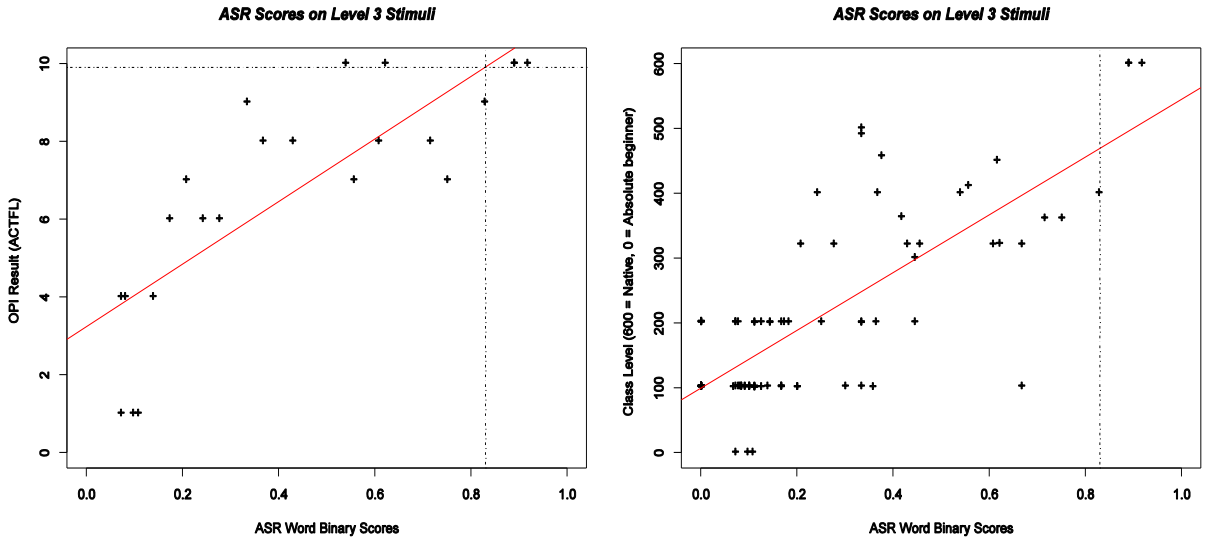


Figure 28: ASR Word Binary Scores and Level 3 Analysis

In the analysis of level 3 sentences, the most interesting result is the rather sharp distinction between native superiors (10) and non-native superiors (also 10). The two graphs in Figure 29: ASR Word Percentage Scores and Level 3 Analysis show this quite well as the three native speakers, listed as 600 in the graph on the right, are clearly separated from the non-native superiors. The two non-native superiors are grouped much more closely with the advanced speakers (7-9), even when evaluating level 3 sentences.

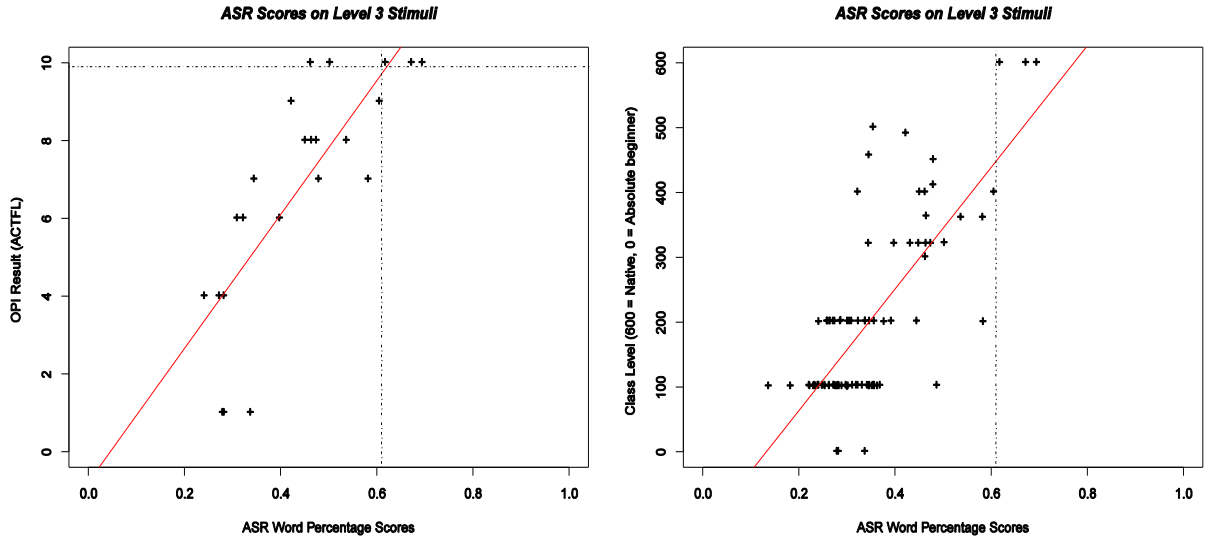


Figure 29: ASR Word Percentage Scores and Level 3 Analysis

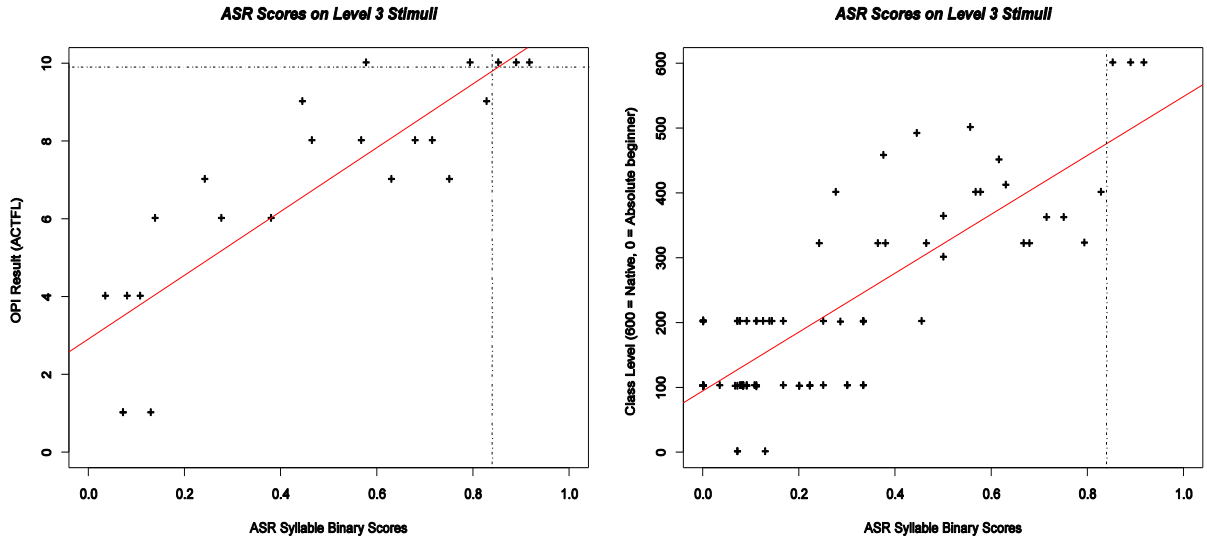


Figure 30: ASR Syllable Binary Scores and Level 3 Analysis

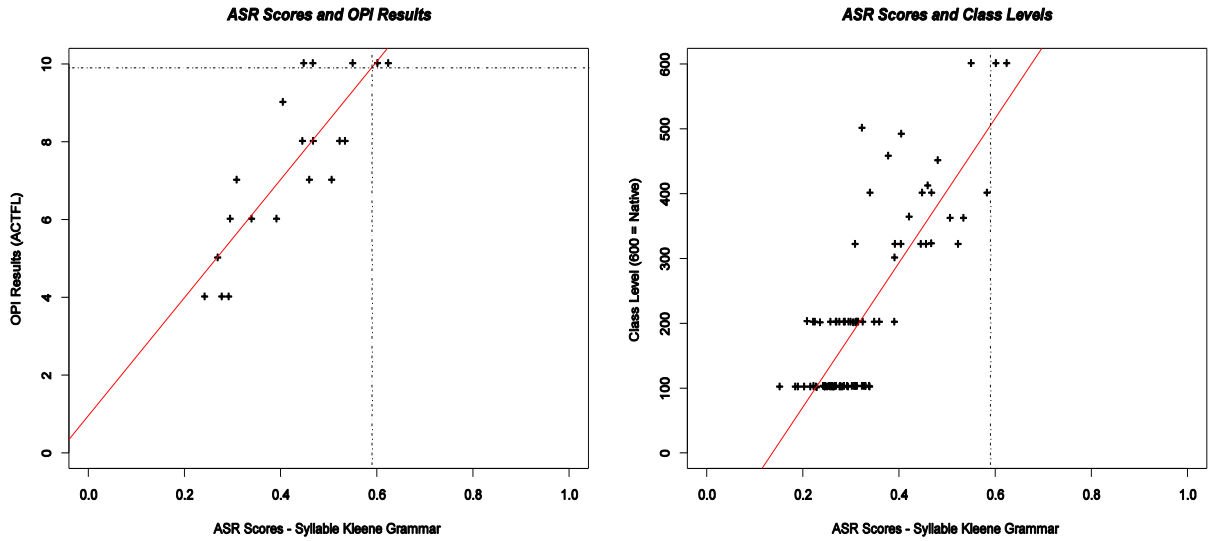


Figure 31: ASR Syllable Percentage Scores and Level 3 Analysis

Figure 31 is an attempt to show how the evaluation of the superior speakers would change if the native speakers were placed in a higher category – level 12 in this case. I use the syllable binary scores as it correlated best to the OPI scores. It appears to do a better job at separating the levels. There is only one advanced high speaker who falls in line with the superiors and one superior who falls below the threshold.

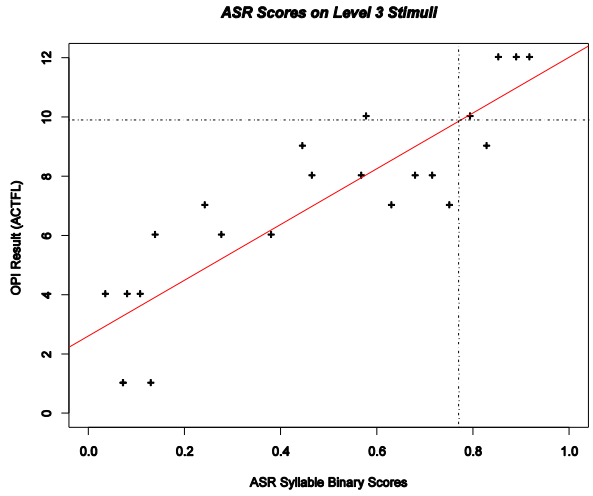


Figure 32: ASR Syllable Binary Scores with Natives Separate as Level 12

5.3 Results Analysis

Several points can be drawn from these results. First, the ASR percentage scoring is too generous at the bottom of the spectrum and too inaccurate for superior or native speakers. The

main problem with ASR percentage scores is derived from the flexible nature of the grammar. The Kleene grammar – which is the only way we are currently able to get to a percentage score – allows for zero or more of each word to be found in the prompt sentence. This flexibility appears to cause the recognizer to fail to recognize words or syllables that were repeated correctly.

The floor and ceiling effect also appear to render some of the results less useful. There may be a need for easier stimulus sentences to further separate intermediate and novice speakers. On the other hand, it seems that it may be necessary in future analyses to separate natives from non-native superior speakers. In almost all of the thresholds set for level 3 speakers (superiors), the non-native superiors consistently fell behind and would have been classed with their advanced speaker counterparts.

Third, the inclusion of criteria to the test development has greatly enhanced the discriminating ability of the test developed for this thesis. With a proficiency association to each sentence, I can now easily and accurately distinguish between the 4 major proficiency groups – novice (0), intermediate (1), advanced (2) and superior (3) – by setting thresholds between them. The accuracy for the syllable binary scores were: level 1 threshold – 91%, level 2 threshold – 96%, level 3 threshold – 91 %.

The results of this thesis have shown that EI testing is able to accurately estimate oral proficiency in French speakers, despite the need for further testing of thresholds, imperfections in the test itself and issues with recognition. This provides an opportunity for schools, businesses, or other organizations to quickly, accurately and frequently test people on their ability to speak French.

Chapter 6: Conclusions and Future Work

In the introduction, I stated that this thesis is to evaluate the validity and reliability of elicited imitation (EI) and automatic speech recognition (ASR) to estimate French oral proficiency. I did so by demonstrating that the processes established for English and Spanish EI testing are directly relevant to French testing. I also documented my specific contributions in the development, implementation and evaluation of an EI test to assess French oral proficiency. I detailed the incorporation of automatic speech recognition to score French EI items. Last, I substantiated with statistical analyses that carefully engineered and automatically scored French EI items correlate with a high degree to French OPI scores and are usable in automatic and adaptive EI tests.

The high correlations to the OPI, .918 for human ratings with the 4-score method, establish that the processes that have been successfully used for English and Spanish EI testing are directly relevant to French EI testing. One the main factors for the high correlations is the inclusion of naturally occurring sentences that have been annotated with the features described in the OPI guidelines. The resulting proficiency level associations help form a test that correlates with a higher degree to the OPI. This is primarily done by assigning score thresholds to each proficiency level. And though there is still a need for validation studies, preliminary results in this area are promising. With high correlations to the OPI - .883 – and an overall threshold accuracy of 92.7% using the syllable binary scores, I assert that this test can be used in conjunction with automatic speech recognition to quickly and accurately estimate proficiency of a speaker into one of the four main categories.

The feasibility of incorporating this into a fully automated test is certain. I show in this thesis that the ASR is currently accurate enough to be placed in a testing environment where it could automatically score speaker level immediately upon completion of their test. Even better, yet not fully explored, is the option of simultaneous ASR scoring while the speaker is taking the test.

Last, the use of thresholds answers the question of feasibility for an adaptive test. Assuming simultaneous scoring is implemented in the near future, adaptive testing for French EI tests becomes quite simplified with the methods and findings established in this thesis. As each

stimulus sentence is associated to a proficiency level, a simple percentage correct tally could be maintained and used to move the speaker through the test and stop them at the appropriate level.

6.1 Future Work

Despite strenuous efforts to make the best test, improvements made in certain areas may lead to better results. One of those areas is in the web-delivery the test. One of the findings from this thesis is that microphone placement can play a surprisingly important role in recognition rates. I showed in the test administration section that I needed to clean up some of the audio files using a Praat script because of microphone breathing. To solve this, it may be beneficial to include a "next" button to allow the participant to move forward once the response has been repeated and they are ready to move forward. Doing this alone would not only decrease the amount of time it took to take the test, but would reduce excessive noises and background noise that disrupts speech recognition. Another possibility is to include a screen that shows the proper microphone placement or have the test proctor ensure that all microphones are correctly placed.

For future testing in French, the findings of this thesis suggest that a syllable binary grammar be used during French EI testing. I found that this grammar rendered the best results in every area of analysis. I also recommend using the parameter settings for the ASR linguist in Sphinx 4 listed in the ASR Configuration section.

In this thesis, I used an out-of-the-box language and acoustic models built on native speaker data. Though these models work fairly well, I would expect recognition rates on learners to increase with the development of learner language and acoustic models. The more interesting of the two is language model development. I mentioned work in this area at the end of section 2.6 of the review of literature by Matushita (2011). I would like to follow a similar path for French by developing possible incorrect responses to increase the recognition. Doing so may significantly increase the time it takes to recognize a test, but it would greatly increase scoring accuracy since it is increasing recognition accuracy.

Work in the area of using fluency measures as a part of oral proficiency evaluation has increased greatly in the last 3 years (Müller et al. 2010). As much of this work uses similar methods to do so, it can be combined with accuracy scores from the EI test to provide a more accurate and fine-grained measure of the participant proficiency.

One of the most promising areas for EI testing is its ability to provide an automatic score at the end of the test. This step in the process will provide something that has not been provided before in oral proficiency testing – instant results that are valid, reliable and accurate. As the ASR improves and the test becomes more and more accurate in its ability to correctly estimate proficiency level to a finer level, this method will be very beneficial. Another possibility that could accompany automatic scoring is automatic feedback. With linguistic features annotated for every sentence, the ability of the test to provide feedback upon completion becomes feasible. Incorrectly reconstructed sentences could be quickly analyzed for the features that were missed and then generalized. This would allow the computerized test to provide generalized feedback to participants as to areas where they may be able to improve. It might say something like "you may want to look more closely at the subjunctive" when they incorrectly reconstruct the syllables of a subjunctive on several occasions or of multiple sentences that have the subjunctive.

With the inclusion of criterion-referenced analysis, the sentence level data gathered in the Results in View of Criterion-referenced Analysis section of this thesis can be used in adaptive testing. Information about how consistent a participant should be at each level, especially consistency thresholds, can be utilized in a number of different adaptive algorithms. Other information such as years of French taken could also be included in the algorithm that determines where to start participants in the test. This would reduce test time and examinee fatigue.

References

Bernstein, Jared, Alistair Van Moere & Jian Cheng. 2010. Validating automated speaking tests. Language Testing 27.355.

Bley-Vroman, Robert & Craig Chaudron. 1994. Elicited imitation as a measure of second-language competence. Research methodology in second-language acquisition, ed. by E.E. Tarone, S. Gass & A.D. Cohen, 245-61. Northvale N.J.: L. Erlbaum.

Brown, James Dean & Thom Hudson. 2002. Criterion-referenced language testing: Cambridge Univ Press.

Carrow, Elizabeth. 1974. A test using elicited imitations in assessing grammatical structure in children. Journal of Speech and Hearing Disorders 39.437.

Chaudron, Craig, Matthew Prior & U. Kozok. 2005. Elicited Imitation as an Oral Proficiency Measure. Paper presented to the 14th World Congress of Applied Linguistics, Madison, Wisconsin, 2005.

Christensen, Carl, Ross Hendrickson & Deryle Lonsdale. 2010. Principled Construction of Elicited Imitation Tests. Paper presented to the Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC), 2010.

Connell, Phil J. & Catherine Myles-Zitzer. 1982. An analysis of elicited imitation as a language evaluation procedure. Journal of Speech and Hearing Disorders 47.390.

De Calmès, Martine & Guy Pérennou. 1998. Bdlex: a lexicon for spoken and written french. Paper presented to the Proceedings of 1st International Conference on Langage Resources & Evaluation, 1998.

Deléglise, Paul, Yannick Estève, Sylvain Meignier & Teva Merlin. 2009. Improvements to the LIUM French ASR system based on CMU Sphinx: what helps to significantly reduce the word error rate?

Erlam, Rosemary. 2006. Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. Applied Linguistics 27.464.

Erlam, Rosemary. 2009. The Elicited Imitation Test as a Measure of Implicit Knowledge. Implicit and explicit knowledge in second language learning, testing and teaching, ed. by R. Ellis, 31. Bristol UK ;Buffalo N.Y.: Multilingual Matters.

Gallimore, Ronald & Roland G. Tharp. 1981. The interpretation of elicited sentence imitation in a standardized context. Language Learning 31.369.

Graham, C. R., Troy Cox, Jeremiah McGhee & Matthew LeGare. forthcoming. Examining the validity of an elicited imitation instrument to test oral language in Spanish. Paper presented to the Language Testing Research Colloquium (LTRC), University of Michigan, forthcoming.

Graham, C. Ray, Deryle Lonsdale, Casey Kennington, Aaron Johnson & Jeremiah McGhee. 2008. Elicited Imitation as an Oral Proficiency Measure with ASR Scoring. Paper presented at the 6th International Language Resources and Evaluation Conference (LREC), Marakech, Morocco.

Graham, C. Ray, Jeremiah McGhee & Benjamin Millard. 2010. The Role of Lexical Choice in Elicited Imitation Item Difficulty. Paper presented at the 2008 Second Language Research Forum: Exploring SLA Perspectives, Positions, and Practices, Somerville, MA.

Grimes, Joseph E. 1992. Calibrating sentence repetition tests. Windows on Bilingualism, ed. by E. Casad, 73. Dallas: Summer Institute of Linguistics and the University of Texas at Arlington.

Hamayan, Else, Joel Saegert & Paul Larudee. 1977. Elicited Imitation in Second Language Learners. Language and speech 20.86.

Hood, Lois & Patsy Lightbrown. 1978. What children do when asked to "say what I say": does elicited imitation measure linguistic knowledge. Reprints from Allied Health and Behavioral Sciences 1.195.

Lamere, Paul, Philip Kwok, Evandro B. Gouvêa, Bhiksha Raj, Rita Singh, William Walker & Peter Wolf. 2003. The CMU SPHINX-4 speech recognition system, 2003.

Liskin-Gasparro, Judith E. 1982. Educational Testing Service Oral Proficiency Testing Manual. Princeton, NJ: Educational Testing Service.

Lonsdale, Deryle, Dan Dewey, Jerry McGhee, Aaron Johnson & Ross Hendrickson. 2009. Methods of Scoring Elicited Imitation Items: An Empirical Study. Paper presented to the Paper presented at the annual conference of the American Association for Applied Linguistics, 2009.

Lonsdale, Deryle & Yvon LeBras. 2009. A Frequency Dictionary of French: Core Vocabulary for Learners New York, New York: Routledge.

Lowe, Pardee. 1982. ILR Handbook on Oral Interview Testing.8-13.

Luoma, Sari. 2004. Assessing Speaking Cambridge, UK: Cambridge University Press.

Malone, Margaret. 2007. Oral Proficiency Assessment: The Use of Technology in Test Development and Rater Training. In *CALdigest*.

Matsushita, Hitokazu, Deryle Lonsdale & Dan Dewey. 2010. Japanese Elicited Imitation: Asr-Based Oral Proficiency Test and Optimal Item Creation.

Matsushita, Hitokazu, Deryle Lonsdale & Dan Dewey. 2011. Language Modeling to Improve Elicited Imitation and Japanese ASR. Paper presented at the American Association for Applied Linguistics.

Menyuk, Paula. 1964. Comparison of grammar of children with functionally deviant and normal speech. Journal of speech and hearing research 7.109.

Millard, Benjamin & Deryle Lonsdale. 2011. Developing French Sentences for Use in French Oral Proficiency Testing. Paper presented at the Linguistic Symposium on Romance Linguistics, University of Ottawa, Canada.

Müller, Pieter F., Thomas R. Niesler & Febe de Wet. 2010. Automatic Oral Proficiency Assessment of Second Language Speakers of South African English.

Naiman, Neil. 1974. The use of elicited imitation in second language acquisition research. Working Papers on Bilingualism 2.1.

New, Boris, Christophe Pallier, Marc Brysbaert & Ludovic Ferrand. 2004. Lexique 2: A new French lexical database. Behavior Research Methods, Instruments, & Computers 36.516.

Okura, Eve. 2011. The Effects of Working Memory on Elicited Imitation Assessments of Second Language Oral Proficiency. Provo: Brigham Young University. MA Thesis.

Ortega, Lourdes. 2000. Understanding syntactic complexity: The measurement of change in the syntax of instructed L2 Spanish learners. Manoa, Honolulu: University of Hawai'i. PhD Dissertation.

Radloff, Carla F. 1991. Sentence repetition testing for studies of community bilingualism. Dallas Summer Institute of Linguistics and the University of Texas at Arlington: Summer Institute of Linguistics and the University of Texas at Arlington Publications in Linguistics.

Radloff, Carla F. 1992. Sentence repetition testing for studies of community bilingualism: An introduction. Notes on Linguistics 56.19.

Slobin, Dan I. & Charles A. Welsh. 1971. Elicited Imitation as a Research Tool in Developmental Psycholinguistic. Language training in early childhood education, ed. by C.B. Lavatelli, 170. Urbana, IL: University of Illinois Press.

Spitze, Kimett & Susan D. Fischer. 1981. Short-term memory as a test of language proficiency. ESL Talk 12.32-41.

Tomita, Yasuyo, Wataru Suzuki & Lorena Jessop. 2009. Elicited Imitation: Toward Valid Procedures to Measure Implicit Second Language Grammatical Knowledge. TESOL Quarterly: A Journal for Teachers of English to Speakers of Other Languages and of Standard English as a Second Dialect 43.6.

Zechner, Klaus, Derrick Higgins, Xiaoming Xi & David M. Williamson. 2009. Automatic scoring of non-native spontaneous speech in tests of spoken English. Speech Commun. 51.883-95.

Appendix 1: OPI Criteria Used in Test Development (Lowe 1982)

Levels for this Study	Database FEATURES	DEFINITION
NOVICE (0)		No real ability to form coherent sentences other than a few memorized phrases. Inconsistency at level 1
INTERMEDIATE (1)	VER_PRES_REG PAST_FREQ FUT_FREQ PAST_NEG POSS_ADJ SIMP_IL-Y-A PRO_SIMP french1k	Regular verbs in present tense Composite past or imperfect in top 1k Future tense verb in top 1k Negation of past tense Possessive adjectives 'Il y a' – existential (there is) Simple pronouns % of vocabulary in top 1k French lemmas
ADVANCED (2)	VER_PRES_IRR IMPF COMPLEX_PAST FUT_IRR IMPT REFX SUBJ_FAL_VOUL CMPV SPRV INT_PROs IDIO_DEPUIS IDIO_ILYA OBJ_PRO RelPRO	Irregular verbs in present Imparfait – imperfect Past tenses other than passé composé & imp Irregular future verbs Imperatives Reflexives Subjunctive w/ falloir / vouloir Comparative Superlative Interrogative pronouns Idiomatic 'depuis' + time Idiomatic 'il y a' + time Pronouns 'y + en' Relative pronouns
SUPERIOR (3)	IDIO_ALL PAST_EVENT FUT_EVENT PRO_ALL INV_ALL SUBJ_ALL PLUPFCT REFX_PASS PAST_FUT MANQUER RESTER NEG_COMP NIMPORTE	Any idiomatic phrase Any past tense usage Any future tense usage Any pronoun usage Any inversion usages Any subjunctive usage Pluperfect Reflexive as passive voice Past future All uses of 'manquer' All uses of 'rester' Any use of complex negation Usage of 'n'import . . .'

Appendix 2: Consent Form

Consent to be a Research Subject

This research study is being conducted by Dr. C. Ray Graham at Brigham Young University to aid in the development of an oral language test that can be administered quickly and yet will reflect accurately the oral language ability of the speaker. As a participant you will be asked to listen to sentences of varying lengths and complexity and to repeat them exactly as you heard them. The task will last approximately thirty minutes. You may also be asked to take a thirty-minute oral proficiency interview (OPI) or some other oral proficiency test. The test will be given in JFSB B161 or one of the humanities computer labs.

There are minimal risks for participation in this study. However, you may feel some stress in trying to listen to and repeat sentences which are beyond your current ability. Just do the best you can and do not worry about making mistakes.

There are no direct benefits to you. However, it is hoped that through your participation test developers will be able to provide an efficient way of assessing the speaking ability of language learners who wish to certify their Speaking Ability.

All information provided will remain confidential and will only be reported as group data with no identifying information. All data will be kept in a secure computer file and only those directly involved with the research will have access to them. After the research is completed, files will be destroyed.

Participation in this research study is voluntary, except in cases where your institution is using it as a part of their placement procedure. You have the right to withdraw at anytime or refuse to participate entirely without jeopardy to your class status, grade or standing with the university. If you have questions regarding this study, you may contact Dr. C. Ray Graham at 422-2208, ray_graham@ byu.edu. If you have any questions regarding your rights as a research participant, and you do not feel comfortable asking the researcher, please contact the BYU IRB Administrator, A-285 ASB; 801-422-1461; irb@byu.edu.

I have read, understood, and received a copy of the above consent and desire of my own free will to participate in this study.

Signature:_____ Date_____