Automated Scoring for Elicited Imitation Tests

Tim Ashwell

Scoring oral tests of second language ability can be a time-consuming and labour-intensive task. Trained assessors can rate learners’ oral proficiency on the spot, but this is an expensive solution to the problem and takes time to administer to a large number of learners. With the advent of more advanced and reliable voice recognition software, however, using computers to do the scoring has now become a possibility. The focus in this paper is on the first stage of a research project which eventually aims to automatically score an oral test of English second language proficiency. This first stage involves building up a database of audio files of Japanese university students recording English elicited imitation test items and performing on elicited imitation tests. This database will then be used in the next stage of the research to see if an open source automatic speech recognition (ASR) tool can be used to reliably score students’ test performance.

This paper reports on the first stage of a research project which ultimately aims to automatically score performance on an Elicited Imitation (EI) test. In a standard EI test, the second language (L2) learner hears and then imitates sentences of varying difficulty. As the length or complexity of items increases, learners find it more difficult to imitate and the kinds of errors they make reveal the characteristics of their grammatical ability. EI tests have attracted attention in recent years as ways of assessing L2 learners’ productive grammatical ability, but the time it takes to score such tests manually severely limits the applications to which they can be put. If automatic scoring becomes possible, immediate feedback becomes a reality, and the usefulness of the tests greatly increases.

The immediate aim of the first stage in this research project is to build a database of audio files of Japanese university students recording English elicited imitation test items and performing on elicited imitation tests. This database will then be used in the next stage of the research to see if an open source automatic speech recognition (ASR) tool can be used to reliably score students’ test performance.

Elicited Imitation Tests

The standard type of elicited imitation test is a deceptively simple procedure for gathering data about an L2 learner’s grammatical ability. Bley-Vroman and Chaudron (1994) describe this procedure as one in which subjects hear an aural stimulus and then imitate what they have heard, recording their responses onto tape or computer. The stimuli are often sentences of varying length. Sometimes ungrammatical as well as grammatical sentences are employed, usually there is no time pressure imposed, and usually no context is provided in the form of pictures or thematically linked items. Test items are carefully constructed to contain target grammatical features and to impose a range of processing loads so that the learner’s mastery of the features can be assessed.

Bley-Vroman and Chaudron (1994) argue that EI tests are reconstructive in nature. The imitation is based on the representation the subject forms of the stimulus on various levels from the lowest phonemic level to the highest interpretive level. L2 learners whose processing of a grammatical form is still not fully automatized on one level of the representation may struggle to imitate accurately because ac-
curate reproduction may depend on a complete and automatized representation on that level. Thus, if the task requires that the learner forms a representation on the interpretive level, but accurate imitation requires a complete representation on the phrasal level, accuracy will be affected if the representation on the phrasal level has not been fully automatized. The main underlying assumption is that memory limitations in processing require that information is chunked into larger units higher up the representational hierarchy in order to facilitate automatic (and accurate) (re)production. When learners are pushed to process at above the lexical level due to the length or complexity of the stimulus, the degree to which these target forms have been automatized through chunking information can be assessed.

Ellis (2005) and Erlam (2006) have argued that the standard EI test procedure is likely to reveal the learner’s explicit knowledge of a grammatical feature and have shown how the standard format can be adapted to create a test which measures implicit knowledge of a grammatical feature. This is achieved by requiring the subject to process the item for meaning at the highest interpretive level before imitating the item. The subject is thus dissuaded from processing the item for form and producing an imitation in an echoic fashion based on a purely lexical representation. This effect can further be increased by placing time pressure on subjects to imitate the item within a limited amount of time and by creating items which are thematically linked. The subject is therefore strongly encouraged to process the item for meaning before producing an imitation. The assumption is that an imitation so produced will draw more on the subject’s implicit knowledge of the grammatical feature than on their explicit knowledge. Implicit knowledge is the fully proceduralized knowledge that represents full mastery of a feature and enables fluent production even under pressured conditions. Explicit knowledge can be drawn on for accurate production if time is available. This is known as monitored production. Information about both types of knowledge is desirable to form a full picture of the state of a learner’s grammatical ability regarding a particular feature.

The Strengths and Weaknesses of EI Tests

Elicited imitation tests are attractive in several ways. Firstly, they require oral production and therefore give some way of measuring spoken, as opposed to written, grammatical ability. Secondly, they allow specific target forms to be elicited. In assessing spoken grammatical ability it can often be difficult to obtain performance on specific features because the task is too open-ended. EI tests allow specific features to be targeted. Thirdly, an EI test can be administered to multiple learners at the same time. Fourthly, it does not require a highly-trained oral interviewer to conduct the test. Finally, as described above, EI tests can be manipulated to provide tests of implicit and explicit knowledge.

EI tests also have their weaknesses. Some learners may have a more developed echoic memory and may be able to repeat items very accurately without processing them at a deeper level. Secondly, EI tests do not assess the learner’s ability to produce longer turns or to perform other aspects of their spoken repertoire. Thirdly, even in the limited sense of measuring spoken grammatical accuracy, one major practical problem with EI tests is that they are extremely labour intensive to score manually. Human scoring of one individual takes nearly as long as it does to sit the test, so scoring of classes of learners is impossible to achieve in a timely fashion. Sometimes this may not be such a problem when, for example, the results are to be used in a research project, but for use as diagnostic tests in an ongoing language program, for example, or as pedagogical tools in themselves, this can be a major drawback.

Research into Automatic Scoring of EI Tests

The desire to dramatically speed up the scoring of EI tests has spawned a number of research projects around the world in recent years (e.g. Christensen et al., 2010; de Wet et al., 2011; Lonsdale and Christensen, 2011). Graham (Graham et al., 2008) reports on the development of a computer-based EI test and subsequent attempts to apply automatic speech recognition (ASR) to the scoring of test performance by L2 learners from a variety of language back-
grounds. The researchers point out several reasons why one might be sceptical about the chances of success in using ASR. Firstly, ASR is still an emerging technology. Secondly, automating scoring with ASR requires integrating already complex systems. Finally, the speakers in these EI test are non-native speakers sometimes with heavy accents which may not be recognized by ASR designed for the recognition of native speakers.

Graham et al (2008), however, point out several reasons for optimism. One is that the expected input for any given test sentence is already known, so the ASR task is much more constrained. ASR systems are often designed to deal with unpredictable input of any kind and therefore have to be very powerful. With EI tests, the input for each test item is already known, so the features the system needs to look for and match are much more tightly defined.

Through successive iterations, Graham et al (2008) were finally able to score non-native EI data with good correlations with human scoring using the Sphinx open source ASR tool.

**Stage 1 of the present study**

Encouraged by the apparent success of these researchers, this research project aims to assess whether automatic scoring of two types of EI test is possible using open source speech recognition software. One type of test is aimed at measuring implicit knowledge of target grammatical features and one is aimed at measuring explicit knowledge. Two versions of each test type were constructed, making a total of four tests.

The first stage of this project involves the following tasks:

- confirming grammatical structures and vocabulary range for university entrants
- principled construction of EI test items taking into consideration length (number of syllables), structural complexity and lexical frequency
- preparing visuals to accompany test items
- recruiting student subjects
- arranging and conducting recording sessions
- cataloguing and storing audio files

**Item Design**

As the tests are aimed at measuring the productive grammatical ability of first year Japanese university students, a list of the grammatical structures these students should know was made by referring directly to school textbooks and to the Ministry of Education, Culture, Sports, Science and Technology English syllabuses. From this list, 13 structures and 4 tense forms were selected as targets. Test items were then written to include the structures using a vocabulary based on a list of common vocabulary contained in school textbooks (Shiomi, 2002). The 3030 words on the Shiomi list were first classified by part of speech and then the most frequent items in each category were selected for use in items. The vocabulary pool used to construct items amounted to 443 words. The first 40 items formed the template for all four tests and contained items ranging between 4 and 16 syllables in length. A total of 160 items were thus created. All items were question forms.
Test Design

Two types of test were created. One was a Timed Elicited Imitation Test and the other an Untimed Elicited Imitation Test. In the Timed EI Test, learners listen to an item through headphones while looking at a slide on a computer. They then have to answer the item (question) using information in the slide and then repeat the question they heard. This must be done before the next slide appears. The learners are thus under time pressure and are being asked to focus on meaning by having to answer the question before repeating it. This type of test is believed to elicit performance based on implicit knowledge. In the Untimed EI Test, learners simply repeat the items they hear. They have control over the speed with which the items are presented and are thus under no time pressure. They can focus on form without the distraction of having to answer the question before repeating it. This type of test therefore allows learners to rely on explicit knowledge. Two interchangeable versions of each test type were made.

Each item on the tests needs to have an accompanying slide to provide the test taker with a context in which to understand the item and, in the case of the Timed EI Test, for the test taker to be able to answer the question. The one hundred and sixty original drawings required were made by a student assistant. He then incorporated these drawings into a video slideshow for each test. The time lapses between slides/items for the Timed EI Tests had to be carefully set and varied depending on the length of items. These videos were then loaded into the PC@LL system in a language lab so that students could watch the video while being able to record their voice onto the computer.

Recording sessions

Two hundred students participated in the recording sessions. Each student was only required to take part in one recording session which lasted between 50 minutes and an hour.

Test performance data is needed in order to assess how good ASR software is at scoring both accurate and inaccurate performance. Each participant was therefore asked to record responses to one of the tests. The participant was then asked to read the same test items from a separate slideshow. This data is needed in order to train the ASR software to recognize fully accurate learner production of the items.

Organizing the data

All student audio files containing test and reading data were catalogued and approximately 20% were then edited into smaller files by student assistants each containing an individual item. A start was also made on transcribing the performance so that there is a written record to use when training or testing the automatic speech recognition software. Test data still needs to be scored by hand, however, so that the ASR software’s ability to score performance accurately can be assessed.

The next stage

The next step is to see if an open source speech recognition program can be used to accurately score learner data. The data used to train the system needs to be scored by hand, but the biggest challenge at present is setting up and learning to use the Sphinx system. Much of the input requires programming in Perl and various components must be available for the system to function properly. As novice programmers, this is the biggest hurdle we face and we need to build a small network of people who can advise us on the process.

The first stage of the project is complete, but now the really hard work begins. Over the coming year, the hope is that we can learn how to use the Sphinx system and that we can begin to manipulate it so that we are able to input our own data. It is anticipated that it will be another two years before we are able to say clearly whether automatic scoring of our data seems feasible. The benefits that would be gained by achieving automatic scoring of these elicited imitation tests make it worth pursuing this long-term goal. For example, one direction the pro-
ject may take in future is in assessing the potential the EI procedure has as a pedagogical technique for helping learners proceduralize the grammatical knowledge they possess. This will only be possible when automatic scoring becomes a reality. Automatic scoring also opens up the possibility of making a test or teaching application which is adaptive, adjusting the level of items to the performance of the learner.

References


De Wet, F., Muller, P., Van der Walt, C., & Niesler, T. (2011). Readability index as a design criterion for elicited imitation tasks in automatic oral proficiency assessment. ISCA.


