DEUCE: A TEST-BED FOR EVALUATING ESL COMPETENCE CRITERIA

George R S Weir
Dept of Computer and Information Sciences
University of Strathclyde
Richmond Street
Glasgow, G1 1XH
UK
Email: gw@cis.strath.ac.uk

Toshiaki Ozasa
Graduate School of Education
Hiroshima University
1-1-1 Kagamiyama
Higashi-Hiroshima, 739-8524
JAPAN
Email: tozasa@hiroshima-u.ac.jp

Abstract

This paper describes work in progress to apply a Web-based facility for evaluating differing criteria for English language competence. The proposed system, Discriminated Evaluation of User's Competence with English (DEUCE), addresses the problem of determining the efficacy of individual criteria for competence in English as a Second Language (ESL). We describe the rationale, design and application of DEUCE and outline its potential as a discriminator for ESL competence criteria and as a basis for low cost mass ESL competence testing.

Key Words

ESL competence, Web and Internet Tools and Applications, Multimedia Tools and Architectures, Educational Multimedia

1. Introduction

English is often the common language of communication between speakers of different native tongues and continues to grow in importance with the expansion of international business, tourism and travel. Non-native speakers learn English as a second language (ESL) and there is an established tradition in TOEFL - the testing of English as a foreign language (cf. www.toefl.org).

For teachers of ESL there is a recurrent need to evaluate the English competence of their students. This may take several dimensions, including oral proficiency, comprehension and written abilities. For students, such procedures can be time consuming and stressful. Many techniques for testing the competence of second language English speakers grade ability according to the subject's grasp of specific language features. For instance, facility with grammar or range of vocabulary, provide means of rating a subject's active knowledge against more or less well-known levels of complexity.

Similar techniques are applied in commercial packages for determining language competence. Perhaps the best known standard test, The Test of English as a Foreign Language (TOEFL) measures the ability of non-native speakers of English to use and understand North American English as it is used in college and university settings. This TOEFL test is available both as a 'pen and paper' and a computer-based assessment through authorised test centres worldwide. The fee for this test is presently $110 per subject.

The 'pen and paper' user application CELT (a Comprehensive English Language Test for learners of English) is a similar system. CELT consists in three components of discrete-point sub-tests: listening (50 items), structure (75 items) and vocabulary (75 items). Each of these test components is presented with a time-limit. CELT assumes that the sum of the knowledge of these three discrete-points measures the general proficiency level of ESL learners [1].

1.1 Need for alternatives

There are several reasons why alternative means of assessing English language competence are desirable. In the first place, the availability of low cost, non-commercial techniques affords an economical basis for

testing. This is especially true if we can provide an on-line mass testing facility. In the second place, such alternative testing methods may be used as a basis for feeding back information on performance to the subjects themselves. By such means, learners may monitor their developing language skills and perhaps identify areas of particular weakness that merit more detailed attention and practice

Such a technique is applied in the grammatical error-detection system, ALEK [2]. 'ALEK detects two types of errors: those that violate basic principles of English syntax (e.g., agreement errors as in *a desks*) and those that show a lack of information about a specific word (e.g., treating a mass noun as a count noun in *a pollution*)' (p. i). This automated system was evaluated by its authors for effectiveness in providing performance-based measures of communicative competence.

In previous work, we have identified several other criteria that reflect language competence. For instance, Zhang [3, 4, 5] empirically researched the learning of the null subject parameter (pronouns) and the head parameter (word order), focussing on Chinese and Japanese learners of English. Hettiarchchige [6] also investigated the acquisition of countable and non-countable parameters by Sinhalese and Japanese learners of English. The criteria derived from these studies can plausibly be applied as measures of ESL learners' developing communicative competence.

Other applicable criteria may be based upon comprehension skills, compositional abilities, or summary generation. In related work, our application of word use analysis allowed comparisons to be made between language learning texts from different eras of ELT in Japan. Coupled to word counts and structural content, we have also employed readability metrics as means of discriminating texts of differing complexity.

For example, Ozasa, et al. [7, 8], made quantitative comparisons of three major historically significant English language teaching texts (*Sander's Union Readers* (1861-3, a set of English textbooks published in U.S.A.), *Jack and Betty* and *New High School English*, a set of middle grade school English textbooks used in the 1950s in Japan, and *Sunshine English Course*, a set of English textbooks currently used in junior and senior high school in Japan).

The comparisons used in the analysis were word frequency count, i.e., token (the number of all the running words), type (different words), type/token ratio (the ratio between types and tokens), the frequency count of passive sentences, Flesch Reading Ease, and Flesch-Kincaid Grade Level.

Such techniques afford additional criteria that may also serve as means of evaluating ESL competence. While there are many potential tests for language competence, there are few clues to the best means of applying these criteria in order to ascertain broadly accurate indications of language competence. Differing criteria test different aspects of language proficiency hence systems such as CELT combine a variety of tests in order to ascertain an overall proficiency measure.

Inevitably, there is a degree of heuristic judgment applied in determining appropriate combinations of criteria. Practical constraints limit the extent of testing that is tolerable for the student so there is considerable mileage in being able to select a minimal set of competence criteria that are known to be incisive and informative in combination.

2. Evaluating Criteria and Implementations

Our focus in the present work is to provide a test-bed facility that will allow the evaluation of alternative criteria for ESL competence. The DEUCE (Discriminated Evaluation of User's Competence with English) system addresses the problem of determining the efficacy of individual criteria for competence in English as a Second Language. This software-based system will be used experimentally to implement and test a variety of competence criteria both individually and in terms of their potential effectiveness in combination. Our longer-term goal is to establish a reliable and low-cost set of ESL competence tests that have proven effectiveness. There is significant potential in developing a small and concise test set that may reflect a broader measure of ESL competence. Only comparative testing of alternative competence criteria can determine such effectiveness, and this is a primary purpose for DEUCE.

2.1. Dual evaluation mode

There are two aspects to consider in the evaluation of any particular criterion for English competence. Assessment of the criterion cannot be achieved without a suitably implemented user test. Inevitably, assessment of the criterion is influenced by the effectiveness of the selected implementation. For this reason, the DEUCE system is designed to facilitate two levels of discrimination. The first level gauges the applicability of individual (or combined) criteria. The second level addresses the evaluation of specific test implementations. In each case, our comparative measure relies upon existing accepted standards for ESL competence assessment.

In operation, we 'load' DEUCE with one or more implementations of a single criterion and run a series of tests on student subjects. The results provide a metric of student performance and directly reflect the effectiveness of both the criterion (as a measure of general ESL competence) and the specific implementations (as instances of the criterion). By comparison with

independent ESL competence measures, we can gauge the efficacy of the criterion/implementation pair as a general indicator of ESL competence.

By determining each student's ESL competence in advance, by means of an independent system (e.g., CELT), we are able to provide DEUCE with a rating for each subject's competence. Against these measures, DEUCE engages with students, applies the tests, and determines its own measure for each student performance. Thereafter, DEUCE evaluates the criteria and the implementations against the predetermined rating for each student. In this fashion, DEUCE is able to discriminate across the criteria and their individual implementations. The three principal domain-specific ingredients that contribute to DEUCE operation are represented in Figure 1, below.

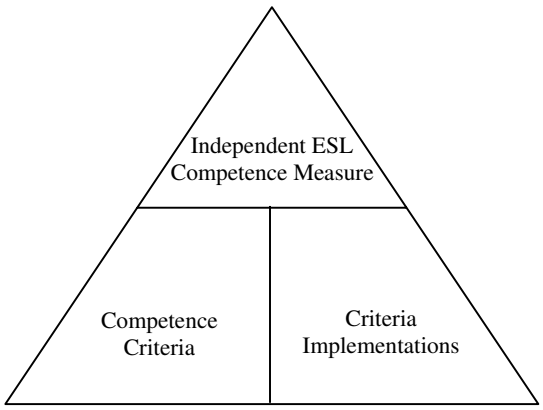

Figure 1: DEUCE components

A significant component in DEUCE operation is its management of individual criteria and implementations. In virtue of its dual-level discrimination (criteria and implementations) and the fact that criteria can be assessed in combination, DEUCE must manage a combinatorial set of factors. For instance, when 'loaded' with only three criteria, each with three implementations, there are nine first order tests (combinations of different implementations applied to single criteria), two combinations of two criteria, each with nine combinations of implementations, and one combination of three criteria with twenty-seven combinations of implementations, yielding a total set of fifty-four test cases.

In such a scenario, DEUCE might be configured to deliver each of these test cases to a statistically significant number of subjects before correlating the results against the predetermined ESL competence levels. We refer to this configuration as a 'three by three' array of tests.

An example may clarify the relationship between criteria and their implementations. For instance, we may be evaluating a criterion based upon the subject's grasp of past tense grammatical forms (C1). In testing the applicability of this criterion as an indicator of ESL competence, we might employ two implementations. The first implementation displays a series of words and requires subjects to re-order the words in order to compose a grammatically correct sentence form.

One possible display is shown in Figure 2. The screen is presented to the subject who is required to manipulate the ordering of the words with their mouse. For this instance, the correct ordering would give the sentence 'It had proved necessary to book in advance'.

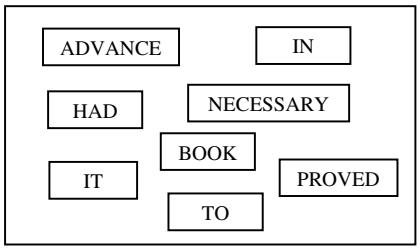

Figure 2: Sample first implementation of C1

Our second implementation of the same criterion may employ a different means of eliciting the subject's response; Figure 3 illustrates a screen that elicits the subject's selection from four alternative English expressions (requiring the user to place a 'tick' against the selected answer). This alternative implementation exhibits the same criterion (C1) as the implementation illustrated in Figure 2. As previously noted, the DEUCE approach to discriminating the effectiveness of ESL criteria also enables us to compare the use of alternative implementations for each criterion.

To book in advance had it necessary proved.	☐
Book in advance had it necessary to proved.	☐
It had proved necessary to book in advance.	☐
Had it necessary proved to book in advance.	☐

Figure 3: Second implementation of C1

There are several aspects of the experimental design that we have yet to determine. For instance, the use of different test cases on individual students has to be carefully managed to minimise any learning or other interference effects. (Thus, the alternative implementations illustrated in Figure 2 and 3 would not both be given to the same subjects.) The important point is that DEUCE yields the prospect of delivering such test cases in a distributed environment to multiple simultaneous subjects. In itself, this promises considerable time and effort savings in experimental operation and may also simplify the correlation process of comparing specific test scenarios against known standards.

3. Software Architecture

Our primary considerations in development of the software test-bed were the wish to accommodate multiple ESL tests simultaneously and the desire for platform independence. Our Web-based distributed solution accommodates these requirements and comprises a client-side delivery system in tandem with a server-side management facility.

The need to deliver diverse ESL competence tests in ways that are both effective and engaging, led us to select animation-based client-side delivery. This provides scope for highly interactive animation-based presentations that engage the subject. For example, the implementations cited above (Figures 2 & 3) require the subject to make selections between sentence forms or to reconfigure sentence components. Our aim is to employ engaging means of user interaction. In turn, the efficacy of these implementations, as enablers of competence criteria, is a component that we also seek to gauge through the DEUCE facility.

A further consideration is the need for certain criteria to determine student performance time on specific tests. In a Web context, this is best achieved through client-side processing. Flash provides this computational facility and also links readily to server-side PHP programs for ease of data transfer. Thus, the client-side tests are driven and managed by a PHP-based server-side facility (the DEUCE engine) that delivers the test information, records subject performance via an SQL database and discriminates between test cases on the basis of subject's performance.

In order to gauge the effectiveness of individual and combined competence criteria, individual criteria are prototyped, applied as student tests and their acuity evaluated against an independent ESL competence measure (CELT). Note that our software system serves a further dual role. Firstly, it acts as a delivery and test vehicle for the competence criteria and their implementations. Secondly, this system provides a means of delivering the validated tests in earnest. In the latter case, we have a fully-fledged facility for ESL competence testing.

The architecture for our evaluation test-bed is shown in Figure 4. This approach builds upon experience gained in developing the 'English Assistant' system [9], and is similar in architecture to our work on a prototype multi-lingual facility for the Greek Finance Ministry [10].

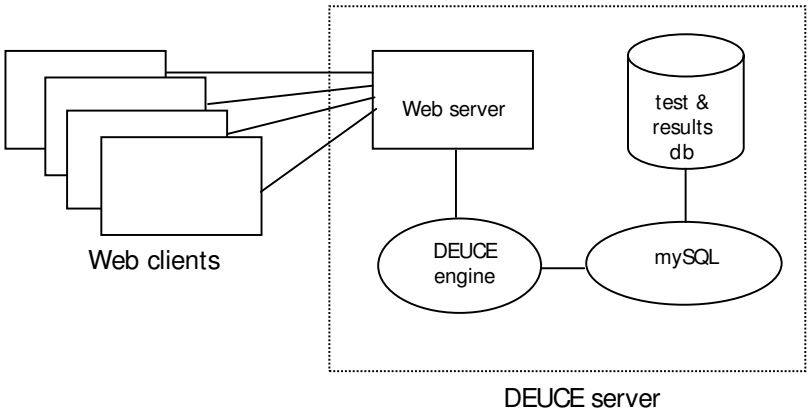

Figure 4: DEUCE software architecture

Subjects receive Flash™ animation-based interactive tests at their Web client. Each test is an implementation for a specific ESL competence criterion. These tests are delivered to the Web client via a Web server by the DEUCE engine. Performance results for each subject are returned to the DEUCE engine, correlated against the student's predetermined ESL competence level, and recorded in the system database.

The DEUCE management facility can automatically deliver and process sets of tests for any given combinations of criteria and their implementations. In operation, the DEUCE engine takes account of the number of criteria being tested and the number of implementations provided for each criterion. Further factors affecting the test delivery are the required number of subjects for each test condition and any constraints on the re-use of subjects for subsequent test conditions. For instance, we might wish to discriminate the effectiveness of two criteria for ESL competence, each having three test implementations. If we are untroubled at the prospect of subjects undertaking multiple test implementations for a single criterion, the DEUCE engine will deliver every test condition to each subject. If desired, an element of randomness can be introduced to affect the order in which subjects meet specific tests. Likewise, we may choose to randomly allocate test implementations across our subjects rather than assign every case to each subject. These are variables that are accommodated by configuration of the DEUCE engine.

4. Further Developments

Trials in which a set of ESL criteria are evaluated in isolation and in combination, are planned for the near future. These tests will also enable us to assess the potential of using DEUCE as an independent facility for mass testing of ESL competence. This system will be tested in two ESL contexts. The first is the use of English by Japanese students, while the second is a similar setting with Greek students of English. Our use of diverse populations of ESL learners aims to provide greater credence to the results and may also reveal population specific insights on competence criteria. For instance, we may determine that grammar-based criteria are stronger indicators of English competence for Greek learners than for Japanese.

A variety of prototype implementations for four sample criteria are under development. The four initial criteria provide two examples that gauge grammatical knowledge, one to determine grasp of compositional complexity and one to assess active vocabulary. The implementations for these criteria will provide three differing tests for each of the four criteria. Thereby, DEUCE will initially operate with a 'four by three' array of tests. Our two site trials will discriminate the effectiveness of our test array as compared to CELT measurements of ESL competence.

References

[1] D.P. Harris & L.A. Palmer, *A comprehensive English language test for learners of English, Form A and B* (New York: McGraw-Hill, 1986).

[2] C. Leacock & M. Chodorow, Automatic assessment of vocabulary usage without negative evidence, Educational Testing Service, Princeton, New Jersey, Research report #: RR-01-21, 2001.

[3] S. Zhang, The null subject parameter and the head parameter in adult SLA: A study of Japanese and Chinese learners of English, *International Journal of Curriculum Development and Practice 3*, 1, 2001, 1-13.

[4] S. Zhang, The learning of pronouns and word order: with a focus on Chinese and Japanese learners of English, *Bulletin of Graduate School of Education*, Part II, 50, Hiroshima University, 2001, 145-152.

[5] S. Zhang, A study on the learning of pronouns and word order: with a focus on Chinese and Japanese learners of English, Ph.D. Dissertation (Hiroshima University), 2003.

[6] R.C.K. Hettiarchchige, The acquisition of countable and non-countable structure by Sinhalese and Japanese learners of English, *International Journal of Curriculum Development and Practice, 5*(1), 2003 (in print).

[7] T. Ozasa et al, A qualitative analysis of New National Readers, The Globe Readers, The Standard Readers (in Japanese), *Historical Studies of English Teaching in Japan, 17*, 2002, 21-40.

[8] T. Ozasa et al, A quantitative measurement analysis of *Sander's Union English readers*, *English and English Teaching, 8,* 2003 (in print).

[9] G.R.S. Weir & G. Lepouras, English Assistant: A support strategy for on-line second language learning, *ICALT 2001, IEEE International Conference on Advanced Learning Technologies*, Madison, USA, 2001.

[10] G. Lepouras, C. Vassilakis, & G.R.S. Weir, Serving enhanced hypermedia information, in F. Crestani, M. Girolami & C. J. van Rijsbergen (Eds.), *Advances in Information Retrieval*, Springer, 2002, 86-92.