# 論文 / 著書情報
# Article / Book Information

| | |
|---|---|
| Title | DEVELOPMENT AND EVALUATION OF ADAPTIVE ENGLISH PRACTICING SYSTEM FROM ERROR-ANNOTATED LEARNER CORPUS |
| Author | Thisayakorn Pakaporn, AKINORI NISHIHARA |
| Journal/Book name | 7th International Technology, Education and Development Conference, , , |
| Issue date | 2013, |
| Citation | P. Thisayakorn, A. Nishihara (2013) DEVELOPMENT AND EVALUATION OF ADAPTIVE ENGLISH PRACTICING SYSTEM FROM ERROR-ANNOTATED LEARNER CORPUS, INTED2013 Proceedings, pp. 6397-6403. |

# DEVELOPMENT AND EVALUATION OF ADAPTIVE ENGLISH PRACTICING SYSTEM FROM ERROR-ANNOTATED LEARNER CORPUS

**P. Thisayakorn, A. Nishihara**

*Tokyo Institute of Technology (JAPAN)*
*noey@nh.cradle.titech.ac.jp, aki@cradle.titech.ac.jp*

## Abstract

This study proposes a computer-assisted language learning (CALL) system for increasing error awareness, aiming for correct English usage and improvement of English proficiency of learners. The system incorporates two principles: learning from negative evidences, and prioritizing communication errors. Negative examples, specifically erroneous sentences, are extracted from learner corpus, to create practice exercises. These exercises constitute an intelligent practicing system, by exposing adaptively to abilities and weaknesses of each user, determined by his responses while using the learning system. Among all type of errors in the exercises, ones that impaired understanding, so-called communication errors, gain priority for being studied. Unlike traditional language instructions, actual frequent errors made by learners in general are emphasized, while individual errors are corrected. Moreover, learners are directed to focus on their weaknesses and unintelligible errors, thus it is considered as a straightforward and effective method. To evaluate, experiments are conducted by target users, which are Japanese learners. After using prototypes, evaluation is done in aspects of adaptive mechanism, learning achievement, and satisfaction of learners. The overall result concludes that the system has more potential to improve English ability than traditional non-adaptive approaches. At the end, recommendations for future improvements are discussed.

Keywords: Intelligent Tutoring System, Educational Software, Computer Assisted Language Learning, Second Language Acquisition, English as a Foreign Language.

## 1 INTRODUCTION

Studying English as a foreign language has been compulsory for people in all over the world. Many people study English obligatorily for several years without being able to use it to convey meaning. In such a way, attaining English skills is still indispensible, however, tedious to start reviewing everything from the beginning. Thus, we propose an English learning system that is a straightforward approach to what a particular learner does not know, or to his (or her) weak points.

## 1.1 Computer Assisted Language Learning (CALL)

Computer-assisted system is a suitable approach for this task, because, unlike traditional paper-based approach, contents can be programmed to be adaptable to different learners. Even for a particular learner, it can adjust to the development of his knowledge. U.S. Department of Education indicates that computer-assisted system engages students to spend their time on lesson over face-to-face approaches [1], so it has a potential to be resolution for afore-mentioned tediousness problem as well. In language learning, this kind of adaptable learning system is often called Intelligent CALL (iCALL), which is currently one of the topics of the most interest.

### *1.1.1 Structure*

To build iCALL that can provide adaptive instructions, system often consists of 3 models: expert model, which stores knowledge that experts have or, in other senses, instructions to be taught, student model, which keeps tracks of student performance while studying, and tutor model, which selects instructions to teach from Expert model that is proper for performances stored in student model.

### *1.1.2 Student Model*

There are several ways to implement student model such as students' knowledge, interests, background, goals, or individual traits [2]. Even in knowledge-based modelling, student knowledge can be represented in several ways, such as by subset of all knowledge domains, called overlay model, set of pairs of user stereotype and value indicating if user is belonging to the stereotype, called stereotype model [3]. In this research, we use overlay model and define domain of knowledge by concepts or categories of errors of English.

## 1.2 Error Treatment

One of the means to decrease weak points is error treatment. Several studies suggest that learning from negative evidences plays important role in second language acquisition [4], and adult learners perform better when learning on conflict [5]. Thus, in addition to valid target language, learning from ill-formed patterns might be a good practice to improve proficiency of students.

### *1.2.1 Corrective Feedback*

Research on corrective feedback for error treatment has been done over time, and there are many controversial issues on its effectiveness and the best way to correct errors. The overall effect is neutral [6], and more investigations still go on. However, we believe that adopting error treatment in our system can give a bit of evidence to support or argue this issue.

### *1.2.2 Adaptive Feedback in iCALL*

Several researches incorporate iCALL in providing corrective feedbacks to fulfil gaps in classrooms where a teacher cannot concentrate and correct errors of every student. Heift [7] provided different types of feedbacks in web-based CALL and argued that providing meta-linguistic with highlighting feedbacks has effect on learners' attempt to correct their mistakes. Later, Gutierrez [8] also add selection model trained by teacher-student dialog interaction to automatically generate proper feedbacks for specific type of errors.

Our research incorporates error treatment in iCALL in a different way from giving feedbacks to students' errors. Instead, students learn from common errors of students with the same native language (in this case, Japanese), from which their individual weaknesses are treated.

# 2 PRACTICING SYSTEM

This section explains process of building the adaptive learning system. Starting from NICT Japanese Learner of English (JLE) Corpus [9], we build a program to extract errors and create practice exercises, from which we create our English learning system.

## 2.1 Practice Exercises

First, we programmatically select the parts that learners say. Then, context tags are processed or removed. Finally, sentences that contain error tags will be chosen to create exercises. Fig. 1 shows an example of an exercise stored in practicing system with resolution and its error category.

```
(0)They (1)enjoyed (2)the (3)dinner (4)very (5)well (6).
Erroneous part: 5-5
Correction: much
Error Category: Adverb Lexis
```

**Fig. 1 Example of an Error Correction Exercise.**

### *2.1.1 Type of Task*

There are 2 types of erroneous parts: single-worded and multiple worded errors, which are respectively transformed into error correction and error identification task. In error correction, a student picks one word which he thinks it is an erroneous part and corrects it. On the other hand, error identification is the task that a student only chooses any part of an error without correcting it, or answers "No error". Error correction task is further divided into 2 types: word error and space error, of which the erroneous part is a word in the sentence and space between words respectively.

### 2.1.2 Completeness and Number of Exercises

Because sentences are automatically chopped out from the context to create exercises, meaning of some sentences depends on previous or following sentences and become incomplete as an exercise. Besides, some categories generate too few number of exercises, thus is not enough for practicing. Therefore, we remove categories that contain a lot of incomplete sentences or too few exercises, remaining 22 from 43 categories with about 5 thousands exercises in total to use in the system.

## 2.2 Design of Practicing System

We design our system like a general adaptive learning system. Exercises in section 2.1 are generated beforehand and stored in the file system before being used as instructions in the system.

### 2.2.1 System Mechanism

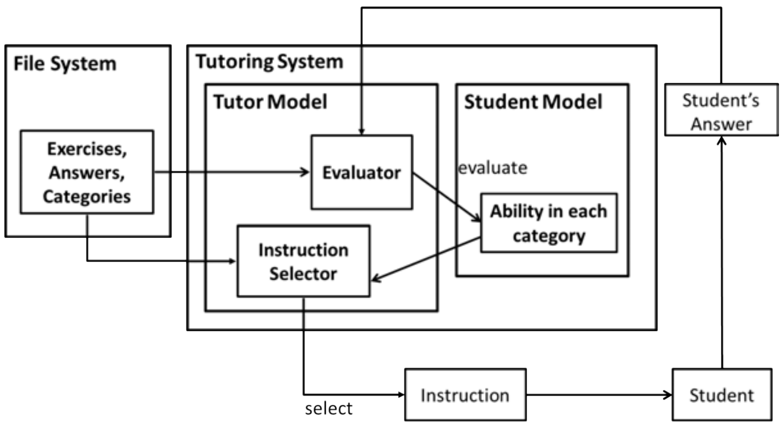


**Fig. 2 System Configuration.**

Our system is contracted from what is explained in section 1.1.1. As illustrated in Fig. 2, it composes of 2 models: Tutor model and Student model. Tutor model selects exercises randomly from file system to present to user. When the user answers, Tutor model evaluates the user's performances and stores them in Student model, which will in turn become a reference indicating student's ability for tutor model. After a specified period of time, Tutor model will start to adapt its instruction to present more of categories that the student is poor at.

### 2.2.2 Scoring

To model ability of a student, we save number of tests done and sum of scores in each category, and calculate using equation 1. Score of each category ranges from 1-100.

$$Ability\ of\ a\ Category\ C = \frac{\sum_{i=1}^{N(C)} S(C_i)}{N(C)} \quad (1)$$

Where $C$ is an error category

$S(C_i)$ is a score of exercise number i in category C.

$N(C)$ is total number of exercises done in category C

In error Identification task, the system gives 1 point for correct position of error answered, while in error correction task, the system gives 0.5 points for correct position, and full 1 point, for both correct position and the right correction.

### 2.2.3 Student Model

As stated in section 1.1.2, we adopt overlay model in student modelling. Our domain is 22 error categories we selected in section 2.1.2. Each student is defined by subset of categories of weakness. Weaknesses of a student are identified by 2 kinds of thresholds. Standard threshold indicates the lowest score to pass. Thus, any categories of which score lower than this point will be regarded as weaknesses. Another threshold is called advanced threshold, which indicates that if score of a category of weakness gets higher than this point, that category will be removed from the set of weaknesses.

### 2.2.4 Tutor Model

Tutor model selects exercises based on categories of weakness specified in overlay model of each student. However, there is a large difference of number of generated exercises in each category, which makes direct randomisation may result in one categories being selected more than others. Thus, we need to find other approaches to deal with this problem.

Izumi [10] examined intelligibility of error categories of the same NICT JLE Corpus, by having native speakers rate errors in the corpus how they think those errors affect correct senses of sentences. From this research, we assign unintelligibility indication of each category to each exercise item as a weight, and use weighted random selection, so that exercises to raise errors that impair communication will gain priority to be taught. Tutor model also memorises which exercises a student got correct or wrong answers, and exercises with wrong answers are repeated again until the correct answer is received.

## 2.3 Interface of Practicing System

Our system simply contains 3 pages: start page, where students choose their profile and start study from saved data, question page, where students work on exercises provided, and solution page, where answer for previous exercise is shown with corresponding correct sentence and error category. Another setting of system is created only to use for experiment in this research. We called it non-adaptive setting, where students choose categories to study themselves.

Mainly, users study on question and solution Page, which screen composition is shown in Fig. 3. Parts number 1 and 2 show status, such as current user, current test mode, and question number. Number 3 is shown only in question page, telling users how to answer current exercise according to types of task. Number 4 displays editable erroneous sentence in question page, and correct sentence in solution page. Number 5 is shown only in solution page, telling whether the user got right or wrong answer in current exercise. Number 6 and 7 are navigation buttons to proceed to next screen, and for non-adaptive setting, back button to go back to menu to choose other categories to study.

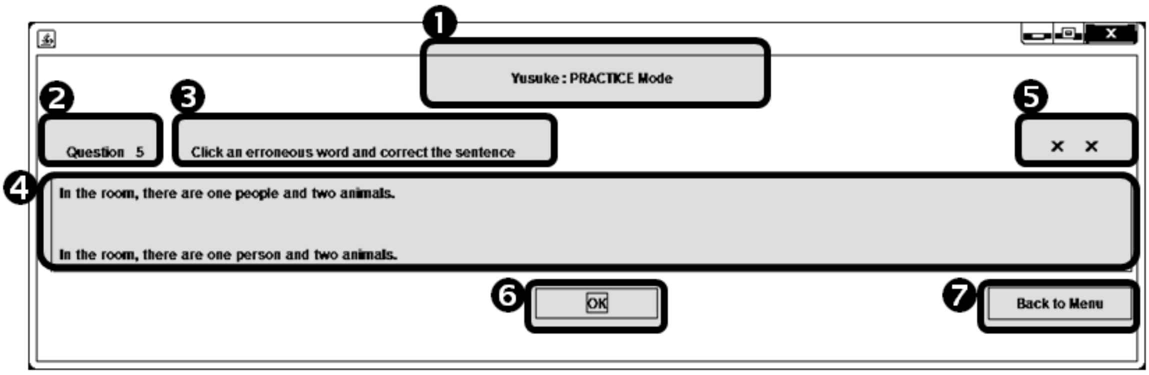


**Fig. 3 Screen Composition of Question Page and Solution Page.**

# 3 EXPERIMENT

We conduct a simple experiment to evaluate our practicing system. We have 18 Japanese students study using our system and collect data saved in student model for system evaluation.

## 3.1 Participant

All of participants are Japanese native speakers, and are studying in forth year of bachelor degree or master degree in Graduate School of Decision Science, Tokyo Institute of Technology. They are considered as intermediate to advanced students with the standard TOEIC score ranged from 530 to 915, and the average score is 732.94.

## 3.2 Participant Group

First, all students work on the system in test environment, which presents exercises one by one as in practice system explained in section 2.3, but no solution provided after each exercise. All students have to finish the same number of test from all categories in the same amount of time. From the test score, we divide students into 2 groups with similar average scores, and assign them as experimental group, studying by our adaptive system, and control group, studying with non-adaptive setting especially built for this experiment. In adaptive setting, students work on exercises automatically selected by the system, while in non-adaptive setting, students are informed to choose categories they think they are their weak points and study.

## 3.3 Experiment Flow

After group assignment, students in each group study by our system, working on exercises only in the categories selected (either by themselves or by the system), for 30-45 minutes for 3 times on 3 consecutive days. We call this practice environment, where they are free to study by any speed they like. We ask them to answer questionnaires about the system after the 3 times of practise, and two weeks later, they work on test environment of all categories again. Scores are collected from student model in 3 stages: pre-test score, collected after the first test, immediate post-test score, collected after the third practise, and delayed post-test score, collected after the last test.

# 4 RESULT AND DISCUSSION

Data collected from the experiment are analysed according to layered evaluation method in 3 aspects [11]: adaptive mechanism, to assure whether adaptive instructions match the needs of the students, learning achievement, to measure improvement of performance of students, and satisfaction of students toward the system, measured by questionnaire. At the end of this section, weak categories of Japanese students as a whole are also presented for future references.

## 4.1 Adaptive Mechanism

Students in adaptive group are set to study only categories that they are poor at, while those in non-adaptive group select some categories to study by themselves. We want to check whether selections of our adaptive mechanism are what students think they are their weaknesses; therefore, we extract data saved in student model of non-adaptive group and compare them with actual selection of students. The result indicates that 61.74% of selected categories by student model match student choices, and among those categories, 39.16% are which students especially spent their time in, comparing to number of total exercises studied. This indicates that our adaptive mechanism is capable to choose instructions that match students' need.

Average number of categories selected by students in non-adaptive group (13.3) is lower than number of selections in student model (16.1). In contrast, number of exercises done in each category of non-adaptive group is higher than that of adaptive group. We can imply that, students in non-adaptive group prefer to spend a lot of time in few categories, rather than probing many categories. This suggests future adaptation of the practicing system to focus on a few categories at a time.

## 4.2 Learning Achievement

We do statistical analysis of pre-test, immediate post-test, and delayed post-test scores to measure improvement of the students.

### *4.2.1 Pre-test and Post-test Score*

First, we investigate the effectiveness of both adaptive and non-adaptive settings, which in both groups, students practice by the same exercises. We do paired t-test between pre-test scores and immediate post-test scores, and between pre-test scores and delayed post-test scores of all students. The result shows that both pairs are difference with 99% confidence. Difference between pre-test and immediate post-test has larger effect size of value 1.891, than between pre-test and delayed post-test, which is 0.775. This means retention loss or that students forget what they learned after a period of time.

### 4.2.2 Post-test Score of Two Groups

We use independent t-test to test hypothesis that post-test score of students who study in adaptive group is higher that those who study in non-adaptive group. The result shows that, when students in two groups perform similarly in pre-test, there is also no significant difference between two groups in either post-test or delayed post-test. Thus, we cannot conclude that students in adaptive group learn better than the other.

We further investigate the result by dividing students by their pre-test scores into high and low-score levels, namely advanced level and intermediate level, each of which has mean score of 1011.13 and 691.63 respectively, and standard distribution of 65.211 and 179.595 respectively. We do t-test again, and the result shows that, for students in advanced level, the adaptive group performs better in immediate post-test with 95% confidence and in delayed post-test with 90% confidence. In contrast, for students in intermediate-level, t-test does not show a significant difference between immediate and delayed post-test of two groups.

## 4.3 Satisfaction

We ask students to answer questionnaire consisting of 2 parts. One is to rate usability, usefulness, and how they feel toward the system, another is to write comments about the system.

### 4.3.1 Ratings

In total, ratings of adaptive and non-adaptive groups are not significantly different, because both group study with similar navigations. The only significant difference is that students in non-adaptive group think they learn new expressions more than adaptive group do.

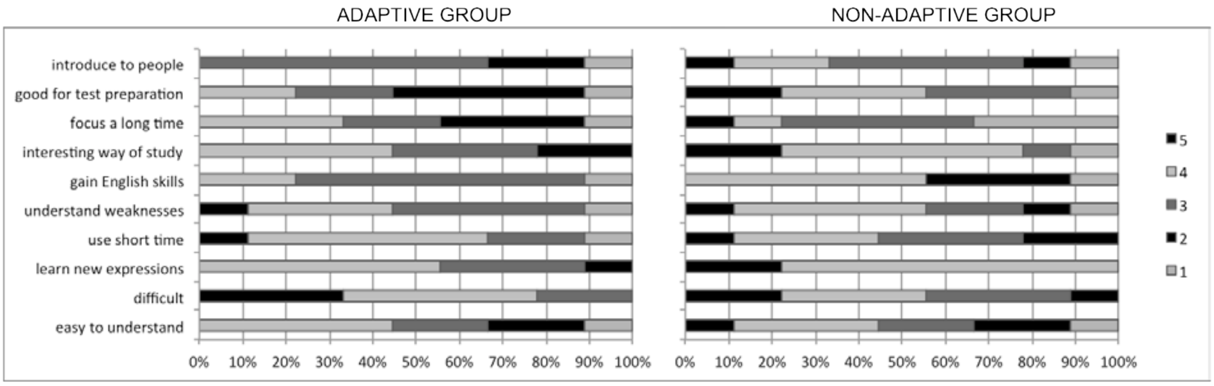


**Fig. 4 Result of Questionnaire by Students in Adaptive and Non-adaptive Group.**

Items that gain the highest rating in adaptive group are a positive opinion that they learn many things in short period of time, and a negative opinion that they think the exercises are difficult for them. For non-adaptive side, the highest are that they learn new expressions and they think the system is an interesting way of study.

From overall ratings, we can see in Fig. 4 that satisfaction of students in non-adaptive group is slightly higher than those in adaptive group. This is thought to be result of our adaptive mechanism that is designed to pick exercises that the particular student is poor at, so students feel that most of the exercises are difficult, and the system is less interesting than the one they can freely choose what to learn by themselves. Anyway, those in adaptive group think they learn many things in short period, which indicates effective selections of the system.

### 4.3.2 Comments on Practicing System

We further investigate more of students' opinions to improve our system. Despite our attempt to decrease incomplete exercise explained in section 2.1.2, students in both groups comment that some problems have more than one possible correct answer and some are not complete enough to understand. Moreover, some students want the system to have more explanations or Japanese translations for difficult vocabularies. From these issues, we consider that to improve our system, we need natural language processing to judge completeness of exercises, and back-translation corpus in NICT JLE corpus might be useful in providing translations.

Other issues are that students need the system to review sentences they have studied, and that it is good that our system repeat exercises when they got wrong answer. This can give suggestions to current problem of retention loss stated in section 4.2.1.

## 4.4 English of Japanese Students

We examine average pre-test scores of all participants to find out which categories Japanese students in total are good or poor at. Five categories with the highest scores are verb form, adjective inflection, word order, finite/infinite verb, and unknown type, and the five with the lowest score are pronoun lexis, adjective lexis, adverb lexis, noun lexis, and Japanese English. This indicates that lexis is the part that most Japanese students are poor at, although it has most effects on sentence intelligibility [10].

# 5 CONCLUSION AND RECOMMENDATION

We construct an adaptive English practicing system for reducing weaknesses of learners, by letting them learn by errors of other students extracted from learner corpus. The experiment shows that our system has potential to be an effective studying approach especially for advanced students to extend their knowledge; however, improvements are left for future researches.

We suggest that difficulty level should be added to each exercise in order to improve efficiency of selection mechanism. System could change from choosing exercises for students, to recommending for them, and letting them decide on selections themselves. Selected categories could be presented a few at a time, alternating with review of sentences studied especially ones that they got wrong answers. Moreover, the system should provide more explanations or translations to facilitate learners, which can be referred from back-translation corpus. Finally, we recommend that future researches give more importance to lexis or vocabulary, which is the weakest but important part of Japanese learners.

# REFERENCES


[1] U.S. Department of Education, Office of Planning, Evaluation, and Policy Development (2009). Evaluation of evidence-based practices in online learning: A meta-analysis and review of online learning studies.

[2] Brusilovsky, P., & Millán, E. (2007). User models for adaptive hypermedia and adaptive educational systems. The adaptive web, 3-53.

[3] Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. User modeling and user-adapted interaction, 6(2), 87-129.

[4] Leeman, J. (2007). Feedback in L2 learning: Responding to errors during practice. In DeKeyser, R. (Ed.), Practice in a second language. Perspectives from applied linguistics and cognitive psychology. New York: Cambridge University Press, 111-137.

[5] MacWhinney, B. (1987). The competition model. In MacWhinney, B. (Ed.), Mechanisms of language acquisition, New Jersey: Lawrence Erlbaum Associates, Inc., 249-308.

[6] Li, S. (2010). The Effectiveness of Corrective Feedback in SLA: A Meta‐Analysis. Language Learning, 60(2), 309-365.

[7] Heift, T. (2004). Corrective feedback and learner uptake in CALL. ReCALL, 16(2), 416-431.

[8] Gutierrez, F., & Atkinson, J. (2011), Adaptive feedback selection for intelligent tutoring systems, Expert Systems with Applications, 38(5), 6146-6152.

[9] Tono, Y., Izumi, E., & Kaneko, E. (2005). The NICT JLE Corpus: the final report. In Bradford-Watts, K., Ikeguchi, C., Swanson, M. (Eds.), JALT 2004 Conf. Proceedings, Tokyo, 345-356.

[10] Izumi, E., Uchimoto, K. & Isahara, H. (2007). Vocabulary usage and intelligibility in learner language. In Davies, M., Rayson, P., Hunston, S. & Danielsson, P. (Ed.), Proceedings of the Corpus Linguistics Conference. Birmingham, UK: The University of Birmingham.

[11] Brusilovsky, P., Karagiannidis, C., & Sampson, D. (2004). Layered evaluation of adaptive learning systems. International Journal of Continuing Engineering Education and Life Long Learning, 14(4), 402-421.