Empirical CALL evaluation: The Relationship Between Learning Process and Learning Outcome

QING MA
Catholic University of Louvain

ABSTRACT
CALL evaluation is important because it is the most efficient means to prove CALL effectiveness. While both learning process and learning outcome should be investigated in empirical evaluation, the precise relationship between the two needs to be examined closely. Only by doing so can we identify useful CALL design features that facilitate relevant user-computer interaction which can lead to an improved learning outcome. This study examined how certain user actions can affect or predict receptive/productive vocabulary retention in a computer-assisted vocabulary learning (CAVL) program. Results showed that only the user actions which contributed to identifiable levels of vocabulary processing—those at which (a) the initial meaning-form mappings of the vocabulary items are established and (b) the vocabulary items are practiced in different contexts—significantly accounted for vocabulary retention. In addition, user actions accounted for receptive vocabulary retention better than for productive vocabulary retention.

KEYWORDS
Empirical CALL Evaluation, Learning Process, Learning Outcome

INTRODUCTION
Rapid technological development has made computer hardware and software cheaper and more readily available; consequently the traditional means to prove CALL effectiveness by comparing CALL with non-CALL instruction is outmoded (Chapelle, 2001, 2003; Jamieson, Chapelle, & Preiss, 2005). What is needed is to improve CALL to make it more effective and efficient. CALL evaluation thus becomes an important issue in current and future CALL research. Only through appropriate evaluation can we demonstrate CALL effectiveness, identify useful CALL design features, select the right programs for specific educational settings, and improve future CALL design. This article first provides a brief summary of how CALL evaluation has been conceptualized in the literature. It then focuses on summative empirical evaluation in which the learning process and the learning outcome were measured. Finally, it reports on a study which examined the relationship between the learning process—a number of selected user actions—and the learning outcome—receptive/productive vocabulary retention—for a computer-assisted vocabulary learning (CAVL) program.

CALL EVALUATION
CALL literature frequently distinguishes between two types of evaluation: formative and summative evaluation (Hémard, 1997, 2004; Levy, 1999; Colpaert, 2004). Formative evaluation is conducted during CALL development to ensure the efficiency of the program and involves prototype testing with CALL experts or with a small number of selected users. Improvements
in various design features are made, and the process is repeated until the best design is found. Summative evaluation is conducted with the final product to test whether the program can achieve its intended purpose; this can be done judgmentally by expert opinion or empirically by data analysis (e.g. testing or rating scores) obtained from the end-users. Ideally, formative evaluation should be combined with summative evaluation. While CALL designers or developers may treat the two as equally important, language teachers or CALL researchers are normally more interested in summative evaluation because they are more often concerned with the finished CALL product than being directly involved in the development process. Given this consideration, this article focuses on summative evaluation.

Summative evaluation can be divided into two types: judgmental and empirical evaluation, a distinction first made explicit by Chapelle (2001). If judgmental evaluation provides the hypotheses about the efficiency of a given CALL application, then empirical evaluation is the process for testing those hypotheses. The two types of summative evaluation also differ in terms of the evaluation methodology employed. Judgmental evaluation is often conducted with the help of checklists or questionnaires based on intuitive, subjective, or expert judgment. A more rigorous type is heuristic evaluation, which involves a group of evaluators who assess the usability of software independently and then integrate the findings to produce the evaluation results (see Nielsen, 2006).

Empirical evaluation of courseware requires empirical data as input but also the appropriate analysis of the data to produce results. While judgmental evaluation primarily helps teachers or educators to evaluate and select appropriate CALL learning materials, empirical evaluation is often conducted by researchers to verify whether the judged value of a given CALL program can be demonstrated through empirical research. Earlier empirical evaluation was typically set within a comparative framework in which the learning outcome of the CALL activities was measured through experimental or quasi-experimental design and compared with non-CALL activities. It is now commonly agreed that this type of evaluation research by itself is not very revealing, largely due to the methodological limitations associated with making comparisons between CALL and non-CALL (Garrett, 1991; Chapelle, Jamieson, & Park, 1996; Jamieson, Chapelle, & Preiss, 2005). As an alternative, Chapelle (2001, 2003) has forcefully argued for a process-oriented approach to empirical evaluation. What is needed is investigation of the learners’ behavior on the computer (e.g., mouse clicks, use of help options, and various interactions with the computer) and to analyze this behavior in the search for evidence for the (potential) acquisition of the targeted linguistic forms. However, we should not abandon measurement of the learning outcome because it is the most convincing way of demonstrating CALL efficiency (or inefficiency), as evidenced by its frequent use by CALL researchers (see below). The methodological limitations of measuring learning outcomes can be at least partially overcome by careful design of the experiment. In the “phenomenographic approach” described by Goodfellow (1999, p. 113), learning process and learning outcome are the two sides of the same coin: learning outcome is the natural result of learners’ learning experience (or learning process). In this sense learning outcome and learning process analysis may be equally important in empirical evaluation for CALL.

EVALUATING LEARNING OUTCOME AND LEARNING PROCESS

Based on the six criteria for CALL evaluation put forward by Chapelle (2001), learning outcome can be interpreted as language-learning potential, that is, how well the linguistic forms are mastered after CALL use. This involves identifying the targeted linguistic objectives (e.g., grammatical/lexical knowledge, reading comprehension, or writing competence) and designing the corresponding tests to measure the learning of these objectives. Often a standard experimental design is used, consisting of a pretest and posttest. The retention of linguistic
knowledge can then be attributed to the language-learning potential of the CALL application: the greater the retention, the greater the learning potential. Sometimes a delayed posttest is used to measure long-term learning potential. Such an experimental design is typically employed to measure vocabulary retention resulting from the use of CALL applications (see Grace, 1998; Groot, 2000; Laufer & Hill, 2000; Jones & Plass, 2002; Yeh & Wang, 2003; Tozcu & Coady, 2004; Smidt & Hegelheimer, 2004). Alternatively, only a posttest is used but previously acquired knowledge of the vocabulary is checked (see Chun & Plass, 1996; Al-Seghayer, 2001; De Ridder, 2002). This kind of approach is essentially quantitative, but a qualitative approach, based on learners’ self-reporting or reflection on what they have learned, can complement the quantitative measure. For example, Jamieson, Chapelle, and Preiss (2005) asked students to fill in a “weekly reflection” sheet to report what they had learned from using a CALL application during the experiment period. Their study provides a good example of how quantitative and qualitative approaches can be combined in the empirical evaluation.

What should be evaluated in the learning process is process data, a definition of which was provided by Jamieson and Chapelle (as cited in Chapelle, 2003, p. 98): “working style data—consistent, observable behavior displayed by students as they worked on L2 tasks,” sometimes called “tracking data” or “computer logs” (Chapelle, 2003, p. 98). Learner actions or behaviors on the computer can be tracked in multiple ways. They can be recorded by the built-in tracking system of the CALL application (Goodfellow, 1994; Hegelheimer & Tower, 2004), by a tracking device in a course management system (Hwu, 2003), recorded on tape or video as in a “walk through” (Hémard, 2004), or simply observed and manually documented by researchers (Desmarais, Duquette, Renié, & Laurier, 1998). Often these actions are collected as quantitative data, for example, the number of mouse clicks or time spent on (each section of) the CALL application. In addition, learners’ reflections or comments (Goodfellow, 1994; Desmarais et al., 1998; Vinther, 2005) about the learning process, often collected by a think aloud protocol, are a way to capture qualitative data.

A recent review on tracking technology by Hwu (2005) showed that it has been widely used in CALL studies, covering a variety of topics: incidental/intentional vocabulary learning, writing, listening, pronunciation, grammar, and exercises and drills, and so forth. Hwu divided tracking data into two types: (a) directly observed mouse-clicks, keyboard activity, or cursor movement and (b) inferred navigation patterns, error rates and time. Chapelle and Mizuno (1989) distinguished between “low-inference” user actions, such as learners’ use of sections for grammatical or lexical help, and “high-inference” user actions, such as learners’ use of strategies to monitor, evaluate, or manage their performance.

While tracking technology provides opportunities to record or infer various user actions taking place in the learning process, how to organize and analyze these data efficiently turns out to be a complicated issue. Ma (2007a) showed that the success of evaluating the learning process largely depends on how the user tracking system is designed; most important is not how many types of user actions are tracked and recorded, but rather which user actions should be tracked and how the data should be recorded. To make the tracking process more efficient, we should decide what data are most meaningful to explain the quality of the learning process; to find out the most relevant user actions we need to understand the cognitive processing of the targeted linguistic item. In terms of vocabulary learning, Ma proposed four criteria to evaluate the level of vocabulary processing:

1. Is there any evidence that learners have understood/accessed the lexical meaning(s) and paid attention to the lexical forms?
2. Is there any evidence that learners have been exposed to the multiple input channels (textual, aural, and pictorial) of the lexical items?
3. Is there any evidence that learners have mapped the meaning with the form and strengthened the link via efficient means such as semantic elaboration or imagery?
4. Is there any evidence that learners have rehearsed the words in different contexts?

The first question checks the basic condition for vocabulary learning to occur, that attention should be paid to both the meaning and the form of the targeted lexical items. The second question checks whether the lexical items are presented in multiple modes such that the multimedia effect enhances the memory trace of the lexical form and of the mapping with the meaning. The third question ensures that the link between the lexical form and meaning will be strong enough to be held in memory so that further learning (e.g., other meanings, usage, syntactic features, and morphological features) will take place. The purpose of the fourth question is similar to that of the third, namely, to increase the form-meaning mapping, not through semantic elaboration but through direct use of the lexical items (comprehension or production). It is assumed that each time a lexical item is comprehended or produced, the form-meaning mapping is strengthened and the item has a greater chance of being stored as a lexical entry in the learner's mental lexicon. In a similar vein, regarding other types of linguistic knowledge such as reading or listening, identification of the essential criteria relevant to the cognitive processing involved will facilitate the selection and organization of the tracking data.

Thus far, in terms of empirical CALL evaluation, researchers have been looking into both learning process and learning outcome. However, very few studies have explored the relationship between the two. Although it is commonly agreed that examining a learning process is a useful way to understand its associated learning outcome, it remains unclear to what extent learning process can account for learning outcome. In particular, do all user actions contribute to the learning outcome? If not, what user actions are particularly helpful and why? The few studies which have examined the relationship between the process and outcome tended to be based on the analysis of qualitative data (see Goodfellow, 1994, 1999); the interpretation of the results may have suffered from a certain amount of subjectivity and the small number of participants involved makes the results difficult to generalize. When the analysis of quantitative data was employed, however, the results either showed no relationship (Liou, 1997) or only a weak, partial relationship (Hegelheimer & Tower, 2004) between the process and outcome. In other words, the question of how particular user actions can be related to the learning outcome has not been clearly answered. The study reported in the following sections attempts to reveal this relationship from a quantitative standpoint.

THE STUDY

Background Information

A CAVL program named WUFUN was developed to help Chinese university students of English learn vocabulary items both receptively and productively, additionally offering useful vocabulary-learning strategies in order to foster learner autonomy. Its design was based on the CALL efficacy model, which brings together four components: theory, computer technology, user actions and learner information. The interaction and integration of these four components determines the quality of the CALL program (for details of the model, see Ma & Kelly, 2006; Ma, 2007a). Learner information is the background data for all learner characteristics and informs the decisions made regarding the other three components. Computer technology is a means to control and guide user actions, particularly those of low-level language learners and those who have little prior experience with computers. User actions should be controlled to some
degree to achieve learning efficiency as well as prepare novice learners to ultimately become autonomous learners. The success of the model lies in the balance between user control of the program and program control of the user.

In the WUFUN program, users first preview the text to be presented later in the form of pictures and accompanying audio/written comments. Then they view some important vocabulary items in a Word Focus (WF) section, a form of a minidictionary. Next, users read the text containing the vocabulary items introduced in the WF section. Afterwards, they study some vocabulary-learning strategies in a Word Memorization Aids (WMA) section to help them to memorize the newly learned items. Then users perform a series of exercises in which the new items are rehearsed in various contexts, receptively and productively. Following the exercises a number of idioms is introduced accompanied by pictures and sounds. Finally, some culture-embedded jokes are presented to make users realize that language learning is often culture bound. Almost all learning activities are accompanied by aural input, which serves to enhance the learning effect and help learners to acquire the habit of listening to the target language. Users are required to perform some basic requirements at each step before proceeding to the next one. For example, they have to look up at least one word in the WF or WMA sections before they are able to proceed further in the program, and they have to listen to the answers to the exercises before accessing the written answers. Learners have limited freedom regarding what to do and how to use the software, but they gradually acquire more and more freedom after the completion of a number of initial steps.

In the studies reported in Ma (2007a, 2007b), a modified version of WUFUN was developed in which all user constraints were removed and users had complete control over the program. The evaluation was thus done within a comparative framework in which the original, highly structured version of the program was compared to a modified, unstructured program on the four dimensions of learning outcome, learning process, learner evaluation, and the relationship between learning outcome and learning process. Results generally showed that the original WUFUN was more efficient than the modified WUFUN from the CALL researcher's point of view.¹

The study reported here was part of the fourth dimension: the relationship between a learning outcome, defined as vocabulary retention, and the learning process, defined as various user actions recorded by the system, in the original WUFUN program. The results of the modified program are not reported here; although they followed a pattern similar to those of the original program, the relationship between the learning process and the learning outcome was clearer in the original WUFUN program. Exercising a degree of control over student learning made it easier to observe the relationship between the learning process and the learning outcome because the user actions tended to take place in a predictable order thereby reducing the effect of unknown variables on the learning outcome.

This study addressed two research questions:

1. What user actions may have an effect on the learning outcome?
2. How do these user actions account for the learning outcome?

**Participants**

The participants (N = 50) in the study were first-year university students at the China Three Gorges University. The students were nonlanguage majors, specializing in the fields of science, engineering, and the arts. Most had studied English for 6 years or more before entering the university.
**Procedure**

The study followed a pretest/posttest design. A receptive vocabulary pretest followed by a productive vocabulary pretest, each lasting 10 minutes, were administered to the students to ascertain their preexisting knowledge of the vocabulary items in question. One week later, the students used the WUFUN program as a self-teaching exercise in a multimedia classroom. The students were given a document to show them how to use the software. No time limit was set, but most students finished using the software by the end of the class (the equivalent of two 50-minute classroom periods). Immediately after the students finished the program, they completed a receptive vocabulary posttest and a productive vocabulary posttest in the same way as the pretests. One month later, a delayed receptive vocabulary posttest and a productive vocabulary posttest were administered to the students, again in the same way as the preceding tests. All three testing sessions took place during normal classroom time.

**Analysis**

Each test had 30 questions, but only 25 items were scored (the remaining 5 items were distracters). Each correctly answered question received 1 point, maximum score of 25. The questions in the receptive and the productive tests were the same in the pretests, immediate posttests, and delayed posttests, but the order of the questions was changed.

The receptive test contained multiple-choice questions in which the English word was given along with five options. The first four options were Chinese translations, and the last one was the statement: “I have never seen this word before.” This last option was included to reduce guessing by chance, particularly in the pretest (see sample below; the English is provided for the reader’s benefit, but obviously the participants had only the Chinese).

**Dump**

1. 丢弃 ‘dump’
2. 沉默 ‘keep silent’
3. 拍打 ‘beat, run against’
4. 敲击 ‘knock’
5. 我从未见过该词 ‘I have never seen this word before’

The productive test consisted of gap-filling questions. The participants were asked to produce a word the first letter of which was given in a context together with its Chinese translation. This followed the pattern of the controlled active vocabulary test (Laufer, 1998) (see sample below; again English is added for the reader’s benefit).

Don’t wash jeans in a washing machine, they will **s__________ (v) 窒小 ‘shrink’**

Both short-term and long-term vocabulary retention were measured; the difference between the pretest and immediate posttest scores represented the short-term vocabulary retention, and the difference between the pretest and the delayed posttest scores represented the long-term vocabulary retention. In this study, the learning outcome was interpreted as the immediate vocabulary retention because the long-term retention could have deteriorated for a variety of unknown reasons during the month between the treatment and the delayed posttest. Such deterioration would have blurred the relationship between the user actions performed in the program and the delayed vocabulary retention.
The built-in tracking system of the software recorded various user actions while the users were interacting with the software. The system automatically produced an XML file for each student upon exiting the software. Not all user action variables were important for the learning outcome. A number of user actions were selected based on the four criteria for the level of vocabulary processing mentioned above and taking practical constraints into consideration (for details of the selection process, see Ma, 2007a, 2007b). In addition, three preknowledge variables were added because they could have affected the learning outcome: proficiency level, receptive pretest score, and productive pretest score. The selected user action variables and the three preknowledge variables were encoded as shown in Table 1.

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Preknowledge Variables and User Action Variables</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Encoded variables</strong></td>
<td><strong>Explanation</strong></td>
</tr>
<tr>
<td>Preknowledge variable</td>
<td></td>
</tr>
<tr>
<td>Pre_recep</td>
<td>receptive pretest score</td>
</tr>
<tr>
<td>Pre_pro</td>
<td>productive pretest score</td>
</tr>
<tr>
<td>Proficiency</td>
<td>participant’s final exam score in the English course completed the semester prior to the experiment</td>
</tr>
<tr>
<td>User action variables</td>
<td></td>
</tr>
<tr>
<td>EX_initial_time</td>
<td>amount of time spent on the initial completion of the exercises</td>
</tr>
<tr>
<td>EX_initial_score</td>
<td>participant’s score on initial completion of the exercises</td>
</tr>
<tr>
<td>EX_final_score</td>
<td>participant’s score on the exercises after receiving feedback on initial completion of the exercises</td>
</tr>
<tr>
<td>EX_answers</td>
<td>number of the target words in participant’s answers to the exercises</td>
</tr>
<tr>
<td>WF_W</td>
<td>number of different words looked up in WF section of the program</td>
</tr>
<tr>
<td>WMA_W</td>
<td>number of different words looked up in WMA section of the program</td>
</tr>
</tbody>
</table>

After collecting the data for both the learning outcome (the receptive/productive vocabulary retention) and the learning process (user actions), the relationship between the two was examined. First, Pearson product-moment correlation tests were performed between the dependent variables (receptive vocabulary retention and productive vocabulary retention) and the independent variables (preknowledge and user actions). Several user actions were identified as being significantly correlated with either the receptive or productive retention or both. Based on the results of the correlation analysis, a series of multiple linear regression tests were then performed to determine how much the user action variables accounted for receptive and productive retention.

**RESULTS AND DISCUSSION**

**Means and Standard Deviations of the Variables**

Table 2 lists the means and standard deviations of the dependent variables (receptive retention and productive retention) and the independent variables (preknowledge and user actions).
As shown in Table 1, after using the software, students gained an average of 9.38 words in receptive vocabulary and 5.16 words in productive vocabulary. On the one hand, the learning results were generally satisfactory: the receptive learning rate was 71% and the productive learning rate 39%. On the other hand, given that the students had already known more than 10 words receptively and productively out of the total of 25 words as measured by the pretests ($M = 11.72$ for receptive vocabulary and 13.44 for productive vocabulary), the results of the posttests may have been subject to a ceiling effect. Many researchers on vocabulary acquisition prefer to ensure that learners’ previous knowledge of target words is kept to a minimum in order to maximize the learning gains attributable to a treatment. However, this is not realistic in a real-world learning situation where vocabulary acquisition is an ongoing, additive process. Although the ceiling effect in the vocabulary retention in this study may have posed certain difficulties in interpreting the results, the study described here documents vocabulary acquisition in an authentic language-learning environment.

The XML file produced by the system showed that the students initially spent an average of 22.46 minutes doing the exercises and obtained a mean score of 52.55 out of 100. The data also showed that the students made a substantial effort when redoing the exercises and obtained a final mean score of 85.94 with a mean of 28.72 correct words out of 35 words in the exercises, a large increase from the initial score. Table 1 also shows that the students looked up more words in the WF section of the program than in the WMA section. This finding is understandable since the WMA section only gave tips on how to study and memorize the target words and other similar types of words. Apparently most learners felt that the information provided in the WF section was more directly relevant to their learning the words and that the information in the WMA section was of secondary importance.

**Correlation Analysis**

Correlation analysis measures the relationship among variables. The dependent and independent variables listed in Table 2 were submitted to Pearson moment-product correlation analysis. Table 3 summarizes the results of this analysis.
Table 3
Pearson Correlation Coefficients Among Selected Dependent and Independent Variables (Two Tailed)

<table>
<thead>
<tr>
<th></th>
<th>Receptive retention</th>
<th>Productive retention</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre_recep</td>
<td>-.739**</td>
<td>-.237</td>
</tr>
<tr>
<td>Pre_pro</td>
<td>-.241</td>
<td>-.410**</td>
</tr>
<tr>
<td>Proficiency</td>
<td>-.455**</td>
<td>-.416**</td>
</tr>
<tr>
<td>EX_initial_time</td>
<td>.178</td>
<td>.026</td>
</tr>
<tr>
<td>EX_initial_score</td>
<td>-.178</td>
<td>-.006</td>
</tr>
<tr>
<td>EX_final_score</td>
<td>.260</td>
<td>.309*</td>
</tr>
<tr>
<td>EX_answers</td>
<td>.336*</td>
<td>.347*</td>
</tr>
<tr>
<td>WF_W</td>
<td>.641***</td>
<td>.362*</td>
</tr>
<tr>
<td>WMA_W</td>
<td>.379**</td>
<td>.223</td>
</tr>
</tbody>
</table>

Note: *p < .05, **p < .01, ***p < .001

The three preknowledge variables—proficiency, receptive vocabulary, and productive vocabulary—had generally negative relationships with the receptive retention and productive retention. The relationship between preknowledge receptive vocabulary and receptive retention was strong ($r = -.739, p < .001$) and the one between preknowledge productive vocabulary and productive retention was moderate ($-.410, p < .01$). The relationship between proficiency and receptive retention and between proficiency and productive retention were also moderate ($-.455$ and $-.416, p < .01$; respectively).

The user action variables generally showed a positive relationship with the two types of vocabulary retention. The variables, initial time (EX_initial_time) and initial score on the exercises (EX_initial_score), did not show any significant relationship with the two types of vocabulary retention. However, most of the remaining four user action variables—final score on the exercises (EX_final_score), number of target words in the answers to the exercises (EX_answers), the number of different words looked up in the WF section of the program (WF_W), and the number of different words looked up in the MWA section of the program (WMA_W)—had moderate to good relationships with one or the other (or both) of the two types of vocabulary retention. Only EX_final-score/receptive retention and WMA_W/productive retention were nonsignificant.

The first research question of the study focused on the user actions that may have an effect on learning outcome. The correlation analysis showed that the three preknowledge variables—proficiency, receptive vocabulary, and productive vocabulary—had a general negative relationship with the two types of vocabulary retention, that is, the more preknowledge a participant had, the less vocabulary retention that occurred. This finding seems counterintuitive and deserves some comment.

Hegelheimer and Tower (2004) showed that language proficiency was positively correlated with learning outcome, but the measures they used for both differed from those used in this study. They defined language proficiency as learners’ scores on the placement test built into the CALL program and learning outcome as learners’ performance in the exercises in the program. In this study, language proficiency was defined as the participants’ scores on the final exam in an English course from the previous semester and the learning outcome as
the receptive or productive vocabulary retention obtained from the pre- and posttests. More important, in this study there were only 25 vocabulary items to be studied, of which the participants already had a considerable preknowledge (see Table 2 above). The possible ceiling effect in the vocabulary retention means that the more preknowledge of the target items the learner had, the smaller the gain in the amount of knowledge acquired. This situation is unavoidable when we look at language learning in authentic, less controlled environments where knowledge is not gained in a vacuum but often subject to other influences.

The correlation analysis also showed that four user action variables were positively related to receptive and/or productive vocabulary retention: (a) final score on the exercises in the program, (b) number of target words in the answers to the exercises, (c) number of words looked up in the WF section, and (d) number of words looked up in the WMA section. The correlations between the user actions and vocabulary retention differ in their degree of strength.

Returning to the four criteria for evaluating the level of vocabulary processing presented above, looking up words in the WF section is related to the first and second criteria, namely, to understand the word meaning and to access available lexical information in multiple channels (e.g., textual, audio, and pictorial). Looking up words in the WMA section is related to the third criterion, that is, to strengthen already established meaning-form mappings of the vocabulary items. The score obtained for the exercises and the number of the target words in the answers to the exercises reflect the fourth criterion: to rehearse the words learned in different contexts. In summary, all four of these user action variables, which are positively related to the learning outcome, reflect the four criteria for the level of vocabulary processing.

**Multiple Linear Regression Analysis**

The correlation analysis showed the relationships among the variables. To measure the degree to which specific variables contributed to the learning outcome, multiple linear regression analysis was used. While correlation analysis allows us to examine the relationship among individual variables, multiple linear regression analysis allows us to examine the relationship between several predictor variables (independent variables) and the criterion variable (dependent variable). In this study, receptive retention and productive retention were treated as separate criterion variables, and the three preknowledge variables and the four user action variables served as predictor variables entered into stepwise regression models.

Tables 4 and 5 summarize the results of the regression analyses.

**Table 4**

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized coefficients</th>
<th>Standard error</th>
<th>Standardized coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pre_recep</td>
<td>-.773</td>
<td>.113</td>
<td>-.587**</td>
</tr>
<tr>
<td>WF_W</td>
<td>.207</td>
<td>.053</td>
<td>.349**</td>
</tr>
<tr>
<td>EX_answers</td>
<td>.073</td>
<td>.033</td>
<td>.182*</td>
</tr>
</tbody>
</table>

*Note: $R^2 = .700$

*p < .05, **p < .001
Table 5
Stepwise Regression: Productive Vocabulary Retention

<table>
<thead>
<tr>
<th></th>
<th>Unstandardized coefficients</th>
<th>Standard error</th>
<th>Standardized coefficients</th>
</tr>
</thead>
<tbody>
<tr>
<td>Proficiency</td>
<td>-.164</td>
<td>.059</td>
<td>-.339**</td>
</tr>
<tr>
<td>EX_final_score</td>
<td>.049</td>
<td>.017</td>
<td>.344**</td>
</tr>
<tr>
<td>Pre_pro</td>
<td>-.304</td>
<td>.116</td>
<td>-.320*</td>
</tr>
</tbody>
</table>

*Note: $R^2 = .338$.
*p < .05, **p < .001

The model for receptive vocabulary retention was generally better than the model for productive vocabulary retention, demonstrated by the larger $R^2$ in Table 4 (.700) than that in Table 5 (.338). For user actions in reception retention, the number of words looked up in the WF section of the program (WF_W, $p < .001$) and the number of target words in the answers to the exercises (EX_answers, $p < .05$) were accepted as predictor variables. For user actions in productive retention, only the final score for the exercises (EX_final_score, $p < .001$) was accepted as a predictor variable.

These results are, in general, consistent with those of the earlier correlation tests. The preknowledge variables negatively predicted vocabulary retention, but several of the user action variables contributed significantly to vocabulary retention. Looking up words in the Word Focus section of the program was the process by which the word meaning was accessed and the meaning was matched to the word form, which should be helpful at least with respect to receptive learning. This was indeed the case. The WF_W variable was accepted as a predictor for receptive vocabulary retention. The two user actions related to performance on the exercises, involving both reception and production, were accepted as predictors for receptive or productive vocabulary retention. Although in the correlation analysis the number of words looked up in the Word Memorization Aids section of the program was moderately correlated with the receptive vocabulary retention, it did not turn out to be a predictor for vocabulary retention in the regression models. This indicates that the Word Memorization Aids section was not a component in the software that directly contributed to the retention of the words but that it should perhaps be treated as a separate component for vocabulary strategy learning (for a detailed discussion of this issue, see Ma 2007a).

To answer the second research question of how these user actions account for the learning outcome, there are two points to be made. First, three user actions—the number of words looked up in WF, the number of target words in the answers to the exercises, and the final score on the exercises—predicted the learning outcome of vocabulary retention. However, they did not alone account for the learning outcome; they were part of a model in which preknowledge negatively predicted the learning outcome. This result is largely due to the fact that the participants already had considerable knowledge of the target words prior to the project. Second, receptive vocabulary retention was more completely predicted than productive vocabulary retention. This may be due to the nature of receptive learning versus productive learning. Receptive learning, generally considered to be easier than productive learning, requires less learning effort. Meara (1990) suggested that the distinction between receptive items and productive items is a marked one rather than a matter of degree of knowledge or familiarity. In this sense, the relationship between user actions and receptive retention seems relatively clear cut and easier to observe than the relationship between user actions and productive retention. This issue certainly deserves further investigation.
CONCLUSION
This study investigated a number of recorded user actions for a CAVL program and examined how they could affect receptive/productive vocabulary retention. Regarding the first research question, *What user actions may have an effect on learning outcome?* correlation analysis showed that four user action variables were identified as being closely related to the learning outcome: final score on the exercises, number of target words in the answers to the exercises, number of words looked up in the Word Focus section of the program, and number of words looked up in the Word Memorization Aids section. These four user actions can be interpreted in accordance with the four levels of vocabulary processing that ensure a successful mapping between word form and word meaning.

As for the second research question, *How do these user actions account for the learning outcome?* linear regression analyses showed that three of these user actions were accepted as predictor variables for receptive or productive vocabulary retention: number of words looked up in the Word Focus section, number of target words in the answers to the exercises, and final score on the exercises. The first user action variable represented the initial meaning-form mapping of the new vocabulary items, and the second two the ways in which vocabulary items were used in different contexts, involving both reception and production. The number of words looked up in Word Memorization Aids did not predict vocabulary retention, suggesting that the Word Memorization Aids section was a separate learning component in the software which promoted vocabulary strategy learning rather than contributing directly to vocabulary retention. The results also showed that user action variables could account satisfactorily for receptive learning, but less satisfactorily for productive learning. Taking all these together, three issues emerged regarding the empirical evaluation of CALL in terms of the relationship between the learning process and the learning outcome.

First, in empirical evaluation, evaluating the learning outcome is as important as evaluating the learning process because the two can often reflect each other. Efficient evaluation of the learning process largely depends on identifying and selecting relevant user actions that contribute to the learning outcome.

Second, in the case of vocabulary learning, user actions that can contribute to levels of vocabulary processing should be selected as a priority. User actions should be able to build up the initial meaning-form mappings of vocabulary items and subsequently strengthen those mappings by using the items in different contexts. Similarly, to evaluate other types of CALL programs, user actions that contribute directly to the targeted linguistic knowledge should be looked into more closely than those that do not. Thus CALL evaluation should not be treated in general, but in the context of a clearly identified linguistic focus of the CALL program.

Finally, *user actions encoded as quantitative data cannot adequately account for productive learning.* Something else may account better for productive learning. Processes which are not directly observed may be responsible for productive learning, for example, deep mental processing or deliberate effort made during learning sessions. This calls for qualitative analysis, such as think-aloud protocols, retrospective group discussions, or retrospective interviews in future studies.
NOTES

1 From the learners' point of view, both versions of the software were successful; the great majority of participants expressed their approval for both versions of the program.

2 The items included 22 individual words (acquaintance, burst, dam, damage, despair, dump, fail, formal, former, fun, funny, jump, leap, ridiculous, shallow, shrink, stretch, suit, utter, utterly, weight, wonder) and 3 idioms (he is in the depths of despair, he is fit to burst, he splits his sides laughing). These items were the focus words studied in the software.

3 For example, the total time spent on the program by the users was to be included as a user action variable. However, these data were incomplete due to unexpected technical problems and were excluded from analysis.

4 Following Brace, Kemper, and Snelgar (2006), in multiple linear regression the independent variable is termed "predictor variable" and the dependent variable "criterion variable."

REFERENCES


**AUTHOR’S BIODATA**

Qing Ma was awarded her doctoral degree in applied linguistics at the Catholic University of Louvain, Belgium. She is currently an assistant professor of English in the Department of General Studies, Macau University of Science and Technology. Her main research interests include second language vocabulary acquisition, research methodology, and CALL.

**AUTHOR’S ADDRESS**

Qing Ma  
A308, Macau University of Science and Technology  
Avenida Wai Long, Taipa  
Macau  
Phone: +853 66990340  
Email: maqingangel@gmail.com