CHAPTER 7

Experimentalized CALL for adult second language learners

Nora Presson, Colleen Davy and Brian MacWhinney
Carnegie Mellon University

Improvements in computer technology have opened up new possibilities for integrating web-based language learning with classroom practice. In particular, experimental computer-assisted language learning (eCALL) methods can make student learning more efficient, while also providing detailed data for second language acquisition theories and models. Studies show that eCALL systems that target basic language skills can lead to significant learning gains after only two or three hours of practice, with gains retained months later. Training in basic skills can be supplemented by online methods for using Internet media, map tours, subtitled video, chat rooms, and learning games. All these systems can be linked to classroom teaching to provide deeper support for second language learning.[1]

1. Introduction

Rapid advances in computer technology have stimulated a wide array of new approaches to computer-assisted language learning (CALL). There have been underlying improvements in bandwidth, connectivity, operating systems, processors, programming languages, and high-resolution touch screens. These improvements have been accompanied by the growth of online resources such as games, dictionaries, grammars, translators, multilingual media, and Wiki pages. Based on these developments and the advent of ubiquitous computing, software for second language learning is now moving away from the desktop to mobile devices such as iOS and Android. Integrating these new capabilities with traditional classroom language teaching poses major new challenges, and opens up fascinating possibilities for researchers, as well as for teachers, developers, and learners. Using these new methods, researchers can gather data in the laboratory and over the web that will help us understand the actual process of second language learning, as it occurs

1. This work was supported by NSF grant SBE-0836012 to the Pittsburgh Science of Learning Center.

in both instructed and naturalistic situations. However, to realize these new potentials, researchers, learners, and instructors must reconfigure their relations and roles in order to work in a new, more collaborative manner.

This chapter examines the ways in which computerized second language (L2) instruction can incorporate established experimental methodology as well as newer methods in dynamic assessment (Cen, Koedinger et al. 2006) to produce demonstrable improvements in learning.This work is important for two reasons. First, it can provide concrete, fully implemented, methods for improving instruction. Second, it can provide detailed data, collected in systematically over the web, that permit the formulation of more precise accounts of the language learning process. We will refer to this method as eCALL (experimentalized computer-assisted language learning). We begin by explaining the logic of eCALL system development. Next, we show how eCALL systems can address issues in SLA regarding stages of development, patterns of error and correct usage, the role of explicit corrective feedback, practice effects, training based on student modelling, and the provision of explicit rule instruction. After that, we examine methods for automating lesson creation and data collection in eCALL. Then, we examine current limitations of eCALL, ways of addressing these limitations, and future possibilities for research and pedagogy in computerized contexts. Our analysis will examine how these interventions can be used to improve student learning, and how these interventions can allow second language acquisition (SLA) researchers to develop more accurate models of second language learning and processing informed by these new types of data.

Work in the CALL tradition has explored ways in which computers can be used to provide language instruction. Comparative studies (Chenoweth, Ushida et al. 2006) have shown that full computer-based courses (see examples at learner.org or oli.cmu.edu) can produce learning outcomes comparable to those obtained from classroom teaching. However, full course systems tend to integrate a wide variety of activities in ways that make it difficult to track details of the learning of particular skills, such as vocabulary, grammar, dictation, and pronunciation. In contrast, eCALL is designed to configure separate modules for training in these basic skills. For these basic skill activities, computers excel in their ability to provide carefully selected learning targets and to score, tabulate, and analyze student responses. In many cases, computers can make accurate diagnoses of student errors and provide proper diagnostic feedback. Computers can store a virtually unlimited trial history from which they can construct full profiles, learning curves, and error analyses. As we will see in Section 4, students can also rely on computers to locate written materials in the second language (L2), listen to L2 radio broadcasts, watch video in L2, correspond in L2 chat rooms, and explore the L2 physical and cultural environment.

However, there are limits to what computers can do. Computer speech recognition is still too imprecise to permit good recognition of learners' oral productions, and this limits the ways in which computers can deliver articulatory training or engage in conversation. Fortunately, instructors excel exactly in those areas where computers are the weakest. Ideally, learners could use computers for what they do best and could interact with instructors for what they do best. However, for this division of labor to operate properly, instructors must interact with the researchers developing online materials to assure that they interleave correctly with classroom materials and textbooks. The eCALL materials we discuss here are configured to link to the specific vocabulary, grammatical forms, or pronunciation skills targeted by the textbook selected by each individual instructor. These methods that place the instructor in control of the coordination of eCALL with other materials to produce an integrated treatment of all linguistic subsystems and areas of content both inside and outside the classroom.

To test the efficacy of educational interventions, one must collect controlled experimental data. However, not all CALL methods generate such data. Often CALL programs fail to achieve random assignment or involve other violations of internal validity, making it difficult to draw reliable conclusions about the efficacy of specific interventions and their implications for L2 processing and learning. Fortunately, experimental evaluation of training interventions is becoming increasingly frequent (Felix 2005), as researchers come to realize that eCALL evaluation serves not only to validate interventions but also to improve future implementations.

In addition to showing the effectiveness of a particular training activity, experimental evaluation can test specific predictions of particular language learning models. MacWhinney (1995) noted that it is possible to configure computerized training in basic skills so that learners are assigned randomly to treatment conditions. However, extracting valid experimental data from online learning systems also requires that students engage fully with the tasks. This engagement can only happen if teachers, students, and researchers have a shared vision of the importance of such an intervention providing computerized training in basic skills. Once this consensus has been achieved, eCALL research and instruction can progress on a solid footing. However, achieving this consensus requires experimenters to consider carefully how eCALL instruction can mesh with the pedagogical goals of classroom practice.

Training through eCALL offers the promise of providing a continually improving baseline of instructional quality. By refining and improving existing models on the basis of ongoing experimentation, we can ratchet up the quality of instruction, even as we continue to collect data that will improve our understanding of the basic cognitive mechanisms underlying second language learning. The iterative process in which theory informs intervention and interventions inform theory can provide a solid basis for improving the science of second language acquisition.

Training with eCALL can be applied in either online instruction or laboratory "pull-out" experiments in which classroom learners visit the controlled laboratory setting for data collection. Examples of the latter include studies of phonological contrasts (Edwards and Zampini 2008), the use of context to interpret prosodic cues (Hardison 2005), or studies with naïve beginners of the type we will discuss later in this paper. The advantage of laboratory eCALL is that there is less need to fit in closely with the classroom syllabus and it is easier to guarantee that subjects are focused clearly on the task at hand. However, integrated or "in vivo" eCALL (Koedinger 2011) has greater relevance to the actual process of instruction, particularly if the eCALL lessons are closely linked to the materials covered in the classroom. Moreover, it is often easier to recruit large numbers of participants for online eCALL training, facilitating the design of more sophisticated comparisons within an experiment.

Current eCALL systems maintain many features of earlier CALL systems, of which they are a natural outgrowth. In the past, "the majority of CALL uses were limited, in form, to drill and practice exercises" (Liu, Moore et al. 2002). Although current eCALL applications continue to emphasize the role of practice and repetition, there have been important additions in terms of improved multimedia support, ubiquitous Internet connectivity, individualized student tracking, logging of responses for instructor tracking, and creation of uniform data sets for analysis by researchers. At the same time, there has also been an upsurge in open-ended computerized methods such as games, subtitled video, and real-world activities provide computerized support systems that go far beyond the CALL systems of the past. It is important to realize that what is novel about eCALL is not the specific methods used in particular tutors or modules, but the ability to extract experimentally valid data that can inform the construction of a system that integrates learning resources, the classroom, and the community through the web. We are clearly not yet there and the studies we will examine here are just first steps in the direction of this very ambitious research program.

2. eCALL examination of SLA principles

From the viewpoint of SLA research, eCALL can be regarded as a method for examining pedagogical principles that can accelerate second language learning. The effect is cyclical: the pedagogical principles used in the field of SLA can be applied to eCALL training programs, which can then in turn refine and test those principles, thus adding to our knowledge base, which can be used to improve eCALL methods. In this section, we consider work that explores the role of four important instructional factors: the provision of immediate explicit corrective feedback, modeling of student knowledge, repeated practice, and explicit instruction. In Section 4, we discuss ways of studying more interactive feedback types, such as negotiation of meaning in chat rooms, usage patterns in mobile computing, or the computerized analysis of conversations recorded in naturalistic settings.

2.1 Corrective feedback in the classroom

Corrective feedback (CF) can be a key engine for learning in the classroom (Bangert-Drowns, Kulik et al. 1991; Ellis 2009). Classroom CF can come in many forms, and recent experiments and meta-analyses (Lyster 2004; Sato and Lyster 2012; Lyster, Saito et al. 2013) have demonstrated the complexity of the effects of these forms of CF. These analyses have distinguished (1) written vs. oral feedback, (2) implicit vs. explicit correction (prompts and recasts vs. metalinguistic correction), (3) instructor feedback vs. peer feedback, (4) high vs low work memory learners, and (5) younger vs. older learners. Analyses of classroom interactions have shown that, within the larger categories of prompts, recasts, and explicit correction there are further variations in terms of conversational structure and instructional impact. These analyses have been informed by the theories of noticing (Schmidt 1993), focus on form (Long 2000), input processing (Robinson 1995), the output hypothesis (Swain 2005), skill theory (DeKeyser 2007), the procedural/declarative distinction (Morgan-Short and Ullman 2011), monitoring theory (Levelt 1989), and retrieval theory (de Bot 1996; Karpicke and Roediger 2006).

2.2 Corrective feedback in eCALL

Given the interactive complexity of these forces, successful delivery of optimal levels of CF in the classroom may require a high level of instructor skill and engagement. Instructors must quickly diagnose the nature of the learner error, choose a feedback method, and then possibly interrupt the flow of classroom interaction, leading to violations of normal conversational practice (Gardner and Wagner 2005). Providing CF through the computer can address some of these problems. When the computer provides CF, learners may be frustrated, but they will not suffer any embarrassment. More importantly, the speed of computer processing makes it possible to quickly compose feedback contingent on the nature of the learner's response, personal learner characteristics, instructional theory, and instructional goals. Moreover, during interactions in a classroom, learners are being constantly exposed to errors that they themselves may never make, whereas the errors being corrected by the computer are ones that are specifically a problem for the individual learner interacting with the program.

Recent literature on cognitive tutors in math and science domains demonstrates the added power of immediate corrective feedback that is targeted at specific student performance errors (Koedinger, Anderson et al. 1997). By integrating a practice interface with an underlying cognitive model of correct and incorrect ("buggy") student knowledge, alternative feedback messages can be constructed to match the type of misconception a student is likely to have given his or her behavioral performance. For example, a student answering the question "2 – 4 =?" with "6" probably does not understand the symbol for subtraction, whereas one answering "2" may not understand that order matters in subtraction.

It is important to consider how the use of CF in eCALL can be informed by the rich literature on classroom feedback methods reviewed briefly in the last section. There are some findings from that literature that cannot apply directly. For example, the contrast between spoken and written feedback in eCALL is different from that in the classroom, because written feedback can be provided immediately by the computer with great diagnostic accuracy. Also, the comparison between instructor feedback and peer feedback would only be relevant in the eCALL context if lessons were configured within a social space. This is an interesting possibility, but not one which has yet been explored. However, there are three findings from classroom CF literature that may well apply to the eCALL context. First, the evidence for better uptake of CF in the form of prompts by younger learners (Lyster and Saito 2010) could suggest that eCALL instruction that targets younger learners should emphasize this form of CF. Second, the relation between high working memory capacity (WMC) and the ability to benefit from recasts (Sagarra 2007; Goo 2012) may also operate in the eCALL context. Third, the potential advantage of prompts over explicit correction in the classroom might possibly extend to eCALL instruction (Sanz and Morgan-Short 2004).

In the eCALL context, the difference between explicit corrective feedback and recasts is not as sharp as in the classroom context. This is because recasting only really makes sense as a conversational move. For example, if the learner translates *you went to the store* into Spanish as *tú fuí al mercado*, the computer's response would be *tú fuiste al mercado* in which the first person verb *fuí* is restated or recast as *fuiste*. The computer's response could be characterized as a recast, but it could equally well be viewed as explicit corrective feedback, particularly if the word *fuiste* is highlighted. Thus, the contrast between recasts and explicit correction may not be as sharp or interesting in eCALL as in classroom-based research.

On the other hand, the implementation of feedback through prompts in an eCALL program could be more interesting. First, recent classroom research has suggested that prompts may be more effective than recasts in promoting learning

(Ammar and Spada 2006; Ellis, Loewen et al. 2006; Ellis 2007; Sheen 2007; Yang and Lyster 2010). Second, implementing prompts through the computer is relatively easy. The standard method here is just to say *please try again*, but the computer can also deliver a simple repetition with a question mark at the end, clarification requests such as *pardon* or partial feedback that repeats the correct segment and asks the student to revise the incorrect segment. Third, there are good theoretical reasons to explore the role of prompts in eCALL. Tutorial design theory (Koedinger, Pavlik et al. 2008) has characterized this issue as "the assistance dilemma". The dilemma here is how to decide when instruction should provide information and assistance to students and when it should request students to generate this information. On the one hand, assistance can function as scaffolding, but on the other hand it can serve as a crutch. Similarly, asking students to generate their own correct forms can function as effective teaching or it can lead to confusion and imposition of an unnecessary cognitive load. Both Koedinger et al. and Long (1997) conclude, reasonably enough, that the answer to this dilemma depends largely on details of the material to be learned, as well as characteristics of the learner.

In practice, the study of contrasting effects of CF in eCALL is still in its infancy. The work we will consider here has only been able to examine a few segments of this general problem space. One area of emphasis has been on the effects of diagnostic CF, as opposed to simple correctness feedback. Zhang examined this contrast in the context of a system called the PinyinTutor (http://talkbank.org/pinyin) that helps learners of Chinese practice the dictation into Pinyin of Chinese words and phrases. Analyses of the effectiveness of this system showed that the experimental group that received diagnostic CF attained an 18% improvement in accuracy, as opposed to the control group subjects who received simple correctness feedback and who attained a 10% improvement across the duration of the study. This advantage for the training group was highly significant. In this task, there are many possible errors: the letters typed could be an illegal sequence in the Pinyin system, the initial or final sound of either syllable could be incorrect, there could be an incorrect number of syllables in the student response, and the tone of one or both syllables could be incorrect. In the diagnostic version of the program, students are given feedback regarding the exact nature of each error type. Moreover, each of these errors leads to different predictions about what the student needs to know, as well as different estimates of his or her current knowledge state (Gordon & Kowalski in press). In this system, response and latency data collected by the Flash program are transferred to servers at Carnegie Mellon University (CMU) for ongoing computation of adaptive feedback. The PinyinTutor data are also sent to the CMU DataShop repository (http://pslcdatashop.web.cmu.edu) for further offline analysis and possible future reanalysis. Providing correctness feedback immediately after the student types an answer makes correcting mistakes easier, and is important for student uptake (Ellis 2009), but this basic feedback can be supplemented by diagnostic CF targeted at the specific component of Pinyin typing that led to the error.

Speech recognition technology also has the potential to provide immediate feedback on pronunciation, a feature that requires a particular type of feedback "that does not rely on the student's own perception" (Ehsani and Knodt 1998). Refinement of pronunciation requires an outside observer to monitor and provide corrective feedback, as speakers are often unable to compare their pronunciation to a model. Cucchiarini, Neri and Strik (2009) further suggest that feedback on pronunciation should occur (a) in a stress-free environment, (b) in real-time, and (c) be individualized for each speaker. Computer software kits like EduSpeak® (Franco, Bratt et al. 2010) use phone-level mispronunciation detectors that, when they are able to produce reliable transcriptions, are comparable to human raters, and can provide feedback in real time and based on individual performance. These kits can be used to replace native speaker human listeners, which are often in short supply for second language learners. Speech recognition can also be used for the automatic scoring of oral fluency by focusing on temporal dynamics of speech, such as word count, length and rate of speech, and so on. The SpeechRater(TM) system uses this information, as well as rough estimates of number of repetitions and corrections, to provide real-time feedback to speakers on their oral performance (Zechner, Higgins et al. 2009). Speech recognition systems can provide feedback on both segmental (phonemes or syllables) and suprasegmental (prosody and intonation) features of the language (see Ehsani & Knodt 1998 for an overview).

2.3 Explicit rule instruction

In addition to varying the content of feedback, computers can implement different levels of explicitness in the presentation of linguistic patterns and rules. The value of explicit metalinguistic information for adult learners is a central and ongoing question in SLA theory. Meta-analyses (Norris and Ortega 2000; Spada and Tomita 2010) have indicated a general positive effect for explicit rule presentation. In an earlier review of this issue, MacWhinney (1997) concluded that explicit rule instruction was most useful when the rule to be learned was quite simple. In such cases, the rule can be kept active in working memory and used to match to incoming positive exemplars, thereby consolidating learning on both the explicit and implicit levels (MacWhinney 2012). For example, in a computerized study of the learning of the Spanish counterfactual conditional, Rosa and Leow (2004) found that the most effective instruction involved explicit corrective rule feedback concurrent with each test trial. The rule governing this construction in Spanish is fairly complex, but Rosa and Leow were able to formulate it on computer screens in a way that was concrete and memorable (p. 196).

A major problem with the conclusions reached in these meta-analyses is that the studies involved have typically involved untimed measures that may not assess the proceduralization of the relevant skills. When evaluated in this way, explicit instruction could be viewed as "teaching to the test." It is clear that research needs to disentangle the effects of explicit rule instruction from the testing of these effects. There are at least three ways in which this can be done. One is to administer posttests that require generalization of the newly acquired knowledge to constructions not involved in the training. For example, training of French gender marking on the article should be able to support improvement in gender marking on the adjective. Generalization tests of this type are rare, but they can prove quite useful in assessing this issue. A second way of testing generalization is to examine changes in speed of processing. We report some initial attempts in this direction below. A third way is to examine retention. Many of the studies discussed here have administered repeated posttests across intervals of weeks and months to assess long-term retention. The assumption here is that explicit rule formulations should be more prone to loss than implicit or proceduralized learning.

To illustrate how we can approach this issue, consider two multi-session eCALL training experiments for novice learners of French learning to categorize nouns by grammatical gender (Presson and MacWhinney in press). In these experiments, the rules for the orthographic cues to gender (e.g., words ending in *-age* are almost always masculine) were presented in an eCALL tutorial. Learners selected the gender and received correctness feedback with either no additional information, explicit orthographic cue statements (e.g., *-age* –> *le*), or highlighting of the relevant ending. The computerized training interface allowed for immediate feedback, randomization to feedback conditions, automatic data logging, and monitoring of participant progress. By computerizing the task, we were also able to test the prediction that, although explicit feedback might lead to better performance and greater retention after delay, this advantage could come at the cost of less rapid performance under time pressure, due to the additional time required to process explicit information. To that end, the program presented post-tests both with and without a time pressure constraint (a response deadline of 1400ms). This additional testing condition clearly showed that explicit cue feedback not only led to better learning and retention with no time pressure, but that explicit cue feedback led to greater accuracy even with a time pressure constraint, and that the addition of time pressure during training did not make a difference in the amount of improvement, suggesting there may be less of a trade-off between learning explicit cues or rules and rapid online behavioral performance than often assumed. This study also showed that, with only 90 minutes of practice, learners' ability to judge the gender of French words rose from 62% accuracy to 78% accuracy. Moreover, this ability was retained two months later, even though these novice learners were receiving no further exposure to French in the interval. The validity of the cues involved in this training was between 90% and 97%. However, this means that there are still some exceptions and the training did not include the exception words. Thus, a fuller study of learning of French gender in future experiments will need to include both training on the valid cue patterns and additional training on exception words to see how well the two types of training can be integrated.

2.4 Repeated practice and student modeling

CALL programs are well adapted to the task of providing practice in component language skills. At the very beginning of experimental psychology, Ebbinghaus learned about the ways in which memory for new items and associations could be promoted by distributed practice that focused on the generation of remembered items. The role for such graduated interval recall and generation in vocabulary learning has been studied frequently since then (Pimsleur 1967; Royce 1973; McNamara and Healy 1995; Barcroft 2007). A particular variant form of eCALL focuses specifically on the optimization of practice scheduling, using graduated interval recall. An example of this is Pavlik and Anderson's (2008) English-Japanese vocabulary program that demonstrated that practice is most efficient when the interval between practice trials starts small and gradually increases. Their program models ongoing advances in student knowledge to optimize the practice schedule and select practice trials accordingly. In this approach, a general group-based model of learning efficiency over spacing intervals is combined with past student data to estimate the current strength of student knowledge. Modelling these processes can determine when to present practice trials in order to lead to the greatest learning gains. The organization of vocabulary learning through graduated interval recall is only one of many features of eCALL design that can potentially lead to faster and better word learning. Barcroft (this volume) provides a thorough analysis of the many other factors that can support this process in terms of his input-based incremental (IBI) approach. These factors can be realized through eCALL systems that emphasize meaningfully resonant (MacWhinney, 2012) links between words in terms of paradigmatic associations, sensory associations, derivation, synonyms, and sentential context.

In addition to diagnosing the learner's knowledge state from behavioral responses, eCALL programs can allow for the integration of data from previous student behavior to improve feedback and trial selection. Zhao (2012) created an eCALL tutor to provide instruction on correct use of the English article system. She first created a list of cues for selecting which article (*a, an*, *the*, or no article) to use in various contexts. In particular, she differentiated between cues that are rule-based (e.g., when introducing new information into the conversation, use the indefinite article *"a"*, as in "I just bought a new car.") and those that are feature-based (e.g., when giving a street name, do not use an article, as in "turn onto Fifth Avenue" instead of "turn onto the Fifth Avenue"). By tracing student performance on practice trials with each type of cue, she found that students could acquire proper use of rule-based cues easily without explicit instruction, but explicit instruction was needed for learning of feature-based cues. After only three hours of using the English Article Tutor, students showed a 23% improvement in accuracy of article selection, moving from 53% accuracy to 76% accuracy (where chance accuracy is 33%).

In addition to feedback based on simple trial-by-trial error diagnosis, recent interventions outside the language domain (most notably Cognitive Tutor Algebra; Koedinger, Anderson, Hadley, & Mark 1997) diagnose student misconceptions with a quantitative model of student knowledge based on past performance. In this approach, immediate feedback can be responsive to specific estimates of a student's knowledge state. It can also take into account prior errors and help-seeking behavior in estimating the probability that a student has a specific wrong idea, thereby making remediation easier and more targeted to the individual student. Of course, the strategy of targeting feedback to specific learner errors and characteristics is not novel in eCALL – rather, it is a foundation of teacher expertise, as instructors track student performance and give appropriate feedback, as they deem necessary. It remains to be seen, however, to what degree adding specifically quantitative model-based individualized feedback can increase the amount of learning in a language task, especially because the quality of any targeted feedback is dependent on the quality of the underlying model. As theoretical descriptions of learner behavior and common errors continue to improve, so too can the effect of implementing individualized feedback in a computer training context.

In addition to training on grammar and vocabulary, computers can also be used to practice speaking. As described above, advances in speech recognition technology have allowed language learners to practice their pronunciation even without a teacher (for an overview of the use of speech recognition in language learning, see Eskenazi 2009). However, even without the use of speech recognition, speaking practice with a computer has shown to lead to improvements in both fluency and accuracy of speech. Yoshimura and MacWhinney (2007) had English learners of Japanese repeatedly read Japanese sentences and recite them from memory. They heard a sentence spoken by a native Japanese speaker, then read the sentence aloud six times, then repeated it from memory three times. Through these repetitions, they not only began to repeat the sentences with greater fluency but showed improvements in phonological accuracy of their repetitions as well.

Davy and MacWhinney (in preparation) used a task in which adult English learners of Spanish listened to a native Spanish speaker and repeat what they hear multiple times,. They found that this procedure leads to greater accuracy and fluency on later sentence production tasks compared to untrained sentences. Learners benefitted more from training that allowed them to practice complex sentences in separate phrasal groups than through complete sentence repetition. The study also showed how fluency training can be controlled through picture stimuli, thereby eliminating any reliance on pure echoic production of sentences (Erlam 2006).

Computerized practice is also useful in longer speaking tasks. De Jong and Perfetti (2011) created a computerized program to guide adult ESL students through the 4/3/2 task, an activity in which students are given a topic to talk about for four minutes, then three minutes, and finally two minutes. Traditionally this task is done with a partner, with the speaker giving the speech to three different people. However, de Jong and Perfetti found that students not only showed increased accuracy and fluency with each practice during the training period, but increased in fluency from pre-test and post-test, where the tests were completely unrelated speaking activities on untrained topics.

3. Automating trial generation and data collection

In addition to their aid in the rigorous control of pedagogically relevant variables, computers can also ease the arduous processes of material creation and data collection and analysis. In this next section we consider the use of computers in automatizing tasks that, when done by hand, can be prohibitively time-consuming: the generation of practice and testing materials, and data logging.

3.1 Automatic generation of materials

Technology can also aid in the generation of eCALL materials, either for practice (Brown and Eskenazi 2004; Heilman, Zhao et al. 2008) or testing (Feeney and Heilman 2008; Pino and Eskenazi 2009). This generation can be a difficult and time-consuming task for teachers and researchers, who must consider elements such as student level, relevant interests, subject matter, and representation of certain target vocabulary items or grammatical elements. Any degree of computerized automation of this task would be a great help to educators as they could then devote the time spent creating tests and training materials to other instructional activities.

For example, the REAP Project (or REAder's Practice) is designed to provide extensive reading practice for ESL learners, with the goal of increasing vocabulary.

This project scrapes the Internet for reading passages containing vocabulary from the Academic Word List (Coxhead 2000). The passages are then analyzed for content information and reading level so that students can read on topics that are both interesting for them and at a level appropriate for them. Whereas some of the passages are screened, either by researchers or by the students, many are simply mined and presented directly to students. The REAP Project also works to semi-automate the process of creating assessments to check whether students were reading the text, and measure vocabulary knowledge. For example, Feeney and Heilman (2008) discovered that simply by generating a list of unique words that appeared in the text and asking students to choose between that list and a similar list that also contains random words, they can see whether the students are actually reading. Student performance on this task correlated significantly with post-reading vocabulary assessments, suggesting that reading the passages leads to vocabulary growth.

Another example of semi-automated generation of practice trials comes from an L2 preposition tutor asking learners to move objects around a virtual room by following instructions presented in the L2 using spatial prepositions (Presson, MacWhinney et al. 2010) – available at http://talkbank.org/SLA/prepositions). For example, a learner can see (in Spanish) "Pick up the ball to the left of the plant and put it on the sink" and would click the object (ball) to move it to the target location (on the sink). The goal of this training intervention was to use spatial and enactment cues to strengthen the gains from practice comprehending preposition words. The tutor itself was created by manually segmenting target areas around each object (e.g., "above" and "under") such that many objects could be arranged into multiple configurations. For example, a fixed object such as a chair, or a movable object such as a ball, could have another object to its left or right, above or below it, or more generally near it. Therefore, using those manual specifications, the program was able to automatically generate a much larger number of room scenes, each containing several fixed and movable objects that could be manipulated when following L2 instructions. This trial generation would have been extremely difficult if each individual trial had to be created separately and by hand, and the degree of automation results in a practically unlimited number of trials from a relatively limited practice set (although a greater variety of objects and locations could be highly beneficial to motivation and generalization).

In addition to generating natural language stimuli for real learners, computerization also aids in the generation of artificial and miniature languages that can be used to conduct tightly controlled laboratory studies of language learning processes (Opitz and Friederici 2007; Morgan-Short, Sanz et al. 2010). These statistically reliable but sometimes complex stimuli can allow for systematic manipulation of language frequency and structure, as well as the processing difficulty for naive learners. For example, de Graaff (1997) used an artificial language in a computerized training paradigm to show that explicit instruction improved L2 grammar learning relative to no explicit instruction, and that this advantage was present for both simple and complex grammatical structures, and for both morphology and syntax. In this case, computerization was used to "optimally control the input and exposure" (p. 253), reflecting the advantage of both artificial grammar and computerized training.

3.2 Data logging

A core feature of eCALL is the capacity to log learner performance, instantaneously and automatically. With Internet connectivity, data from remote access to training materials can be logged to a central repository, and can be easily exported, manipulated, and analyzed. Without automatic logging, an experimenter often must code and log all responses manually, and sometimes must use coarse aggregate data (e.g., time to finish taking a test) instead of fine-grained trial- or student action-level data (e.g., time to finish each problem independently), which provide more detailed information about student behavior. Earlier, we noted how data from the PinyinTutor system (Zhang 2009) is transmitted continually to servers at CMU. This system is now in use at 40 locations internationally, yielding semester-long learning records from 3844 students. As these data come in, they are collected into web pages that allow each instructor to monitor the usage and progress of each of their students. These scores can be used as a component of the class grade, thereby freeing the instructor from the task of grading pinyin dictation assignments. At the same time, the computer provides students with summary scores that allow them to track their own progress.

The benefit of automatic logging is especially clear in the emerging field of the neuroscience of second language learning. In running a neuroimaging experiment, brain data must be matched to the behavioral events that trigger or reflect the activation being recorded. This correspondence allows researchers to analyze the brain correlates of behavioral and perceptual events, which can provide key evidence into the mechanisms of second language learning (Hernandez, Hofmann et al. 2007; Abutalebi 2008; Osterhout, Poliakov et al. 2008; van Hell and Tokowicz 2010). Without a fine-grained temporal log of behavioral events and stimulus presentation, used to tag events on the resulting brain data, it would be difficult or impossible to associate changes in brain activity to events in the world. For example, Morgan-Short and colleagues (2010) showed different event-related potential (ERP) responses in high and low proficiency L2 learners for noun-article and noun-adjective agreement errors in a computer-generated artificial grammar, which was possible because of the ability to match automatically generated time-stamped stimulus presentation logs (thereby differentiating signals from article and adjective trials) to corresponding ERP data.

3.3 Limitations to computerization

Although eCALL expands the scope of interventions and data collection beyond what is possible with non-computerized training, there are important limitations in current technology that make it difficult or impossible to shift all language training to the computer.

First, although speech recognition technology continues to advance, it is not yet reliable enough to be used alone as an automated feedback mechanism for second language learners. Systems that use speech recognition must anticipate a certain rate of failure, either in terms of *misses* (saying the speech is correct when it's not) or *false alarms* (saying the speech is incorrect when it's not). To minimize frustration on the part of the user, programmers typically will err on the side of maximizing misses. Even given this bias, there will still be some frustrating false alarms to sounds that are acceptable, but which are judged as unacceptable. As a result, speech recognition cannot always be a reliable source for pronunciation training. In addition, speech recognition cannot yet support the interpretation of full sentence input. There have been a few programs that have attempted to create immersion-like speaking environments, but in order to make speech recognition possible, they must use highly constrained tasks, such as providing a question for the student to answer and a list of different possible answers to choose from. The speech detection mechanism can then choose which possible answer was uttered and give feedback. However, the answers must be different enough that the speech detection can choose between them, limiting the usefulness of this program (for an overview of spoken dialogue systems, see Eskenazi 2009). These limitations in speech recognition technology mean that training activities that improve comprehension (which does not require but can include speech output) are easier to computerize with automatic feedback than interventions targeted at oral production (which is necessarily dependent on speech output).

Second, the interactional aspects of language acquisition are difficult to replicate on the computer. It is generally accepted (Long 1983; Long 1996; Gass, Mackey et al. 1998) that there are clear benefits of conversational interaction with native speakers. In the Unified Competition Model, MacWhinney (2012) argues that, in order to maximize second language learning progress, we want to couple the benefits of eCALL with the benefits of conversational interaction. There are many benefits from conversational interaction for both comprehension (VanPatten 2011) and production . Moreover, social participation can support acculturation (Pavlenko and Lantolf 2000; Firth and Wagner 2007), the growth of the use of L2 for inner speech or "thinking in the second language", and improved motivation for learning (Dörnyei 2009). However, providing socially supportive contexts for learners can be a challenge. Chat sites like LiveMocha (http://www.livemocha.com) or systems such as Second Life (http://secondlife.com/destination/chinese-island) provide one solution to this problem, using computer access as a way to facilitate learner connection to human interlocutors rather than attempting to automatize or simulate naturalistic conversation. There is evidence that realtime computer-mediated communication (CMC) in chat rooms leads to higher levels of learner output. It has also been suggested that CMC may provide opportunities for negotiation of meaning of the type observed in face-to-face interactions (Varonis and Gass 1985). However, Peterson (2006) found that these systems produce only low levels of meaning negotiation. In terms of the overall effects of chat rooms, there have been no controlled studies, and student reactions to these systems are often mixed (Yao 2009; Peterson 2010).

Finally, the cost of setup and implementation of computerized training serve as limitations to the rapid spread of this technology. During the early stages of development, programming and equipment costs are a central obstacle to constructing eCALL exercises. The REAP Project (Heilman, Juffs et al. 2007), for example, requires the scanning of potential reading activities for appropriateness and student interest. Although the program can scrape the Internet for content, at this stage in its development the intervention is dependent on human readers to screen that content to ensure its appropriateness for the study. Creating models of student knowledge, such as the ones used in Cognitive Tutor Algebra (Koedinger, Anderson, Hadley & Mark 1997) requires a relatively large up-front investment to develop, test, and refine a usable model to track student learning. These models are domain-specific, meaning that each skill requires its own model, which must be developed from basic principles. In the language domain, this means that each language requires its own model and that transfer between language models may be limited. In addition, speech recognition programs also often require the initial loading in of expected speech to use for comparison to the inputs (Eskenazi 2009). However, there are also costs to traditional classroom or interactive activities, namely the human capital required to create interactive contexts or administer those activities. In the long run, computerized training may save time and human capital, but the up-front investment continues to be a consideration for educators and researchers alike.

4. Future directions in eCALL

As we learn how to take advantage of these ongoing technological advances, we can begin to see how to configure eCALL in interestingly novel ways. In particular, we can develop eCALL methods for (1) studying large data sets, (2) tracking usage patterns in non-experimental settings with mobile devices, and (3) recording data from naturalistic interactions.

4.1 Experiments with large data sets

One key opportunity provided by eCALL exercises is the ability to collect massive amounts of data from a broad learner population. Advances in data mining and model discovery algorithms can improve the scope of analysis possible with existing data. In addition, the possibility of collecting data from a large number of learners (e.g., releasing publicly usable mobile training activities) gives leverage to test not only the main effects of instructional variables on learning effectiveness, but also the interactions, both simple and higher order, that require substantially more statistical power to effectively test than is often available in traditional lab or classroom experiments. This can allow researchers to expand language models beyond the positive effect of, for example, explicit instruction, to include the interaction of those effects with other instructional properties and individual differences.

For example, in two experiments using a training intervention for typing conjugated Spanish verbs at the very early stages (first and third semester) of undergraduate classroom instruction, data were collected from over 1000 learners across multiple training and testing sessions, resulting in a massive database of written production accuracy data. Learners completed 90 minutes of training (three 30-minute sessions over one month) typing conjugated verbs. Learners in first semester Spanish benefitted the most from this training, improving their accuracy rate by about 18%. Learners were provided with correctness feedback and the correct target response. The training task included several different verb contrasts, designed to test the prediction that accuracy in conjugation was predictable from properties of the verb (e.g., type frequency) and instruction (e.g., instructional sequence): present compared to past (preterite) tense, fully regular compared to subregular (stem and spelling change) verbs, and default *-ar* compared to non-default *-er* and *-ir* verbs in a fully crossed design. The large dataset showed higher-order interactions reflecting the fact that difficulty factors (e.g., preterite tense is harder than present tense) were sometimes additive (e.g., preterite tense non-*ar* verbs were harder than preterite tense *-ar* verbs, which in turn were harder than present tense non-*ar* verbs), and revealed that both baseline accuracy and the amount of improvement after training can be predicted by a combination of these verb properties.

By collecting trial-level data, and by automatically logging student performance, the marginal cost of collecting data from additional learners is greatly reduced, and collecting larger and systematically sampled datasets becomes easier. By collecting over 75,000 observations of student production of conjugated Spanish verbs, this computerized training intervention provided data painting a detailed picture of the interaction of verb properties in increasing not only the baseline difficulty of conjugation, but also the size of gains after training. The large sample size meant that such a dataset was possible using a test that took only 5–10 minutes per student. The pattern of data in this case both emphasizes the need for an explicitly developmental model of production of verb morphology and makes future predictions about which verbs should be most targeted in production training.

In addition, large datasets have the statistical power to begin to make more productive inferences from null effects than is typically possible. In a comparison of explicit feedback and analogy to a familiar example, three separate experiments with varying sample size showed a null difference in amount of improvement for all verb types (Presson, Sagarra et al. in press). In a typical classroom or laboratory sample, it would be difficult to make inferences from this null effect; however, given both the large sample sizes and replication across samples in the computerized, large-scale training context, these effects can be taken as suggestive evidence that the two interventions were equally effective (i.e., the true effect size of the difference is negligible). This result suggests that the mechanism of improvement after training is the provision of correctness feedback itself, and that the additional information of an example or rule statement was either not useful or equally useful across these samples, and this suggestion is only meaningful in the context of large samples and multiple replications.

4.2 Mobile computing and usage patterns

Increasingly, the future of CALL and eCALL will be linked to the use of mobile devices. With the introduction of devices such as the iPhone, the iPad, and the Android tablets, the very meaning of "computer" has begun to shift. As computing devices become smaller, more mobile, and more ubiquitous, there will be more and more ways to use these technologies in socially flexible contexts. One of the largest disadvantages faced by adult second language learners is the greatly reduced time on task for a classroom learner compared to a child first language learner. This difference in time on task may be a major force producing positive outcomes for study abroad programs (DeKeyser 2007). By providing more such opportunities to practice L2 skills, ubiquitous computing can also improve learning outcomes. Currently, learners can use the iPad to access gamelike applications for vocabulary training such as uTalk or MindSnacks. Links to many of these applications can be found at (http://talkbank.org/SLA/ipad.html). If the tablet is linked to the Internet through either wireless or a cellular network, it can be used to listen to audio or watch television on L2 websites. Beyond that, teachers can configure contextualized lessons using systems such as Google Earth tours or map-based games systems (Hoshino, Saito et al. 2009) that direct the learner through specific interactions in the community. At this point, the focus of eCALL program development and research will shift to careful coordination of basic skills training with support of mobile computing by the classroom instructor. This coordination will require an increasingly close collaboration between researchers, programmers, and instructors.

There are many strategies that commercial training programs use to increase motivation: game-like features such as scoring mechanisms and social collaboration, prioritizing graphic design, adaptive difficulty, and de-emphasis of explicit grammar instruction, to name a few. However, we must be cautious in applying all of these features indiscriminately, as it remains to be seen whether such manipulations of the training environment can improve learner compliance to a training regimen, and whether they improve or harm learner outcomes in various language skills.

In addition to locally stored computer software, the Internet also allows for the creation of websites (therefore accessible on multiple hardware systems) that can provide instruction or practice. Websites like Memrise (http://www.memrise.com) or Anki (http://ankisrs.net) are centered around vocabulary drills, allowing users to review vocabulary at any time on their computers or mobile devices, either by using provided "decks" of words or by creating their own custom lists. Memrise in particular leverages research on gaming features (Aleven, Myers et al. 2008), practice scheduling (Pavlik and Anderson 2008), and encoding elaboration (Craik and Lockhart 1972; Ellis 1995; Barcroft 2002; Barcroft and Sommers 2005) to teach and practice vocabulary in a number of languages. The social collaborative nature of the program also allows any user to create a vocabulary list, allowing learners of any language to use the site. Websites that are free to use and readily available provide ample opportunities for language learners to receive instruction and practice their target language at any time and place, and have the potential to become a huge source of data for further research on computer-supported language learning.

Given these new technological possibilities, we can begin to think in terms of an extended eCALL system that incorporates all of these learning methods within a single platform. Let us refer to this integrated platform as the Language Partner. This system can be programmed using HTML5 and the Google Web Toolkit (GWT) so that it can run equally well on mobile devices or desktop computers, both relying on information transmitted over WiFi connections to the Internet. This system will include many of the components described earlier: basic skills and vocabulary tutors, games, interactive media access, and situated learning activities. The various facilities will continually record student usage patterns and responses that will be logged into DataShop files for subsequent student modelling and linkage to classroom activities. Figure 1 illustrates the general shape of a Language Partner system.

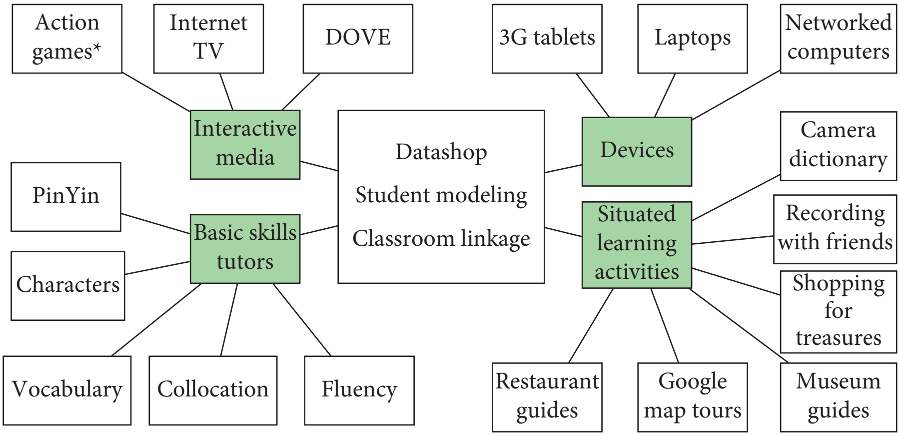

Figure 1. The Language Partner System

4.3 Computerized studies of naturalistic interactions

An important component of the Language Partner is the use of mobile devices to obtain photos, videos, and recordings of interactions in the community. The goal here is for the student to record real-life interactions of the type promoted in the Språkskap Project (Clark, Wagner et al. 2011) for later analysis in the classroom. For example, a learner of Icelandic recorded her interactions in a bakery in Rejkjavik. These data were transcribed and the transcripts were linked to audio records. The resulting corpus, called IceBase, was used by Guðrún Theodórsdóttir as the basis for her dissertation in the Conversation Analysis framework. These data are available to researchers from http://talkbank.org. By using applications such as *Recorder* in the iPad, or recording H.264 movies, learners can record interactions with native speakers in sites such as restaurants, museum tours, excursions, or homes. Like the Icelandic corpus, these records can then be analyzed either for pedagogical or research purposes. In the classroom, these materials could help students understand conversational practices, pragmatic norms, linguistic forms, and methods for negotiating meaning. For researchers, the corpora can be analyzed by programs such as CLAN (MacWhinney, Fromm et al. in press) for automatic lexical and morphosyntactic analysis or Praat (Boersma and Weenink 1996) and Phon for phonological analysis (Rose and MacWhinney in press). Within Praat, researchers (de Jong and Wempe 2009) are developing methods for linking transcripts to audio at the word level. Once these methods are available, we will be able to conduct increasingly powerful analyses for fluency and phonological accuracy.

5. Conclusion

Technology has always expanded the realm of the possible in human learning and education. Advances such as the printing press, the telephone, the motion picture, and the television have all had their impact on second language learning. The desktop computer, the Internet, and tablet computing are now further transforming how we can learn second languages. Constant access to mobile training exercises, complex algorithmic modelling of student behavior, immediate feedback, subtitled video, and automated trial selection and randomization, among other advances, can help adult learners compensate for the often challenging and difficult circumstances that confront them when beginning to learn a second language. Here, we have presented the results from studies examining the use of computer tutors for both novice learners and in real course contexts. We have shown that these methods can lead to rapid increases in competence for features such as French gender assignment, English article usage, Spanish fluency, Spanish verb conjugation, and Pinyin dictation. However, the full power of this approach will not be realized until a much fuller set of resources is created for each relevant language. Once these resources become available, they can lead to a revolution in SLA research and theory, as we automatically log data, adapt student feedback, and conduct large-scale experiments with relatively low cost. Ideally these advances in theory, experimentation, and training can be coordinated with best practices from classroom pedagogy to lead to breakthroughs in the learning of second languages.

References

Abutalebi, J. (2008). Neural aspects of second language representation and language control. Acta Psychologica, 128, 466–478.

Aleven, V., Eben Myers, E., Easterday, M., & Ogan, A. (2008). Toward a framework for the analysis and design of educational games. In G. Biswas, D. Carr, Y.S. Chee, & W.Y. Hwang (Eds.), Proceedings of the 3rd IEEE Conference on Digital Game and Intelligent Toy-Enhanced Learning (pp. 69–76). Los Alamitos, IEEE Computer Society.

Ammar, A., & Spada, N. (2006). One size fits all? Recasts, prompts, and L2 learning. Studies in Second Language Acquisition, 28, 543–574.

Bangert-Drowns, R., Kulik, C.-L., Kulik, J.A., & Morgan, M.T. (1991). The instructional effects of feedback in test-like events. Review of Educational Research, 61: 213–238.

Barcroft, J. (2002). Semantic and structural elaboration in L2 lexical acquisition. Language Learning, 52, 323–363.

Barcroft, J. (2007). Effects of opportunities for word retrieval during second language vocabulary learning. Language Learning, 57, 35–56.

Barcroft, J., & Sommers, M. (2005). Effects of acoustic variability on second language vocabulary learning. Studies in Second Language Acquisition, 27, 387–414.

Boersma, P., & Weenink, D. (1996). Praat, A system for doing phonetics by computer. Tech. Rep. 132. Amsterdam: Institute of Phonetic Sciences of the University of Amsterdam.
Brown, J., & Eskenazi, M. (2004). Retrieval of authentic documents for reader-specific lexical practice. InSTIL/IOCALL Symposium, Venice.
Cen, H., Koedinger, K., & Junker B. (2006). A general method for cognitive model evaluation and improvement. Intelligent Tutoring Systems, 164–175.
Chenoweth, A., Ushida, E., & Murday, K. (2006). Student learning in hybrid French and Spanish courses: An overview of language online. CALICO Journal, 24, 115–145.
Clark, B., Wagner, J., Lindemalm, K., & Bendt, O. (2011). Språkskap: Supporting second language learning in the wild. INCLUDE 11. Available on: <http://soda.swedish-ict.se/5446/1/F514_1578.PDF>
Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34, 213–238.
Craik, F.I.M., & Lockhart, R.S. (1972). Levels of processing: A framework for memory research. Journal of Verbal Learning and Verbal Behavior, 11, 671–684.
Cucchiarini, C., Neri, A., & Strik, H. (2009). Oralprodiciency training in Dutch L2: The contribution of ASR-based corrective feedback. Speech Communication, 51, 853–863.
Davy, C., & MacWhinney, B. (In preparation). Fluency training through picture description. Language Learning.
de Bot, K. (1996). The psycholinguistics of the output hypothesis. Language Learning, 46, 529–555.
de Graaff, R. (1997). The eXperanto experiment: Effects of explicit instruction on second language acquisition. Studies in Second Language Acquisition, 19, 249–275.
de Jong, N., & Perfetti, C. (2011). Fluency training in the ESL classroom: An experimental study of fluency development and proceduralization. Language Learning, 61, 533–568.
de Jong, N., & Wempe, T.O.N. (2009). Praat script to detect syllable nuclei and measure speech rate automatically. Behavior Research Methods, 41, 385–390.
DeKeyser, R. (Ed.). (2007). Practice in a second language: Perspectives from applied linguistics and cognitive psychology. Cambridge: CUP.
DeKeyser, R. (2007). Study abroad as foreign language practice. In R. DeKeyser, Practice in a second language: Perspectives from applied linguistics and cognitive psychology (pp. 208–226). Cambridge: CUP.
Dörnyei, Z. (2009). The psychology of second language acquisition. Oxford: OUP.
Edwards, J., & Zampini, M. (Eds.). (2008). Phonology and second language acquisition. Amsterdam: John Benjamins.
Ehsani, F., & Knodt, E. (1998). Speech technology in computer-aided learning: Strengths and limitations of a new CALL paradigm. Language Learning and Technology, 2, 45–60.
Ellis, N.C. (1995). The psychology of foreign language vocabulary acquisition and its implications for CALL. International Journal of Computer Assisted Language Learning, 8, 103–128.
Ellis, R. (2007). The differential effects of corrective feedback on two grammatical structures. In A. Mackey (Ed.), Conversational interaction in second language acquisition: A collection of empirical studies (pp. 339–360). Oxford: OUP.
Ellis, R. (2009). Corrective feedback and teacher development. L2 Journal, 1, 3–18.
Ellis, R., Loewen, S., & Erlam, R. (2006). Implicit and explicit corrective feedback and the acquisition of L2 grammar. Studies in Second Language Acquisition, 28, 339–368.
Erlam, R. (2006). Elicited imitation as a measure of L2 implicit knowledge: An empirical validation study. Applied Linguistics, 27, 464–491.

Eskenazi, M. (2009). An overview of spoken language technology for education. Speech Communication, 51, 832–844.
Feeney, C., & Heilman, M. (2008). Automatically generating and validating reading-check questions. Ninth International Conference on Intelligent Tutoring Systems.
Felix, U. (2005). Analyzing recent CALL effectiveness research: Towards a common agenda. Computer Assisted Language Learning, 18, 1–32.
Firth, A., & Wagner, J. (2007). Second/foreign language learning as a social accomplishment: Elaborations on a reconceptualized SLA. Modern Language Journal, 91, 800–819.
Franco, H., Bratt, H., Rossier, R., Venkata Rao, G., Shriberg, E., Abrash, E. & Precoda, K. (2010). EduSpeak®: A speech recognition and pronunciation scoring toolkit for computer-aided language learning applications. Language Testing, 27, 401–418.
Gardner, R., & Wagner, J. (2005). Second language conversations. London: Continuum.
Gass, S., A. Mackey, & Pica, T. (1998). The role of input and interaction in second language acquisition: An introduction. Modern Language Journal, 82, 299–307.
Goo, J. (2012). Corrective feedback and working memory capacity in interaction-driven L2 learning. Studies in Second Language Acquisition, 34, 445–474.
Hardison, D. (2005). Contexutualized computer-based L2 prosody training: Evaluating the effects of discourse context and video input. CALICO Journal, 22, 175–190.
Heilman, M., Juffs, A., & Eskenazi, M. (2007). Choosing reading passages for vocabulary learning by topic to increase intrinsic motivation. 13th International Conference on Artificial Intelligence in Education. Marina del Rey, CA.
Heilman, M., Zhao, L., Pino, J. & Eskenazi, M. (2008). Retrieval of Reading Materials for Vocabulary and Reading Practice. The 3rd Workshop on Innovative Use of NLP for Building Educational Applications, Association for Computational Linguistics.
Hernandez, A.E., Hofmann, & J., Kotz, S. (2007). Age of acquisition modulates neural activity for both regular and irregular syntactic functions. NeuroImage, 36, 912–923.
Hoshino, J., Saito, T., & Kazuto, S. (2009). Task-based second language learning game system. In Proceedings of the 8th International Conference on Entertainment Computing (pp. 323–324). Paris: Springer.
Karpicke, J., & Roediger, H. (2006). Repeated retrieval during learning is the key to long-term retention. Journal of Memory and Language, 57, 151–162.
Koedinger, K.R. (2011). Bridging the lab and the field with /in vivo/ experimentation. Society for Research on Educational Effectiveness (SREE).
Koedinger, K.R., Anderson, J., Hadley, W., & Mark, M. (1997). Intelligent tutoring goes to school in the big city. International Journal of Artificial Intelligence in Education, 8, 30–43.
Koedinger, K.R., Pavlik, P., Mclaren, B., & Aleven, V. (2008). Is it better to give than to receive? The assistance dilemma as a fundamental unsolved problem in the cognitive science of learning and instruction. In B.C. Love, K. McRae & V.M. Sloutsky (Eds.), Proceedings of the 30th Annual Conference of the Cognitive Science Society, Austin, TX, Cognitive Science Society (pp. 2155–2160).
Levelt, W.J.M. (1989). Speaking: From intention to articulation. Cambridge, MA: The MIT Press.
Liu, M., Z. Moore, Graham, L., & Lee, S. (2002). A look at the research on computer-based technology use in second language learning: A review of the literature from 1990–2000. Journal of Research on Technology in Education, 34, 250–273.
Long, M. (1983). Native speaker/non-native speaker conversation and the negotiation of comprehensible input. Applied Linguistics, 4, 126–141.

Long, M. (1996). The role of the linguistic environment in second language acquisition. In W.C. Ritchie & T.K. Bhatia (Eds.), Handbook of second language acquisition (pp. 413–468). New York, NY: Academic Press.

Long, M. (1997). Problems in SLA. Mahwah, NJ: Lawrence Erlbaum Associates.

Long, M. (2000). Focus on form in task-based language teaching. In R. Lambert & E. Shohamy (Eds.), Language policy and pedagogy: Essays in honor of A. Ronald Walton (pp. 179–192). Amsterdam: John Benjamins.

Lyster, R. (2004). Differential effects of prompts and recasts in form-focused instruction. Studies in Second Language Acquisition, 26, 399–432.

Lyster, R., Saito, K., & Sato, M. (2013). Oral corrective feedback in second language classrooms. Language Teaching, 46, 1–40.

Lyster, R. & Saito, T. (2010). Oral feedback in classroom SLA. Studies in Second Language Acquisition, 32, 265–302.

MacWhinney, B. (1995). Evaluating foreign language tutoring systems. In V.M. Holland, J.D. Kaplan & M.R. Sams (Eds.), Intelligent language tutors: Theory shaping technology (pp. 317–326). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (1997). Implicit and explicit processes. Studies in Second Language Acquisition, 19, 277–281.

MacWhinney, B. (2012). The logic of the Unified Model. In S. Gass & A. Mackey (Eds.), The Routledge handbook of second language acquisition (pp. 211–227). New York, NY: Routledge.

MacWhinney, B., Fromm, D., Forbes, & Holland, A. (In press). AphasiaBank: Data and Methods. In N. Mueller & M. Ball (Eds.), Research methods in clinical linguistics and phonetics (pp. 268–287). New York, NY: Wiley.

McNamara, D.S., & Healy, A.F. (1995). A generation advantage for multiplication skill training and nonword vocabulary acquisition. In A.F. Healy & L.E. Bourne (Eds.), Learning and memory of knowledge and skills: Durability and specificity (pp. 132–169). Thousand Oaks, CA: Sage.

Morgan-Short, K., Sanz, C., Steinhauer, K., & Ullmann, M.T. (2010). Second language acquisition of gender agreement in explicit and implicit training conditions: An event-related potential study. Language Learning, 60, 154–193.

Morgan-Short, K., & Ullman, M.T. (2011). The neurocognition of second language. In S. Gass & A. Mackey (Eds.), The Routledge handbook of second language acquisition (pp. 282–300). New York, NY: Routledge.

Norris, J.M., & Ortega, L (2000). Effectiveness of L2 instruction: A research synthesis and quantitative meta-analysis. Language Learning, 50, 417–528.

Opitz, B., & Friederici, A. (2007). Neural basis of processing sequential and hierarchical syntactic structures. Human Brain Mapping, 28, 585–592.

Osterhout, L., Poliakov, A., Inoue, K, McLauglin, J., Valentine, G., Pitkanen, I., Frenck-Mestre, C., & Hirshensohn, J. (2008). Second-language learning and changes in the brain. Journal of Neurolinguistics, 21, 509–521.

Pavlenko, A., & Lantolf, J. (2000). Second language learning as participation and the (re)construction of selves. In A. Pavlenko & J. Lantolf, Sociocultural theory and second language learning (pp. 155–178). Oxford: OUP.

Pavlik, P.I. & Anderson, J. (2008). Using a model to compute the optimal schedule of practice. Journal of Experimental Psychology: Applied, 14, 101–117.

Peterson, M. (2006). Learner interaction management in an avatar and chat-based virtual world. Computer Assisted Language Learning, 19, 79–103.

Peterson, M. (2010). Computerized games and simulations in computer-assisted language learning: A meta-analysis of research. Simulation and Gaming: An Interdisciplinary Journal, 41(1): 72–93.

Pimsleur, P. (1967). A memory schedule. Modern Language Journal, 51, 73–75.

Pino, J., & Eskenazi, M. (2009). Semi-automatic generation of cloze question distractors' effect of students' L1. SLaTE Workshop on Speech and Language Technology in Education. Pittsburgh, PA.

Presson, E., Sagarra, N., MacWhinney, B., & Kowalski, J. (In press). Compositional production in Spanish second language conjugation. Bilingualism: Language and Cognition.

Presson, N., & MacWhinney, B. (In press). Learning grammatical gender: The use of rules by novice learners. Applied Psycholinguistics.

Presson, N., MacWhinney, B., & Heilman, M. (2010). A simulated environment for the embodied learning of Spanish prepositions. Second Language Research Forum, University of Maryland.

Robinson, P. (1995). Attention, memory, and the noticing hypothesis. Language Learning, 45, 285–331.

Rosa, E.M., & R. P. Leow (2004). Computerized task-based exposure, explicitness, type of feedback, and Spanish L2 development. Modern Language Journal, 88, 192–216.

Rose, Y., & MacWhinney, B. (In press). The Phon and PhonBank initiatives.

Royce, J.M. (1973). Memory effects for test-like events during acquisition of foreign language vocabulary. Psychological Reports, 32, 195–198.

Sagarra, N. (2007). From CALL to face-to-face interaction: The effect of computer-delivered recasts and working memory on L2 development. In A. Mackey (Ed.), Conversational interaction in second language acquisition: A collection of empirical studies (pp. 229–248). Oxford: OUP.

Sanz, C., & Morgan-Short, K. (2004). Positive evidence versus explicit rule presentation and explicit negative feedback: A computer-assisted study. Language Learning, 54, 35–78.

Sato, M., & Lyster, R. (2012). Peer interaction and corrective fedback for accuracy and fluency development. Studies in Second Language Acquisition, 34, 591–626.

Schmidt, R. (1993). Awareness and second language acquisition. Annual Review of Applied Linguistics, 13, 206–226.

Sheen, Y. (2007). The effects of corrective feedback, language aptitude, and learner attitudes on the acquisition of English articles. In A. Mackey (Ed.), Conversational interaction in second language acquisition: A collection of empirical studies (pp. 301–322). Oxford: OUP.

Spada, N. & Tomita, Y. (2010). Interactions between type of instruction and type of language features: A meta-analysis. Language Learning, 60, 263–308.

Swain, M. (2005). The output hypothesis: Theory and research. In E. Hinkel (Ed.), Handbook of research in second language teaching and learning (pp. 471–484). Mahwah, NJ: Lawrence Erlbaum Associates.

van Hell, J., & Tokowicz, N. (2010). Event-related brain potentials and second language learning: Syntactic processing in late L2 learners at different L2 proficiency levels. Second Language Research, 26, 43–74.

VanPatten, B. (2011). Input processing. In S. Gass & A. Mackey (Eds.), The Routledge handbook of second language acquisition. New York, NY: Routledge.

Varonis, E., & Gass, S. (1985). Non-native/Non-native conversations: A model for negotiating meaning. Applied Linguistics, 6, 71–91.

Yang, Y., & Lyster, R. (2010). Effects of form-focused practice and feedback on Chinese EFL learners' acquisition of regular and irregular past tense forms. Studies in Second Language Acquisition, 32, 235–263.

Yao, T.-c. (2009). The current status of Chinese CALL in the United States. Journal of Chinese Language Teachers Association, 44, 1–23.

Yoshimura, Y., & MacWhinney, B. (2007). The effect of oral repetition in L2 speech fluency: System for an experimental tool and a language tutor. SLATE Conference (pp. 25–28).

Zechner, K., Higgins, D., Xi, X. & Williamson, D.M. (2009). Automatic scoring of nonnative spontaneous speech in tests of spoken English. Speech Communication, 51, 883–895.

Zhang, Y. (2009). A tutor for learning Chinese sounds through pinyin. Modern Languages. Pittsburgh, Carnegie Mellon University.

Zhao, Y. (2012). Explicitness, cue competition, and knowledge tracing: A tutorial system for second language learning of English article usage. Modern Languages. PIttsburgh, Carnegie Mellon University.