

Technology Enhanced Assessment Proposal: Technology-Based Formative
Assessment in Second Language Learning

Elizabeth Dalton

Boston College

A) Definition of problem with support from the literature

Learning a second language, particularly in a largely monolingual culture like that in the United States, is a daunting task, but a rewarding one. One essential component of success is regular, high-quality formative assessment. (Ehsani & Knodt, 1998; Zhang, 1998). Contemporary theories of language learning focus on the communicative properties of language (as opposed to "surface" linguistic properties such as grammar or morphology), and it would therefore seem that the most appropriate kind of assessment (formative or summative) would be communication tasks with rating and feedback provided by someone fluent in the language and in methods of language teaching (Richards & Rodgers, 2003; McNamara, 1996, 2000). However, human raters in language assessment have several drawbacks. One is their inter-rater reliability, which varies in a well-documented way, sufficiently to confound measurements of learner skill (McNamara, 1996). Contemporary statistical methods such as Rasch Factor Analysis can aid in the rater reliability problem, but require sophisticated analytical tools and skill to implement, not always readily available to classroom teachers. Another more critical issue is experienced rater availability. For rapid progress in language learning, learners need regular, frequent practice opportunities. But language instructors or tutors are often only available for a few hours a week, and, as described below, may be entirely unavailable in some communities.

Computer-based learning might provide the accessible, tailored, structured environment needed to support language learning -- if it can provide adequate opportunities for true practice, with feedback, of the four essential language skills: listening, speaking, reading, and writing. The issue of providing reliable, quality speaking practice with feedback, in particular, is the subject of this proposal.

Listening and reading practice materials are easy to present with contemporary multi-media machines, even when non-Latin scripts are involved (Ehsani & Knodt, 1998). For languages with concise alphabets or syllabaries (e.g. European and African languages), keyboard overlays allow for practice in writing the language, or at least inputting constructed answers. Handwriting recognition packages are also now available for most major languages of interest, which may be used to support practice especially in writing non-Latin alphabets and syllabaries as well as more extended ideographic systems as are used in many Asian languages (Ge, 2003; Twinbridge Software Corporation). But until recently, the most difficult of the four skills to support technologically was also one of the most basic: speaking. Fortunately, recent technological improvements may have finally caught up with this difficult challenge.

Software-based speech recognition has improved tremendously over the past decade, and preliminary research has begun to show that the quality may now be sufficient to provide useful feedback to learners in areas such as pronunciation and prosody (Eskenazi, 1999; Hardison, 2004). In fact, commercial products have appeared which claim to do just that. One such example is the *Rosetta Stone* line of language learning software. The *Rosetta Stone* software gives learners sample phrases in the form of images (usually photographs),

text, and audio clips. Learners are then asked to produce the phrases when cued by the images or text. (Other non-speaking activities in which learners match audio to images or written text are also provided.) Learners are provided with a "meter" which indicates how close they have come to a native speaker's pronunciation. They can also record and hear their own voices producing the phrases, and compare these recordings with recordings of native speakers of the target language. (Fairfield Language Technologies a).

The problem with this approach is that although the *Rosetta Stone* designers have been thorough in their language coverage, the activity provided by the software itself is neither engaging nor motivating (Discovery Channel, 2003). The repetitive four-panel flashcard approach soon bores even the most enthusiastic and committed learners, and is especially tiresome for young children, who, as described below, are well suited to the acquisition of a second language. The *Rosetta Stone* designers cite the need for engaging communicative activities in their research documentation (Fairfield Language Technologies b), but have missed the mark in their execution, because the activities offer no meaning for the learner. Additionally, and even more critically, the "meter" approach does not give specific guidance on the error being made by the learner, or feedback tailored to the error being made, but only a score indicating how close the learner has come to "perfect" native speech -- more of a summative type of feedback than a formative one. Often the error is one of rhythm difference (prosody) rather than pronunciation, but a beginning learner, still working on mastering the different phonemes of a new language, may need more specific guidance. This is especially true in languages with tones, when being learned by speakers of non-tone languages. Finally, the "record and compare" function is virtually useless for beginning speakers, who often do not have the listening skills to distinguish between the quality of their own speech and that of the native speaker, especially when key phonological differences between the learner's own language and the target language exist (again, tones are a good example) (Ehsani & Knodt, 1998). The graphic displays of base frequency and spectrographs are potentially more useful, but only with sufficient training (Sokolik, 1999; Zhang, 1998), which is not provided in the software program. (Similar problems with poor and non-specific feedback exist with the other learning modes of The Rosetta Stone, negatively affecting learners in the areas of reading, for example. See Ahmed, 2003, Cheapshop, and Kaiser, 1997 for criticisms of the pedagogy and feedback mechanisms in The Rosetta Stone.)

What is needed is an engaging way to involve learners in communicative tasks which provide the necessary specific feedback to help learners. For beginning learners, this will often be phonological feedback, gradually broadening to prosodic, semantic, grammatical, and pragmatic feedback for more advanced learners. But the technological means of providing formative assessment, while important, must also be paired with sound design of activities which will be meaningful to the learners.

This situation may present an opportunity for the effective use of computer-assisted language learning (CALL) in the specific area of language speech acquisition, if it can be provided at low enough cost to be made available in public schools, public libraries and other venues which offer free or low-cost computer access to the public.

B) Significance of the problem with support from the literature

Numerous advantages to bilingualism (even relatively limited bilingualism) have been supported by research, including cognitive advantages such as enhancements to early reading acquisition, creativity and problem-solving abilities, cultural advantages such as greater tolerance and deeper understanding of other cultures, and economic advantages such as wider employment opportunities. Second languages also often serve as ties to cultural heritage, whether to religious communities, in the case of Hebrew or Arabic schools, or to older family members, in the case of the children of immigrant families. Bilingualism can support a positive cultural identity, especially in immigrant and adopted children. (For a review of bilingualism and its advantages, see Baker, 2000.)

However, as noted above, the acquisition of a second language can be daunting. Young children may acquire second-language fluency (speaking and understanding) in a relatively painless way, by immersion (especially with other young speakers of the target language), but this is often not a speedy or efficient process, even when an immersion environment is available. Older children or adults tend to have better study and organizational skills which facilitate *learning* (as opposed to *acquiring*) a second language more quickly than younger children can, provided they are able to spend sufficient time on the task. Unfortunately, older children and adults also tend to have far more demands on their time than younger children, which can inhibit the attention and effort they are able to give to learning a second language. Some evidence also shows that children who acquire a second language at an early age may achieve better pronunciation than older children or adults (Baker, 2000). There is also some research evidence to support an advantage bilingualism provides to children learning to read, who seem to be able to more easily accept learning additional (written) symbols for concepts than monolingual children.

With all this in mind, it seems that it would be highly beneficial to try to provide more opportunities for bilingualism through second language instruction, especially in the early years, when the natural facility for language acquisition may be amplified by careful construction of language learning opportunities. Second language educators have developed ambitious programs to start second language instruction as early as kindergarten (Massachusetts Department of Education, 1999), but with funding for even "basic skills" education under constant attack and with the distractions of statewide and national testing programs which do not include second languages as a priority, inclusion of second language instruction as a regular part of public school curriculum from the early grades seems unlikely in the near future. Some parents will find ways to provide instruction in second languages outside the public school system, especially those with strong ties to ethnic or religious communities that value a heritage language. But many communities are too small (or too monocultural) to be able to offer resources to families interested in pursuing a second language. Private sources may also exclude lower income families.

C) Proposed solution

The system proposed here will provide practice and feedback of speaking skills to young (ages 5-10 years) learners of Mandarin Chinese as a second language, using two engaging game-based formats. Mandarin Chinese was chosen for this project because of the scarcity of second-language instruction programs for children (CFOC, 2000-2004), the strong motivation of parents of Chinese children to provide language learning for the development of positive ethnic identity and other reasons, and the lack of acceptable CALL software for Chinese students of this age group. For the initial phase of this proposal, it will be assumed that the first language of the target learners is English. Reading fluency in English will not be assumed.

The core components of a computer-based system to support the speech component of language learning are the model of learner ability and the feedback system. These two systems would be used in all learner activities and will be described in more detail here.

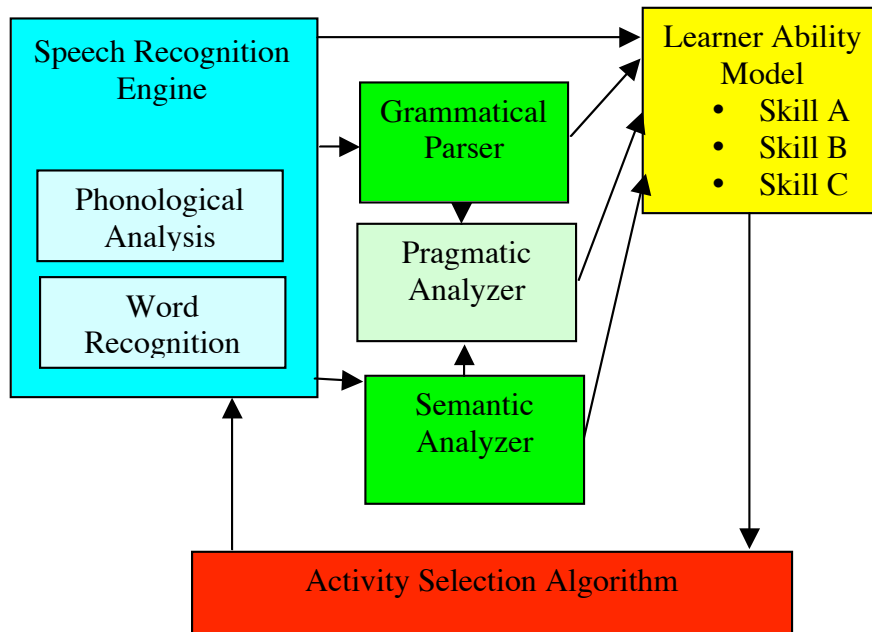


Figure 1 : Extended Learner Ability Model

First, the model of learner ability will function as an ongoing assessment repository. Every activity in the system is treated as an assessment of one or more pre-identified language skills, which the learner is attempting to acquire and demonstrate. For example, the correct pronunciation of a range of vocabulary words are each treated as scorable skills, along with the correct use of grammatical, semantic, or pragmatic rules. Individual phonemes of interest are also scored separately as aggregations of the pronunciation scores of words containing those phonemes (either by using factor analysis or by using the speech recognition software to score individual phonemes as well as words). In particular, aggregate skills relating to the scoring categories of COPI and ELLOPA (see “**Validation**,” below) will be developed. The scoring component of the learner ability model will require models of expected responses related to the skills in each of the

activities, against which the learner input will be scored (and feedback provided; see below). Partial scoring will be used when appropriate, e.g. if a learner produces the correct consonants and vowels for a word, but mispronounces the tone, they will still receive some credit for their performance. Following scoring, the next activity will be selected based on the current learner model state, using IRT-based Computer Adaptive Testing algorithms. The parameters of the selected activity will be fed into the Speech Recognition Engine to inform the recognition process. (This allows the engine to recognize learner speech – and predict likely learner errors – more efficiently.)

The second main component is the feedback system. Through field trials and analysis of prior research, a pool of expected learner errors for each skill will be developed. Appropriate learner feedback will then be developed for each expected error. To keep this task manageable, the learner errors will be grouped into categories, as will the feedback types, and the relationships between errors and feedback will be stored in a database. For example, phonological errors might include: wrong consonant, wrong vowel, and wrong tone. Feedback for wrong consonant might include playing the consonant sound in isolation, displaying an animation of lip movements or sagittal sections, as appropriate, or offering verbal descriptions such as “place your tongue as you would to make an /s/ sound, but move it a little further back in your mouth.” Field trials will be used to prioritize the kinds of feedback offered for each category of learner error, with other forms of feedback also available to learners if initial feedback is not effective. Learner response to feedback types may also be tracked and used to select feedback type for future errors. In this way the error-correction component of the feedback will be made as specific and helpful as possible.

The two systems described above would offer a significant improvement over available CALL systems in the nature and quality of their feedback. However, they do not, by themselves, address the issue of providing meaningful and engaging practice to learners.

A common and powerful method for engaging learners of all ages meaningfully in communicative tasks is the use of learning games. Most if not all of these games can be adapted to online use, and configured to incorporate speech recognition with the appropriate feedback mechanisms. Additionally, games originally designed for the computer may be adapted to provide language practice. Games allow the incorporation of powerful motivating factors such as narrative, suspense, competition, and modeling. Equally importantly, regular positive feedback will be emphasized in the system by the completion of target tasks in the games, allowing the learner to progress to a satisfying conclusion to the activity.

For this study, two games will be developed: a visual communication game, and an “interactive fiction” role-playing game. The first will center around the use of single vocabulary words and short phrases, and will be used primarily to provide phonological and basic content area vocabulary feedback to early language learners. The second will require the production of longer phrases and sentences by the learner, and will be used to provide more comprehensive syntactic, semantic, pragmatic, and prosodic feedback to more advanced learners.

The visual communication game is a computer-based adaptation of a classic language game (Wright et al., 1984). In the original game, learners work in pairs. One learner (A) has a picture, which the other learner (B) cannot see. A describes the picture to B using the target language, watching B and giving ongoing feedback, again in the target language.

This activity is easily adapted to a computer-based model. In the proposed system, the learner helps a cartoon character construct an image from objects on the screen (either simple geometric shapes, or more natural shapes in a content area, such as rocks or branches), to match a sketch outline. Here, words such as "short", "tall", colors, shapes, directions, etc. can be used by the learner in effect as "magic words" to identify or change the objects on the screen to match the sketch and complete the image. One advantage this game offers over the traditional two-student version is that the "blind" computer player can also act as a friendly language coach. The "coach" may throw out suggestions if the learner seems hesitant or stuck, and can offer guidance and feedback on learner utterances which are not close enough to correct pronunciation. An animation sequence with the cartoon character provides a reward when the task is completed. The algorithm and code of this game can also be easily adapted to other vocabulary areas simply by changing the graphics and the target image to reconstruct. Adding a timing element at higher levels can add a self-competitive element to this game. Consistent use of the same or a small group of cartoon coaches, an overall setting, and the addition of a simple plot connecting the activities can also add the power of narrative to this set of learning games. The computerized format also helps to avoid the common problem during activities in language classrooms of breakdown into the dominant language rather than the target language, as the computer system can be set to ignore phrases in the dominant language, or to offer suggestions in the target language and require correct repetition if preferred.

The proposed development tool for this activity is Macromedia Flash, using server-side extensions to link to the speech recognition, learner ability model, and feedback generation components via XML. This development environment will allow for rapid prototyping, delivery across a broad range of platforms, and flexibility in reuse of developed code.

The second game system is a variation of role-play. Role-play is a specific example of a popular game format used in language learning, and it has its online equivalents as well (Mullin, 2004). Online roleplaying allows advanced learners to benefit from an "interactive fiction" approach to language learning, again with speech recognition and guidance. Using this approach, a parser accepts learner input as speech, and provides descriptions of the results of learner actions in synthesized or recorded speech. (Optionally, the game provides voice-to-text for the learner and text output to correspond with the spoken output, for learners who are also studying written Chinese.) The games themselves are "puzzle" type games in the style of the old "Infocom" products (e.g. the popular *Zork* and its sequels). The puzzles are scored as they are solved, and there is a turn limit to encourage active engagement in the game. The cognitive advantage to these games in their classic form lies in their lack of graphics or other visual cues -- the

learners must pay attention to the text (spoken or not) of the game to construct the world in their minds, requiring truly active listening as well as full sentence production. However, this does also limit these games to more advanced learners. A simpler version with limited illustrations and more constrained vocabulary might also be developed. (One interesting hybrid game which used a text menu system layered onto a graphical interface was the Windham Classic title, *Below the Root*. See “Underdogs,” 2004 for more information.) Such games can be configured to cover any desired topic area, vocabulary content, or grammatical structures, and as the parser needs to determine the grammar of the input, it can also be configured to provide hints when it determines that grammatical rules have been violated, as well as the pronunciation feedback discussed above. If necessary, for less advanced learners, the parser could also provide some translation assistance to help the learner “come up to speed” in the necessary points of the target language. These sorts of games hold the interest power and suspense of a narrative, and provide opportunity for more complete sentence production as well as the inclusion of various pragmatic (situational) aspects of language.

The proposed development tool for this game is Perl, and specifically the online interactive system “PerIMUD” (Boutell, 2004). This system was chosen because of its openness and ease of modification, low cost, and established stability. In addition to the PerIMUD server, which will be accessed via telnet or http protocols, a custom client will be developed which will enhance the learner’s ability to request additional feedback or to control the game parameter settings.

D) Vignettes that provide a description of how the proposed solutions would work

Vignette 1: Visual Communication Game

In this example of a visual communication game, the learner is presented with an outline image of a train engine. The learner interacts with the cartoon monkey to identify objects on the screen and tell the monkey where to place them. The target areas of language are simple shapes, colors, and directional words (up, down, left, and right). To aid the learner, the monkey asks questions which can be answered in single words or short phrases. If the learner seems to be having difficulty understanding the question, the monkey will repeat the phrase, and after two repetitions, will explain the phrase in English. The monkey also offers assistance when waiting for the learner to respond to a phrase by offering suggestions. At first, the monkey only offers verbal suggestions (“Circle? Square?”) but if the learner delays in response, the monkey will eventually begin to point to items on the screen and provide their names or perform other actions to give visual cues to vocabulary words.

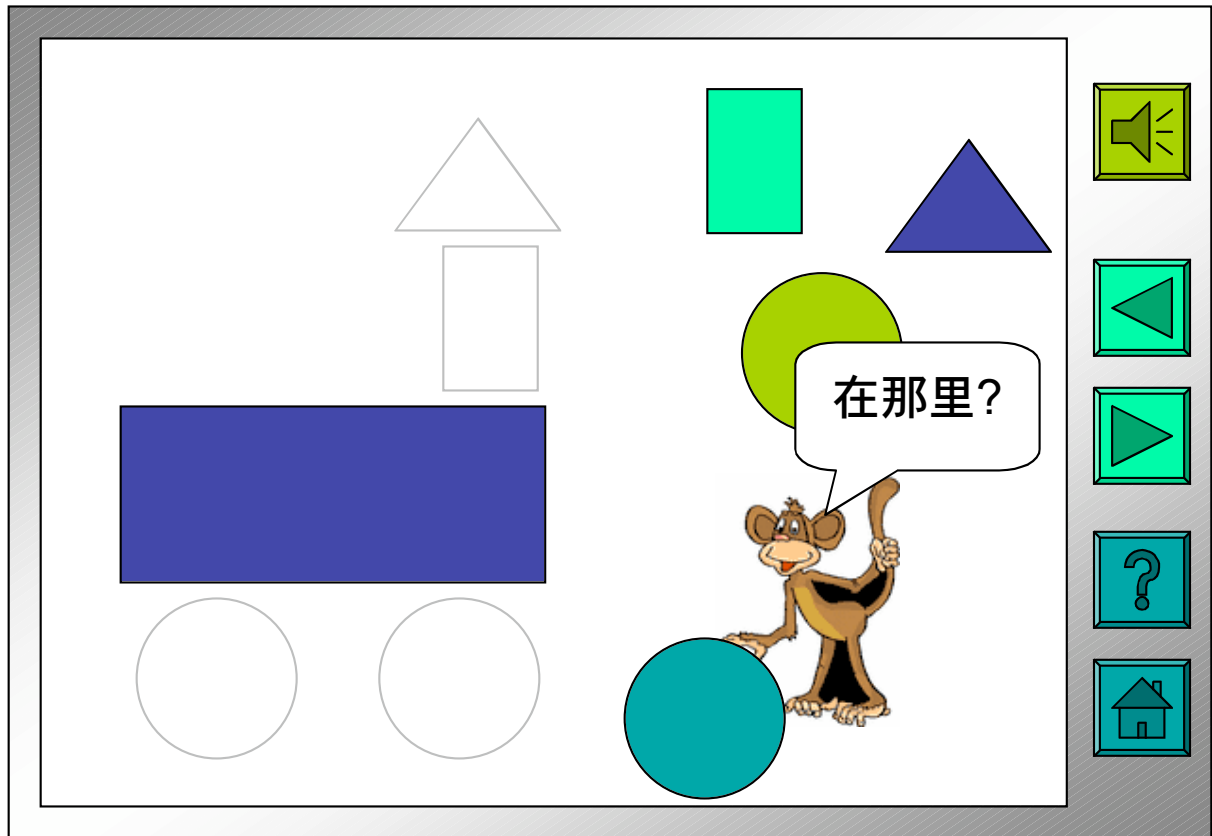


Figure 2: Visual Communication Game

For each successful placement of the object, the monkey provides a short animation (an acrobatic flip, eating a banana, etc.) and gives praise in Chinese. When the entire image is complete, the monkey performs a longer animation with music.

In the background, the system is using learner responses (including fluency of response) to build the learner ability model. This model is used to decide when to offer coaching and how much coaching to offer, as well as which activity to select from a pool of activities ranging in difficulty and spanning a number of content areas. Again, this game is meant to provide practice in pronunciation of isolated words and short phrases, so the learner model is fairly simple at this point, and a more limited number of sub-components are needed in the scoring system:

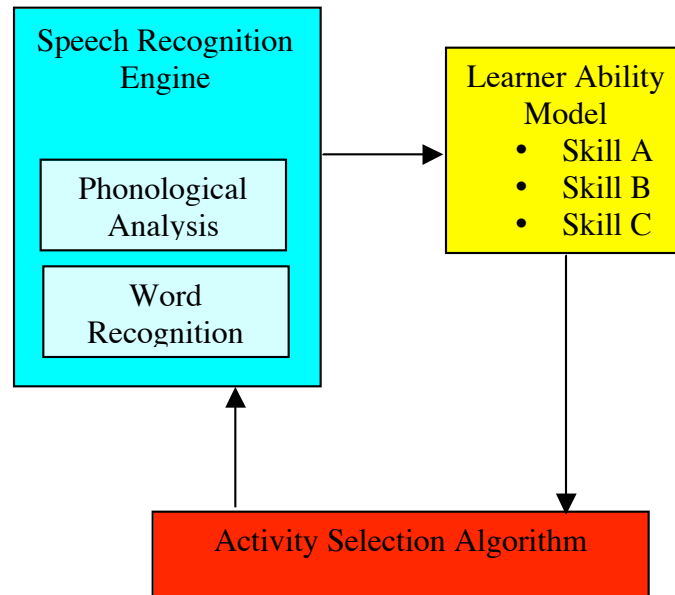


Figure 3: Limited Learner Ability Model

Vignette 2: Interactive Fiction Role Playing Game

In this learning game, the learner is presented with auditory (and optionally, text or graphic) prompts as part of a narrative. The learner provides input using spoken phrases and sentences. This example is taken from a classic interactive game, “Adventure.” The transcript has been modified to illustrate an example of a correction. For sample purposes only, the text of this example appears in English. The auditory cues would be in recorded Mandarin Chinese, and if the text option was enabled, the text would appear in either Chinese characters or pinyin, depending on the level of the learner according to the learner model.

You are standing at the end of a road before a small brick building.
 Around you is a forest. A small stream flows out of the building and down a gully.
 > go into the building
 You are inside a building, a well house for a large spring.
 There are some keys on the ground here.
 There is a shiny brass lamp nearby.
 There is food here.
 There is a bottle of water here.
 > Get the lamp
 I think you meant to say, “Get the lamp”, but it sounded like “Bet the lamp.” Could you try again, please?
 >

Figure 4: Roleplaying Game

As with the Visual Communication game, the system is sensitive to hesitations on the part of the learner, and based on the current Learner Ability Model, may repeat a prompt, offer suggested responses, or translate a prompt, as needed. When the learner does respond, the system is able to use the programmed scenario of the game to anticipate the response, aiding in the system's ability to recognize the intended utterance and provide appropriate feedback. This game provides more comprehensive feedback and requires the full extended learner ability model shown in Figure 1. The system provides pronunciation feedback similar to that in the Visual Communication game, but also provides syntactic, semantic, pragmatic, and prosodic feedback, again tailored to actual user errors. For minor errors, the system has the option of repeating a corrected version of the learner's utterance and moving on, rather than asking the learner to repeat the utterance. This decision is based on the current learner ability model and the severity of the error. This flexibility allows the system to behave more like a human conversation partner might, generously interpreting some slightly incorrect utterances, but offering more direct feedback and requiring practice on more severe errors.

E) Rationale for proposed solution with support from the literature

Colin Baker (2000) makes a compelling case for the advantages of bilingualism in general and the advantages of early bilingualism in particular. He reports, "Young children learn languages as naturally as they learn to run and jump, paint and play. Young children are not worried by their language mistakes, nor about not finding the exact words.... Language acquisition is a by-product of playing and interacting with people." He also writes, "When children are very young, they pick up accurate pronunciation quickly." Later on, he cites research by Yelland, Pollard, & Mercuri (1993): "The children in the research were aged four to six. One group were English monolinguals; the other group had only one hour of Italian language instruction each week. After just six months of such lessons, the 'marginal bilingual' children were much more aware of the notion of words (metalinguistic awareness) than their monolingual counterparts. This advantage carried through to when the children started to learn to read. The 'marginal bilinguals' showed quicker word recognition skills than the monolinguals." These statements clearly support the selection of the target audience, young learners of a second language. They also link the learning of languages with natural child activities such as play.

Regarding the model and methodology of language learning being used, Richards and Rodgers (2001) report: "Mainstream language teaching... opted for Communicative Language Teaching (CLT) as the recommended basis for language teaching methodology in the 1980s and it continues to be considered the most plausible basis for language teaching today...." They describe the essential principles of CLT as follows:

- Activities that involve real communication promote learning.
- Activities in which language is used for carrying out meaningful tasks promote learning.

- Language that is meaningful to the learner supports the learning process.

Richards and Rogers go on to describe Task-Based Language Teaching (TBLT) as a logical development of CLT. They report the key assumptions, as summarized by Feez (1998):

- The focus is on process rather than product.
- Basic elements are purposeful activities and tasks that emphasize communication and meaning.
- Learners learn language by interacting communicatively and purposefully while engaged in the activities and tasks.
- Activities and tasks can be either:
 - Those that learners might need to achieve in real life;
 - those that have a pedagogical purpose specific to the classroom.
- Activities and tasks of a task-based syllabus are sequenced according to difficulty.
- The difficulty of a task depends on a range of factors including the previous experience of the learner, the complexity of the task, the language required to undertake the task, and the degree of support available.

They then describe a “task”: “Although definitions of task vary in TBLT, there is a commonsensical understanding that a task is an activity or goal that is carried out using language, such as finding a solution to a puzzle, reading a map and giving directions, making a telephone call, writing a letter, or reading a set of instructions and assembling a toy.”

This understanding of a “task” and task-based language teaching nicely supports the proposed solution, in which task-based games, with support, are used as the primary learning mechanism.

Further support for the use of games in language learning is prevalent throughout the literature. Ehsani & Knodt (1998) write: “Apart from being more fun and interesting, games and task-oriented programs implicitly provide positive feedback by giving students the feeling of having solved a problem solely by communicating in the target language.” Wright, et al. (1984) write: “Games help and encourage many learners to sustain their interest and work. Games also help the teacher to create contexts in which the language is useful and meaningful. The learners *want* to take part and in order to do so must understand what others are saying or have written, and they must speak or write in order to express their own point of view or give their information.” Wright et al. document over one hundred language games for use in the classroom, including the visual communication game in this proposal (“Describe and Draw a Picture”).

Yao & McGinnis, in their Chinese-specific handbook *Let’s Play Games in Chinese* (2002) offer the following observations: “Perhaps the best thing that can be said about these games for learning Chinese is that they are an extremely entertaining way to use the language. It is the word ‘use’ that is of paramount importance. There are numerous ways by which we can assess a student’s ability to use the target language, be it memorized dialogues or written compositions. Yet these very means of purportedly promoting a

student's progress in fact inhibits a portion of the student population. It is that portion that, intimidated by the pressures of a *graded* activity, may find the chance to freely practice without pressure, and to practice very well indeed, while playing these games." Yao & McGinnis devote an entire chapter of their book to reconstruction/drawing games, and another to role-playing games. Visual and role-playing games are also described in Iacofano (1997).

Having established the value of games in learning a second language, we turn to the question of the use of technology to support language learning, and especially speech learning. Ehsani & Knodt (1998) note that improvements in speech recognition technology offer many possibilities for second-language speech training, provided the limitations of the existing technology are kept in mind. They suggest that activities which offer limited or predictable domains of vocabulary and other linguistic features will be most successful. Hardison (2004) provides support for the use of technology in improving prosody, and includes an extensive bibliography of research supporting the use of technology in language learning. Eskenazi (1999) supports the use of technology in pronunciation training, but notes that existing feedback systems are often inadequate and unhelpful to learners, especially the common, but simplistic "record and compare" systems. Zhang (1998) provides a review of Chinese language learning software packages available at the time, and is generally supportive of the capability of software to aid in learning speech skills, though he also criticizes existing feedback mechanisms. He reports that feedback to learners is most helpful when it is specific, explanatory, and non-judgmental, and notes that existing systems tend to offer general, rather than specific feedback, do not explain the feedback mechanisms being used (e.g. spectrograms), do not offer explanations of how to correct the errors detected, and tend to emphasize errors rather than success. These are all aspects of feedback that it is hoped this proposal will improve on.

The nature and quality of feedback will be especially critical in this system, because of its key importance in formative assessment. The New Zealand Ministry of Education cites Sadler's work in formative assessment in their 2003 review "The Effects of Curricula and Assessment on Pedagogical Approaches and on Educational Outcomes":

A further commentary on formative assessment comes from Sadler (1998). Key points made in this paper are:

- Formative assessment is effective in all educational settings.
- Grades do not deliver as much formative effectiveness as tailored comments.
- Quality is crucial.
- Students need to be trained in how to interpret feedback.
- Any movement towards feedback-enhanced learning conditions must be carried out long enough for the new procedures to be viewed by the learners as normal and natural.
- Teachers can only be effective at formative assessment if they know both sides of the operation - how students learn and the subject area.

The points of tailored comments and student training in how to interpret feedback, in particular, echo the comments of Zhang, above.

The issue of quantity and appropriateness of feedback itself is a complex one. Brisk & Harrington (2000) write, “There is a delicate balance between promoting language accuracy and discouraging students with correction. Constant correction interrupts the thinking process and overtaxes short-term memory. On the other hand, students developing a language need feedback to approach accuracy.” With this in mind, the feedback component needs to be able to judge when to offer feedback, and how much feedback to offer, based on the current learner ability model and success or repeated errors on the current task. This functionality will need to be tuned in field trials.

McNamara (2000) specifically addresses the issue of reliability in computer-based testing of language: “Already these automatic measures of pronunciation or writing quality are being used in place of a second human rating of performances, and have been found to contribute as much to overall reliability as a human rating.” McNamara also discusses the issues of learner or candidate comfort with technology as a testing medium at length. Given the general comfort of young learners with computers and computer-based games, it is likely that this will not present as much of an issue as it sometimes does with older learners.

F) Proposed validity study

Because formative assessment is highly integrated with learning, learning measurements will be one appropriate form of validation (discussed below). But formative assessment is still, at its heart, assessment, and the validity of the assessment methods used must be demonstrated. The learning system in which the formative assessment is embedded must have the ability to maintain a representation of the learner's proficiency, in order to appropriately select and gauge the learning activities to the learner's present level and present appropriate feedback. This representation needs to be calibrated and its accuracy verified. The nature and quality of the feedback to the learner must also be evaluated.

Content validation of the task system should be attempted by matching the areas of formative assessment included in the activities against an established curriculum framework, e.g. the Massachusetts Foreign Language Curriculum Framework (Massachusetts Department of Education, 1999). For the purposes of an initial trial, a limited portion of the Framework might be defined and represented.

Criterion validation must also be addressed. The goal of providing formative feedback is to enhance learning, so a measure of learning will be needed following the use of this proposed software. Additionally, the learner ability model of the system needs to be validated as an assessment. Because the learning goal in this case involves the ability to communicate using spoken language, an existing, respected interview-based assessment of spoken language proficiency should be used to establish concurrent validity. Oral exams developed for use with younger learners include the Early Language Learning Oral Proficiency Assessment (ELLOPA), the Classroom Oral Competency Interview

(COCI), and the Stanford Foreign Language Oral Skills Evaluation Matrix (FLOSEM). All are intended to be usable with any language, in grades K-12. (National K-12 Foreign Language Resource Center (Iowa State University) and CAL, 2002) More specifically, the Computerized Oral Proficiency Instrument is being extended for use with Chinese (Bourgerie, 2002, Malabonga, 2002) and would be directly applicable to this proposal. The COPI uses CAT to provide audio prompts to learners and records speech segments for later scoring by human raters. This is quite similar to the learner ability model subsystem being used, though it is a summative assessment system, rather than a formative one. Use of the COPI would also minimize potentially confounding differences in the medium of the two assessment systems. Existing research describing the criterion validity of the COPI should be used as a base, and then the learner proficiency model used in the software application calibrated against this exam. If necessary, more sensitive oral interview protocols may need to be adapted from the above or other oral exams, which can allow for the use of faceted Rasch modeling to adjust for task difficulty and rater severity. (McNamara, 1996) (These modeling techniques may also be used in the learning software itself to aid in activity selection and difficulty setting.)

Finally, construct validation should be used to look for possible confounding factors such as computer experience and comfort and first language proficiency (including first language literacy or lack thereof). Controls for prior background in the target language, and especially alternate acquisition opportunities with the target language (e.g. members of the family who speak the target language) also need to be established.

In addition to evaluating the validity of the software's learner proficiency model, the quality of the system feedback to the learner also needs to be evaluated. As noted above, the quality of the feedback system will be a critical factor in the overall success of the formative assessment system.

This can be done systematically, by asking language teaching experts to review the decision trees and scripts used, or in a "simulation," by having expert instructors use the software and deliberately make common linguistic errors to evaluate the results, and by recording the efforts of actual users and having the performance of the system evaluated by experts in "playback" mode. An instrument will need to be developed for the experts to use to evaluate the feedback, unless an existing instrument can be found. Feedback should be evaluated on accuracy, specificity and helpfulness. It would also be useful to solicit learner input on the quality and value of the feedback. (One possible approach would be to effectively provide a "Turing test" of the feedback system: sample learners or expert instructors would use the system and receive feedback from either the system, or an expert instructor. The testers would then attempt to determine whether they were being helped by a computer system or a live instructor. The interfaces required for this sort of test could be used during initial development to identify feedback points and content, as well.)

Once the software's learner proficiency model has been calibrated, it may also be helpful to compare its effectiveness to other means of learning a second language, e.g. full bilingual or immersion language programs, after-school or weekend language programs,

competing software (*Rosetta Stone* or more game-formatted software without speech recognition functionality), informal immersion playgroups, commercial audio or video tapes intended for children, etc. Gains in learning for comparable time investment could be compared using the oral exams described above. While it is doubtful that a technology-based approach could completely replace the daily interactive experience of a well-run classroom environment with an experienced teacher, many community language schools are unable to retain experienced teachers, or offer classes infrequently enough that regular use of a software application might improve upon -- or at least complement -- available programs. And as noted above, such programs are not readily available in all communities. It would be helpful to understand the extent to which a technological solution can in fact provide a replacement.

G) Work plan/timeline including additional expertise required and sequence of actions required to accomplish proposed work

To aid in managing complexity, a phased approach will be used in the development of the systems:

- Phase I: Student model and feedback system in isolation
- Phase II: Visual Communication Game
- Phase III: Interactive Fiction Roleplaying Game

Major activities for each phase are as follows:

Phase I: 6 Months

- Record/Acquire Native Speaker Samples
- Speech Recognition Tuning
- Learner Ability Model construction
- Feedback System Development
- System Integration
- Beta Testing/Preliminary Validity Study

Phase II: 6 Months

- Communicative Task Selection
- Visual Communication Game Design
- Visual Communication Game Programming
- Graphic, Animation, and Non-Speech Sound Creation
- Integration with Speech Recognition, Learner Ability Model, and Feedback System
- Beta Testing/Validity Study
- Extended Field Trial
- Analysis and Reports

Phase III: 6 Months

- Communicative Task Selection
- Roleplaying Game Design

- Roleplaying Game Programming (PerlMUD)
- Custom Client Development (Feedback tools)
- Integration with Speech Recognition, Learner Ability Model, and Feedback System
- Beta Testing/Validity Study
- Extended Field Trial
- Analysis and Reports

Expertise required will include:

- Project Manager
- Instruction/Assessment Designer
- Graphic Artist (Phase II only)
- Flash Programmer (Phase II only)
- Perl Programmer
- Sound Engineer
- Speech Recognition System Integrator
- Mandarin Chinese Native Speakers
- Mandarin Chinese Language Instructors

References

Ahmed, I. (2003). Rosetta Stone (Arabic Explorer). *The Reading Matrix* 3(3). Retrieved May 9, 2004 from the World Wide Web:

http://www.readingmatrix.com/software_reviews/ahmed/software_review3.pdf

Baker, C. (2000). *A Parent and Teacher's Guide to Bilingualism* (2nd ed.). Clevedon: Multicultural Matters Ltd.

Bourgerie, D. (2003). Computer Aided Language Learning for Chinese: A Survey and Annotated Bibliography. *Journal of the Chinese Language Teachers Association* 38(2), 17-47. Retrieved May 9, 2004 from the World Wide Web: http://clta.osu.edu/reviews/files/Bourgerie_2003.pdf

Boutell, T. (2004) PerlMUD 3.0. Retrieved May 10, 2004 from the World Wide Web: <http://www.boutell.com/perlud/>

Brisk, E. & Harrington, M. (2000). *Literacy and Bilingualism: A Handbook for All Teachers*. Mahwah, NJ: Lawrence Erlbaum Associates.

Cheapshop. Fairfield Language Technologies - Rosetta Stone Arabic Level 1 & 2 Personal Edition. Retrieved May 9, 2004 from the World Wide Web: <http://software.cheaps.us/B000077DCY.html>

CFOC (Chinese for Our Children). (2000-2004). [Online Discussion Group]. Retrieved May 9, 2004 from the World Wide Web: <http://groups.yahoo.com/group/CFOC/>

Discovery Channel. (2003). Rosetta Stone Spanish Explorer/Rosetta Stone French Explorer. *discoveryschool.com Review Corner [Online Review Archive]*. Retrieved May 9, 2004 from the World Wide Web: <http://school.discovery.com/parents/reviewcorner/software/rosettastonespanishexplorer.html>

Ehsani, F and Knodt, E (1998). Speech technology in computer-aided language learning: strengths and limitations of a new CALL paradigm. *Language Learning & Technology* [Online serial]. 2(1), 45-60. Retrieved May 9, 2004 from the World Wide Web: <http://llt.msu.edu/vol2num1/article3/index.html>

Eskenazi, M. (1999). Using automatic speech processing for foreign language pronunciation tutoring: some issues and a prototype. *Language Learning & Technology* [Online serial]. 2(2), 62-76. Retrieved May 9, 2004 from the World Wide Web: <http://llt.msu.edu/vol2num2/article3/index.html>

Fairfield Language Technologies (a). Rosetta Stone Language Software Product Tour. Retrieved May 9, 2004 from the World Wide Web: http://www.rosettastone.com/edu/tour/classroom_edition8

Fairfield Language Technologies (b). Research Basis for The Rosetta Stone Dynamic Immersion Method. Retrieved May 9, 2004 from the World Wide Web: <http://a1664.g.akamai.net/7/1664/51/294f758f47c6c3/www.apple.com/education/k12/curriculumsolutions/accelerate/pdf/RSResearchBasis.pdf>

Ge, J. (2003). Online Chinese Handwriting Recognition. Retrieved May 9, 2004 from the World Wide Web: <http://www.cs.princeton.edu/ugprojects/projects/j/jge/senior/ChineseHandwritingPaper.doc>

Hardison, D. (2004). Generalization of computer-assisted prosody training: quantitative and qualitative findings. *Language Learning & Technology* [Online serial]. 8(1), 34-52. Retrieved May 9, 2004 from the World Wide Web: <http://l1t.msu.edu/vol2num2/article3/index.html>

Iacofano, J. (Ed.). (1997). *101 Terrific Tips for Language Teachers*. Auburn Hills, MI: Teacher's Discovery.

Kaiser, M. (1997). Review: The Rosetta Stone for Russian. CALL@Chorus [Online Review Archive]. Retrieved from the World Wide Web on May 9, 2004: http://www-writing.berkeley.edu/chorus/call/reviews/rosetta_russian/

Malabonga, V. (2002). Computerized Oral Proficiency Instrument. *CAL Project Archives*. Retrieved May 9, 2004 from the World Wide Web: <http://www.cal.org/archive/projects/copi.html>

Massachusetts Department of Education (1999). Massachusetts Foreign Language Curriculum Framework. Boston: Massachusetts Department of Education. Retrieved May 9, 2004 from the World Wide Web: <http://www.doe.mass.edu/frameworks/foreign/1999.pdf>

McNamara, T. (1996). *Measuring Second Language Performance*. New York: Addison Wesley Longman.

McNamara, T. (2000). *Language Testing*. Oxford: Oxford University Press.

New Zealand Ministry of Education (2003). The Effects of Curricula and Assessment on Pedagogical Approaches and on Educational Outcomes. Wellington: New Zealand Ministry of Education.

Mission to Arabic: It's Not Your Father's Language Lab. (2004). University of California. Retrieved May 9, 2004 from the World Wide Web: <http://www.isi.edu/stories/78.html>

- Mullin, E. (Ed.) (1995-2004) *XYZZYnews: The Magazine for Interactive Fiction Enthusiasts* [Online Serial]. Retrieved May 9, 2004 from the World Wide Web: <http://www.xyzzynews.com/>
- National K-12 Foreign Language Resource Center (Iowa State University) and CAL. (2002). Retrieved May 9, 2004 from the World Wide Web: <http://www.cal.org/k12nflrc/>
- Richards, J. & Rodgers, T. (2003). *Approaches and Methods in Language Teaching* (2nd ed.). Cambridge: Cambridge University Press.
- Sokolik, M. (1999). Real English. CALL@Chorus [Online Review Archive]. Retrieved from the World Wide Web on May 9, 2004: http://www-writing.berkeley.edu/chorus/call/reviews/real_english/page_2.html
- Twinbridge Software Corporation. Chinese Pen. Retrieved May 9, 2004 from the World Wide Web: <http://www.twinbridge.com/>
- Underdogs (1998 – 2004). Home of the Underdogs [Online Review Archive]. Retrieved May 9, 2004 from the World Wide Web: <http://www.the-underdogs.org/game.php?id=113>
- Wright, A., Betteridge, D., & Buckby, M. (1983). *Games for Language Learning* (2nd ed.). Cambridge: Cambridge University Press.
- Yao, T. & McGinnis, S. (2002). *Let's Play Games in Chinese*. Boston: Cheng & Tsui Company.
- Yelland, G.W., Pollard, J. & Mercuri, A. (1993). The metalinguistic benefits of limited contact with a second language. *Applied Psycholinguistics* 14, 423-444.
- Zhang, Z. (1998). CALL for Chinese-issues and practice. Retrieved May 9, 2004 from the World Wide Web: http://www.csulb.edu/%7Etxie/learn_online/issues.htm