

Speech-enabled Card Games for Incidental Vocabulary Acquisition in a Foreign Language

Ian McGraw, Brandon Yoshimoto and Stephanie Seneff

MIT Computer Science and Artificial Intelligence Laboratory

32 Vassar Street, Cambridge, Massachusetts 02139, USA

{imcgraw, byoshimo, seneff}@csail.mit.edu

Abstract

In this paper, we present a novel application for speech technology to aid students with vocabulary acquisition in a foreign language through interactive card games. We describe a generic platform for card game development and then introduce a particular prototype card game called *Word War*, designed for learning Mandarin Chinese. We assess the feasibility of deploying *Word War* via the Internet by conducting our first user study remotely and evaluating the performance of the speech recognition component. It was found that the three central concepts in our system were recognized with an error rate of 16.02%. We then turn to assessing the effects of the *Word War* game on vocabulary retention in a controlled environment. To this end, we performed a user study using two variants of the *Word War* game: a speaking mode, in which the user issues spoken commands to manipulate the game cards, and a listening mode, in which the computer gives spoken directions that the students must follow by manipulating the cards manually with the mouse. These two modes of learning were compared against a more traditional computer assisted vocabulary learning system: an on-line flash cards program. To assess long term learning gains as a function of time-on-task, we had the students interact with each system twice over a period of three weeks. We found that all three systems were competitive in terms of the vocabulary words learned as measured by pre-tests and post-tests, with less than a 5% difference among the systems' average overall learning gains. We also conducted surveys, which indicated that the students enjoyed the speaking mode of *Word War* more than the other two systems.

Key words: speech recognition, intelligent computer assisted language learning, computer aided vocabulary acquisition

1 Introduction

It is estimated that the average high school senior knows around 40,000 words [45]. To have a rudimentary grasp of a language at the conversational level, a vocabulary of around 5,000 words is necessary [39]. Children who are immersed every waking hour in an environment ideal for acquiring their first language learn at a rate of around ten new words per day [6]. These facts and a back-of-the-envelope calculation reveal that vocabulary acquisition accounts for years of a language learner's time.

It is no wonder that the first few years of language learning are often characterized as having a focus on language at the lexical level [29]. It is surprising, however, that relatively little attention is paid to the problem of vocabulary acquisition in the foreign language classroom. Many curricula leave the problem to the student as homework, and little instruction is given regarding effective acquisition techniques. Teachers are often hesitant to waste valuable class time directly teaching individual words when the impact of such teachings seems negligible relative to the sheer number of words a student needs to know [2].

As a result, students often resort to *explicit* memorization to prepare for classroom activities. Proponents of communicative language teaching criticize *intentional* vocabulary learning through word lists or flash cards, citing the lack of linguistic context provided by such techniques [36]. Still, the efficiency of such methods with respect to retention is hard to deny [33]. It is certainly possible to learn thousands of words by explicit memorization; however, this might be a discouraging prospect for students who find these methods tedious [24].

While *implicit* acquisition techniques, in which new words are learned through reading and conversation, might be more palatable for the student, for beginners with little foundation from which to infer the meaning of new words from context, the process can be quite slow. Acquiring new words through reading is complicated with languages such as Chinese that have ideographic scripts, because learning to read can take years. Implicit vocabulary acquisition through conversation is fraught with a different set of problems. As one well-respected Second Language Acquisition (SLA) theorist notes, beginners are often quite hesitant to expose their inexperience to a native speaker [25]. Even in the absence of timidity, for many in the United States, opportunities to practice speaking a foreign language outside of the classroom are rare.

We believe that speech technology is particularly well suited to providing a highly-accessible, non-threatening environment for vocabulary acquisition. Such systems might even resolve the tension between acquisition efficiency and enjoyment with carefully constructed *incidental* acquisition tasks, which may contain small components of explicit learning, but where the focus of the

task is independent of the memorization goals. We explore this assertion in the context of speech-enabled card games designed for incidental vocabulary acquisition in Mandarin Chinese. The goals of the card games presented in this paper only *indirectly* require the student to consciously commit new words to memory. In this way, we attempt to strike a balance between the efficiency of explicit memorization and the appeal of both the communicative nature of implicit acquisition techniques and the relative enjoyment of the student.

The role of Internet technologies in our systems is central to their accessibility. With the recent ubiquity of the Web 2.0 paradigm, and the widespread adoption of Voice over IP (VoIP), one can imagine a day when students will routinely interact with educational services that depend critically on audio capture and transmission of speech over the Internet. Combining the emerging technologies of automatic speech recognition (ASR) and VoIP, we have developed a framework for Web-based games which allows learners to talk to their computers in Chinese from an ordinary Internet browser. The work presented in this paper builds upon this framework to provide a platform for the development of card games. We maintain these card games, and a Web 2.0 card-creation interface through which they can be customized, at the following publicly deployed Web site: <http://web.sls.csail.mit.edu/chinesecards>.

We evaluate the framework in the context of a particular prototype card game called *Word War*. Although the game is designed for two players, we restrict our attention to a single player practice mode in our experiments. Two forms of this game are presented. The first is a speaking mode, in which the student issues spoken commands to the computer. Audio is streamed to a Mandarin speech recognizer, and the student observes the computer carry out his or her directions. The second is a listening mode, in which we use a speech synthesizer to provide oral instructions to the student, who is then required to manually manipulate a set of cards accordingly.

Two user studies presented in this paper evaluate *Word War* in terms of speech recognition performance and vocabulary retention respectively. The first experiment illustrates the feasibility of deploying the system on the Internet and allowing users to interact with the game using their own resources. The second user study examines whether a trade-off exists between efficiency in ensuring long term vocabulary retention and the enjoyment of the system's users.

The remainder of this paper is organized as follows: In section 2, we review a number of the currently available speech-enabled systems for language learning and discuss some related work in the field of computer assisted vocabulary learning (CAVL). Section 3 introduces the framework for customizable card games and section 4 describes the *Word War* game in detail. We provide user-based evaluations of *Word War* in sections 5 and 6, for two variants of the game, a “speaking” mode and a “listening” mode. The experiments

described in section 5 measure the performance of the speech recognizer, while the study discussed in section 6 attempts to assess the students' retention of words learned over a three week time window. Section 7 briefly describes a new game recently developed in our group, which strives towards improved replay value for sustained engagement of student interest over repeated episodes. Section 8 concludes with some suggestions for further improvements and the future directions of our work.

2 Previous Work

This section provides an introduction into the area of computer assisted vocabulary learning (CAVL) as well as an overview of work in automatic speech recognition (ASR) as applied to foreign language learning. The fact that there is little overlap between these fields suggests that this research is in new territory with respect to intelligent computer aided language learning (ICALL).

2.1 Computer Assisted Vocabulary Acquisition

Although CAVL systems are quite pervasive, they vary in terms of their pedagogical grounding and the complexity of the technology employed. Such systems range from simple on-line flash card programs promoting intentional memorization techniques to intelligent reading environments, e.g. [16], which give the student a myriad of tools to deal with vocabulary items in an incidental acquisition setting.

The degree to which these systems can be classified as Artificial Intelligence in Education also varies. On the flash card side, a small community is quite interested in optimal scheduling algorithms [7]. The reading environments, on the other hand, sometimes include natural language processing components to provide a morphological analysis of the text [34].

Clearly these systems also have very different audiences. Flash cards can be used by learners with a range of proficiencies, but are more often found in the hands of beginners trying to learn their first few thousand words in a foreign language. The intelligent reading systems typically target a far higher skill level, and rely on the learner to understand a large degree of context to pick up new words incidentally, or with the help of natural language tools.

Interestingly, the problem of providing an environment for *incidental* vocabulary acquisition to the *beginning* language student remains largely unsolved. Unfortunately, this is precisely where such systems are sorely needed, since lex-

ical acquisition is often the most difficult task for an adult learning a language from scratch [39].

Arguably the most successful effort in developing a well-motivated CAVL system is the commercially available software package, Rosetta Stone [40]. Using images as context, this software package requires the student to choose from a set of pictures associated with spoken descriptions that get progressively longer. While this immersion in comprehensible *input* is appealing, opportunities for the user to *speak* using this software come mainly in the form of pronunciation assessment rather than more substantive tasks.

Aside from being prohibitively expensive for many institutional settings, one of the largest drawbacks of commercial software is the lack of customizability. This brings the discussion back to freely available, easily personalizable flash cards. Although flash cards can be tailored to an individual's learning needs, they also rarely require the student to speak. While some in the SLA theory community would not regard this as a negative characteristic [27], many if not most SLA researchers agree that spoken output is not simply the *result* of learning a foreign language, but an important component of its *acquisition* [41].

2.2 Automatic Speech Recognition for Second Language Acquisition

Over the last decade automatic speech recognition (ASR) has become reliable enough to be considered for use in computer systems for language learning [15]. To overcome the difficulties inherent in processing learner speech, researchers find it necessary to place heavy constraints on the spoken input accepted by the system. It is not surprising then, that the earliest successes in applying ASR to SLA came in the form of pronunciation evaluation, where the exact input is known in full [9]. While some attempts were made to keep the experience engaging [8,13], such systems rarely convince the learners that they are *using* the language to communicate, a concept that many SLA researchers feel is central to acquiring a foreign language [26,28].

As speech recognition technologies became more robust, researchers began to relax the constraints on their systems, allowing for some variation in the student's utterances. This relaxation typically manifested itself in the form of multiple-choice questions that prompt the user with the possible responses [20,22]. While systems that employ multiple-choice spoken responses have begun to make their way into the commercial realm [42], the majority of commercial systems still ensure that there is a single correct user utterance for a given prompt.

The research community has since moved on to creating small context free

grammars (CFGs) [4,3,23,32], whose low perplexity ensures robust recognition. Such systems allow the user to have short conversations in small domains, thus providing environments for language learning grounded in current theories of second language acquisition.

The success of the speech-enabled systems described above can be largely attributed to the restrictions placed on the allowable input. To this day, however, dialogue systems that give the learner a large degree of freedom in both sentence structure and vocabulary remain beyond the reach of even cutting edge speech and language technology. Perhaps due to this limitation, ASR systems that target vocabulary acquisition on a *large* scale are exceedingly rare. Still, one could imagine that, if a large number of these narrow domain systems could be developed and introduced into the classroom, they might collectively be capable of making a meaningful impact on language education.

Although increasing attention is being paid to SLA theories, few of the currently available applications of ASR to SLA have been assessed with respect to their educational value in practice. While some researchers have begun studying the effects of well-established systems for pronunciation assessment on learning [14,21,35], it is important that the recent, highly interactive speech systems for SLA begin to follow this lead. Though the resources to perform such studies are not easy to come by, such experiments give a project both the credibility and exposure that might facilitate their widespread adoption.

2.3 Implications

Providing a well-motivated system for vocabulary acquisition is clearly a delicate balance. While flash cards are highly customizable, they typically take the lexical items out of any meaningful context. Intelligent reading environments have the potential to provide large quantities of comprehensible input, but rarely offer support for the beginner. Neither of these applications require that the student practice speaking. Some of the newer spoken dialogue systems for SLA show great promise from a pedagogical perspective, but are difficult to deploy, have very limited lexical domains, and often lack user-customizability.

In the next section, we introduce a system that attempts to strike this balance in a different way. At the cost of the conversational nature of the task, the system retains the high degree of customizability that flash cards offer; however, through interactive card games, the system is able to convert the explicit memorization task implied by flash cards into one where the vocabulary acquisition is incidental to the game goals. Moreover, the integration of a Mandarin speech recognizer requires the user to manipulate the cards via *spoken* commands to complete the task.

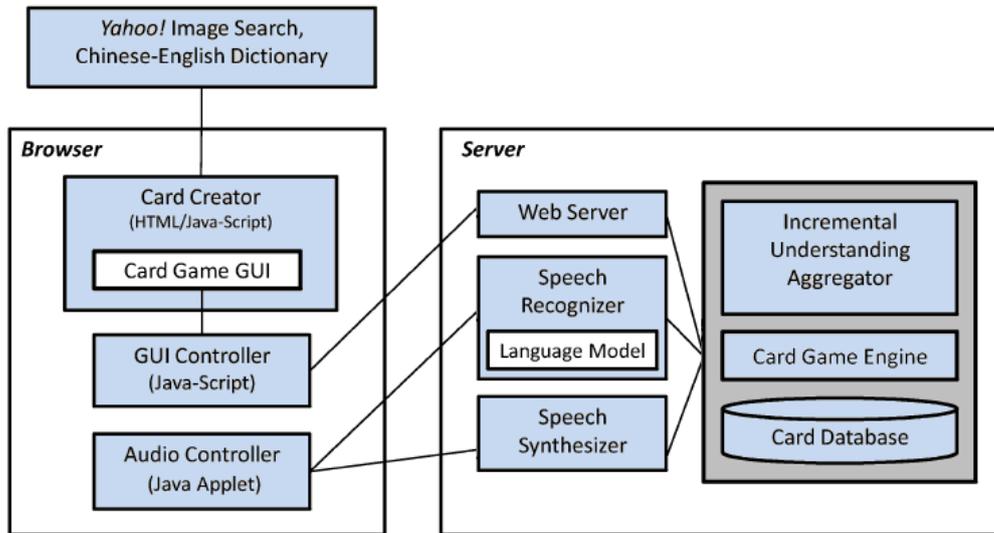


Fig. 1. Card game platform architecture.

3 Card Game Platform

The platform described in this section provides the core infrastructure for developing and deploying speech enabled card games. This platform is built upon a generic framework for Web-based multimodal interfaces that has been used in a number of recent speech-enabled systems at MIT [18].

Figure 1 depicts a block diagram of the platform architecture. A large component of the underlying technology is Asynchronous Java-Script and XML (AJAX), which allows the browser and Web server to communicate freely, and provides an infrastructure for developing highly interactive browser-based user interfaces. Audio is captured by a Java applet embedded into the Web site and streamed over the Internet directly to a speech recognizer sitting server-side. Synthesized speech can be generated on the server, streamed to the client, and played through this applet as well. These client-server connections seamlessly integrate MIT's speech technology into any modern browser.

One of the advantages to deploying a speech-enabled system on the Internet is that it allows the programmer to make use of the vast array of third-party Web-services already developed and deployed. As shown in the architecture diagram, our card game platform takes advantage of two such services. Using typical Web-programming techniques, we integrate a Chinese-English language learning dictionary¹ and Yahoo's image search² directly into a front-end card creation Web site. These tools enable teachers and students to quickly create entire decks of image-based vocabulary cards in just minutes. Figure 2

¹ <http://www.xuezhongwen.net>

² <http://images.yahoo.com>



Fig. 2. Web-based card creation tool.

shows a card being created for the Chinese word for ‘frog’. A few quick clicks saves the card into the back-end database, also shown in the architecture diagram in figure 1.

All of the system components mentioned thus far, as well as the incremental understanding aggregator described later, are available to any card game developed with this platform. The *white* boxes in the architecture diagram in figure 1 indicate the system-specific components that have been abstracted away from the core architecture. The language model for recognition, card game engine, and card game GUI must be provided as needed for each game developed. These components will be described further in the context of the *Word War* card game presented in the following section.

4 Word War

Figure 3 shows a typical computer aided vocabulary learning system: an on-line flash cards program. We provide this interface for students to use if they so choose, however, the main purpose of the card creation front-end described in the previous section is to allow students to choose vocabulary to load into a speech-enabled card game such as *Word War*.

4.1 Speaking Mode

We begin our discussion of *Word War* with the single-player mode in which the student speaks commands in Mandarin Chinese. Once a user has created a set of cards with the tools previously described, the player can load these cards into the *Word War* game grid, such as the one depicted in figure 4. Single-player mode is a simple picture-matching task. A student faced with the grid depicted in this figure might utter the Mandarin equivalent of the following commands: 1) “*Select the snake,*” 2) “*Drop it into the fifth square,*” and 3) “*Shift it to the left.*” As these commands are spoken, the computer



Fig. 3. On-line flash-cards.

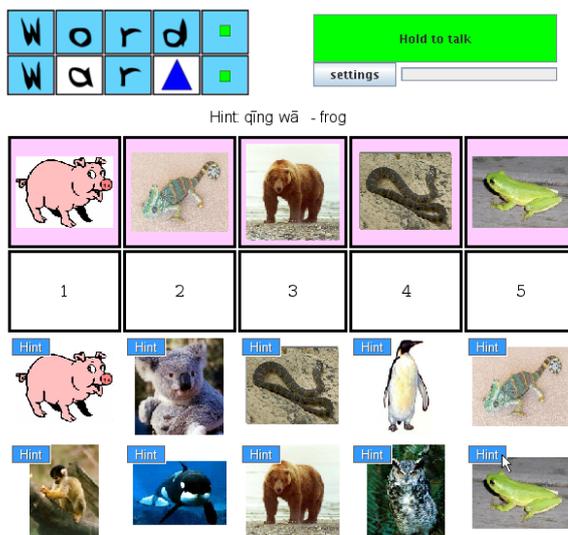


Fig. 4. *Word War* game grid.

carries out the corresponding *select*, *drop*, and *shift* actions in the browser. If the student is unsure of the pronunciation of a word, they are able to click and hold the mouse button while the cursor is hovering over the appropriate **Hint** button, as shown in the figure. Once all of the pictures are placed into the appropriate slots, the game is over and the player has the option to re-deal the cards.

From a technological perspective, the speaking mode of *Word War* required the development of a grammar, a GUI, and a game engine. The GUI makes use of standard Java-Script techniques, while the context free grammar (CFG) adheres to the Java Speech Grammar Format. Since the contents of the game are not known a priori, the grammar is generated dynamically when the game is loaded. Thus, the CFG generated contains only the vocabulary necessary for that particular game. This allows us to keep the grammar for a particular game small, while ensuring that the lexical domain for the system as a whole remains unlimited.

Once the grammar is generated, it is sent to the recognizer running on the server. We use the SUMMIT landmark-based recognizer [17] with acoustic models trained on *native* Mandarin speech. Although Mandarin is a tonal language, we avoid using tone-related features directly at the recognition stage, as it was felt that this would frustrate language learners with poor tone production skills. In future versions of our games, we hope to have a post-processing phase which analyzes fundamental frequency contours to provide feedback regarding tone, so that students may improve in this respect.

Given small grammars, key-value commands can be directly embedded into the CFG to enable the recognizer to then output hypotheses augmented with these understanding annotations. An example utterance in an English grammar for

Word War might be the following:

```
place [command=drop] it in slot five [number=5] and move [command=move] it to the left [direction=L] three [number=3] squares
```

One advantage of incorporating understanding in this way is that the recognizer can generate partial-hypotheses mid-utterance, which can then be understood by our incremental understanding aggregator (IUA). The IUA extracts the key-value pairs *as* they are produced, allowing the game engine to determine the appropriate actions for the client to display on-the-fly. The result is that the student sees a graphical interpretation of their utterance *while* they are speaking it. This is in contrast to typical speech-enabled systems which require the user to wait until after the entire utterance is spoken before receiving a response.

We use incremental understanding to add a degree of visual feedback unparalleled even in human-human interaction. Imagine a *Word War* grid containing geometric shapes of various sizes and colors. Saying the phrase “*Choose the large blue triangle,*” will result in the following sequence of visual reactions: 1) *all* shapes will be selected, 2) only the *large* shapes will be selected, 3) only the *large blue* shapes will be selected, 4) only the *large blue triangle* will be selected. In this way, the incremental understanding allows the user to practice speaking fluently while continuously checking that their words are indeed being understood. We have even encoded false starts and disfluencies into our grammar so that, if the student or system makes a mistake, the user can provide an *immediate* correction.

4.2 Multi-player mode

In this section, we describe how the connectivity of the Web can be harnessed to turn the single-player picture matching task into a more exciting two-player race. In multi-player mode, the initial configuration of the *Word War* game grid is similar to that of the single player mode. Each player independently chooses the vocabulary that they would like to load. The goal of *Word War* in multi-player mode is still to use spoken commands to move images from the bottom two rows of the grid into the numbered slots in the second row, so that they match the images along the top. However, the players now compete over shared access to the second row.

Figure 5 shows a snapshot of two players competing on the five-column game grids of multi-player *Word War*. When an image is matched, the slot is *captured* and the matching image appears on *both* players’ game grids. Notice that, in the figure, Player 1 has captured the third and fourth slots, while Player 2 has only captured the first slot. On the five-column game grids shown, the first player to fill three of the five slots is declared the winner.

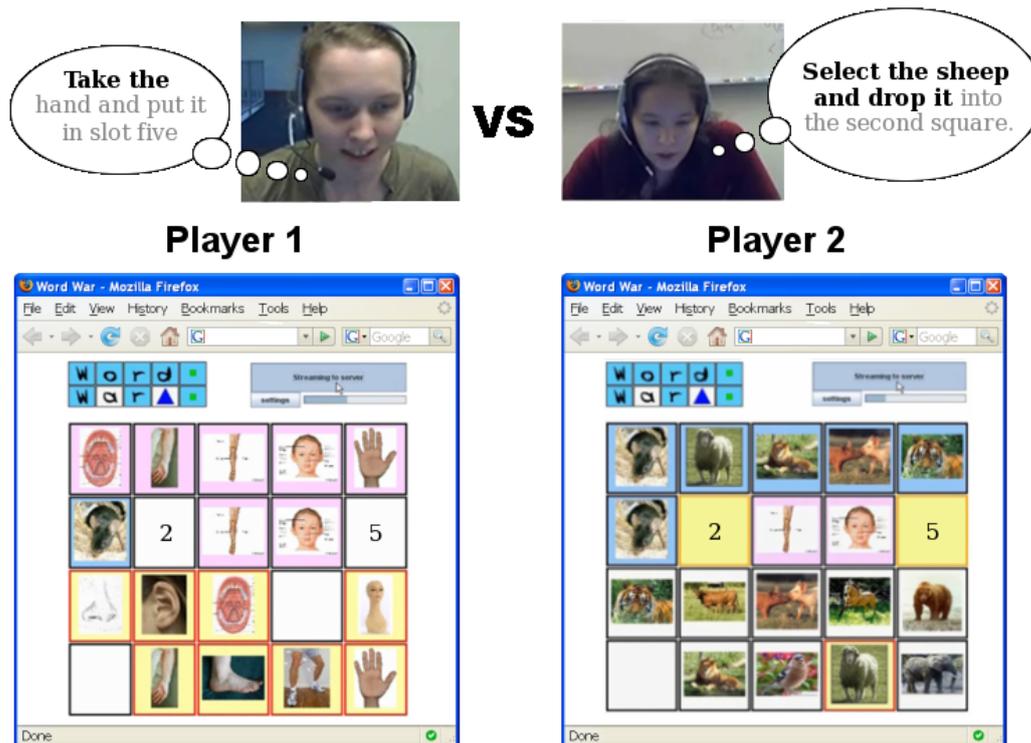


Fig. 5. Two-player *Word War*. Students compete to fill in the shared image space on the second row.

In figure 5, each game grid depicts the state of the incremental understanding according to the *partial* utterance, emphasized in bold text, of the corresponding player. Thus, by the time Player 2 has said the words “*select the sheep and drop it...*”, the incremental understanding engine had sent messages to the browser instructing it to highlight the image of the sheep and the slots into which it might be dropped. Similarly, Player 1 has said only the words “*Take the...*”, and the computer has selected all the images in preparation of narrowing the selection down to a single item.

In multi-player mode, the incremental understanding nature of the speech architecture becomes particularly important with respect to game strategy. The constant visual feedback allows adept students to issue multiple commands in a row, enabling them to speak fluently without pause, confident that they are being understood. To win the game, students must place multiple words in short term memory and balance speed with articulation while speaking recognizable commands quickly to accomplish their goal.

4.3 *Listening Mode*

Discussions with teachers led us to implement a listening mode of *Word War*, in which it is the computer who gives the oral directions. From the student’s perspective, listening mode is the reverse of speaking mode. Here the student is able to *manually* manipulate the cards by clicking the images and dragging them to the appropriate squares. This would be far too easy if the target images were left visible, so in this mode we hide the top row of target images from sight. Instead, the computer speaks commands in Mandarin, which the student must then follow. When the student has attempted to place all of the images in their appropriate locations, the first row is exposed to reveal the aligned target images.

At a minimum, implementing listening mode only requires that pre-recorded sound files be associated with each source image (e.g. “Select the big red square”), and each target slot (e.g. “Place it in slot five”.) In keeping with the theme of customizability, however, we do not wish to require a native speaker to record every single utterance the system will need to say. Thus, our card games rely on powerful speech synthesis technology to generate the necessary audio automatically. For the experiments described later in this paper, we chose to use Envoice[46], an in-house, concatenative speech synthesizer. In this fashion, we recorded a tiny corpus of template sentences and each of the vocabulary words. Envoice then spliced the relevant portions of the sound files together on-the-fly when the system needed to utter a new sentence. We have since integrated WISTON [43], a large corpus-based synthesizer for Mandarin, which circumvents the need to create an audio corpus for individual games.

4.4 *Pedagogical Considerations*

Critical in illuminating the pedagogical underpinnings of *Word War* is understanding what it means to *know* a word. Clearly there are many aspects of knowing a word including, but not limited to, its meaning, pronunciation, written form, morphology, and grammar collocations. Orthogonally, a distinction exists between *receptive* knowledge, which refers to the ability to recognize these aspects of a word, and *productive* knowledge which entails their expression, either in oral or written form.

Since our focus is on how speech technology can play a role in vocabulary acquisition, we will not focus the written form of a word. Instead, we posit that our games might enhance both productive and receptive *oral* vocabulary knowledge.

One criticism that might be leveled at *Word War* as described thus far, is

that, though it uses images to avoid the source language, the game does not do enough to provide a variety of contexts for vocabulary items. This line of argument would suggest that using a single picture as context encourages internalization of picture-word associations rather than a word's underlying meaning. Perhaps more importantly, additional linguistic context is necessary to provide a pedagogically sound environment for acquiring knowledge of word use. The question of linguistic context raises a second, related, criticism which is that the range of vocabulary items *Word War* can support seems quite limited.

A little creativity can go a long way towards increasing the pedagogical power of the card game and addressing the question of context. A given vocabulary word might, of course, have more than one image associated with it, chosen randomly at game time, to ensure that the learner's understanding goes beyond a single picture-word association. Perhaps more interestingly, however, a given picture might warrant a description containing many words, allowing the game to go far beyond simple concrete nouns. Administrators of our Web-site even have the ability to associate entire context free grammars with a single image-based card. Combining multi-word cards with personal photos uploaded via the card creation interface provides enormous opportunity for the provision of meaningful linguistic context. Imagine playing a game of *Word War* with family vacation pictures, e.g. "Choose the photo in which my brother has finally forgiven me for stealing his green beach towel."

A third criticism of *Word War*, is that it does not adhere strictly to the tenants of a particular teaching approach. Proponents of communicative language teaching, for instance, might not see opportunities for students to "negotiate meaning" in *Word War*. In previous work, our lab has indeed attempted to create full-fledged dialogue systems for vocabulary acquisition in narrow domains [30]. In this work, negotiation of meaning played a crucial role. We found, however, that such systems are expensive to develop and limited in lexical scope. Following this strategy prevented us from providing content tailored to the needs of the individual student or teacher, which we eventually determined to be of higher priority.

With these caveats, we believe that *Word War* still falls under the umbrella of task-based learning and by extension a communicative approach to vocabulary acquisition. Unlike simple flash-cards, *Word War* provides an environment where students can both understand and speak the foreign words in a meaningful context. Notice, in particular, that the listening and speaking modes of *Word War* together represent two sides of a language teaching paradigm that is almost always present in well-implemented communicative curricula: *the information gap* [28]. Put simply, an information gap exercise is one in which a meaningful exchange of information must take place in order to complete the task. While these tasks are easy to implement with small groups, they are



Fig. 6. Map indicating points-of-access.

almost never assigned as homework for the following very simple reason: there is rarely anyone at home with whom to exchange information in a foreign language.

Although properly motivating an ICALL system is important, it cannot replace studies that examine their application in practice. In the next two sections, we will discuss two such studies in which we invited learners of Chinese to interact with *Word War*. We begin with an assessment of speech recognition performance for a pilot experiment conducted with users who interacted *remotely* over the Web. Subsequently, after reviewing previous experiments conducted by Rod Ellis that provide motivation for our experimental design, we describe in detail a second *laboratory*-based study intended to measure user retention of the practiced vocabulary after a delay period, as well as to solicit feedback from users as to their impressions of the games.

5 User Study 1: Assessing Recognition Performance

The Web can be a wonderful environment for collecting authentic user interactions from a variety of users across the world. Figure 6 places pins on a map at each location where at least one user has started a game of *Word War*. At the same time, however, data harvested from arbitrary interactions can pose problems during evaluation. Upon listening to one session of *Word War* collected in this manner, it became apparent that a father and a daughter were practicing Mandarin together, at times sounding out words simultaneously!

In order to obtain realistic interactions, while still retaining some degree of control over the nature of the data, we conducted a remote user study that re-

SER by Task (%)						
t1	t2	t3	t4	t5	t6	All
17.5	17.4	15.9	16.1	13.1	15.9	16.02

Notion	Count	AER (%)
<i>select</i>	777	12.23
<i>drop</i>	778	17.56
<i>shift</i>	35	20.00
total	1590	15.01

Fig. 7. Error rate breakdown.

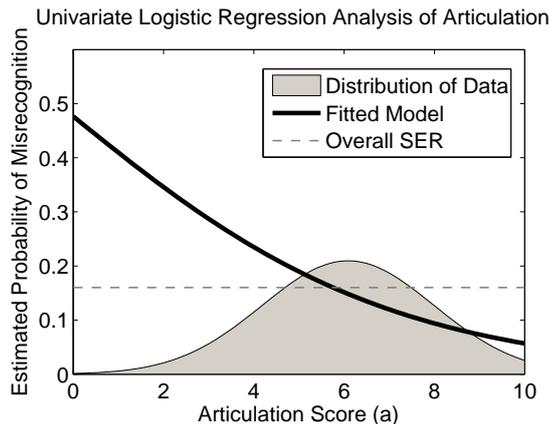


Fig. 8. SER as a function of articulation.

quired users to register at a publicly deployed version of our Web site. Twenty participants with at least one year of Mandarin experience signed up and completed the user study from their own homes. Thirty words were used in the study: 10 were animals, 10 were plants, and 10 were food items. After a short tutorial, each participant was required to complete two *Word War* games in each category.

We collected 1543 utterances from the six non-tutorial tasks, and asked a native Chinese speaker to annotate them. About 5% of the utterances were unintelligible to the annotator – usually short aborted attempts at speaking a command in Mandarin. We eliminated these utterances from further analysis, leaving us with 1467 fully annotated utterances. For each utterance, our annotator determined the action that it represented. With the semantics of each sentence labeled, we were able to determine the accuracy of the speech recognizer in terms of *sentence error rate* (SER). The overall sentence error rate of the 1467 utterances was 16.02%. Figure 7 shows a breakdown by task. From this figure and by replaying user-computer interactions once the study was complete, it became clear that users tended to have fewer recognition problems as they become more comfortable with the speech modality of the system.

Also shown in figure 7 are error rates in terms of individual actions, e.g. “*Select the dog.*” Recall that the incremental understanding aggregator allows a given sentence to contain more than one action command. We extracted 1590 actions from the 1467 annotated utterances. Action error rates (AER) are reported for the *select*, *drop*, and *shift* actions. Note that since recognition performance was fairly accurate, the students rarely had to resort to a *shift* action, which is typically used to correct an error in recognition.

Our annotator also rated the speech along four metrics: articulation (a), tone (t), speed (s), and sound quality (q). These metrics were rated on a scale from

1 to 10. For example, after listening to a number of speakers to get a feel for the average speed of a learner’s speech, she was able to give a rating (s) for each utterance. Since the speech was recorded in varying acoustic environments, the audio quality metric (q) attempted to assess the degree to which microphone problems and background noise were present in a recorded utterance. Because Chinese is a tonal language, (t) measured the accuracy of the learners tones. Finally, articulation (a), was a measure of the non-nativeness of a student’s speech independent of tone. Our annotator said that for this metric she would listen to the utterance, repeat it to herself correcting for tone and speed, and then determine a score.

In [31], we provide a detailed error analysis in terms of the speech metrics rated by our annotator. Using these metrics as independent variables, and a sentence error as the dependent variable, we performed four univariate and one multivariate logistic regression analyses. Here, we show only the effects of the most significant independent variable, articulation (a), and refer the reader to our previous paper for the full discussion. Figure 8 shows the univariate regression model for articulation, providing an estimated probability of misrecognition as a function of the articulation score. It is interesting to note that utterances with high articulation scores ($a > 8$) achieve on average over 90% accuracy.

6 User Study 2: Assessing Educational Value

While recognition error is an excellent metric for assessing usability of the system, a critical concern is whether the games lead to retention of the vocabulary items on the part of the student. Thus, we designed experiments involving three distinct systems, which were introduced in section 4: the image-based flash card system shown in figure 3, the single-player *speaking* mode of *Word War*, and the single-player *listening* mode of *Word War*. In a carefully controlled study, 13 first and second year students of Chinese interacted with all three systems. Tests were given to measure prior knowledge and learning gains over the course of a three week interval.

In our experiments, we chose to examine the acquisition of declarative knowledge, e.g. word meanings, as opposed to procedural knowledge such as word use. In particular, the words chosen for our experiments are concrete nouns. Fortunately the *relatively* straightforward transfer of concrete nouns into procedural knowledge is made easier in Chinese due to the lack inflectional morphology, e.g. plural forms.

The acquisition of word meanings has been examined extensively in applied linguistics literature; thus, there exist rigorous methodologies for testing their

acquisition. In the next section, we describe two such studies performed by Rod Ellis, one of the leading researchers in Second Language Acquisition, and go on to use one of these studies as a model for our experimental setup. After presenting the preliminary findings, their implications are discussed with respect to the utility of these three systems as vocabulary acquisition aids.

6.1 Previous Studies in Vocabulary Acquisition

Most studies involving vocabulary acquisition only analyze the effects of various strategies of *intentional* learning [5,19]. Those studies that do research the effects of *incidental* acquisition on vocabulary retention most often focus on learning through reading rather than oral input or output. Two studies performed by Rod Ellis, [11,12], break the mold.

In 1994, Ellis examined the role that interactionally *modified* oral input plays in the acquisition of word meanings in a large scale user study involving Japanese learners of English [12]. The setup of this study is remarkably similar to the listening mode of *Word War* presented in section 4.3. In this study, the students are given a set of vocabulary items in picture-form and the teacher directs them in the target language to place these pictures into a number of possible positions. In this case, however, the vocabulary items are all kitchen-related and they must be placed in the appropriate spot on a picture of a kitchen. In the task, the teacher might give the following instruction: “*Please put the broom on the floor in front of the stove.*” The researcher can then check for comprehension by examining the contents of the picture after the teacher has completed the list of instructions.

Ellis splits the participants of his study into three groups and gives each group one of the following treatments: *baseline* input, *modified* input, and *premodified* input. The baseline input is in the form of directions that a native speaker might give to another native speaker to accomplish the same task. Modified input is in the form of directions similar to the baseline input, but in which the learner has the opportunity to ask clarifying questions, e.g. “what is a broom?” Premodified input is read more slowly and the directions are augmented a priori with the sorts of paraphrases and definitions that one might find in modified input, e.g. “We have the broom. A broom is a long stick with some kind of brush and you use it to clean the floor. I’d like you to put the broom on the floor in front of the stove.”

The directions in the listening mode of *Word War* fall somewhere in between the baseline and premodified input types defined above. The directions are given at a speed slightly slower than a native speaker might typically speak them, but they do not contain the sorts of paraphrases and definitions as-

sociated with the premodified input in the Ellis study. They are, however, simplified in that only one possibly unknown vocabulary item appears in a given instruction.

At first glance, the results of the Ellis study suggest that interactionally modified input is superior to either premodified or baseline directions in terms of vocabulary acquisition as measured by the post-tests. Surprisingly, however, this study and many similar studies in Second Language Acquisition ignore one variable that is particularly important to a second language learner: time. In [11], Ellis reflects on the results of his 1994 study and previous studies of a similar nature, “A problem arises in interpreting the results of these studies both with respect to comprehension and acquisition. The tasks that supplied interactionally modified input took longer than those based on premodified input. We cannot tell, therefore, whether the interactionally modified input works best because it enables learners to sort out misunderstandings and construct a shared mental model of the task at hand [...], or because learners have more time to process the input.”

Reexamining the data in the 1994 study, it becomes clear that the picture is strikingly different when time-on-task is accounted for. A quick computation reveals that in terms of the mean number of words acquired per minute, the premodified and the baseline input groups were almost identical, while the interactionally modified group was two to three times *slower*. In [10], Ellis notes that the differences in these rates are significant, and discusses in detail the factors that distinguish premodified and modified input as they relate to vocabulary retention.

In a subsequent study performed in 1999, Ellis is more careful to control for time and replaces the baseline group with a new treatment: *modified output* [11]. The modified output group required students to work in pairs, each taking turns giving directions. Each group in this study was given exactly 45 minutes to complete their tasks. In this study, Ellis is not able to show significant differences in rates of acquisition between the premodified and interactionally modified input groups, but *is* able to show that the *modified output* group performs significantly better than either of the two groups that use input alone.

While the research questions addressed in this section are slightly different from those posed by the work of Rod Ellis, his 1999 study exemplifies a rigorous methodology for assessing various treatments on the vocabulary acquisition process. Whereas he is focused on the distinction between modified and premodified input and output *within* the realm of incidental vocabulary acquisition, the user study described below attempts to examine the relationship between *intentional* and *incidental* vocabulary acquisition with respect to the three computer assisted vocabulary acquisition systems previously described.

Although SLA theory might suggest that incidental vocabulary acquisition offers pedagogical advantages, it is not clear whether, when time is taken into account, methods that do not focus explicitly on the memorization task will be as efficient as *intentional* vocabulary learning. Clearly the more communicative approaches in the Ellis studies were not always the most efficient. This is not to say that a methodology that requires more time for vocabulary acquisition is necessarily less valuable. Perhaps it is indeed the case, for instance, that the most *efficient* manner in which a student can internalize new word meanings is through brute force memorization. If the student does not enjoy this task, however, the words-per-minute memorized may be of little value, since the student is unlikely to want to spend much of their time on this task in the first place. In general, we would like to be able to quantify the efficiency with which a given method leads to long term retention of lexical items and, as best we can, assess whether a tradeoff exists between this efficiency and the level of interest of the student. Put more succinctly: does the appeal of *incidental* learning methods come at a cost of the student's time? And also, is this a price the student would willingly pay if such methods are more enjoyable?

6.2 *Experimental Design*

In this user study, a preliminary attempt is made at answering these questions in the context of the three applications for computer aided vocabulary learning already described. Although short term memory effects are measured, they are not the focus of this study. After all, when learning a language, memorizing 50 words in five minutes is of no practical value if the student forgets them all in ten. The following research questions are to be addressed via the subsequent experimental design:

- (1) What effect does the speaking-mode of *Word War* have on vocabulary retention in the long term?
- (2) What effect does the listening-mode of *Word War* have on vocabulary retention in the long term?
- (3) How do the effects of the systems above compare with the retention rates of students who are given an explicit memorization task.

Initially, 15 participants from local universities signed up to participate in a three week laboratory-based study with the promise of receiving two \$50 gift certificates for attending all three weeks. At least one semester of Chinese experience was required for the subjects of this study, since they needed to be familiar with Mandarin's basic pronunciation rules. Five of the students were drawn from a second year Chinese course at Harvard University, four were just finishing up a first semester course in Chinese at MIT, and the remaining

	pillow zhěntou		lottery ticket cǎiquàn		necklace xiàngliàn
	bomb zhàdàn		leopard bàozi		cleaning rag mǒbù
	finger print zhǐwén		batteries diànchí		pond chítáng
	baby yīngér		diamond zuànshí		playing card zhǐpái
	jellyfish shuǐmǔ		cell xìbào		honeybee mìfēng
	matches huǒchái		buttons kuàzi		butterfly húdié
	comicbook mànhuà		eraser xiàngpí		soap féizào
	tongue shétou		bathtub yùgāng		iron yùndǒu
	kite fēngzhēng		earthworm qiūyǐn		rose méiguī
	penguin qǐ-é		eggplant qiézi		bat biānfú

Fig. 9. The 30 vocabulary words used in the learning study.

six were from a second year Chinese course at MIT. Unfortunately the data we obtained from two of the second year MIT students was unusable due to technical problems in the initial phases of the study. This left 13 students with a variety of backgrounds who successfully completed the three week study.

Unfortunately, our limited sample size prevents us from splitting our participants into three independent groups in a manner similar to the Ellis studies. However, given that our systems are computer-based, we have the luxury of distributing the words across our various systems at will. Thus, for each student, we instead shuffle a fixed set of word cards, deal them to the three systems, and compare retention levels for the words conditioned on the mode by which they are learned. The study contained 30 Chinese words (W), shown in figure 9 that the students would try to learn over the course of the three week period.

An attempt was made to choose words of roughly equal difficulty. We avoided words that were likely to be taught in the first two years of university Chinese. To ensure that the words were cognitively comparable, we required that they be exactly two syllables, preventing students from making picture-word associations based solely on word length. While nouns with two or more syllables are more common in Mandarin [1], it should be noted, that the many monosyllabic words could, in general, adversely affect recognition results. To mitigate this, a monosyllabic noun could easily be supplemented with a monosyllabic adjective to avoid possible recognition problems, e.g. “Select the white dog.”

In two episodes spaced one week apart, each student learned three sets of ten words each, under three different conditions: (1) Image based flash cards, (2) *Word War* Student-Listening mode, and (3) *Word War* Student-Speaking mode. A picture matching test, similar to those employed in [11,12], was used to assess retention of the word meanings, and a survey was conducted at the end of the first two sessions.

The listening and speaking modes of *Word War* were kept as similar as possible. Hints consisting of the pin-yin pronunciation and English translation were available in either system; no Chinese characters were shown and no corresponding audio was played when the hint was in view. Although we feel that the student would benefit from hearing the word spoken correctly we disallowed this in the experiments in order to assure a clear distinction between listening mode and speaking mode in *Word War*.

The flash card system allowed a student to review words as many times as they please, in any order. Students could therefore focus specifically on the words they found most troublesome. By contrast, the *Word War* game automatically selects a pseudo-random subset of the words to be used in each episode of the game. Naturally, some words will require more repetitions to ensure acquisition; thus, we designed a simple algorithm to monitor students’ previous games and to bias selection of the words for each game on the basis of the number of times the word had previously been played, weighted against the number of times the student clicked the  button for that word. A simple “Re-deal Cards” button automatically reshuffled the cards and restarted the game with a new set of images, tending to favor the ones that needed more attention. Future versions of the software could incorporate more involved algorithms to perform optimal short-term and long-term scheduling based on cognitive models of retention [37].

All evaluations were performed using the picture matching test. In this test, a list of pictures and the English words they represent were to be matched with the pin-yin transcription of the corresponding Chinese word. Chinese characters were not present in any of the systems or tests as the focus of this study was on spoken vocabulary retention. The picture matching test format

was chosen because it was felt that it would provide the most sensitive measurement of vocabulary acquisition, given that the students would have very limited exposure to each word. One drawback to such a test is that it appears to favor the flash card system where the written form of the pronunciation takes a more central role in the learning process. A similar test was used in parts of the Ellis studies described in the previous section.

6.3 Procedure

Since section 5 extolled the virtues of ensuring that the *Word War* system was easily deployable over the Web, it may seem incongruous that, in this user study, the 13 subjects were required to participate from within the confines of our laboratory. However, when attempting to determine the effects of the individual systems on vocabulary retention, control is absolutely paramount. In a remote user study, there could be any number of hidden variables (poor microphone setup, cheating, etc.) that factor into the measured learning gains. Thus, the 13 participants were required to attend three hour-long sessions held in our laboratory, each spaced one week apart.

The first two sessions of the study required that *each* participant interact with *all* of the acquisition aids in succession, encountering 10 words in each system. Since we are primarily interested in the long-term learning gains of the system, the portion of the last session relevant to this study consisted only of the picture matching test. A diagram of the activities scheduled for each of the sessions is given in figure 10. The first two sessions were broken up into an initial test, followed by 10 minutes with the flash cards and 10 minutes with each mode of *Word War*, interspersed with quick short-term memory quizzes. The 10 minute *Word War* tasks required the user to complete as many games as they could before their time expired.

A picture matching test involving all 30 words was administered at the beginning of each session. When student i signed up for a new account during his or her *first* session, the 30 words were dealt randomly into three piles: $W = \{W_F^i, W_S^i, W_L^i\}$. Once the piles were created for a student i , they remained the same when that student returned for the second session. As indicated in figure 10, each pile was also associated with a system so that, when user i loaded that system they always saw the same cards: W_F^i was associated with the flash cards, W_L^i with the listening mode, and W_S^i with the speaking mode. Thus, in weeks one and two each user encountered the same 30 words; however, a word that appeared in the flash card system for one user might have appeared in the listening mode of *Word War* for another user. Notice also that weeks one and two require each student to perform the same tasks, with the same words, except that the order in which the systems are encountered

Session 1:

Activity:	Test 0	Flash cards	Quiz	Listening	Quiz	Speaking	Quiz
Time:	∞	10 min.	∞	10 min.	∞	10 min.	∞
Words:	W	W_F^i	W_F^i	W_L^i	W_L^i	W_S^i	W_S^i

0 min. $\xrightarrow{\hspace{15em}}$ ~60 min.

Session 2 (one week later):

Activity:	Test 1	Listening	Quiz	Speaking	Quiz	Flash cards	Quiz
Time:	∞	10 min.	∞	10 min.	∞	10 min.	∞
Words:	W	W_L^i	W_L^i	W_S^i	W_S^i	W_F^i	W_F^i

0 min. $\xrightarrow{\hspace{15em}}$ ~60 min.

Session 3 (one week later):

Activity:	Test 2
Time:	∞
Words:	W

Fig. 10. The setup of the user study to assess learning gains on three systems: flash cards (F), *Word War* listening mode (L), and *Word War* speaking mode (S). The 30 words W contained in the study were shuffled and dealt into three piles W_F^i , W_L^i , and W_S^i for each student $i = 1, 2, \dots, 13$.

is altered.

After a user encountered a given pile of words in a particular system, a short-term memory *quiz* was given. These quizzes were in the format of the picture matching test previously described, but only contained the words just seen in the interaction with the most recently used system. The *tests*, given at the start of each session, were used to measure long term vocabulary retention. Test 0, given before any student-system interaction took place, was therefore used to assess a-priori knowledge of the vocabulary items. Tests 1 and 2, given a full week after sessions 1 and 2 respectively, were designed to measure the effects of the three systems on vocabulary retention over a longer period of time. On all tests, students were discouraged from guessing randomly.

Considerable effort was made to minimize the possibility that a user simply did not understand the user interface, and was thus not able to use the system efficiently. Not shown in figure 10 are a set of tutorial activities for each system that were given just prior to the student's encounter with that system. The tutorials were prefaced by a video demonstrating the actions that a user would take, and consisted of a short period during which the user was able to interact with the system, which was initialized with a set of 10 tutorial words that the users were never tested on.

For the flash cards and listening game, these tutorials were sufficient to ensure that the users were accustomed to the interface. Given that previous ex-

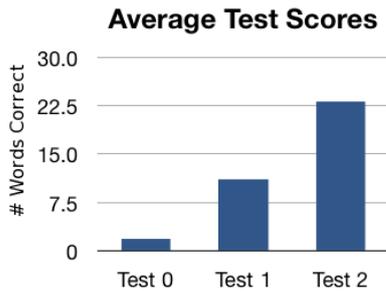


Fig. 11. Average test scores for the picture matching test given at the start of weeks 1, 2, and 3.

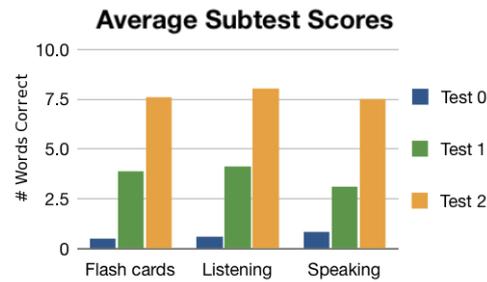


Fig. 12. Average subtest scores for each system computed by grading the words studied in a particular system.

periments with the speech-enabled *Word War* game have yielded noticeably different behaviors depending on how long a given user has been playing [31], the tutorial mode for the speaking system was particularly important. Unfortunately, since each session was only 1 hour in length, the time spent on the tutorial session had to be limited to just a few minutes.

The tasks that required time limits included a Java-Script timer built into the Web page. When the time had expired the students were automatically brought to the next task. Tasks that did not require time limits were completed when the user indicated that they had finished the task by pushing a button on the Web page.

At the end of sessions 1 and 2, students were asked to fill out a short survey. Some questions were open-ended inquiries into the student’s previous experience studying Mandarin, others asked about their study habits, and still others elicited quantitative answers regarding their experience using our three vocabulary-building systems.

6.4 Learning Gains

Figure 11 shows the scores for tests 0, 1, and 2 given at the start of each session. The scores from test 0 indicate that the words we chose were relatively unknown to the study participants beforehand. Five students received a score of 0 out of 30, six students had a score between 1 and 3, and the two remaining students both answered 7 questions correctly.

When grading the quizzes, which were administered immediately after the user studied the words contained therein, it became apparent that all three treatments ensured that these words entered the students’ short-term memories. Of the 13 students, 11 got perfect scores on all three of the first session’s quizzes. The two that did not were both first year students. The first of these students missed questions on all three quizzes, while the second answered 3

	Flash Cards	Listening	Speaking	Full
g_1	0.36	0.38	0.29	0.344
g_2	0.69	0.71	0.70	0.702
g_B	0.76	0.79	0.75	0.768

Fig. 13. Mean learning gains for each system across all 13 users.

questions incorrectly on the quiz following the *Word War* speaking mode.

Although from the student’s perspective the full tests contained all 30 words W , we can define a notion of a *subtest* for each of the three systems and grade these individually. For student i , the subtest associated with system X would be graded by scoring only those words in the test that appeared in the set W_X^i . Since the words that a single student saw across weeks one and two were the same for a particular system, we can also compare subtest scores across weeks. In this way, one can deduce the relative effectiveness of each system in teaching the student a particular set of words. For a single student, the words in each subtest are different, so the results are more meaningful when averaged across all of the students in the user study. Figure 12 plots the average subtest scores for each of the three systems across all three weeks.

A more refined analysis would compare not just absolute test scores, but individual learning gains across the tests. Since some students were able to achieve the maximum score, we use the notion of normalized learning gains, which is defined as follows: $g = (S - R)/(T - R)$ where, R is a pre-test score, S is a post-test score, and T is the total number of questions. In the context of our vocabulary tests, the gain is the number of previously unknown words that the student answers correctly on the post-test, divided by the total number of previously unknown words. In this way, learning gain takes into account prior knowledge without penalizing those who cannot learn more simply because they are nearing the maximum score of the test (or subtest).

Learning gains were computed individually for each student and then averaged to produce the values in figure 13. We compute a learning gain g_1 from week 1 to week 2, a gain g_2 from week 2 to week 3, and a gain g_B from week 1 to week 3. By applying the learning gain equation to the *full* test scores, we produce the results shown in the column labeled “Full”. The gains associated with each system are obtained by applying the equations to the words that appeared in that system, i.e. to the *subtest* scores. Figure 13 displays the results of these calculations as well.

Treating the student and the system as independent variables, a two-way ANOVA reveals that there were no significant differences among means ($F < 1, p > 0.3$). Paired t-tests do indicate, however, that the learning gains achieved

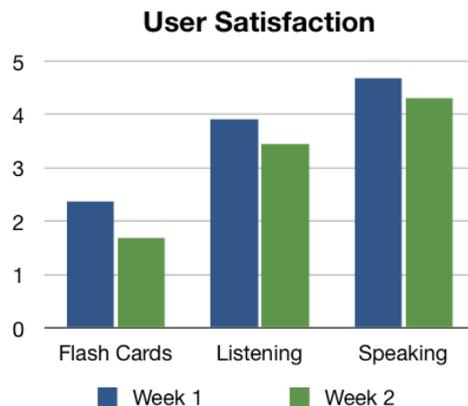


Fig. 14. The average response to the following question: “On a scale from 1 (not at all) to 5 (very much), to what degree did you find interacting with this system enjoyable?”

by each system individually improved significantly between weeks one and two, with $p < 0.01$ for all three systems.

6.5 Survey Results

The surveys given at the end of the first two sessions contained a variety of questions. The questions with answers that can be summarized easily across the participants are presented here.

To estimate the extent to which each of our three systems kept the participants engaged, the survey asked students the following question: “To what degree did you find interacting with this system enjoyable?”. The students were required to respond using a Likert scale from 1 to 5, where 1 was used to indicate “not at all”, and 5 was used to indicate “very much”. Since we asked this question after both sessions one and two, we can compare the responses across both systems and weeks. Figure 14 shows the mean responses received for each of the system/session combinations.

In this case, a two-way ANOVA for each week indicates that there are significant differences among the means of the systems with $p < 0.001$ in both cases. An a posteriori Tukey-Kramer HSD test reveals that in week 1, the mean score for the flash card system is significantly different from both of the other two means, however the speaking and listening mode means are not significantly different from one another. It should be noted, however, that nine of the thirteen individuals chose to give the speaking system a score of five during the first week, causing a ceiling effect. During the second week, no such effect was present, and all differences were found to be statistically significant. When comparing the means for each week *within* a system the differences trend to-

wards significance with $p < 0.1$ in all three cases, indicating that users found all the systems less enjoyable the second time around.

A second set of questions answered using a Likert scale attempted to ascertain whether people felt comfortable interacting with the speech recognizer relative to when they were required to speak in their classes. The questions were posed as follows: 1) “To what degree did you feel nervous/embarrassed when interacting with the speech-enabled system?” and 2) “To what degree did you feel nervous/embarrassed when you are asked to speak in class?” Again the scale was from 1 (not at all) to 5 (very much). The mean response to the first question was 1.38 with a standard deviation of 0.96, while the mean response to the second was 2.38 with a standard deviation of 0.87. A paired t-test reveals that this difference is also statistically significant.

A final question regarding the use of the flash card system is also worth noting: although we did not observe their behavior first-hand, we asked users whether or not they spoke the words aloud when using the flash card system. All 13 participants answered “yes” to this question.

6.6 Discussion

First and foremost it should be noted that, as figures 12 and 13 indicate, all of these systems are extremely competitive with respect to vocabulary retention. There is less than a 5% difference between the average overall learning gains g_B achieved by each system, and just a 1% difference between the speaking mode of *Word War* and the flash cards system. This indicates that, at least with respect to these three systems, there was not a significant loss in efficiency when using the *incidental* acquisition methods over the *intentional* one. Indeed, the listening mode of *Word War* actually performed slightly better on all measures of learning gains.

The relatively low gain achieved by the speaking mode during the first session has a number of possible explanations, both pedagogical and technical. First, it may be the case that placing a word in one’s productive memory, so that it can be spoken, is simply more difficult than placing a word in receptive memory, where it can be understood. Second, it is quite plausible that the tutorials for the speaking mode were not sufficient to ensure that the users were accustomed to this relatively novel user interface. More than once, when a confused participant was unable to navigate the speaking-mode tutorial, a study administrator had to tell that individual that the microphone on their headset needed to go *in front* of their mouth rather than folded behind their ear. Lastly, it could be that some students were unable to correct for pronunciation problems given that no *audio* hints were allowed in the speaking

mode. This might have led to a certain amount of wasted time repeating a command containing a single troublesome word.

It is also interesting to note that not all of the users made use of the incremental understanding feature during their first session. That is, although students were told that they could speak multiple commands in a row, only 6 of the 13 individuals attempted this during the first session. Of those individuals, 3 made heavy use of this feature, at times matching all five target images with a single utterance. The remaining 3 used this feature sporadically. Examining the subtest scores of these individuals reveals that, for those students who made use of this feature, the speaking mode of *Word War* typically outperformed the flash cards system. Unfortunately, it is difficult to determine whether this is a causal effect or simply correlated.

Finally, we have not yet taken the time to manually annotate the thousands of utterances collected in this study; thus, for the moment, the role that recognition errors play in the vocabulary acquisition process of these systems will remain unknown. When reviewing the user interactions with the speaking mode of *Word War* it became apparent that a number of students had difficulty pronouncing the pin-yin solely from the written form alone. In future versions of the system, we certainly plan to add the synthetic speech used in the listening mode to the hints in the speaking mode, so that the student has a model on which to base their speech.

In summary, the analysis of the learning gains seems to suggest that the choice of system is not a significant factor in determining whether or not a word will be learned. This is particularly good news when we take into account the results of the survey, which indicate a strong preference for our incidental vocabulary acquisition games, and for the speaking mode in particular. As the survey results indicate, the students felt little discomfort when interacting with the speech recognizer, especially when compared with their experiences in the classroom.

7 Beyond Word War

The results of this study hint strongly at the possibility of devising new, more interesting card games that make use of similar implementation techniques. Clearly the speech-enabled aspect of *Word War* is valuable in terms of the enjoyment of the student. The slightly greater learning gains achieved by the listening mode are also noteworthy. The obvious conclusion is that a card game should be developed that combines the two modes in some fashion.

While it is clear that the *Word War* modes held a certain appeal in our user



Fig. 15. Rainbow Rummy screenshot showing the student’s hand (top row of three cards) and the board. The student can, for example, play the two penguins from their hand into a new slot, by stealing a penguin from slot 2. The adept player, however, can find a winning move from this configuration!

study, they still suffered a decrease in student-enjoyment in the second week. Furthermore, in some of the answers to the open-ended survey questions, it became clear that users desired a slightly more complicated interaction, both in terms of the sentence structures used and with respect to the task required of them.

These facts motivate the development of more complex card games, perhaps with rules akin to the traditional card games that many people grow up playing. With this in mind, we have completed development of a new card game based on manipulations of vocabulary items treated as playing cards. The game, which we call Rainbow Rummy, is similar to the card game *Gin Rummy* or the multi-player Chinese tile-game of *Majong*. Unlike in *Word War*, the student is able to experience *both* the listening and speaking modes of learning in a single Rainbow Rummy game due to its turn-taking dynamics.

In Rainbow Rummy, two “hands” are dealt from a deck of cards, and each player is required to play at least one card from their hand to the board before relinquishing their turn. Each card encodes both a color and an image representing a word to be learned. The goal of the game is to build “sets” based either on matching color or matching the image. Each set must contain at least three cards. Players are allowed to either create new sets or build on existing sets, displayed on the board. A screen-shot from Rainbow Rummy is shown in figure 15. An aspect that makes the game much more challenging is that players can move cards already on the board from one set to another, for example “stealing” the purple penguin from the second set in the figure to match the yellow and red penguins in the user’s hand. This makes game

strategy quite interesting, and can lead to situations where a complex rearrangement of the board is required before a given card can be successfully played from the hand. Whenever the player is unable to play a card they must draw new cards from the deck until they can play. The first player to empty their hand wins the game.

We have created an implementation of this game where the computer exhibits a fair amount of intelligence using AI strategy. We have designed the game such that, during the student's turn, the student speaks instructions to the computer, and the computer executes their moves according to the instructions. During the computer's turn, the roles are reversed. In this way the student experiences both listening and speaking modes of communication. The game has been made available in both English and Chinese at <http://web.sls.csail.mit.edu/chinesecards/>.

8 Conclusions

The studies presented in this paper are a first step towards understanding the advantages and disadvantages of incidental vocabulary acquisition in the context of speech-enabled card games for computer aided language learning. It is our belief that these card games have the potential to make a meaningful impact on language education because they are highly customizable, and because a variety of interesting games can be created using the same generic framework. Furthermore, due to their replay value, card games offer an advantage over narrow domain dialogue systems which might be of interest to a first or second year Chinese student for only a brief period of time.

A number of interesting ideas come to mind about how we might improve our existing *Word War* system. First, it is clear that the two-player game mode has the potential to be much more interesting than the single-player picture matching variant. Without a marketing strategy, however, the currently deployed system will only rarely have more than one user at a time. One simple extension that we have not yet implemented would be to *simulate* a two-player mode. In Rainbow Rummy, we pit the student against an artificially intelligent opponent that plays with its own hand of cards. The *Word War* system could also choose a set of cards and play them on the student's game grid at random intervals. It is conceivable that the system could even keep track of the user's proficiency and ensure that both the system and human are evenly matched.

Indeed, there are a number of opportunities to model the knowledge of the student from within the context of a card game. We might, for instance, expand upon the algorithm, described briefly in section 6, that uses hint-clicking

behavior to ascertain which words the student does not yet know. As previously described, some standard flash card systems employ spaced-repetition, a process by which vocabulary retention for each word is measured over a long period of time, and the computer estimates the optimal schedule for each word's review. For example, using the hint-clicking behavior as a guide, *Word War* might keep track of the vocabulary items that the user seems to already have memorized. If the user is willing to relinquish control of choosing the study material to the system, this algorithm might be used to choose which cards to load into the user's next game of *Word War*.

With respect to recognition, we would eventually like to integrate pronunciation assessment, particularly automatic tone evaluation, into our games. Ideally, we might be able to leverage the work on tone assessment already taking place in our laboratory [38], to create an adjustable garbage model which tosses out poorly pronounced utterances, and requires the user to repeat them.

Finally, to ensure that a user is never in a position where they are unable to continue with a *Word War* game due to pronunciation problems or recognition errors, we could implement a manual back-off mechanism. That is, if the user is unable to use speech to properly place an image in its target location after a fixed number of failed attempts, the system could then allow the user to instantiate a move manually by clicking a card or slot. The system would then execute that move *and* speak the corresponding phrase describing the move, e.g. "select the necklace." We could even choose to hide the click-able hints entirely, instead relying on this back-off mechanism to kick in if a user is unsuccessful in verbally issuing a command. By guessing at the word a few times, the user would finally be allowed to hear the audio associated with that word.

When designing card games for language learning it is necessary to balance the enjoyment of the end user, the pedagogical advantages of a given implementation, and technological feasibility. The hope is that, with a properly designed card game, a language learner might one day be able to log onto a Web site, start playing, and soon forget that they are even learning a foreign language.

Luis von Ahn has made a career around harnessing what he calls *human computation* [44], where the hours that a person spends playing online games are given a purpose. In his lectures, he notes that 9 billion person-hours were spent playing Microsoft's solitaire game in 2003 alone. This colossal waste of time, he reasons, could be put to better use, and so he devises clever Web-based games, such as the ESP game, in which the game play has the side effect of performing some useful task, such as labeling an image. For the individual player, however, the game is *still* a colossal waste of time. Now, imagine that we instead add value to these hours *for the player* through online, speech-

enabled games for language learning. How might the world be different if these 9 billion person-hours were spent inadvertently studying a foreign language? Admittedly, this might take away from von Ahn's effort to label images on the Internet, but the benefits of a linguistically and by extension culturally enlightened global population are impossible to deny.

9 Acknowledgments

This research is supported by ITRI, the Industrial Technology Research Institute, Taiwan. Ming Zhu and James McGraw provided valuable discussions that contributed to the analysis of the studies presented in this work. The thoughtful comments of two anonymous reviewers are also greatly appreciated.

References

- [1] *Chinese: An Essential Grammar*. Essential Grammars. Routledge (Taylor and Francis), New York, 2006.
- [2] R.C. Anderson and W.E. Nagy. The vocabulary conundrum. *American Educator*, 16:14–18;44–47, 1992.
- [3] Eric Atwell, Dan Herron, Peter Howarth, Rachel Morton, and Hartmut Wick. Recognition of learner speech, isle project report d3.3, <http://nats-www.informatik.uni-hamburg.de/~isle>, 1999.
- [4] Jared Bernstein, Amir Najmi, and Farzad Ehsani. Subarashii: Encounters in japanese spoken language education. *Computer Assisted Language Instruction Consortium*, 16(3), 1999.
- [5] Thomas S. Brown and Fred L. Perry. A comparison of three learning strategies for ESL vocabulary acquisition. *Teachers of English to Speakers of Other Languages (TESOL) Quarterly*, 25(4):655–670, 1991.
- [6] Susan Carey. The child as word learner. *Linguistic theory and psychological reality*, pages 264–293, 1978.
- [7] R.E. Cooley. Vocabulary acquisition software: User preferences and tutorial guidance. In *AIED 2001 Workshop Papers: Computer Assisted Language Learning*, pages 17–23, 2001.
- [8] Jonathan Dalby and Diane Kewley-Port. Explicit pronunciation training using automatic speech recognition technology. *Computer Assisted Language Instruction Consortium*, 16(3), 1999.

- [9] F. Ehsani and E. Knodt. Speech technology in computer aided language learning: Strengths and limitations of a new CALL paradigm. *Language Learning and Technology*, 1998.
- [10] Rod Ellis. Modified input and the acquisition of word meanings. *Applied Linguistics*, 16:409–441, 1995.
- [11] Rod Ellis and Xien He. The roles of modified input and output in the incidental acquisition of word meanings. In *Studies in Second Language Acquisition*, volume 21, pages 285 – 301, 1999.
- [12] Rod Ellis, Yoshihiro Tanaka, and Atsuko Yamazaki. Classroom interaction, comprehension and the acquisition of word meanings. *Language Learning*, 44:449–491, 1994.
- [13] Maxine Eskenazi. Using automatic speech processing for foreign language pronunciation tutoring. *Language Learning & Technology*, 2(2):62–76, 1999.
- [14] Maxine Eskenazi, Angela Kennedy, Carlton Ketchum, Robert Olszewski, and Garrett Pelton. The NativeAccent pronunciation tutor: measuring success in the real world. In *SLaTE Workshop on Speech and Language Technology in Education*, pages 124–127, 2007.
- [15] Johann Gamper and Judith Knapp. A review of intelligent CALL systems. In *Computer Assisted Language Learning*, 2002.
- [16] Johann Gamper and Judith Knapp. Tutoring in a language learning system. In *Proceedings of 6th World Multiconference on Systemics, Cybernetics and Informatics (SCI)*, 2002. Similar to Glosser1999, this is a way of providing lots of comprehensible input and tools to understand it.
- [17] J. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 2003.
- [18] Alexander Gruenstein, Ian McGraw, and Ibrahim Badr. The WAMI toolkit for developing, deploying, and evaluating web-accessible multimodal interfaces. In *IMCI '08: Proceedings of the 10th international conference on Multimodal interfaces*, pages 141–148, New York, NY, USA, 2008. ACM.
- [19] M.M. Gruneberg and R.N. Sykes. Individual differences and attitudes to the keyword method of foreign lanugage learning. *Language Learning Journal*, 4:60–62, 1991.
- [20] William G. Harless, Marcia A. Zier, Michael G. Harless, and Robert C. Duncan. Virtual conversations: An interface to knowledge. *IEEE Computer Graphics and Applications*, 23(5):46–52, 2003.
- [21] Rebecca Hincks. Speech technologies for pronunciation feedback and evaluation. *ReCALL*, 15(1):3–20, 2003.
- [22] Melissa M. Holland, Jonathan D. Kaplan, and Mark A. Sabol. Preliminary tests of language learning in a speech-interactive graphics microworld. *Computer Assisted Language Instruction Consortium*, 16(3), 1999.

- [23] W. Lewis Johnson, Carole R. Beal, Anna Fowles-Winkler, Ursula Lauper, Stacy Marsella, Shrikanth Narayanan, Dimitra Papachristou, and Hannes Högni Vilhjálmsson. Tactical language training system: An interim report. In James C. Lester, Rosa Maria Vicari, and Fábio Paraguaçu, editors, *Intelligent Tutoring Systems*, volume 3220 of *Lecture Notes in Computer Science*, pages 336–345, 2004.
- [24] Stephen Krashen. Why support a delayed-gratification approach to language education? *The Language Teacher*, 28:3–7.
- [25] Stephen Krashen. *The Input Hypothesis: Issues and Implications*. London: Longman, 1982.
- [26] Stephen Krashen. *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon, 1982.
- [27] Stephen Krashen. The input hypothesis and its rivals. In Ellis, N. (ed) *Implicit and Explicit Learning of Languages*, pages 45 – 77. Academic Press, London, 1994.
- [28] Michael H. Long. Input, interaction and second language acquisition. pages 259 – 278, 1981.
- [29] Geoff Brindly Manfred Pienemann, Malcolm Johnston. Constructing an acquisition-based procedure for assessing second language acquisition. *Studies in Second Language Acquisition*, 10:217–243, 1988.
- [30] Ian McGraw and Stephanie Seneff. Immersive second language acquisition in narrow domains: A prototype ISLAND dialogue system. In *SLaTE Workshop on Speech and Language Technology in Education*, 2007.
- [31] Ian McGraw and Stephanie Seneff. Speech-enabled card games for language learners. In *Proc. of AAAI*, 2008.
- [32] Wolfgang Menzel, Daniel Herron, Rachel Morton, Dario Pezzotta, Patrizia Bonaventura, and Peter Howarth. Interactive pronunciation training. *ReCALL*, 13(1):67–78, 2001.
- [33] I.S.P. Nation. *Learning Vocabulary in Another Language*. Cambridge University Press, 2001.
- [34] John Nerbonne, Duco Dokter, and Petra Smit. Morphological processing and computer-assisted language learning. In *Computer-Assisted Language Learning (CALL)*, pages 543–559, 1998.
- [35] Ambra Neri, Catia Cucchiari, and Helmer Strik. ASR-based corrective feedback on pronunciation: does it really work? In *Interspeech*, 2006.
- [36] R. Oxford and D. Crookall. Vocabulary learning: a critical analysis of techniques. *TESL Canada Journal*, 7:9–30, 1990.
- [37] P. I. Pavlik and J. R. Anderson. Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, 29:559–586, 2005.

- [38] Mitchell Peabody and Stephanie Seneff. Towards automatic tone correction in non-native mandarin. In *Chinese Spoken Language Processing, 5th International Symposium, ISCSLP*, pages 602–613, 2006.
- [39] Paul Pimsleur. *How to learn a foreign language*. Heinle & Heinle Publishers, Inc, 1980.
- [40] Rosetta Stone. <http://www.rosettastone.com>. Last accessed: May 1, 2008.
- [41] Merrill Swain. Communicative competence: Some roles of comprehensible input and comprehensive output in its development. In S. Gass and C. Madden (Eds.), *Input in Second Language Acquisition*, pages 235 – 253. Rowley, MA: Newbury House, 1985.
- [42] Talk To Me. <http://www.auralog.com/>. Last accessed: May 1, 2008.
- [43] Jianhua Tao, Jian Yu, Lixing Huang, Fangzhou Liu, Huibin Jia, and Meng Zhang. The WISTON text to speech system for Blizzard 2008. In *Proc of the Blizzard Challenge*, 2008.
- [44] Luis von Ahn. Games with a purpose. *IEEE Computer*, 39(6):92–94, 2006.
- [45] P.A. Herman W. E. Nagy. Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. *The Nature of Vocabulary Acquisition*, pages 19–35.
- [46] J. Yi, J. Glass, and I. Hetherington. A flexible, scalable finite-state transducer architecture for corpus-based concatenative speech synthesis. In *ICSLP*, 2000.