Web-based, Speech-enabled Games for Vocabulary Acquisition in a Foreign Language

by

Ian C. McGraw

B.S., Stanford University, California, USA (2005)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2008

Author . . . . . .
Department of Electrical Engineering and Computer Science
July 7, 2008

Certified by . . . . . .
Stephanie Seneff
Principal Research Scientist
Thesis Supervisor

Accepted by . . . . . .
Arthur C. Smith
Professor of Electrical Engineering
Chairman, Department Committee on Graduate Students

Web-based, Speech-enabled Games for Vocabulary Acquisition in a Foreign Language

by

Ian C. McGraw

Submitted to the Department of Electrical Engineering and Computer Science
on July 7, 2008, in partial fulfillment of the
requirements for the degree of
Master of Science

Abstract

In this thesis, I present two novel ways in which speech recognition technology might aid students with vocabulary acquisition in a foreign language. While research in the applied linguistics field of second language acquisition (SLA) increasingly suggests that students of a foreign language should learn through meaningful interactions carried out in that language, teachers are rarely equipped with tools that allow them to provide interactive environments outside of the classroom. Fortunately, speech and language technologies are becoming robust enough to aid in this regard. This thesis presents two distinct speech-enabled systems to assist students with the difficult task of vocabulary acquisition in Mandarin Chinese. At the core of each system is a Mandarin speech recognizer that, when connected to a web-based graphical user interface, provides students with an interactive environment in which to acquire new Mandarin vocabulary.

Thesis Supervisor: Stephanie Seneff
Title: Principal Research Scientist

Acknowledgments

First and foremost, I would like to thank my advisor, Stephanie Seneff. Without our conversations and her unwavering support, this thesis would not have been possible. I would also like to thank Victor Zue, who had the difficult job of convincing me to put my ideas down on paper in a timely manner. I also owe a great debt to the many researchers whose years of work prior to my affiliation with the Spoken Language Systems group provided the solid foundation of robust speech technology on which I built my applications. Specific technical contributions to this thesis work in particular include a web-based framework for multimodal dialogue systems developed by Alex Gruenstein and a Mandarin language model for a simple card game provided by Chao Wang. Also, Brandon Yoshimoto was essential in ensuring the success of the user study described in chapter 5. Lastly, Ming Zhu and James McGraw provided greatly appreciated support, both technical and familial.

This research was supported by the Industrial Technology Research Institute (ITRI).

Contents

List of Figures

Chapter 1

Introduction

This thesis examines the role that speech technology currently plays in educational software for language learning, and report on two novel applications that aid in the task of vocabulary acquistion in Mandarin Chinese. Both systems represent a departure from traditional methods of integrating speech technology into software for foreign language learners; however, each system has taken a distinct approach to providing an environment in which a student can acquire new words. These systems are evaluated along a number of dimensions: the satisfaction of the end-user, the performance of the speech recognizer, and finally in terms of learning gains with respect to vocabulary retention.

1.1 Motivation

Learning to speak a second language as an adult is an enormous task; the rewards of its achievement, no less monumental. For some, it is a job requirement in today's global economy. For others, fluency in a foreign tongue represents a window into another culture. Motivations aside, proficiency in a second language is often recognized as necessary for personal and even national development. To the adult learner, however, the gap between this recognition and its realization can at times seem insurmountable.

Regrettably, in the United States adult learners often do not progress further than their introductory language courses. A recent survey conducted by the Modern Language Association of America confirms that, of every six students enrolled in an introductory foreign

language course at the university level, only one will take the language beyond its second year [15]. This is especially unfortunate because the benefits that come with acquiring a foreign language often begin to appear far later in the learning process.

These unsettling realities raise questions about which aspects of language learning are discouraging students and how we might alter or augment our current curricula to retain their interest. One explanation posited by Stephen Krashen, a respected researcher in the applied linguistics field of second language acquisition (SLA), is that curricula currently employ what he pejoratively refers to as the *delayed gratification* approach to language teaching [30].

To put an example of delayed gratification into the context of vocabulary acquisition, imagine that a student is given a list of 50 vocabulary words to memorize. This tedious explicit memorization task is assumed to be prerequisite to the far more gratifying experience of comprehending these words in stories or using them in conversation. Krashen would argue that encountering unknown vocabulary words in a meaningful context should be *part* of the process of internalizing them, rather than a reward delayed until after explicit mastery through memorization.

Of course, internalizing new vocabulary is only one aspect of language learning in which our current curricula might inadvertantly encourage delayed gratification. Rote memorization of grammar rules and repetitive pronunciation drills also embody this approach to language learning, which many SLA theorists believe to be misguided.

In this thesis, however, I focus solely on the vocabulary acquisition task. Fortunately, the first few years of language learning are often characterized as having a similar focus on language at the lexical level [37]. After all, a word cannot be pronounced, let alone embedded into a grammar rule, if it is not known. The scale of the task is also undeniable. It is estimated that the average high school senior knows around 40,000 words [61]. To have even a rudimentary grasp of a language at the conversational level, a vocabulary of around 5,000 words is necessary [47]. Even children who are immersed every waking hour in an environment ideal for acquiring their first language learn at most 10 new per day. These facts and a back-of-the-envelope calculation reveal that vocabulary acquisition accounts for years of a learner's time.

Unfortunately, however, in many curriculums the internalization of new words is left as a problem for the student to tackle, 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 is negligible relative to the sheer number of words a student is expected to know [45]. The result is that students often find themselves resorting to explicit memorization techniques at home which take the words out of any meaningful linguistic context and can be classified squarely under Krashen's definition of a delayed gratification approach to language learning.

Krashen's solution to the general problem of delayed gratification is to provide students with *comprehensible input* by which they can immerse themselves in the target language without being overwhelmed. This comprehensible input is often in the form of reading material appropriately tuned to the learner's proficiency. Indeed, many SLA theorists posit that much of one's vocabulary in a second language is acquired incidentally through reading [27]. For the beginner with little foundation from which to infer new words, however, reading in hopes of "picking up" new vocabulary is relatively inefficient. The issue is exacerbated with respect to Chinese because learning to read can take years, and as a result, finding authentic material at an appropriate level is difficult.

A second solution one might propose is to inundate the learner with comprehensible input in the form of spoken conversation. Incidental vocabulary acquisition through conversation, however, is fraught with a different set of problems. As Krashen notes, beginners are often quite hesitant to expose their inexperience to a native speaker [31]. Furthermore, for many in the United States, opportunities to practice speaking a foreign language outside of the classroom are rare. As a result of these inconveniences, beginning and intermediate language classes are often ill-equipped to break free from the paradigm of *delayed gratification*, especially with respect to the homework they assign.

One could argue that technology is poised to fundamentally transform the capabilities of the language teacher to provide comfortable environments in which learners can interact in a foreign language at a level specifically tuned to their proficiency. Furthermore, 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. In fact, the transformation is already taking place outside of our curriculums. Skype, a popular VoIP service for making calls from a computer, is often used as a platform for language exchange.

Still, there is relatively little in the way of structured materials that are available to teachers in a form that can be conveniently included in a typical curriculum. It is my belief that automatic speech recognition (ASR) technology can provide tools to address these needs. This thesis explores this assertion in the context of the task of vocabulary acquisition in Mandarin. Combining the emerging technologies of ASR and VoIP, we have developed two experimental Web-based games which allow learners to talk to their computers in Chinese from an ordinary Internet browser.

1.2 Chapter Summary: Prototype Systems

Each of the vocabulary acquisition systems presented in this thesis has a Mandarin speech recognizer at its core, and both provide non-threatening, interactive environments in which a student of Chinese can acquire new words while performing meaningful tasks. The fundamental differences between the two systems lies in the scope of their respective lexical domains and in their reliance on very different underlying natural language technologies. These distinctions have several ramifications, both practical and pedagogical, which will be discussed in later chapters.

The first system, called the *Family ISLAND*, is an application built around a set of principles for constructing dialogue systems for **I**mmersive **S**econd **L**anguage **A**cquisition in **N**arrow **D**omains. These principles are designed to circumvent many of the pitfalls of dealing with non-native speech while still providing a pedagogically grounded environment for acquiring basic conversational competence, in this case, on the topic of family.

Chinese Cards, the second system introduced, is a flexible framework for creating customizable, speech-enabled card games for vocabulary acquisition. Although in this thesis I will only be discussing the first prototype card game we built upon this platform, our group is already nearing completion of the development of a second game, and other potential

games are in the planning stage.

The remainder of this thesis is layed out as follows: In Chapter 2, I will review a number of the currently available speech-enabled systems for language learning and discuss the relevant pedagogical underpinnings. Chapter 3 will discuss the *Family ISLAND*, a simple dialogue game designed for students who have little or no previous experience with Mandarin. Chapter 4 introduces the *Chinese Cards* framework for customizable card games and the first prototype card game called *Word War*. The chapter then goes on to evaluate this game in terms of recognition accuracy and then provides a detailed error analysis. Chapter 5 extends the work of the previous chapter by evaluating two variants of *Word War* and a simple flash-card system in a longitudinal user study which examines the effects of the three systems on vocabulary retention over a three week period. Finally, a short concluding chapter will discuss future directions of this work and make some final remarks.

Chapter 2

Background

This chapter begins with a short section on Second Language Acquisition (SLA) theory and terminology. I then provide an overview of work in automatic speech recognition as applied to foreign language learning and an introduction into the area of computer assisted vocabulary acquisition. 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 Pedagogical Trends and Terminology

Before detailing the short history of ICALL, it is necessary to briefly describe some of the recent trends in SLA theory, to put the software into pedagogical context. While the rich field of second language acquisition theory reveals many insights into how teaching methodologies *might* be improved, it rarely provides concrete answers. Even seemingly simple conjectures, e.g. that requiring students to speak is beneficial to their acquisition of a foreign language, spark animated debates [58, 33, 36].

Since the 1980's, western theories of SLA increasingly suggest that a second language is best acquired through the transmission and comprehension of meaningful messages in the target language: that is, through its use [32, 51]. Krashen even makes a careful distinction between *acquisition* and *learning*. Acquisition, he says, is the development of an ability in a language through subconscious processes. Learning, on the other hand, is a conscious

process that results in the accumulation of formal knowledge *about* a language.

Although the *acquisition-learning* distinction is more often applied to the internalization of grammar rules, there exists analogous terminology for vocabulary acquisition: *explicit* vocabulary learning is the conscious process of storing new words in memory, while *implicit* vocabulary acquisition is the entirely unconscious process of "picking up" words through context. In [10], Nick Ellis provides a detailed review of research that attempts to assess the roles that both implicit and explicit processes play in vocabulary acquisition.

While the terms *explicit* and *implicit* are used above to describe awareness in the cognitive processes that accompany learning, the terms *intentional* and *incidental* vocabulary acquisition are defined relative to the task carried out during which the learning is taking place. Unfortunately, these sets of terms are sometimes used interchangeably, and a confusion arises when studies do not make clear their distinctions [50]. Under the definitions provided here, a task designed for *incidental* vocabulary acquisition may contain small components of explicit learning, however, by and large the focus of the task is not on *intentional* vocabulary learning, where a student's *main* task is to consciously commit new words to memory.

Both systems described in this thesis represent incidental vocabulary acquisition tasks. The decision to rely solely on implicit acquisition of the words in the tasks or employ additional explicit memorization techniques is left to the student. To place these systems in the broader context of computer aided vocabulary acquisition (CAVL), the following section summarizes a number of example systems in this area of research.

2.2 Computer Assisted Vocabulary Acquisition (CAVL)

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 online flash card programs promoting intentional memorization techniques to intelligent reading environments, e.g. [17], 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 Edu-

cation (AI-ED) 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 (NLP) components to provide morphological analysis of the text [46].

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 lexical acquisition is often the most difficult task for an adult learning a language from scratch [47]. Meanwhile, systems used at the beginner levels, which are typically designed for intentional vocabulary memorization, inherently employ a delayed gratification approach to this difficult task.

Arguably the most successful effort in developing a well-motivated CAVL system is the commercially available software package, Rosetta Stone [52]. Using images as context, this software package requires the student to choose from a set of pictures by listening to descriptions that get progressively longer. While this immersion in comprehensible *input* is appealing, opportunities for the user to *speak* using this software still come 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 too rarely require the student to speak. While some in the SLA theory community would not regard this as a negative characteristic [33], 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* [57]. In fact, a study consisting of activities very similar to the tasks that will be presented in chapter 5 was

even able to show advantages of speech production with regard to the task of vocabulary acquisition in particular [12]. A detailed discussion of these studies is deferred to chapter 5.

2.3 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 [16]. To overcome the difficulties inherent in processing learner speech, researchers find it necessary to place 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, 14], 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 [35, 32].

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 [25, 26]. In the commercial realm, this is the state-of-the art [59], while the vast majority of systems still ensure that there is a single correct user utterance for a given prompt [52].

The research community has since moved on to creating small context free grammars (CFGs) [1, 41, 28], 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 *Family ISLAND*, presented in the next chapter, follows in the footsteps of these conversational systems. One striking difference, however, is that the *Family ISLAND* is designed to be usable by students who have never spoken the target language, in this case Mandarin.

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 virtually non-existent. Still, one could imagine that if a large number of these narrow domain systems could be developed and introduced into the classroom, they might be capable of making a meaningful impact on language education.

Regrettably, although increasing attention is being paid to SLA theories, very few of the currently available applications of ASR to SLA have been field-tested. In fact, Conversim [25] is the only ASR system presented in this section thus far that was evaluated empirically in terms of learning gains. With the goal of teaching children to read, Project LISTEN [43], though not strictly for *foreign* language learning, is a model example of how well-executed classroom experiments can give clear evidence that ASR technology has educational value in practice. In this work, Mostow et. al. rely on large-scale, carefully controlled userstudies to assess learning gains in a classroom setting [42]. Though the resources to perform such studies are not easy to come by, it is disconcerting that systems for foreign language learning have not followed this lead, since these experiments give a project both the credibility and exposure that would facilitate their widespread adoption.

2.4 Chapter Summary: 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 dialogue systems for ASR show great promise from a pedagogical perspective, but are difficult to deploy, have very limited lexical domains, and often lack user-customizability.

In chapter 4, I 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 turn the explicit memorization task flash cards typically imply 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 speech commands to complete the task. Chapter 5 then details our nascent efforts to evaluate the educational value of this system.

Chapter 3

The Family ISLAND

As described in the previous chapter, much of the second language acquisition (SLA) scholarship suggests that conversational skills are best acquired through *communication* in the target language. Although in recent decades communicative approaches to language teaching have seen widespread adoption in the classroom, it remains exceedingly difficult to assign conversational *homework* with the tools currently available. This reality has created a gap between the way in which foreign language courses are often implemented and pedagogical methods that the SLA theory community might recommend. This chapter describes how current technology in spoken dialogue systems is capable of closing this gap.

Of specific interest to the spoken dialogue systems community is the development of Communicative Language Teaching (CLT) as a widely adopted approach to foreign language instruction [51]. The fundamental tenet of CLT is that the basic unit of learning is the communication of a message in the target language. That is, the learner ought to focus on the meaning of their words as uttered in the target language.

Attempting to elicit meaning from human speech is precisely the problem that spoken dialogue systems have been grappling with for some time. The Spoken Language Systems group at MIT has carried out extensive research on dialogue systems in domains such as weather [19] and flight [56] information. Leveraging this research, we have in recent years begun to build dialogue systems targeting the second language learner [55].

This chapter describes one recent effort in particular[1]: the *Family ISLAND*. This system

[1]Portions of this chapter were published in [38].

allows individuals with *no* previous Mandarin experience to communicate with their computers in Chinese on the topic of family. Since this system is completely immersive there is no explicit instruction in the vocabulary, pronunciation, or grammar structures. Thus, the dynamics of the system must be such that the relevant language properties are deducable from context alone.

Section 3.3, lays out the design principles applied to the system's development, describing how they attempt to minimize the effects of common problems in dialogue systems. Then, section 3.4 describes an initial testing and data collection iteration and presents some early but promising results. Of particular interest from a CAVL perspective are the results that indicate that the system is able to convey information regarding the meaning of vocabulary words without resorting to explicit translations.

3.1 Dialogue Systems for Language Learning

The critic might argue that dialogue systems already pose a number of unsolved problems, and that applying them towards language learning merely exacerbates one in particular: non-native speech. Indeed, the limitations of applying speech recognition technology to language learning have been explored thoroughly in [9].

In this chapter, I argue that, due to the special nature of the language learner as a user, certain techniques can be applied to overcome obstacles in dialogue system design. I found that language learners can be far more tolerant than native speakers with respect to recognition errors in dialogue systems. Furthermore, I identify a number of other common complications in spoken dialogue systems, and show how their negative repercussions can be mitigated without sacrificing the goals of a dialogue system for second language learners.

Incorporating these insights into a set of design principles, I have developed a new type of dialogue system to support Immersive, Second Language Acquisition in Narrow Domains (ISLAND). To test my assumptions I have designed and implemented an ISLAND dialogue system in Mandarin Chinese. The ISLAND is *immersive*, in that no content information whatsoever is given to the user in his or her source language. I refer to language *acquisition*, as opposed to language *learning*, as I do not incorporate a formal discussion

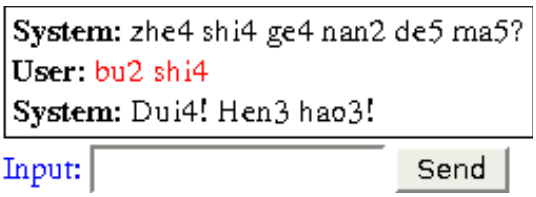


Figure 3-1: *Dialogue panel as presented on web page. Notice the optional "Input" text field, that the language learner can use when recognition problems occur. The recognition result itself is highlighted in red to draw the user's attention to potential mistakes.*

of grammar into the dialogue. Finally, the scope of our dialogue is limited to the *narrow domain* of gender and family relationships.

3.2 Family ISLAND

Dialogue systems for the second language learner, especially systems that make heavy use of natural language processing and automatic speech recognition, often target users at an intermediate level [16]. In contrast, despite the fact that the content is entirely in the target language, I envision this system spanning the pre-beginner to late-beginner stages of students of Mandarin Chinese.

The dialogue system consists of four levels. The first three cover the topics of gender, proper names, and family relationships respectively. The fourth level is an open dialogue about an individual's family tree. The basic building block of each level is the *task*. Tasks are mutually independent segments of the dialogue in which some meaningful exchange takes place.

The dialogue is presented to the user on a web page divided into two sections. The first is the dialogue panel, shown in Figure 3-1, where the user can monitor the conversation and, in particular, the recognition performance. Secondly, a family panel is displayed (see Figure 3-2) to give the user the content and context of the conversation. At any given time, some of the family members' images will have a thick blue border. These are the family members on which the user can click to start recording. Their utterance is then processed in the context of the family member from which it was recorded.

The first level of the ISLAND is about gender. Each task in this level begins by showing

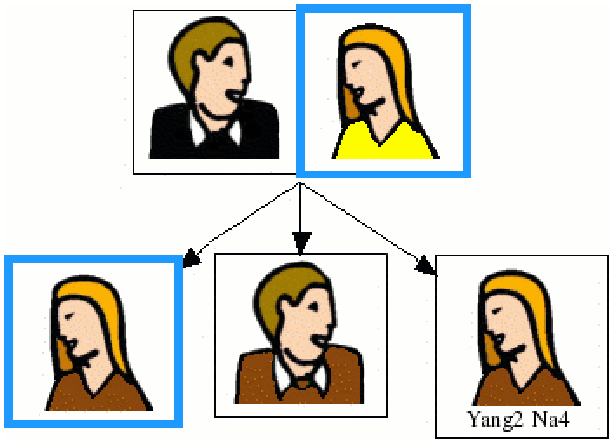


Figure 3-2: *A particular task as presented to the user in the form of a family tree. A user can click on the family members with thick blue borders to record an utterance.*

an image of either a man or a woman. The system then asks the question "Is this a man?" in Chinese. The user is then given the opportunity to respond. Should the user be unable to communicate the appropriate response, hints will appear in the form of possible answers shown in Figure 3-3. In this case, the Chinese equivalent of "Yes" and "No" assist the user in accomplishing the task. These hints may be played so that the user can hear how a native speaker would pronounce the words.

The second level covers proper names. One task of note in this level presents the user with several people with their names displayed below the images, and asks the user to name each person. The user has control over the particular order in which to name the displayed individuals. Users can simply click on the person they are going to name and say something to the effect of "This is Yang Na."

The third level is a system-initiated dialogue about relationships. The system might show a family tree as in Figure 3-2, and ask (in Chinese) the question "Which person is Yang Na's mother?" For a pre-beginner, the word "mother" may not be associated with a particular relationship. The relationship can be deduced, however, by saying "This is her mother" while clicking on various family members. A musical cue accompanied by the Chinese equivalent of "Correct! Great job!" indicates that the user has accomplished the task of finding the mother.

The fourth level begins by displaying a single family member labeled along with an age. An English prompt suggests that they ask about this person's family tree. The user

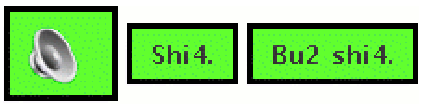


Figure 3-3: *Hint buttons for the question "Is this a man?" Pressing these buttons will play the corresponding text as a native speaker might say them. The speaker button on the far left is always present, as it repeats the system's most recent question or response.*

may ask simple questions such as "Does she have a brother?" or "Does she have a child?". The system answers verbally as well as by displaying the previously hidden family member. The student can also ask about complex relationships such as "Is his wife's older sister married?" Slowly the user is able to uncover the entire family tree of the specified individual.

Out-of-domain questions in all levels are answered with the Chinese equivalent of, "I do not understand." Should the system fail to understand the student more than a certain number of times, the hints will begin to appear in the form of possible responses. In this way the system is able to keep all domain-specific content in the target language. By observing the context as given in the family panel and by exploring different options via the gradually exposed hints, even a user with absolutely no background in Mandarin can progress through the system in its entirety. Section 3.4 describes a set of experiments with users who were able to accomplish this feat.

3.3 ISLAND design

Typical spoken dialogue systems are composed of the following components: speech recognition, speech synthesis, natural language understanding and generation, and dialogue management. ISLAND dialogue systems are no different. The components of this system are integrated using the Galaxy architecture [53], which allows communication among a set of servers that perform each of the aforementioned tasks.

Within each of these components, however, it is my belief that the ISLAND designer can make use of techniques often unavailable to dialogue systems with more standard applications. In this section, I discuss issues commonly thought to be problematic for dialogue systems, particularly as geared towards language learning. I mention how they are dealt

with in the system, and how these techniques can be applied generally to future ISLAND systems.

3.3.1 Speech recognition & synthesis

Speech technology is the core of the dialogue system's framework for conversational communication in a narrow domain.

High quality speech synthesis is crucial to ISLAND design because users model their speech on the spoken output of the system. The speech synthesizer used by the system is Envoice [64]. In an effort to come as close as possible to native speech with minimal recording requirements, Envoice uses a small corpus of pre-recorded utterances from which to splice together new utterances.

The recognition component of the ISLAND utilizes the SUMMIT landmark-based system [18] with acoustic models trained from native Chinese speakers [63]. Given the scarcity of large amounts of learner-speech data [49], most ASR systems for SLA do not employ acoustic models trained from non-native speakers. Although some researchers have also begun to investigate creating hybrid models using data from both the source and target language [40], this approach is not taken for Word War. Tones are not explicitly modeled although they can be inferred by the language model given the system's narrow domain. This is intentional, as many nonnative speakers have poor knowledge of tone, and modelling it explicitly would only lead to enhanced recognition errors. Aside from disregarding tones, the models are in no way biased to non-native speech.

Fortunately, dialogue systems for language learners are different from other applications in that even misrecognized utterances have the potential to be valuable. These can provide pronunciation practice, and the user may even be able to pin-point portions where she might improve by watching recognition output. To this end, the system never attempts to hide recognition results from the user. In fact, the user's hypothesized utterances are highlighted to draw the learner's attention to them in case of an error.

3.3.2 Natural language understanding & generation

The natural language understanding component of the system makes use of a syntax-based grammar, along with a probabilistic model that can be trained on an untagged corpus of synthetic utterances [54]. Language generation is provided via an in-house generation system [2]. The system was first implemented in English and then translated into Mandarin Chinese.

Language portability is made particularly easy in ISLAND dialogues, since they are immersive. As a result, the recognition, synthesis, and natural language processing components need only be implemented in the target language. I was able to port the fully operational English system into Mandarin in one week.

Domain portability is a somewhat trickier issue; however, much has been done to push the domain specific components of the system into the fringes of the code base. The bottlenecks with respect to domain portability are largely the dialogue management and graphical user interface, as the speech and NLP components can be reconfigured for a new domain relatively easily.

3.3.3 Dialogue management

Most commercially deployed dialogue systems today fall within the category of directed dialogues in which the user is taken down a predetermined dialogue path. For a language learner in the early stages, this is not an unreasonable restriction. Ideally, however, the user would be given free range to speak in the manner he or she chooses. Researchers are currently exploring mixed-initiative dialogue systems to allow more flexible interaction. The ISLAND system can be thought of as a mixed-initiative system with a directed dialogue back-off mechanism. If the system is having difficulty understanding the user in the mixed initiative setting, it will offer directed hints in an effort to get the user back on track.

One extraordinarily difficult problem in dialogue systems is managing recognition uncertainty over multiple-turn dialogues. The fact that this problem remains unsolved for native speakers does not bode well for applications geared towards language learners; an inappropriate system response might leave the student confused and unable to continue.

In some instances, dialogue systems researchers are able to employ design techniques that either prevent these errors from occurring or reduce harmful effects if they do. I believe that dialogue systems targeting the language learner are particularly well suited to these advantageous design methodologies.

Pre-fabricated communication

One essential difference between ISLAND and standard dialogue system design is that an ISLAND designer *decides* what message the user should communicate to the system. This is in stark contrast to applications such as the Mercury flight reservation system [56], in which people are trying to reserve real flights. A language learning application in that same domain would likely fabricate the flight information that the user ought to communicate and even assist them in conveying this information back to the system in the target language. This information can therefore be incorporated in the dialogue management and even the recognition components of the system.

In this system, the dialogue manager is certainly aware when there is a single *appropriate* answer for a given task. Until this message has been effectively communicated to the system, the dialogue will not progress. In this way the system can keep the user on track even when it is necessary to keep track of conversation history over multiple turns.

Although conceivable, at the moment, the family ISLAND does not incorporate this kind of information into its language model. In fact, precisely the same recognizer is used at all levels of the dialogue. It is likely that recognition performance could be improved if the language model was based on the context of a specific dialogue task, or even the specific response expected.

Multi-modal dialogue grounding

The family dialogue does not rely solely on users' speech to convey meaning. If the system asks "Who is this person's father?", the user is given images of people on which they can click to record their reply. This grounds the dialogue turn in an absolute truth: the user clicked on X. In this example, if the user clicks on the mother and records an utterance, the dialogue manager is able to confidently say "Incorrect", regardless of recognition output.

In general, multi-modal interfaces can be used in this way to give the dialogue manager guidance when it comes time to perform some sort of semantic evaluation of an utterance in addition to providing the learner with a more engaging environment.

History on display

These techniques aside, recognition errors are inevitable. To minimize their impact on learning, I assert that it is essential to draw the user's attention to them. In the fourth level of the dialogue, the user is asking about a person's family tree. If a user asks "Do you have a brother?", but the system recognizes "Do you have a mother?", the system simply responds "Yes, I have a mother," and the rest of the conversation proceeds as if the user had truly said "mother".

Thus, the burden is shifted to the user to realize that a misrecognition has occurred. In many dialogue system applications, this technique is not available since a misstep in a dialogue can prevent the user from getting or giving some essential piece of information. The worst that can happen in an ISLAND system is that a user might not realize that an odd system response is due to a misrecognition. I attempt to avoid misunderstanding in the system, however, by both highlighting the user's utterance in the dialogue panel and showing the misrecognized relative, effectively putting the dialogue history on display. Though not ideal, I believe that this solution is far more likely to yield effective language learning tools than attempting to hide recognition errors from the user.

3.4 User study

Although there are many aspects of the system that deserve thorough analysis, I focused my initial user study on interactions with pre-beginners, individuals without any formal Mandarin experience. The goal was to discern whether the system enabled them to disambiguate the content vocabulary words solely from context clues.

The study consisted of 17 pre-beginners, each of whom interacted with the system alone for around one hour. A set of general instructions guided them through the use of the system, and they were made aware of each level's general domain. However, they

were never explicitly informed of the English meangings of the Chinese words they were required to speak. They then progressed through levels 1 through 4 of the dialogue. The first three levels each contained between 10 and 15 tasks, varying slightly depending on the correctness of the student's responses. The fourth required them to discover 7 relatives in a person's family tree.

I devised a simple matching test consisting of 16 vocabulary words and their English translations. I allowed the users to take notes as I was not interested in the short-term memory effects of the system. I analyze the test results with respect to the 12 users who took notes as I suggested. Half of these individuals had perfect scores on the translation quiz. The mean score was 14.75 out of 16 with a standard deviation of 1.4.

This indicates that the pre-beginners were capable of extracting the content words from the immersive environment using context clues alone. Extrapolating these results, it is reasonable to assert that we can target language learners at all levels with immersive systems, provided appropriate design principles are employed.

To judge recognition performance, one would normally use word error rate (WER). For this study, however, WER is not an appropriate metric because, at the pre-beginner level, utterances may contain segments without intelligible words as users explore the acoustic space of the target language. Nevertheless, one can infer performance information from the test scores in combination with usage statistics, as summarized in Figure 3-4.

From this table, it is clear that the pre-beginners were able to successfully base their pronunciation on the synthesized speech via the hint buttons. Each of the 17 users was able to progress through all of the tasks in the system, and the majority of the students did so without resorting to text input. Those who did, typically only used the option a few times.

To incorporate user feedback into the development cycle, I provided a survey filled out by each user. The following questions were asked, and answers were given on a 1 (least) to 5 (most) point scale. To what degree...

Q1. ...did recognition errors affect your ability to learn?
Q2. ...did you wish there had been more English to guide you?
Q3. ...was it easy to tell when recognition errors occurred?

	Mean	Std. Dev.
# Hints Played	37.0	22.3
# Times Used Text Input	1.5	2.8
# Utts. Heard	188.6	47.6
# Utts. Spoken	116.2	29.1
% Correct Utts	48.5	13.3
% Incorrect Utts.	21.5	7.4
% Not Understood Utts	30.0	12.5

Figure 3-4: *Usage statistics as averaged over the 17 participants.*

	Q1	Q2	Q3	Q4	Q5	Q6
Mean	1.82	2.00	4.06	2.59	3.41	4.18
Std. Dev.	1.01	1.22	0.83	1.18	1.41	0.88

Figure 3-5: *Survey results for questions Q1-Q6. All questions were rated on a scale of 1 (least) to 5 (most).*

Q4. ...do you think the recognition errors were the system's fault?

Q5. ...do you think your pronunciation caused recognition errors?

Q6. ...would you want to use this system in a language you study?

The user responses are summarized in Figure 3-5. It is exciting to note that neither the lack of English nor recognition errors prevented the pre-beginners from wanting to use such a system in the future.

3.5 Chapter Summary: Future Work

In this chapter, I have described a new tool for the second language learner called an ISLAND dialogue system. An ISLAND system can target a range of abilities by offering assistance incrementally based upon student performance. An initial user study on the family ISLAND has shown that such systems can provide an immersive environment in which even pre-beginners can practice conversational skills.

I have also described the set of principles employed when designing this system for language learners. In addition to alleviating many of the difficulties in dialogue system development, I believe the system has many properties congruent with the precepts of the

second language acquisition theory community.

It remains to be seen, however, if the particular implementation of these ideas has educational value in practice. Thus, in addition to performing system analysis on components such as the speech recognizer, I believe it is crucial to deploy the system in a setting more consistent with the educational environment for which it is designed.

Chapter 4

Chinese Card Games

> *"When students travel, they don't carry grammar books, they carry dictionaries."* Krashen in [34]

While the *Family ISLAND* represents a nice example of how pedagogical and practical design principals can be applied to create an engaging conversational system that can be used at the earliest levels of language learning, there remain a number of serious drawbacks to such an approach. The first is that the topic of family represents a miniscule portion of the vast landscape of human language. Ideally, the user could choose to visit an ISLAND that was built around whichever topic he or she desired to study. This brings us to a second drawback: development time. While every effort was made to ensure language portability, the GUI and dialogue management components were inherently tied to the topic of family. To make a meaningful impact on language education, a large company or institution would need to churn out ISLANDs on a multitude of topics, each requiring interactive tasks tied directly to its domain.

This chapter takes a different approach. Here, I present a generic framework for developing user-customizable card games specifically designed to aid learners in the difficult task of vocabulary acquisition[1]. I then describe a prototype game built on this framework that, using the same Mandarin speech recognizer introduced in chapter 3, provides a student of Chinese with opportunities to practice a variety of vocabulary items in a meaningful

[1]Portions of this chapter were published in [39].

context. The system dynamically loads only the necessary vocabulary for each game in an effort to maintain robust recognition performance without limiting the lexical domain.

Using the Web-based remote user study framework detailed in Appendix A, I demonstrate ease of deployment by conducting a user study remotely. Furthermore, this framework allows students and teachers to create their own content to be loaded into a prototype game. While there do exist a few recent examples of Web-based ASR systems for learning Chinese [6, 62], these systems do not specifically target vocabulary acquisition, nor do they offer the user the ability to personalize their learning experience.

The central research question addressed by this chapter is one of feasibility: given that both learner speech *and* diverse acoustic environments will be encountered by a Web-based, speech-enabled system for language learning, can the recognition be performed robustly enough to provide a worthwhile user experience? To assess the Sentence Error Rate (SER) of the prototype, I asked college-age students from various universities in the United States and beyond to participate in a Web-based user study. The three central concepts in the game were recognized with a SER of 16.02%, illustrating the feasibility of deploying this system in a university curriculum via the Internet. Finally, to ensure that the recognizer is behaving appropriately with regard to learner speech, I perform a rigorous analysis of the recognition errors to determine their underlying causes.

The remainder of this chapter is organized as follows. I first describe the *Card Creator*, an online tool that allows teachers to build the contents of the card games. I describe how this tool is built upon a generic framework for creating speech enabled card games. I then present *Word War*, the first speech-enabled card game built upon this framework, and expound upon a couple of variants of this system. My first experiments in introducing this system to remote users are then recounted in a long section on data collection, evaluation, and error analysis. Finally, I provide a short summary and a look to the future.

4.1 Card Game Framework

Before describing the prototype card game on which the user study was based, I briefly discuss the card creation tool and emphasizes the customizability of the framework. Using

Figure 4-1: *A single card created with the online tool.*

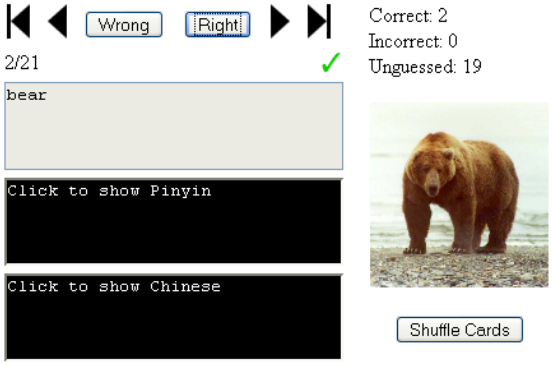


Figure 4-2: *An image based flash card player.*

Web 2.0 technology, an online Chinese language learning dictionary[2] and Yahoo image search[3] have been directly integrated into a card creation web site.

With these tools, students and teachers can quickly build entire categories of image-based cards and store them in the database. A set of public categories is available, or users can choose to sign up for a private account. Figure 4-1 shows an example of a single card formed directly from a dictionary entry and a Yahoo image search of the English word 'frog'. Note that students of Mandarin often use *pinyin*, a romanization of the Chinese characters, as a pronunciation guide. This eases the task of producing a language model for the recognizer, as I will describe later.

On its own, this web site is nothing more than a customizable flash-card database, albeit with a few helpful extra search features built in. In fact, a flash-card player is provided that

[2]http://www.xuezhongwen.net
[3]http://images.search.yahoo.com

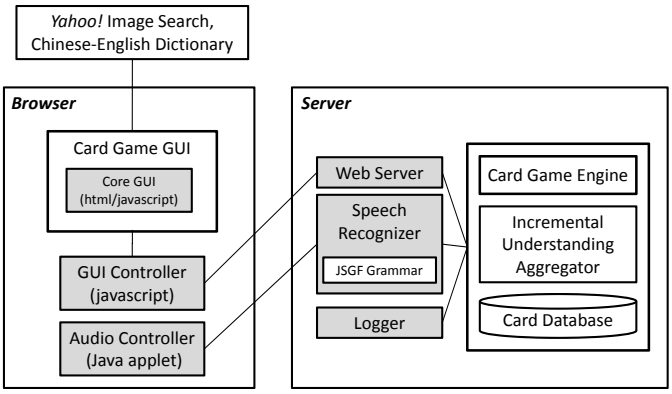


Figure 4-3: *Chinese Cards* Architecture; Generic toolkit modules shaded.

allows students to review vocabulary this way if they so choose (see Figure 4-2). There is no shortage of flash-card systems already available on the Web, and though they vary in ease-of-use and number of features, few have grounding in the rich field of SLA theory. Though highly customizable, an attribute that would be lauded by proponents of learner-centered classrooms [5], flash-cards encourage students to take words out of any meaningful context, not to mention their inherent tediousness.

The card creator's primary function is to enable users to customize one of the card games that is built upon the generic platform which I will now describe. Figure 4-3 shows a block diagram of the client-server configuration used for the card games developed. Large portions of this architecture are kept generic to ensure maximum flexibility for future game developers. To date, two very different card games have been developed on this framework, though only the first is described in this thesis. The Web-based speech platform upon which this card game architecture is based has already been used for a variety of systems within the Spoken Language Systems group [21, 22]. Specific to the card game architecture are: a back-end database of images and cards, a framework for declaring grammars that can be dynamically instantiated and sent to the Mandarin speech recognizer, as well as an incremental understanding aggregator, described later.

4.2 Word War

Word War was the first customizable card game that made public to students of Mandarin. Although this prototype game is simple, it demonstrates well the methods by which more entertaining card games could be developed and used to teach vocabulary in an interactive manner. In fact, the name "Word War" is better suited to the multi-player mode of the game, which makes play far more interesting.

4.2.1 Single-player Speaking Mode

In single-player mode, each game begins by loading a category of cards into the "game grid". An example of a five-column game grid initialized with the "animals" category is depicted in Figure 4-4. The goal of Word War is to use voice commands to move the images in the bottom two rows, subsequently referred to as *source* images, into the slot directly underneath the matching *target* image on the top row. Notice that, when the cursor is over an image a hint appears above the game grid telling the student the pinyin pronunciation of the word.

There are three types of commands understood by the system: *select*, *drop*, and *shift*. Each command type, subsequently referred to as a *notion*, can be instantiated with vocabulary words, numbers, or right/left directions respectively to form an *action* which the system will interpret to make a change on the game grid. The English equivalents of a few actions are exemplified in the following three sentences: 1) Choose the snake. 2) Drop it into slot one. 3) Move it three squares to the right. The game is complete once all of the source images have been appropriately aligned with their targets. Note that a *shift* action is typically used for error corrections. For example, a recognition error in a *drop* action might cause a source image to be placed under a non-matching target image.

Each notion can be expressed a number of ways in Mandarin, so the system maintains a template context free grammar (CFG), found in Figure 4-5, that captures many of them. When the game is loaded, pinyin is extracted from the cards and used to automatically instantiate the grammar with the personalized vocabulary. Before a game begins, this grammar is sent to the recognition component running server-side to be used as the

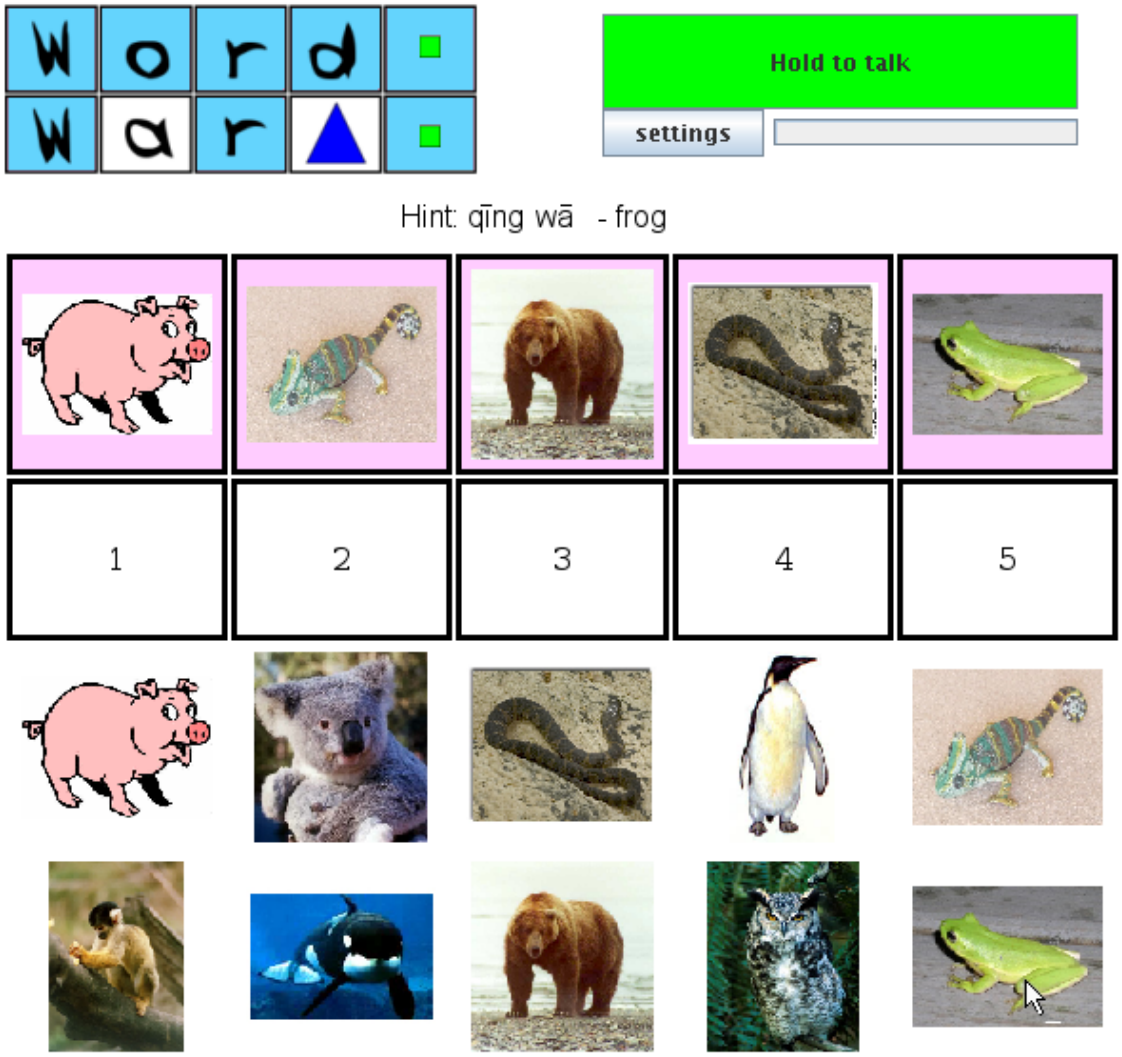


Figure 4-4: *The five-column version of the card game. The user can mouse-over an image to see the pronunciation hint.*

language model.

Word War uses the SUMMIT landmark-based recognizer [18] configured with acoustic models trained on *native* Chinese speakers. Importantly, since Chinese is a tonal language, the features used in these models do not include information about pitch. It is likely that including this information directly in the recognition stage would render the system unusable for many non-native speakers, and instead the *correct* tones are inferred from the language model. Eventually I hope to include feedback on a user's tones, but I must do so in a manner that does not cripple usability for beginners in Mandarin.

```
#JSGF V1.0;
grammar WordWar;

public <top> = (<command> [bing4 [qie3]])+;

<command>       = <select> | <drop> | <move>;

<select>        = (xuan3 [ze2] | na2 qi3) {[command=select]} (<card> {[card]})+;
<drop>          = [ba3 ta1] fang4 (zai4 | dao4) {[command=drop]} (di4 <numberE> <slot>)+;
<move>          = wang3 {[command=move]} <direction> nuo2 (<numberL> <slot>)+;

<slot>          = ge4 (wei4 zhi4 | di4 fang1 | kong4 ge2 | ge2 zi3);

<direction>     = (zuo3 {[direction=left]} | you4 {[direction=right]}) bian1;

<numberE>       = (<numberwo2> | er4 {[number=2]})+;
<numberL>       = (<numberwo2> | liang3 {[number=2]})+;

<numberwo2>     = yi1 {[number=1]}
| san1  {[number=3]}
| si4   {[number=4]}
| se4   {[number=4]}
| wu3   {[number=5]};

<card> = <<<< dynamically generated >>>>
```

Figure 4-5: *The template grammar used in the* Word War *speaking modes. The "card" tag is filled in dynamically depending on the category loaded into the game.*

Once the recognizer and the Web-based interface are initialized, and the user is ready, he or she can speak by pressing and holding the large button above the game grid. Audio is then streamed from the user's machine directly to the recognizer, processed, and the results are passed along to a game manager also residing on a server. Commands are then extracted from these utterances and reactions are sent to the client and executed in Java-script. The code infrastructure that makes this possible is based on AJAX technology and has been used in a number of unrelated projects in the Spoken Language System's group, e.g., [21].

In Word War, the visual reactions available to the browser as a response to an utterance are *highlighting* one or more source images and *moving* an image into the specified slot. A feature of the system is that it provides constant visual feedback in real time *while* the user is speaking. For example, a student can say "Select the cell phone and put it in the third square." Using the incremental understanding aggregator, our system will keep track of and carry out intermediate responses (e.g., highlighting the cell phone) before the utterance is completed. Indeed, it is possible to string multiple sentences together in such a manner that, even with a five-column game grid, the adept student can place all five source pictures into their proper locations without lifting the hold-to-talk button.

A final feature of the system that deserves mention is the play-back mode. During each student's game, the system maintains a log file of the interaction, and also captures all their recorded utterances for later processing. Once a student has completed an entire game, a

teacher or researcher can watch and listen to a replay of their session via the very same game grid interface described above. This feature should be quite valuable to teachers who might want to review a student's game. Play-back mode was also indispensable for the system evaluation procedure described later.

There are a number of aspects of the system that are pleasing from a pedagogical perspective. The first is that images are used to avoid prompting the user in English. Secondly, notice that it is impossible to both ask for a hint *and* speak an utterance at the same time, since the user must remove the mouse from its location over the image in order to press the record button. This intentional limitation requires the user to memorize the pronunciation, if only for a short time, and use it in a sentence that gives the word a meaningful context. Lastly, the real-time visual feedback makes it possible for the student to practice speaking fluently, while checking in real time that they are being understood.

4.2.2 Single-player Listening Mode

Discussions with teachers led us to implement a listening mode of *Word War*, in which it is the computer who gives the directions. From the student's perspective the listening mode is the reverse of the 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 I 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 target images, and thus their true locations, will be revealed.

At a minimum, implementing the listening mode only required that pre-recorded sound files to 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, our group is in the process of developing an interface that we will expose to users where they can record their own sounds. This would likely be most useful for teachers who wish to create listening games for their students, as well as for researchers who wish to generate user studies quickly.

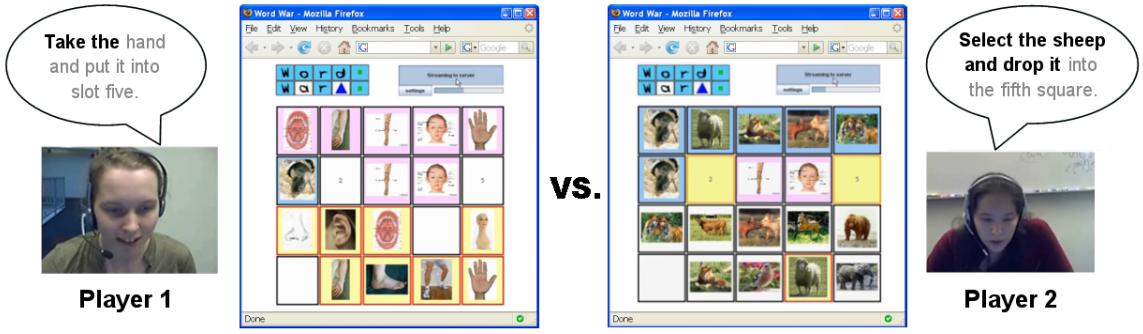


Figure 4-6: Two-player Word War

Choosing the more customizable approach necessitates more powerful speech synthesis technology, since, for more complex games, we do not wish to require a user to record every single utterance the system is required to say. Thus, we use Envoice[64], an in-house, concatenative speech synthesizer, to process the sound files. In this fashion, a teacher could record a tiny corpus of template sentences and each of the vocabulary words. Envoice will then splice the relevant portions of the sound files together on-the-fly when the system needs to utter a new sentence.

Notice that both modes of Word War together represent two sides of a language teaching paradigm that is almost always present in well-implemented communicative curriculums: *the information gap* [35]. 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 almost never assigned as homework for the following very simple reason: there is rarely anyone at home with whom to exchange information.

4.2.3 Multi-player Speaking Mode

A final enhancement to the speaking-mode of word war tackles the problem of making vocabulary acquisition exciting. In this section, I describe how I take the single-player picture matching task of section 4.2.1, and turn it into a multiplayer race.

Figure 4-6 shows a snapshot of two players competing on the five-column game grids of multiplayer wordwar. The goal of *Word War* in multiplayer mode is still to use spoken commands to move images from the bottom two rows of the grid into the set of numbered

slots in the second row, so that they match the images along the top. However, although the players can choose to load different sets of vocabulary, they compete over shared access to the numbered slots. When an image is matched, the slot is *captured* and the matching image appears on *both* players' game grids. Notice that, in Figure 4-6, 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.

Notice that in this configuration two recognizers are necessary, as both players will be speaking simultaneously. Just as in single-player mode, when a player's grid is initialized, a just-in-time language model is dynamically generated and sent to that player's recognizer. Since the grammars are tailored to each player's vocabulary, they remain relatively small, ensuring that recognition is robust with respect to the non-native speech.

In multiplayer mode, the incremental understanding nature of the speech architecture becomes particularly important. In Figure 4-6, 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 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 on the game grid. In addition to providing a more natural interaction and faster game-play, this incremental understanding and feedback mechanism also encourages students to issue multiple commands at once, ensuring that the vocabulary words are placed in their short term memory.

4.3 Data Collection and Evaluation

In this section, I present the findings from a pilot Web-based user study of the speaking mode of single-player *Word War*. The focus is on the evaluation of the technology itself. Assessment of students' learning gains will be deferred to the next chapter. I first describe in detail the experimental design. After explaining the method for human annotation of the data, I present an analysis to determine the relationship between system understanding error and various contributing factors.

Figure 4-7: *Locations around the world where users have interacted with single-player speaking mode of* Word War.

Experimental Design

I considered multiple methods of data collection when the decision to evaluate the recognition performance of *Word War* was made. As the system is publicly deployed, it would have been possible to simply harvest the data collected from these interactions and examine their recognition accuracy. Figure 4-7 shows a map of the world with indicating locations where a user has accessed the *Word War* speaking mode system. When listening to these interactions, however, it became apparent that a slightly more controlled environment was necessary. While replaying one *Word War* interaction, for example, it became apparent that a father and daughter were practicing Chinese together, at times sounding out words simultaneously!

Thus, to measure robustness in a realistic setting I administered a remote user study from the publicly available version of the system with the user study management tools detailed in Appendix A. I invited college-age individuals who had between one and four years of experience studying Mandarin to complete a series of eight tasks from their own computers. As an incentive for finishing all the tasks the users received a $15 Amazon.com gift certificate.

Each of the eight tasks in the study was in the form of a Word War game. With the card

creator I constructed three categories, each with 10 cards complete with characters, pinyin, and images. The categories were: animals, plants, and food. The first two tasks assigned were tutorials constructed from four of the animal cards (very much like those in Figure 4-4.) These tutorials ensured that their audio settings were correct and taught them the basics of the game. The remaining six tasks were assigned the following order: two animal games, two plant games, and two food games, where each game was on a five-column game grid. The target images were selected randomly each time upon initialization of the game. Example sentences for the various commands were always available.

In the week and a half that the study was open 27, users signed up and attempted the first task. Seven of the users did not progress beyond the first tutorial due to technical difficulties relating either to browser incompatibility or misconfigured audio settings. The 20 individuals who *did* finish the first tutorial also finished the remainder of the study, indicating that the recognizer was never an obstacle to task completion.

In all, over 1500 utterances were collected from 5 female and 15 male participants. While most were from the United States, at least two were from the UK, and one actually took the study from China.

Error Rate Evaluations

To evaluate the system I asked a native Chinese speaker to annotate each of the 1543 utterances from the six non-tutorial tasks. I did not require her to transcribe every utterance word for word, as some utterances contained sounds that could not actually be classified as a Chinese syllable. Hiring a professional phonetician to annotate at a lower level was prohibitively expensive. Instead, I devised an interface similar to the play-back mode described earlier. In the standard play-back mode, one can see the visual reactions to the utterances as they are being played. In the annotator-mode, however, the system hid these from view and paused while the native speaker annotated the utterance.

Each sentence was labeled with one or more actions. The annotator also had the option to toss out utterances that she did not understand. Of the 1543 utterances that were recorded, 1467 were fully understood and annotated by the native speaker. Using the human as ground truth I found that the system successfully responded to the sentences 83.98%

Notion	Count	AER (%)	NER (%)
select	777	12.23	0.92
drop	778	17.56	3.66
shift	35	20.00	0.0
total	1590	15.01	2.24

SER by Task (%)						
T3	T4	T5	T6	T7	T8	All
17.5	17.4	15.9	16.1	13.1	15.9	16.02

Figure 4-8: *Error rate breakdown by action, notion, and task.*

of the time. The ability for all the users to complete the exercises suggests that a sentence error rate (SER) of 16.02% is adequate.

Despite the fact that the tutorial walked the user through the game one action at a time, some users realized they could compose a single sentence out of two actions, and proceeded to do so throughout the game. 123 sentences contained the *select* notion followed by a *drop*. Thus, I come up with two other metrics by which I evaluate the system. The first is an *action error rate* (AER), which is similar to SER except that sentences are broken up into independent actions. The second is *notion error rate* (NER) where the human and recognizer agree on the utterance representing either *select*, *drop*, or *shift*, but not necessarily on the instantiation of that notion. Figure 4-8 shows the breakdown. The action error rate is necessarily higher than that of its corresponding notion. Note also that the *shift* notion was rarely used, since the *drop* AER was relatively low. The high *shift* AER is likely due to the students' lack of practice in its use.

I also looked at the individual vocabulary words that instantiated the *select* notion. Reporting error rates for individual vocabulary words would unfairly bias the poor performance to those words that happened to be given to the less proficient users. Indeed, because not all users had the same target images, I can only present a crude analysis of which words caused problems for the recognizer. According to the annotations, a given word was spoken on average by over half of the users. It seemed that many users would practice selecting words even when they were not in their target set. Interestingly, only

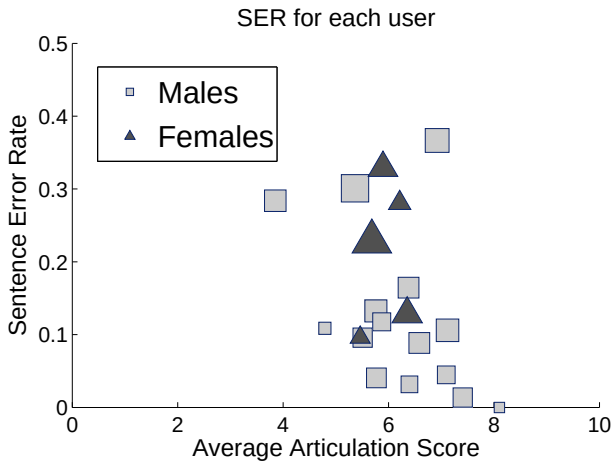


Figure 4-9: *Sentence error rate for individual users as a function of their average articulation score. The size of the shapes is roughly proportional to the number of utterances they used to complete the study.*

six words were misrecognized by more than one of the participants. The most commonly misrecognized vocabulary item was *nì jǐ jīng*, meaning *killer whale*. In addition to being the most obscure word in the study, causing a number of false starts and mispronunciations, it appeared that microphone quality had a large effect on its proper recognition.

Lastly, I also found evidence that users improved in SER as they progressed in the study. The experiments were not designed with a rigorous analysis of such a trend in mind, and I make no claims about what the causes of such a tendency might be. I do, however, report the SER as broken down by task in Figure 4-8.

Error Analysis

When a speech recognizer is used for second language learning, it is of vital importance that the mistakes it is making are in some sense the *correct* ones. Many language learners have undoubtedly already had the frustrating experience of being penalized for a properly uttered sentence while using currently available ASR systems. Thus, I would like to ensure that a sentence uttered proficiently by a learner has a lower probability of misrecognition by the system than one that contains pronunciation or grammar mistakes.

To test this, the annotator also evaluated each utterance against four metrics with values

ranging from 0 to 10: tone accuracy (t), articulation (a), speech speed (s), and sound quality (q). The tone and speed metrics are self-explanatory; sound quality refers to properties of the audio independent of the sentence uttered (e.g., background noise, microphone quality, etc.), and articulation is a measure of the non-nativeness of the Chinese speech independent of tone. The annotator tells us that to measure a she would often repeat the sentence back to herself correcting for tone and speed, then determine a score for the proficiency with which it was uttered. It is this metric that, if the recognizer functions as expected, should inversely correlate with the probability of a recognition error.

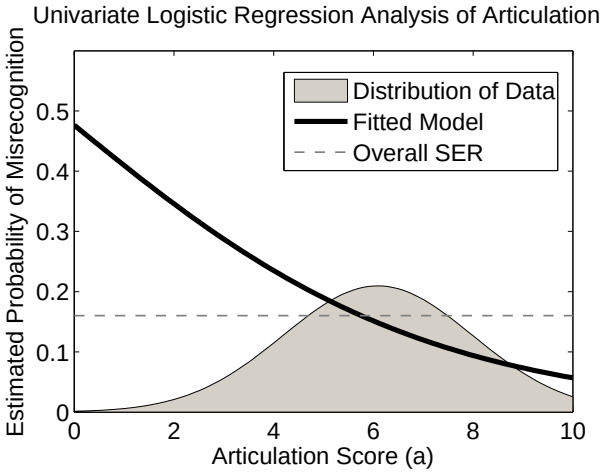
A coarse analysis is given in Figure 4-9 where sentence error rate for a given user is plotted against that user's average articulation score. Even here one can see hints of the correlations one expects; however, outliers, such as the male user towards the top-right of the plot, cloud the picture.

To carry out a more rigorous investigation, the four metrics are treated as continuous variables and a single sentence error (e) as a binary response variable. A multivariate and four univariate logistic regression analyses [24] are performed to measure the influence of the metrics on e. Statistical software [48] enabled us to compute the coefficients and intercept terms in the standard logistic regression model, reproduced below:

$$y(x) = \frac{1}{1 + exp(-\alpha - \beta x^T)}$$

Given the annotations of the $n = 1467$ utterances in the form $y_i = e_i$ and $x_i = [m_i]$ for each metric m, Coefficients β_0 (and intercept) for four univariate models are computed. Each model then estimates the probability of a misrecognition as a function of the associated metric (albeit independent of the other three.)

Figure 4-10 shows a plot of the univariate model for a. The curve clearly shows that the probability of a misrecognition inversely correlates with articulation proficiency. This model estimates that sentences spoken with a high articulation score ($a > .8$) will be recognized correctly over 90% of the time. Although the data at this end of the spectrum are sparse, Figure 4-9 corroborates this with more evidence: the most proficient participant had no recognition errors at all. Coefficients for the remaining metrics appear in the table



	Univariate		Multivariate	
metric	β_0	p value	β_i	p value
(t)one	-0.0358	0.2720	0.0074	0.8383
(a)rticulation	**-0.2770**	**<0.0001**	**-0.2613**	**<0.0001**
(s)peed	**-0.1537**	**0.0006**	-0.0801	0.1145
(q)uality	**-0.1443**	**0.0037**	-0.0901	0.0927

Figure 4-10: *Logistic regression analysis of the four metrics on misrecognition ($n = 1467$). Coefficients for a single multivariate and four univariate regressions are given. A plot of the fitted model for (a) illustrates how articulation score inversely correlates with the probability of misrecognition.*

of Figure 4-10. Not surprisingly, since the recognizer does not include tone information, the slope of the model for t does not significantly differ from zero, even in the univariate case.

To evaluate the effects of the four metrics when considered in combination, a multivariate regression analysis on the data is performed in the form $y_i = e_i$ and $x_i = [t_i, a_i, s_i, q_i]$. The computed coefficients $\beta = [\beta_0, \beta_1, \beta_2, \beta_3]$ and their statistical significance can be seen in Figure 4-10. The multivariate analysis suggests that the recognizer is fairly robust to the speed of one's speech. Sound quality had a slight association with misrecognition; however, overall it appeared that the users were able to interact with the system adequately given their respective recording resources and environments. In the multivariate model, articulation – that is non-nativeness independent of tone – was still the best predictor of a misrecognition, with $\beta_1 = -0.26128$ at $p \ll 0.001$.

4.4 Chapter Summary: Iterative Improvements

To summarize, I have publicly deployed a personalizable, speech-enabled system to aid students of Chinese with vocabulary acquisition in a non-threatening environment[4]. A Web-based user study and a subsequent analysis confirms that the Mandarin recognizer serves quite well as a model for human perception of learner speech in this restricted setting. In reviewing the user sessions with the play-back mode, it was clear that users were willing to experiment with various ways of saying the commands. Some grammatically correct utterances were not covered by the system, inspiring us to augment the grammar.

In talking with teachers and users, many suggestions have been made about how this system might be improved. Students who have already mastered the sentence patterns provided in Word War seem to desire more complicated interaction. To this end, the group is currently working on a card game that goes beyond the simple *select*, *drop* and *shift* notions. Others see the use of images as a limitation. I do provide a version where the target images are replaced with characters; however, the bottom rows of the game grid are always restricted to pictures.

[4] http://web.sls.csail.mit.edu/chinesecards

Of all the parts of speech, concrete nouns are the easiest to associate with images, and it is much more difficult to come up with images representing abstract nouns, verbs, adjectives, etc. That said, even a generic system for nouns is a notable achievement. Studies examining transcripts of both adult and child speech suggest that common nouns make up around 40% of the word tokens [3]. Furthermore, nouns are often the *content* words of a sentence: misunderstanding a noun can often render a sentence meaningless.

Even so, it is important to consider ways of extending this system beyond concrete nouns. To this end, I suggest that users consider placing more than a single word on a card. I have provided a feature where users can upload their own images. Imagine playing a game of Word War with your vacation photos, e.g., "Choose the picture of my sister jumping across the stream." As the success of the image-based Rosetta Stone application indicates, with a little creativity a large amount of daily language can be associated with pictures. *Chinese Cards* now provides the additional benefit that the customizability allows anyone to choose the content, and turns the Rosetta Stone activities inside out, requiring the student to practice speaking in addition to listening. Moreover, I am pleased to offer this system to the public free of charge, so that students of Mandarin may practice speaking from the comfort of their own computers.

4.5 Chapter Acknowledgments

I would like to thank Ming Zhu for providing the annotations, Alex Gruenstein and James McGraw for technical contributions, and the study subjects for their participation. I would also like to thank Brandon Yoshimoto: he was instrumental in the development of the listening mode of *Word War*.

Chapter 5

Learning Gains

> "The reader already knows that I consider vocabulary harder to learn than grammar or pronunciation. To become a fairly fluent speaker of a language, with 5,000 words at his command, a person would have to learn ten new words a day, day in and day out, for a year and a half. Few people can keep up such a pace [...]" [47]

The work in this chapter supplements the analysis of recognition performance with preliminary results regarding *Word War*'s effects on vocabulary retention. Three systems, all of which were presented in the previous chapter, are compared: the image-based flash card system, the single-player *speaking* mode of *Word War*, and the single-player *listening* mode of *Word War*. A carefully controlled study required 13 first and second year students of Chinese to interact with all three systems. Pretests and posttests were given to carefully measure learning gains over the course of a three week interval.

This chapter begins with a description of two studies of vocabulary acquisition performed by Rod Ellis, one of the leading researchers in Second Language Acquisition. Using one of these studies as a model, this chapter then describes the experimental setup designed to examine the learning gains achieved by the three systems mentioned above. After presenting the preliminary findings, their implications are discussed with respect to the utility of these three systems as vocabulary acquisition aids.

5.1 Studies in Vocabulary Acquisition

Most studies involving vocabulary acquisition only analyze the effects of various strategies of *intentional* learning [23, 4]. 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, [13, 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 [13]. The setup of this study is remarkably similar to the listening mode of *Word War* presented in section 4.2.2. 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 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 associated 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 posttests. 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 [12], 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. The amount of time that each of the three groups required to complete their respective tasks varied greatly. The interactionally modified group took around 45 minutes to complete the task, while the premodified group took 20 minutes and baseline group used only 10 minutes. A quick computation reveals that in terms of the mean number of words acquired per minute, the premodified and baseline input groups were almost identical, while the interactionally modified group was two to three times *slower*. In [11], Ellis notes that the difference 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* [12]. 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 chapter 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 vocabulary 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 avoiding the delayed gratification of *intentional* learning methods come at a cost of the student's time? And also, is this a price the student would willingly pay?

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.

5.2 Experimental Design

This section provides the experimental setup for the study that was performed to assess learning gains. 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 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.

5.2.1 Instruments

The following instruments were used in the study:

- Word War: Student-Speaking mode
- Word War: Student-Listening mode
- Image-based flash cards
- Picture matching test
- Survey

The hint mechanism for the speaking and listening modes was slightly altered to ensure that the two modes were as similar as possible. Rather than showing a hint when the user

placed the mouse cursor over an image, the student was required to click and hold the mouse on a small Hint button in the top-right corner of each image. A hint consisted only of the pinyin pronunciation and English translation; no Chinese characters were shown and no corresponding audio was played when the hint was in view. Although I feel that the student would benefit from hearing the word spoken correctly we disallowed this in the experiements in order to assure a clear distinction between listening mode and speaking mode in *Word War*. Also, this way hints could be accessed at any time, except when the mouse needed to be elsewhere: e.g. when recording an utterance in speaking mode or when moving an image in listening mode.

Although the flash cards allow a student to review words as many times as they please and even choose specifically the most troublesome words, the *Word War* modes were not initially built with these study habits in mind. To ensure that, with a given set of cards, *Word War* could easily be restarted, a "Re-deal Cards" button was added to the page, so that the category loaded would be reshuffled, new *target* images would be chosen, and the game would start over. Furthermore, an initial attempt was made at choosing these *target* images intelligently based on the number of times the Hint button was pressed in previous games. Although it is unlikely that this algorithm performs on par with a student's own assessment of their vocabulary knowledge, it allows the student to enjoy the game without worrying specifically about which words they are retaining.

All evaluations were performed using a picture matching test. In this test, a list of pictures and the English words they represent were to be matched with the pinyin 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. A similar test was used in parts of the Ellis studies described in the previous section.

5.2.2 Procedure

Since chapter 4 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 our laboratory walls. 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.

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, we instead choose a fixed set of words and compare retention levels for each word conditioned on the mode by which it is learned. The study contained 30 Chinese words (W) 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 ensured that all words were precisely two syllables long, and also avoided using words that were likely to be taught in the first two years of university Chinese courses.

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 5-1. The first two sessions were broken up into an initial test, 10 minutes with the flash cards, 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

Week 1	Activity:	Test 1	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

Week 2	Activity:	Test 2	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

Week 3	Activity:	Test 3
	Time:	∞
	Words:	W

Figure 5-1: 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 $i = 1, 2, \ldots, 13$.

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 subsequent sessions. As indicated in Figure 5-1, 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 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 1 was treated as a pretest and was therefore used to assess a priori knowledge of the vocabulary items. Tests 2 and 3 were posttests 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 5-1 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, as well as a short session in 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 experiments with the speech-enabled *Word War* game have yielded noticeably different behaviors depending on how long a given user has been playing[39], 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, the students had proceeded through all of the tasks for a given session they were asked to fill out a short survey. Some questions were open-ended inqueries into the student's previous experience studying Mandarin, others asked about their study methods, and still others attempted to elicit quantitative answers regarding their experience using our three vocabulary-building systems.

5.3 Experimental Results

This section reports on the learning gains and survey results obtained when the 13 volunteer participants recruited for the study had completed all three sessions.

5.3.1 Learning Gains

Figure 5-2 shows the scores for tests 1, 2, and 3 given at the start of each session. The scores from test 1 indicate that the words we chose were relatively unknown to the study

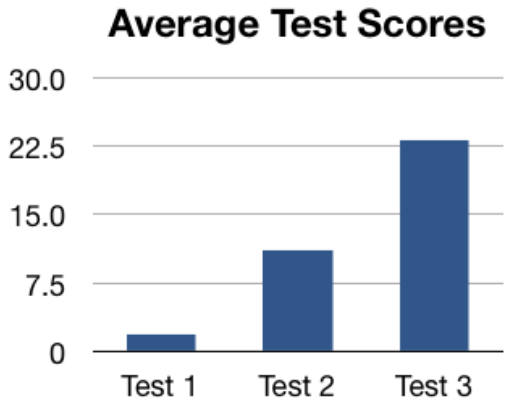


Figure 5-2: Average test scores for the picture matching test given at the start of weeks 1, 2, and 3.

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 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 5-3 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

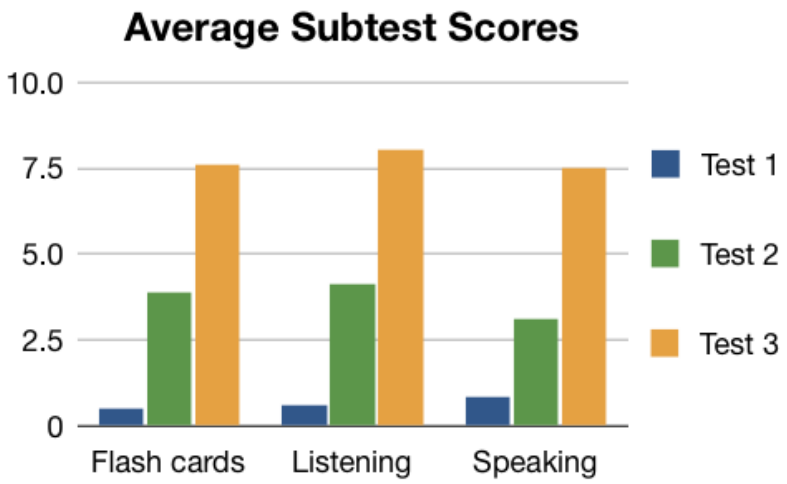


Figure 5-3: Average subtest scores for each system. A subtest score for a system X is computed by taking the words from a test (1, 2, or 3), and grading only the words that were seen in that particular system.

	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

Figure 5-4: Mean learning gains for each system across all 13 users.

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 pretest score, S is a posttest 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 posttest, 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 5-4. 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". Perhaps more interestingly, we can divide the gains into the pieces that are associated with each system by applying the equations to the words that appeared in that system, i.e. to the *subtest* scores. Figure 5-4 displays the results of these calculations as well.

Given the small sample size and the relatively large variance between individual test takers, it is difficult to ascertain statistical significance between these means. A paired t-test indicates that the only difference in learning gains across systems that tends toward significance, with $p < 0.1$, is found when comparing the g_1 scores between the listening and speaking modes of *Word War*. It can be noted, however, that the learning gains achieved by each system individually improved significantly, with $p < 0.01$, between weeks one and two.

5.3.2 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 indicated *"very much"*. Since we asked this question after both sessions one and two, we can compare the responses both across systems and weeks. Figure 5-5 shows the mean responses received for each of the system/session combinations.

A paired t-test reveals that the means between systems are significantly different, with $p < 0.01$, for both weeks one and two. Moreover, when comparing the means for each week within a system the differences trend towards significance with $p < 0.1$ in all 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

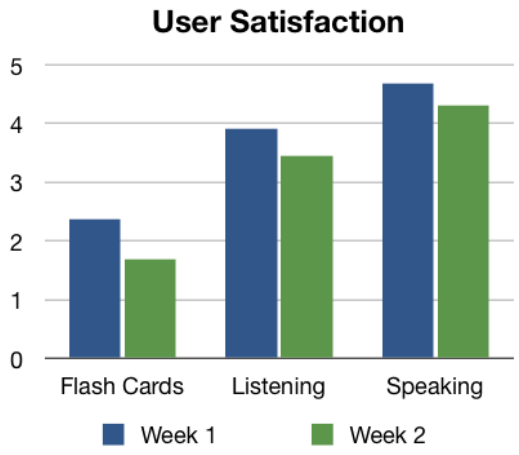


Figure 5-5: *The average response to the following question: "On a scale from 1 (not at all) to 5 (very much), to what degree did you find interacting with this system enjoyable?"*

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 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.

5.4 Discussion

First and foremost it should be noted that, as Figures 5-3 and 5-4 indicate, all of these systems are extremely competitive. 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 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.

It is also interesting to note the relatively low gain achieved by the speaking mode during the first session. There are a few possible explanations for this, 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 lead 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 out-performed the flash cards system. Unfortunately, it is difficult to determine whether this is a causal effect or simply correlated.

It is also unclear whether recognition errors had a large effect on learning gains. When reviewing the user interactions with the speaking mode of *Word War* it became apparent that a number of students had difficulty pronouncing the pinyin 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. It is clear that this will improve recognition given

that, as presented in chapter 3, even prebeginners are able to successfully improve their pronunciation based on synthetic speech.

Somewhat paradoxically, recognition errors *could* contribute positively to retention rates, since they might allow difficult words to be practiced multiple times. One application of ASR technologies for child literacy found learning gains caused precisely for this reason [29]. 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.

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 both the speaking and listening modes. In this case, it is clear that the incidental vocabulary acquisition methods successfully avoid the delayed gratification Krashen warns us against. What might surprise Krashen, however, is that the students particularly enjoyed the speaking mode of *Word War*. As the survey results indicate, the students felt little discomfort when interacting with the recognizer, especially when compared with their experiences in the classroom.

5.5 Chapter Summary: Future Directions

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.

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. The importance of ensuring that the game has high *replay* value is indicated by the fact that user enjoyment

dropped between sessions one and two of our study.

These facts appear to motivate the development of more complex card games, perhaps with rules akin to the traditional card games that many people grow up playing. Indeed, our group is now putting the finishing touches on a game called Rainbow Rummy, which combines both speaking and listening into an enjoyable 1-on-1 strategy game with an artificially intelligent opponent.

Chapter 6

Discussion and Future Work

The studies presented in this thesis are a first step towards understanding the advantages and disadvantages of incidental vocabulary acquisition in the context of computer aided language learning. Two very different approaches to assisting with vocabulary acquisition were presented. The *Family ISLAND* took the approach of a mixed initiative dialogue system with a directed dialogue back-off mechanism to provide prebeginners with an environment in which they could infer the meanings of new words solely from context clues provided by the graphical user interface. A prototype card game called *Word War* maintained tight restrictions on the underlying grammars, but opened the lexical domain to user-created content through *Chinese Cards*, a card creation Web site.

The results of studies on both systems seem to indicate that speech technology for Mandarin Chinese is ready to handle the non-native speech of language learners, provided appropriate measures are taken to manage misunderstandings. It will be important in the coming years to extend this work to other languages, as each language has unique characteristics which pose a variety of challenges to its robust recognition. For Chinese in particular, robust pitch tracking and tone scoring algorithms are needed to assess and provide feedback on tones in learner speech.

At the same time, one should not be tempted to toss out an interesting idea or system simply because a particular component of the necessary speech or language technology seems unreliable. A common theme in this thesis is that there are inherent uncertainties in these technologies, and thus design techniques to deal with them need to be employed. If

care is taken to ensure that the learner is never confused and that system misunderstandings that might cause frustrations are kept to a minimum, the development of enjoyable speech-enabled games for language learning is already possible.

With respect to the two system presented in this thesis in particular, it seems wise to prioritize work on the card game architecture presented in chapter 4. In my estimation, the card games have the best chance of making a meaningful impact on language education in the near future both because they are highly customizable, and because a variety of interesting games can be created using the same generic framework. Furthermore, while the *Family ISLAND* might be of interest to a single first or second year Chinese student for the brief period that they are studying that topic, a well-designed card game might hold the attention of a user over many years.

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 their 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 chapter 5, 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 al-

gorithm might be used to choose which cards to load into the user's next game of *Word War*.

Finally, to ensure that a user is never in a position where they are unable to continue with a *Word War* game due to 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 within a certain number of utterances, 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." When a user is playing with a set of cards that they have previously studied, we might even choose to hide the hints entirely, and rely on this back-off mechanism to kick in if a user has forgotten a word. By guessing at the word a few times, the user would finally be allowed to hear the audio associated with a forgotten card.

When designing card games for language learning it is necessary to balance both the enjoyment of the end user, and the pedagogical advantages of a given implementation. 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* [60], 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 labelling 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.

Appendix A

User Study Management

One method of data collection is to simply deploy a Web-based application, publicize its URL, and analyze whatever user interactions occur. This is certainly appropriate for beta testing, and public system deployments. However, it is often useful for researchers to have a greater amount of control over the types of users recruited and the tasks performed. This appendix describes a web-based user study management interface that allows researchers to manage studies with remote users. Several controlled user studies have already successfully been conducted using this tool.

In keeping with the theme of Web based interfaces, all the user study management tools presented in this appendix are accessible from an ordinary Internet browser. The underlying technology is Asynchronous Java-Script and XML (AJAX), as generated by the Google Web Toolkit [20]. The look and feel of the interface was achieved with the open-source MyGWT libraries available online [44].

This tool supports a number of operations helpful when conducting a user study. Study administrators can create and manage user groups, assignable tasks, and generic web forms through a set of *views* described below. Furthermore, once a user's interaction with the system is complete, a study administrator can replay the interaction just as it appeared in the user's browser. The replayed interaction even outputs the utterances spoken by the user at the appropriate times.

A screen-shot of the web-based *group management view* is shown in Figure A-1. In this view, a study administrator can add or remove users. The easiest way to add users to a

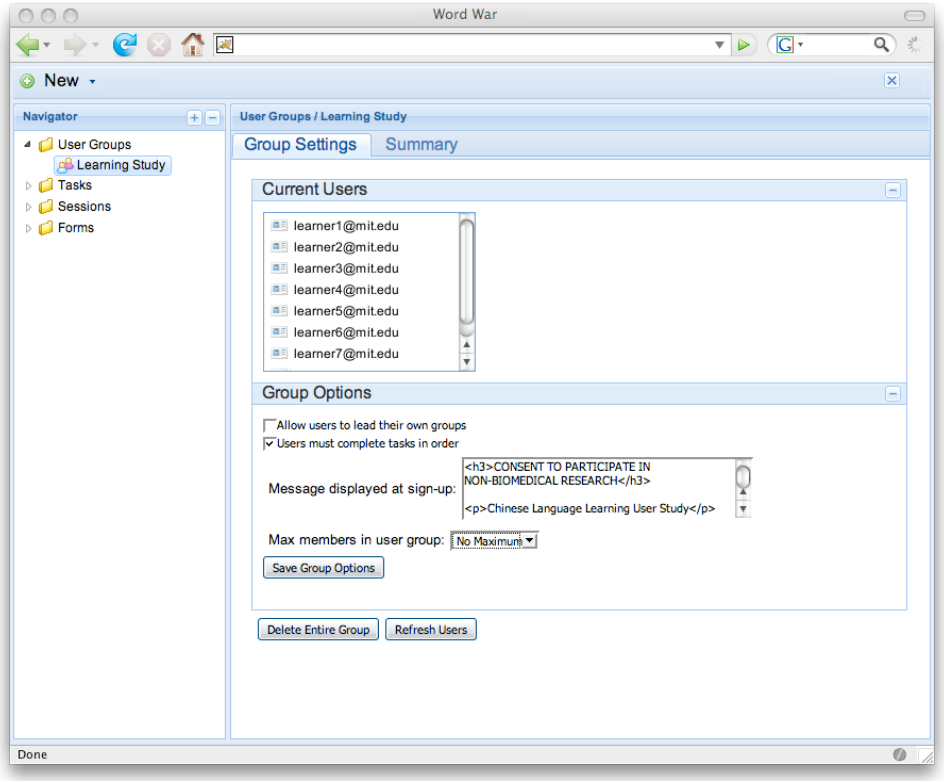


Figure A-1: Group Management View.

group is to have them browse to a special Web site set up for that group and sign up for an account. Study administrators are also able to edit a set of options for each user group. The options include setting the maximum number of members, showing a particular message when the user signs up, etc.

In order to fully specify a new user study, a set of tasks must also be defined. This is done in the *task management view* shown in Figure A-2. Typical tasks might include: reading instructions, completing a warm-up exercise, solving a problem, and filling out a survey. Defining common tasks, such as surveys, is particularly easy, as the core framework provides support for creating and displaying web forms and recording the results to the database. Application-specific tasks can also be defined, in which case the application is responsible for displaying the appropriate content and monitoring a user's progress. Figure A-2 shows the management view for the *Word War* listening mode task described in chapter 5. Within this view, a task can be assigned to a group and web forms can be associated with a task and used either as surveys for the study participants, or as parameters

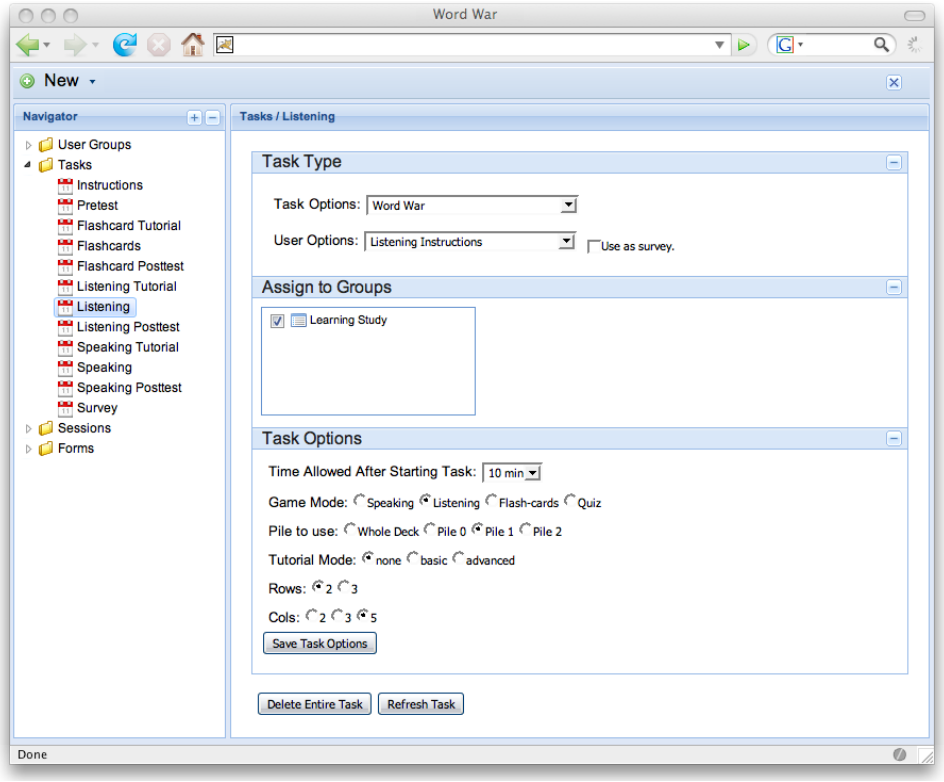


Figure A-2: Task Management View.

passed to the underlying speech-enabled system when the user begins the task.

When a member of a user group logs into an application, he or she is presented with a sequence of assignments to complete, as in Figure A-3. Notice that, although the layout of the tool is similar, when a *non*-administrator logs into the system they do not have any of the functionality associated with administrator privileges. Instead, they are restricted to completing the assigned tasks in the prescribed order. In Figure A-3, the example task displayed requires the user to click on a link and watch a demonstration video. A subsequent task, for example, required the user to interact with *Word War* for 10 minutes, completing as many games as possible within the allotted time.

In a number of instances, a web form is needed to accompany a piece of the user study management architecture, but the contents of the form are not known a priori. One example where this is particularly useful is for creating surveys. For this reason, the *form management view*, shown in Figure A-4, offers administrators a visual form editor, allowing them to choose a form's fields dynamically. Web forms created in this way plug into various

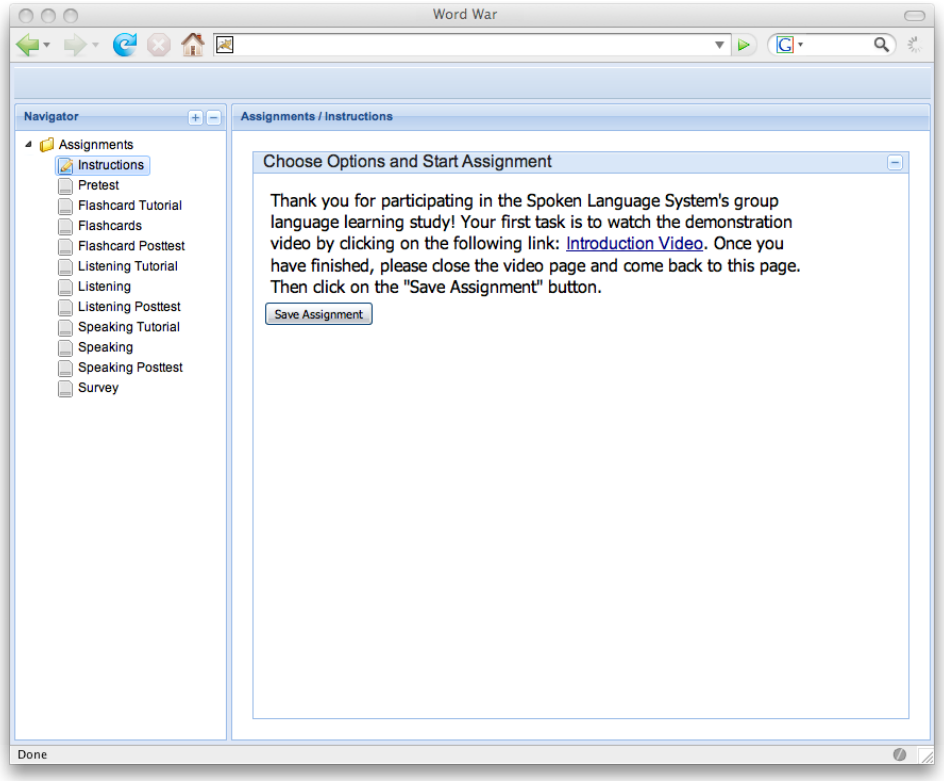


Figure A-3: Assignment View.

components of this user study management architecture when needed. In addition being useful for surveys, such forms can be used to pass options to the application logic, or even used during session playback to annotate each utterance.

Lastly, the *session management view* allows the administrator to review user interactions in real time. Figure A-5 depicts the session manager poised to play back three sessions of the flash cards task. It is particularly interesting to watch interactions that involve speaking, such as the *Word War* speaking-mode task of the user study presented in chapter 5. With this tool we were able to watch learners interact with *Word War* as if we had a video camera trained directly on their browser while they were using it.

All of the aforementioned views combine into a powerful user study management tool that can be accessed from anywhere in the world. This even makes it possible to manage a remote user study... remotely. Indeed, the human-annotator of chapter 4 was a few hundred miles from the user study site while she was replaying the user interactions to perform the annotation task.

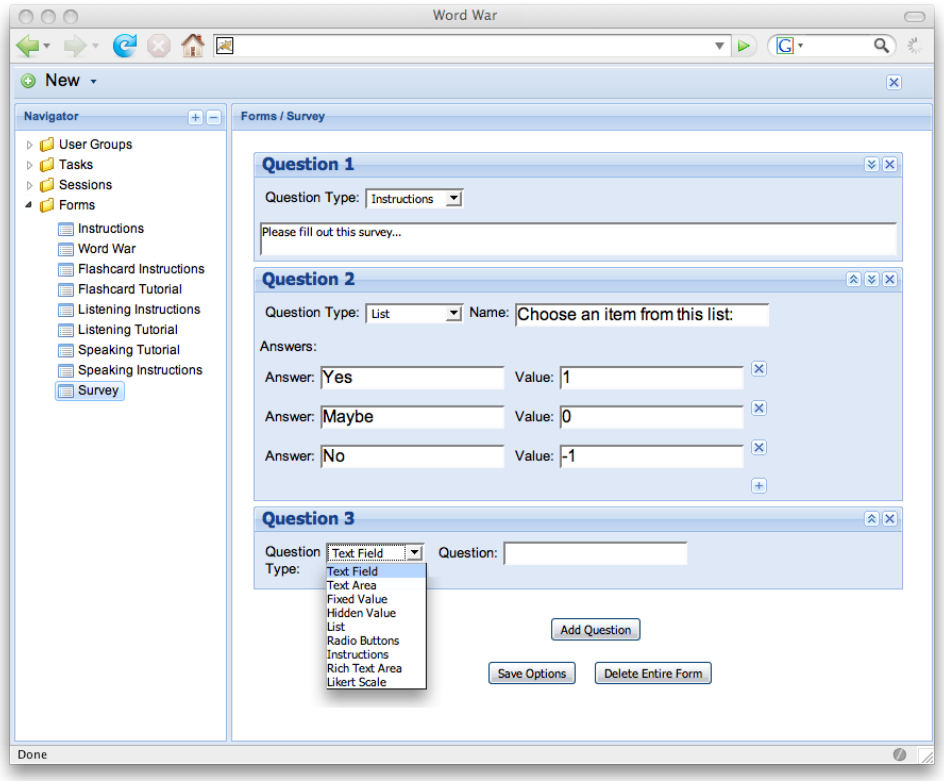


Figure A-4: Form Management View.

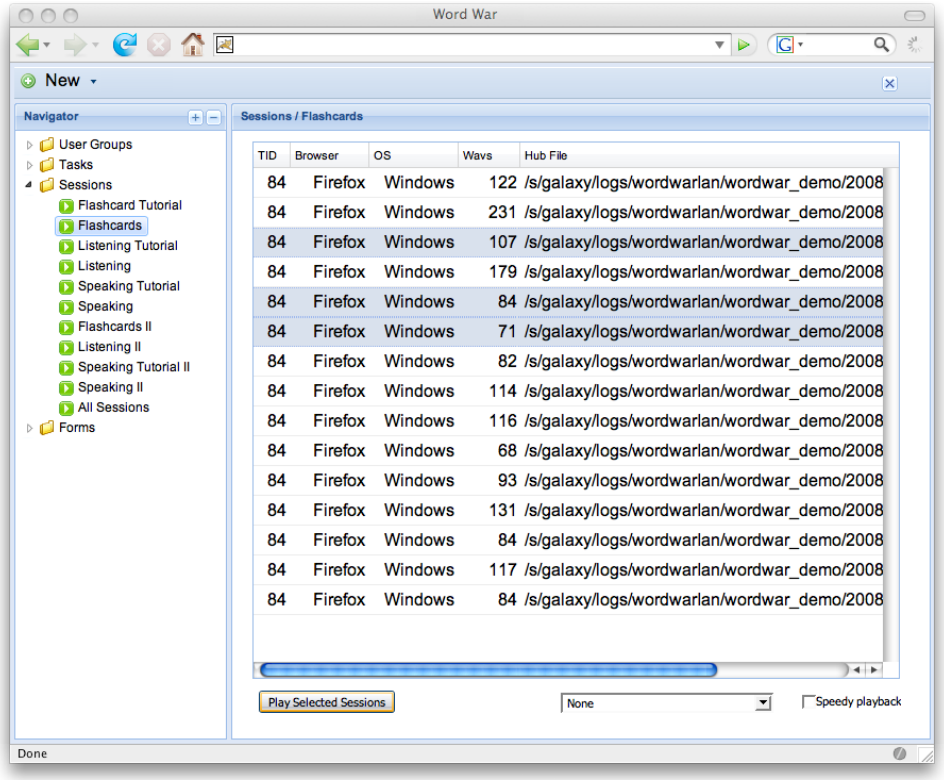


Figure A-5: Sessions Management View.

Bibliography

[1] 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.

[2] L. Baptist and S. Seneff. Genesis-II: A versatile system for language generation in conversational system applications. In *ICSLP*, 2000.

[3] Lois Bloom. The intentionality model. In Marc Marschark, editor, *Becoming a Word Learner*, pages 19 – 50.

[4] 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.

[5] Colin Campbell and Hanna Kryszewska. Oxford University Press, 1992.

[6] Chengo Chinese. E-language learning system, http://www.elanguage.cn. Last accessed: January 30, 2008.

[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] Nick Ellis. The implicit ins and outs of explicit cognitive mediation. In Ellis, N. (ed) *Implicit and Explicit Learning of Languages*, pages 211 – 281. Academic Press, London, 1994.

[11] Rod Ellis. Modified input and the acquisition of word meanings. *Applied Linguistics*, 16:409–441, 1995.

[12] 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.

[13] Rod Ellis, Yoshihiro Tanaka, and Atsuko Yamazaki. Classroom interaction, comprehension and the acquisition of word meanings. *Language Learning*, 44:449–491, 1994.

[14] Maxine Eskenazi. Using automatic speech processing for foreign language pronunciation tutoring. *Language Learning & Technology*, 2(2):62–76, 1999.

[15] Nelly Furman, David Goldberg, and Natalia Lusin. Enrollments in languages other than English in the United States institutions of higher eduction, fall 2006. 2007.

[16] Johann Gamper and Judith Knapp. A review of intelligent CALL systems. In *Computer Assisted Language Learning*, 2002.

[17] 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.

[18] J. Glass. A probabilistic framework for segment-based speech recognition. *Computer Speech and Language*, 2003.

[19] J. Glass and T. Hazen. Telephone-based conversational speech recognition in the Jupiter domain. In *ICSLP*, 1998.

[20] Google Web Toolkit (GWT). http://code.google.com/webtoolkit/. Last accessed: May 1, 2008.

[21] Alex Gruenstein and Stephanie Seneff. Releasing a multimodal dialogue system into the wild: User support mechanisms. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, pages 111–119, Antwerp, Belgium, September 2007.

[22] Alexander Gruenstein, Bo-June (Paul) Hsu, James Glass, Stephanie Seneff, Lee Hetherington, Scott Cyphers, Ibrahim Badr, Chao Wang, and Sean Liu. A multimodal home entertainment interface via a mobile device. In *Proc. of the ACL Workshop on Mobile Language Processing*, 2008.

[23] 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.

[24] Shelby J. Haberman. *Analysis of Qualitative Data: Volume 2, New Developments*. Academic Press, New York, 1979.

[25] 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.

[26] 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.

[27] Thomas Huckin and James Coady. Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition*, 21(02):181–193, 2000.

[28] 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.

[29] I. Kantrov. Talking to computer: A prototype speech recognition system for early reading instruction, 1991.

[30] Stephen Krashen. Why support a delayed-gratification approach to language education? *The Language Teacher*, 28:3–7.

[31] Stephen Krashen. *The Input Hypothesis: Issues and Implications*. London: Longman, 1982.

[32] Stephen Krashen. *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon, 1982.

[33] Stephen Krashen. The input hypothes and its rivals. In Ellis, N. (ed) *Implicit and Explicit Learning of Languages*, pages 45 – 77. Academic Press, London, 1994.

[34] Michael Lewis. Language Teaching Publications, 1993.

[35] Michael H. Long. Input, interaction and second language acquisition. pages 259 – 278, 1981.

[36] Michael H. Long. The role of linguistic environment in second language acquisition. pages 413 – 486, 1981.

[37] 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.

[38] Ian McGraw and Stephanie Seneff. Immersive second language acquisition in narrow domains: A prototype ISLAND dialogue system. In *SigSLaTE*, 2007.

[39] Ian McGraw and Stephanie Seneff. Speech-enabled card games for language learners. In *Proc. of AAAI*, 2008.

[40] Helen Meng, Yuen Yee Lo, Lan Wang, and Wing Yiu Lau. Deriving salient learners' mispronunciations from cross-language phonological comparisons. In *ASRU*, pages 437–442, 2007.

[41] Wolfgang Menzel, Daniel Herron, Rachel Morton, Dario Pezzotta, Patrizia Bonaventura, and Peter Howarth. Interactive pronunciation training. *ReCALL*, 13(1):67–78, 2001.

[42] Jack Mostow, Gregory Aist, Cathy Huang, Brian Junker, Rebecca Kennedy, Hua Lan, DeWitt Talmadge Latimer, IV, R. O'Connor, R. Tassone, Brian Tobin, and A. Wierman. 4-month evaluation of a learner-controlled reading tutor that listens. In F. N. Fisher V. M. Holland, editor, *The Path of Speech Technologies in Computer Assisted Language Learning: From Research Toward Practice*, pages 201 – 219. Routledge, New York, 2008. in press.

[43] Jack Mostow, Steven F Roth, A. G. Hauptmann, and M. Kane. A prototype reading coach that listens. In *Proceedings of the Twelfth National Conference on Artificial Intelligence (AAAI-94), American Association for Artificial Intelligence*, pages 785 – 792, August 1994. Recipient of the AAAI-94 Outstanding PaperAward.

[44] MyGWT. http://mygwt.net/. Last accessed: May 1, 2008.

[45] I.S.P. Nation. *Learning Vocabulary in Another Language*. Cambridge University Press, 2001.

[46] John Nerbonne, Duco Dokter, and Petra Smit. Morphological processing and computer-assisted language learning. In *Computer-Assisted Language Learning (CALL)*, pages 543–559, 1998.

[47] Paul Pimsleur. *How to learn a foreign language*. Heinle & Heinle Publishers, Inc, 1980.

[48] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2007.

[49] Martin Raab, Rainer Gruhn, and Elmar Noeth. Non-native speech databases. In *ASRU*, pages 413–418, 2007.

[50] Angelika Reider. Implicit and explicit learning in incidental vocabulary acquisition. *VIEWS*, 12(2):24–39.

[51] Jack C. Richards and Theodore S. Rodgers. *Approaches and Methods in Language Teaching*. Cambridge University Press, 2001.

[52] Rosetta Stone. http://www.rosettastone.com. Last accessed: May 1, 2008.

[53] S. Seneff, E. Hurley, R. Lau, C. Pao, P. Schmid, and V. Zue. Galaxy-II: A reference architecture for conversational system development. In *ICSLP*, 1998.

[54] Stephanie Seneff. Tina: A natural language system for spoken language applications. *Computational Linguistics*, 1992.

[55] Stephanie Seneff. Interactive computer aids for acquiring proficiency in Mandarin. In *ISCSLP*, 2006.

[56] Stephanie Seneff and Joseph Polifroni. Dialogue management in the Mercury flight reservation system. In *ANLP/NAACL Workshop on Conversational systems*. Association for Computational Linguistics, 2000.

[57] 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.

[58] Merrill Swain. Three functions of output in second language learning. In Guy Cook and Barbara Seidlhofer, editors, *For H.G. Widdowson: Principles and practice in the study of language*, pages 125 – 144. Oxford: Oxford University Press, 1995.

[59] Talk To Me. http://www.auralog.com/. Last accessed: May 1, 2008.

[60] Luis von Ahn. Games with a purpose. *IEEE Computer*, 39(6):92–94, 2006.

[61] 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.

[62] Chao Wang and Stephanie Seneff. Automatic assessment of student translations for foreign language tutoring. In *Proceedings of HLT/NAACL 2007*, pages 468–475. Association for Computational Linguistics, April 2007.

[63] H.C. Wang, F. Seide, C.Y. Tseng, and L.S. Lee. Mat2000-design, collection, and validation of a Mandarin 2000-speaker telephone speech databases. In *ICSLP*, 2000.

[64] J. Yi, J. Glass, and I. Hetherington. A flexible, scalable finite-state transducer architecture for corpus-based concatenative speech synthesis. In *ICSLP*, 2000.