Speech-enabled Card Games for Language Learners

Ian McGraw and Stephanie Seneff
MIT Computer Science and Artificial Intelligence Laboratory
32 Vassar Street, Cambridge, Massachusetts 02139, USA
{imcgraw, seneff}@csail.mit.edu

Abstract

This paper debuts a novel application of speech recognition to foreign language learning. We present a generic framework for developing user-customizable card games designed to aid learners in the difficult task of vocabulary acquisition. We also describe a prototype game built on this framework that, using a Mandarin speech recognizer, provides a student of Chinese with opportunities to speak vocabulary items in a meaningful 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. To assess the Sentence Error Rate (SER) of our prototype, we 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 of 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 our recognizer is behaving appropriately with regard to learner speech, we perform a rigorous analysis of the recognition errors to determine their underlying causes.

Introduction

For most, learning to speak a foreign language is both frustratingly difficult and incredibly rewarding. For some, it is a job requirement in today's global economy. Regardless of motivations, proficiency in a second language is increasingly being 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.

Why is mastering a foreign language so difficult? Dr. Paul Pimsleur, a respected figure in applied linguistics and a household name in language learning, provides some insight into the relative difficulty of the three main components of second language acquisition (SLA):

> "[...] 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; as the vocabulary begins to pile up, one may find oneself forgetting old words almost as fast as one learns new ones." (Pimsleur 1980)

Obviously the difficulty in acquiring new vocabulary depends on both the source and target languages. For instance, it is likely to be a great deal more difficult for a native speaker of English to internalize new words from Chinese than French, due to the lack of shared cognates.

Independent of language choice, the task of explicitly memorizing new vocabulary is uninteresting at best, and downright tedious at worst. It is not surprising that SLA theorists posit that much of one's vocabulary in a second language is acquired incidentally, through reading (Huckin & Coady 2000). Still, an adult learner starting from scratch has little or no foundation from which to infer new vocabulary from written context in the way a more advanced student might. The problem is exacerbated in Chinese where learning to read can take years.

Incidental vocabulary acquisition through conversation is fraught with a different set of issues. For many in the United States, opportunities to practice speaking a foreign language outside of the classroom are rare. Even when presented with such opportunities, beginners are often uneasy about exposing their inexperience (Krashen 1982).

It is our belief that automatic speech recognition (ASR) technology is capable of providing a comfortable environment in which language learners can use new vocabulary in a meaningful way. With the recent rapid emergence of Web 2.0 technologies, and the widespread adoption of Voice over IP (VoIP), one can easily imagine a day when students will routinely interact with educational services that depend critically on audio capture and transmission of speech from an ordinary Internet browser. With this in mind, we have developed a Web-based experimental system which allows learners of Chinese to talk to their computers.

This paper introduces a new framework whose goal is to help learners of Mandarin acquire basic vocabulary by playing Web-based games involving manipulations of images related to the vocabulary. We describe a game that requires the student to speak commands to achieve a simple objective while the computer reacts according to the student's verbal instructions.

The remainder of this paper is organized as follows. After a brief section on related work, we provide a description of

our card creator and the online game. Our first experiments in introducing this system to remote users are then recounted in a long section on data collection, evaluation, and error analysis. Finally, we provide a short summary and a look to the future.

Related Research

Speech recognition as applied to SLA has recently enjoyed increased attention in the research community (Gamper & J.Knapp 2002). Prototype spoken dialogue systems, e.g., (Wik, Hjalmarson, & Brusk 2007; Raux & Eskenazi 2004), attempt to provide pedagogically grounded environments for language acquisition. Some effort has been put into making these systems engaging: one system even provides a 3D video game interface (Johnson *et al.* 2004).

Developing full fledged dialogue systems for SLA is both difficult and time consuming. To ensure a robust system, it is often necessary to greatly restrict the domain (McGraw & Seneff 2007). While research in such systems often produces promising prototypes, they rarely make it out of the laboratory.

The work presented in this paper differs from previous research in a number of respects. First, we adopt a Web-based model, and demonstrate ease of deployment by conducting a user study remotely. Furthermore, our framework allows students and teachers to create their own content to be loaded into our prototype game. While there do exist a few recent examples of Web-based ASR systems for learning Chinese (Chengo Chinese 2004; Wang & Seneff 2007), these systems do not specifically target vocabulary acquisition, nor do they offer the user the ability to personalize their learning experience.

Card Creator

Before describing the prototype card game on which we based our user study, we briefly discuss an additional tool that emphasizes the customizability of our framework. Using Web 2.0 technology, we have integrated an online Chinese language learning dictionary[1] and Yahoo image search[2] directly 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 our database. A set of public categories is available, or users can choose to sign up for a private account. Figure 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 we 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, we do provide a flash-card player that allows students to review vocabulary this way if they so choose. 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, flash-cards encourage students to take words out of any meaningful context, not to mention their inherent tediousness.

[1] http://www.xuezhongwen.net

[2] http://images.search.yahoo.com

Figure 1: *A single card created with our online tool.*

Another common failing of flash-cards is that they do not require the user to speak. While some in the SLA theory community would not regard this as a negative characteristic (Krashen 1994), 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* (Swain 1985). A study consisting of activities very similar to the game we are about to present was even able to show advantages of spoken output towards the task of vocabulary acquisition in particular (Ellis & He 1999).

Word War

The card creator's primary function is to enable users to personalize a card game such as *Word War* to their individual language learning needs. 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. We will not, however, be elaborating on multi-player mode in this paper.

In single-player mode, each game begins by loading a category of cards into the "game grid". A small example of a two-column game grid initialized with the "animals" category is depicted in Figure 2. Typically the game is played on a grid with at least five columns. 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 *concepts* understood by our system: *select*, *drop*, and *shift*. The concepts can be instantiated with vocabulary words, numbers, or right/left directions respec-

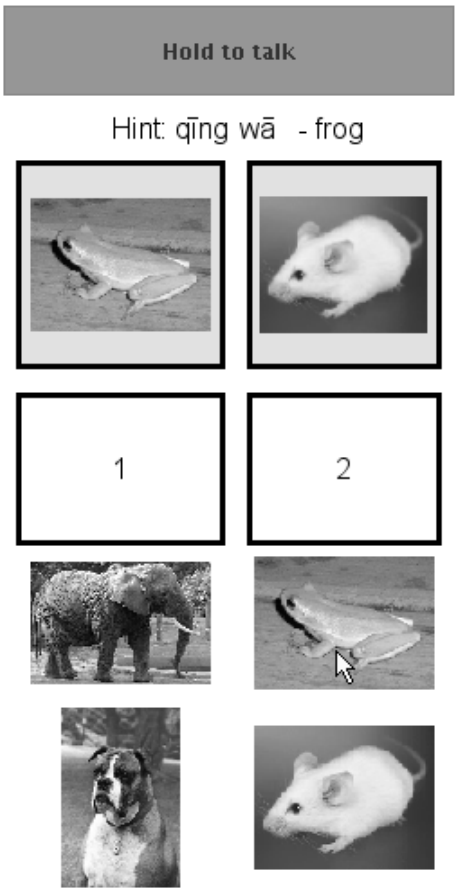


Figure 2: *The two-column version of the card game. The user can mouse-over an image to see the pronunciation hint.*

tively 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 frog. 2) Drop it into slot two. 3) Move it one square to the left. The game is complete once all of the source images have been appropriately aligned with their targets. Note that the *shift* action is typically used for error corrections. For example, a recognition error in the *drop* command might cause a source image to be placed under a non-matching target image.

Each concept can be expressed a number of ways in Mandarin, so we maintain a context free grammar (CFG) 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 speech recognition component running server-side to be used as the language model.

We use the SUMMIT landmark-based recognizer (Glass 2003) 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 we infer the *correct* tones from the language model. Eventually we hope to include feedback on a user's tones, but we must do so in a manner that does not cripple usability for beginners in Mandarin.

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 our 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., (Gruenstein & Seneff 2007).

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 our system is that it provides constant visual feedback in real time as the user is speaking. For example, a student can say "Select the cell phone and put it in the third square." Our system will 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 our system that deserves mention is the play-back mode. During each student's game, the system maintains a detailed log file of the user 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 our system that are pleasing from a pedagogical perspective. The first is that we are able to use images 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 that they are being understood.

Data Collection and Evaluation

This section presents our initial findings from a pilot Web-based user study. Our focus is on the evaluation of the technology itself. Assessment of students' learning gains will be deferred to future research. We first describe in detail our experimental design. After explaining our method for hu-

Concept	Count	AER (%)	CER (%)
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
17.53	17.43	15.86	16.10	13.06	15.86	16.02

Figure 3: *Error rate breakdown by action, concept, and task.*

man annotation of the data, we present an analysis to determine the relationship between understanding error and various contributing factors.

Experimental Design To measure robustness in a realistic setting we administered a user study from the publicly available version of our system[3]. We 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 we gave users a $15 Amazon.com gift certificate.

Each of the eight tasks in the study was in the form of a Word War game. With our card creator we 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 2.) 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. An example sentence for each concept was 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 appeared to have technical difficulties relating either to browser incompatibility or misconfigured audio settings. These users did not progress beyond the first tutorial. The 20 individuals who *did* finish the first tutorial also finished the remainder of the study.

In all, we collected over 1500 utterances 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 our system we asked a native Chinese speaker to annotate each of the 1543 utterances from the six non-tutorial tasks. We 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, we 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 our 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 speaker as ground truth we found that the system successfully responded to the sentences 83.98% 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 we walked the user through the tutorials 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* concept followed by a *drop*. Thus, we come up with two other metrics by which we evaluate our 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 concept error rate (CER) where the human and recognizer agree on the utterance representing either *select*, *drop*, or *shift*, but not necessarily on the instantiation of that concept. Figure 3 shows the breakdown. As we expect, the action error rate is necessarily higher than that of its corresponding concept. Note also that the *shift* concept was rarely used, since the *drop* AER was relatively low. The high *shift* AER is likely due to the student's lack of practice in its use.

We also looked at the individual vocabulary words that instantiated the *select* concept. 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 we are only able to present a crude analysis of which words caused problems for our 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 six words were misrecognized by more than one of our participants. The most commonly misrecognized vocabulary item was *nǐ jǐ jīng*, meaning *killer whale*. In addition to being the most obscure word in our study, causing a number of false starts an mispronunciations, it appeared that microphone quality had a large effect on its proper recognition.

Lastly we also found evidence that users improved in SER as they progressed in the study. Our experiments were not designed with a rigorous analysis of such a trend in mind, and we make no claims about what the causes of such a tendency might be. We do, however, report the SER as broken down by task in Figure 3.

Error Analysis When a 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, we would like to ensure that a sentence uttered proficiently

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

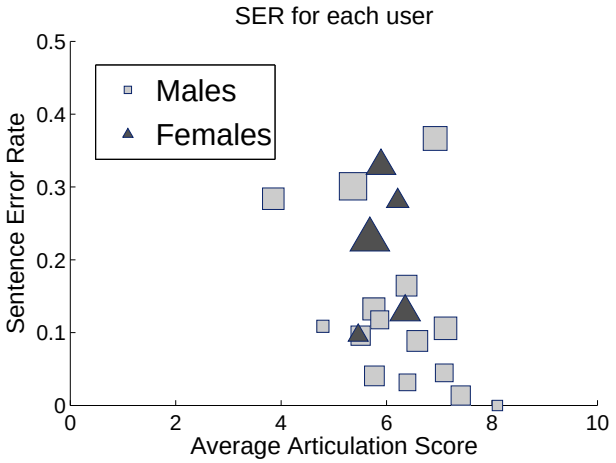
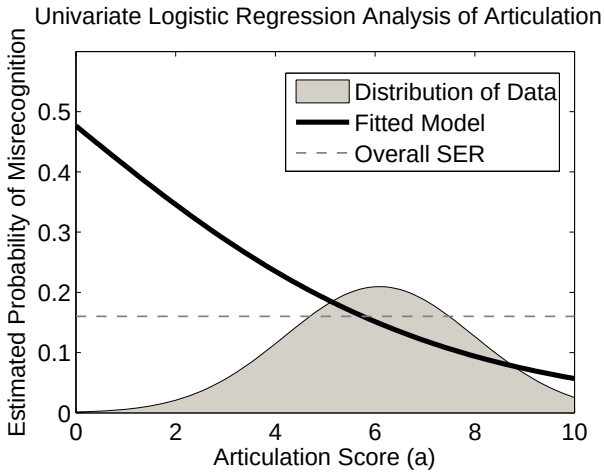


Figure 4: *Sentence error rate for individual users as a function of their average articulation score. The size of the shapes is roughly proportional to the number of utterances they used to complete the study.*



	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
audio (q)uality	**-0.1443**	**0.0037**	-0.0901	0.0927

Figure 5: *Logistic regression analysis of our 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 shows that articulation score inversely correlates with the probability of misrecognition.*

by a learner has a lower probability of misrecognition by our system than one that contains pronunciation or grammar mistakes.

To test this, our 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 (q) refers to properties of the audio independent of the sentence uttered (e.g., background noise, microphone quality, etc.), and articulation (a) is a measure of the non-nativeness of the Chinese speech independent of tone. Our 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 our recognizer functions as expected, should inversely correlate with the probability of a recognition error.

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

To carry out a more rigorous investigation, we treat our four metrics as continuous variables and a single sentence error (e) as a binary response variable. We perform a multivariate and four univariate logistic regression analyses (Haberman 1979) to measure the influence of the metrics on e. Statistical software (R Development Core Team 2007) enabled us to compute the coefficients and intercept terms in the standard logistic regression model, reproduced below:

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

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

Figure 5 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 our data at this end of the spectrum are sparse, Figure 4 corroborates this with more evidence: our most proficient participant had no recognition errors at all! Coefficients for the remaining metrics appear in the table of Figure 5. Not surprisingly, since our 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 our four metrics when considered in combination, we next perform a multivariate regression analysis on the data 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 5. The multivariate analysis suggests that our recognizer is fairly robust to the speed of one's speech. Sound quality had a slight association with misrecognition; however, overall it appeared that our 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 $p \ll 0.001$.

Conclusions and Future Work

To summarize, we have publicly deployed a personalizable, speech-enabled system to aid students of Chinese with vocabulary acquisition in a non-threatening environment. A Web-based user study and a subsequent analysis confirms that our 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 three concepts. Some grammatically correct utterances were not covered by our system, inspiring us to augment our grammar.

In talking with teachers and users, many suggestions have been made about how we might improve our system. Students who have already mastered the sentence patterns provided in Word War seem to desire more complicated interaction. To this end, we are currently working on a card game that goes beyond the simple *select*, *drop* and *shift* concepts. Teachers note that a listening mode, where it is the computer who gives the directions, would be of practical value as well. Thus, we are also exploring the possibility of using synthetic speech to generate commands that the student can follow manually.

Others see the use of images as a limitation. We do provide a version where we replace our target images with characters, however the bottom rows of the game grid are always restricted to pictures. 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. We suggest, however, that users consider placing more than a single word on a card. We have provided a feature where users can upload their own images. Imagine practicing a game of Word War with your vacation photos: "Select the picture of my sister jumping across the stream."

Finally, in the near future, we hope to assess learning gains in terms of vocabulary retention. It is our belief that applications of ASR to SLA need not merely draw from the rich field of applied linguistics, but can contribute to it as well. In his paper, "The roles of modified input and output in the incidental acquisition of word meanings," Ellis suggests that it is *modified* spoken output that explains the higher rates of vocabulary acquisition in one of his experimental groups. We plan to bring the Word War game into the classroom and perform a longitudinal experiment in a similarly controlled setting. Measuring long term memory effects induced by our system could help to evaluate whether the relatively *unmodified* output of the system's users also produce similar learning gains.

In the meantime, we are pleased to offer our system to the public so that students of Mandarin may practice speaking from the comfort of their own computers.

Acknoledgments

The authors would like to thank Ming Zhu for providing the annotations, Alex Gruenstein and James McGraw for technical contributions, and all our anonymous subjects for their participation in the experiments.

References

Chengo Chinese. 2004. E-language learning system: http://www.elanguage.cn. Last accessed January 30, 2008.

Ellis, R., and He, X. 1999. The roles of modified input and output in the incidental acquisition of word meanings. In *Studies in Second Language Acquisition*, volume 21, 285 – 301.

Gamper, J., and J.Knapp. 2002. A review of intelligent CALL systems. In *Computer Assisted Language Learning*.

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

Gruenstein, A., and Seneff, S. 2007. Releasing a multimodal dialogue system into the wild: User support mechanisms. In *Proceedings of the 8th SIGdial Workshop on Discourse and Dialogue*, 111–119.

Haberman, S. J. 1979. *Analysis of Qualitative Data: Volume 2, New Developments*. Academic Press, New York.

Huckin, T., and Coady, J. 2000. Incidental vocabulary acquisition in a second language. *Studies in Second Language Acquisition* 21(02):181–193.

Johnson, W. L.; Beal, C. R.; Fowles-Winkler, A.; Lauper, U.; Marsella, S.; Narayanan, S.; Papachristou, D.; and Vilhjálmsson, H. H. 2004. Tactical language training system: An interim report. In Lester, J. C.; Vicari, R. M.; and Paraguaçu, F., eds., *Intelligent Tutoring Systems*, volume 3220 of *Lecture Notes in Computer Science*, 336–345.

Krashen, S. 1982. *Principles and Practice in Second Language Acquisition*. Oxford: Pergamon.

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

McGraw, I., and Seneff, S. 2007. Immersive second language acquisition in narrow domains: A prototype ISLAND dialogue system. In *SigSLaTE*.

Pimsleur, P. 1980. *How to learn a foreign language*. Heinle & Heinle Publishers, Inc.

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

Raux, A., and Eskenazi, M. 2004. Using task-oriented spoken dialogue systems for language learning: potential, practical applications and challenges. In *InSTIL/ICALL*.

Swain, M. 1985. Communicative competences: Some roles of comprehensible input and comprehensive output in its development. In S. Gass and C. Madden (Eds.)*, Input in Second Language Acquisition*, 235 – 253.

Wang, C., and Seneff, S. 2007. Automatic assessment of student translations for foreign language tutoring. In *Proceedings of HLT/NAACL 2007*, 468–475. Association for Computational Linguistics.

Wik, P.; Hjalmarson, A.; and Brusk, J. 2007. Deal, a serious game for CALL, practicing conversational skills in the trade domain. In *SigSLaTE*.