Immersive Second Language Acquisition in Narrow Domains: A Prototype ISLAND Dialogue System

Ian McGraw and Stephanie Seneff
{imcgraw,seneff}@csail.mit.edu

MIT Computer Science and Artificial Intelligence Laboratory

Abstract

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 the manner in which the SLA theory community might recommend. It is our belief that automatic speech recognition technology in general and spoken dialogue systems in particular have the potential to bridge this gap. In this paper, we lay out some principles behind dialogue system design in this context, and introduce a prototype language learning dialogue system in Mandarin Chinese.

1. Introduction

It almost goes without saying that learning a language is an extremely difficult endeavor. If one's aim is to acquire conversational fluency in the target language, opportunities to practice speaking outside of the classroom are paramount to success. Regrettably, such opportunities do not always exist. This alone seems to present a niche best filled by automatic speech recognition as geared towards the language learner.

Of specific interest to the spoken dialogue systems community is the development of Communicative Language Teaching (CLT) as a widely adopted approach to teaching a foreign language [1]. 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 [2] and flight [3] information. Leveraging this research, we have in recent years begun to build dialogue systems targeting the second language learner [4].

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 [5].

In this paper, we 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. We found that language learners can be far more tolerant than native speakers with respect to recognition errors in dialogue systems. Furthermore, we 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.

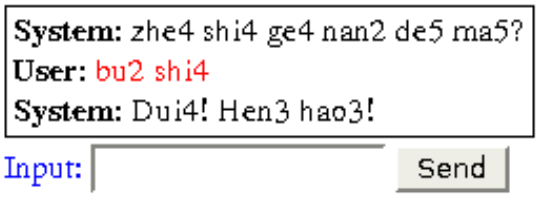

Figure 1: *Dialogue panel as presented on web page. Notice the optional "Input" text field, that the language learner can use when recognition problems occur. The recognition result itself is highlighted in red to draw the users attention to potential mistakes.*

Incorporating these insights into a set of design principles, we have developed a new type of dialogue system to support Immersive, Second Language Acquisition in Narrow Domains (ISLAND). To test our assumptions we have designed and implemented an ISLAND dialogue system in Mandarin Chinese. Our ISLAND is *immersive*, in that no content information whatsoever is given to the user in his or her source language. We refer to *Language Acquisition*, as opposed to "language learning," as we do not incorporate a formal discussion of grammar into our dialogue. Finally, the scope of our dialogue is limited to the *narrow domain* of gender and family relationships.

This paper is organized as follows: In section 2, we describe our family ISLAND in detail. We then lay out the design principles we applied to our system in section 3, describing how they attempt to minimize the effects of common problems in dialogue systems. Then, in section 4, we describe an initial testing and data collection iteration, and present some early but promising results.

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 [6]. In contrast, despite the fact that the content is entirely in the target language, we envision our system spanning the prebeginner to late-beginner stages of students of Mandarin Chinese.

Our 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

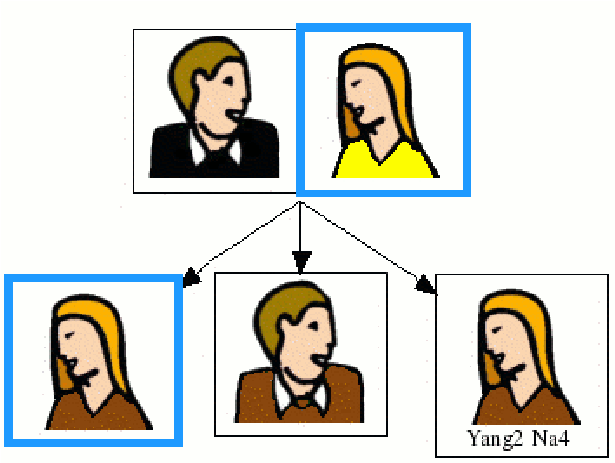

Figure 2: *A particular task as presented to the user in the form of a family tree. A user can click on the family members with thick blue borders to record an utterance.*

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 1, where the user can monitor the conversation and, in particular, the recognition performance. Secondly, we display a family panel (see Figure 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 our ISLAND is about gender. Each task in this level begins by showing 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. 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 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 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.

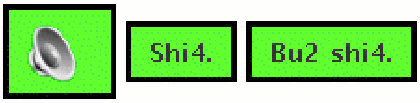

Figure 3: *Hint buttons for the question "Is this a man?" Pressing these buttons will play the corresponding text as a native speaker might say them. The speaker button on the far left is always present, as it repeats the system's most recent question or response.*

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 we are 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. In section 4, we describe a set of experiments with users who were able to accomplish this feat.

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 our system are integrated using the Galaxy architecture [7], which allows communication among a set of servers that perform each of the aforementioned tasks.

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

3.1. Speech recognition & synthesis

Speech technology is the core of our 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 our system is Envoice [8]. In an effort to come as close as possible to native speech with minimal recording requirements, Envoice uses a small corpus of prerecorded utterances from which to splice together new utterances.

The recognition component of our ISLAND utilizes the SUMMIT landmark-based system [9] with acoustic models trained from native Chinese speakers [10]. Tones are not explicitly modeled although they can be inferred by our language model given the system's narrow domain. Aside from disregarding tones, our 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, our system never attempts to hide recognition results from the user. In fact, the user's utterances are highlighted to draw the learner's attention to them in case of an error.

3.2. Natural language understanding & generation

The natural language understanding component of our system makes use of a syntax-based grammar, along with a probabilistic model that can be trained on an untagged corpus of synthetic utterances [11]. Language generation is provided via an in-house generation system [12]. Our 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. A single developer was able to port our 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 our system into the fringes of our 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. 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. Our 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. We believe that dialogue systems targeting the language learner are particularly well suited to these advantageous design methodologies.

3.3.1. 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 [3], 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 at the dialogue management and even the recognition components of the system.

In our domain, 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.

3.3.2. 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.

3.3.3. History on display

These techniques aside, recognition errors are inevitable. To minimize their impact on learning, we assert that it is essential to draw the user's attention to them. In the fourth level of our 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. We attempt to avoid misunderstanding in our 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, we believe that this solution is far more likely to yield effective language learning tools than attempting to hide recognition errors from the user.

4. User study

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

Our study consisted of 17 pre-beginners, each of whom interacted with our system alone for around one hour. A set of general instructions guided them through the use of our system, and they were made aware of each level's general domain. They then progressed through levels 1 through 4 of our 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

	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 4: *Usage statistics as averaged over our 17 participants.*

required them to discover 7 relatives in a person's family tree.

We devised a simple matching test consisting of 16 vocabulary words and their English translations. We allowed the users to take notes as we were not interested in the short-term memory effects of our system. We analyze the test results with respect to the 12 users who took notes as we 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. 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, we are able to infer performance information from the test scores in combination with usage statistics as summarized in Figure 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 our 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 our development cycle, we 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?
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 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.

5. Conclusions and future work

In this paper, we 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 our family ISLAND has shown that such systems can provide an immersive environment in which even pre-beginners can practice

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

conversational skills.

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

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

In the long term, while a single ISLAND could be deployed in a multitude of classrooms, it merely covers a small area in the vast land of language. To truly affect second language education, one could imagine a suite of ISLANDs covering various domains: an archipelago.

6. References

[1] J. C. Richards and T. S. Rodgers. *Approaches and Methods in Language Teaching*. Cambridge University Press, 2001.

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

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

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

[5] 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.

[6] J. Gamper and J.Knapp. A review of intelligent CALL systems. In *Computer Assisted Language Learning (CALL)*, 2002.

[7] 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.

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

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

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

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

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