Language Learning Dialogue Systems: Lessons in Proving Yourself

Sean Robertson
Cosmin Munteanu
Gerald Penn
University of Toronto
Toronto, ON M5S 2E4 Canada
sdrobert@cs.toronto.edu
cosmin@taglab.ca
gpenn@cs.toronto.edu

Abstract
With examples from Wizard-of-Oz experiments performed at the University of Toronto, we argue that the principal determinant of success in speech-based spoken dialogue systems for language learning today is users' interactions with and perception of the system in question, making this topic more the provenance of HCI than pedagogy or engineering. Examples include challenges to the application's authority, the complexities of evaluation, participants' cheating, and cultural differences between participants. Some recommendations are provided where appropriate.

Author Keywords
Computer Assisted Language Learning; CALL; Dialogue Systems; Second Language Learning; SLL; Pronunciation Error Detection; Speech Recognition; Wizard of Oz; Human Subjects Experiment.

ACM Classification Keywords
H.5.2. User Interfaces: User-Centered Design

Introduction
Computer-Assisted Language Learning (CALL), depending on the application, either facilitates the practice of some aspect of a second language or supplants traditional teaching entirely, from micro learning sessions [8] to full curricula offered by consumer products like Duolingo [1].

CALL applications must balance pedagogical needs – to learn the language, and to learn *well* – with the perceptions of the user himself. Unlike traditional language courses whose successes are measured by language teachers and the grades they assign, CALL

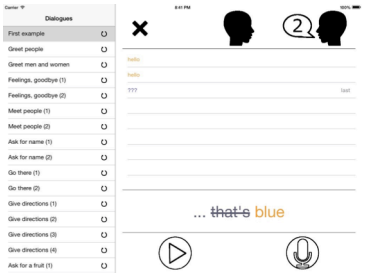

Figure 1. Final experiment user interface on iPad. Word-level feedback were made explicit.

applications require both teacher and user to be satisfied.

This paper focuses on some qualitative findings of two Wizard-of-Oz experiments performed at the University of Toronto. The experiment extrinsically evaluated the performance of various pronunciation error detectors embedded in a prototype CALL dialogue system. Subarashii [2], DEAL [16], and DISCO [15] are examples of CALL dialogue systems which record user's conversational turns. Participants were paired and asked to finish as many scenarios as they could within an hour. Scenarios were dialogues designed by an expert in French second language curricula to gradually introduce absolute beginners to the language. Though participants recorded their speech into and received feedback from the mobile dialogue system, the dialogues otherwise adhered to the popular communicative paradigm of learning [14]. The primary goal of the experiment was to explore the efficacy of various pronunciation error detectors, but most of the interactions between participant and application were observed and, unbeknownst to the participants, partially controlled by a French language expert with a background in second language teaching.

As the experiment proceeded, a number of design decisions were made that were unique to the domain of speech-related CALL. The problems they addressed lay neither in the domain of second language research nor in speech engineering. They included challenges to the authority of the application, difficulty establishing construct validity, subversive user behaviour, and differing cultural backgrounds. Each decision involved changing the parameters of user interaction to manage their perception of the application, making the decisions relevant to HCI.

Against Authority

User perspective on the sophistication of an application can influence their behaviour [13]. This is especially important to learning applications that must portray expertise. Authority can either be granted by an appropriate external institution, such as the CEFRL [18], or is earned. Hence, any feedback presented by a CALL dialogue system must be reliable so as not to undermine itself.

This problem was illustrated in one of the Toronto experiments' dialogue systems. The system could only provide binary feedback per utterance (accept or reject), since real conversation was unlikely to elicit linguistic corrective feedback. At this point, the pedagogy would expect negotiation of meaning [14], in this case between partners. Without the pressure to perform and the assurance that something was indeed incorrect implicit in the presence of a live expert, participants would often repeat the exact same utterance until the wizard capitulated. This served to solidify the mistake and prove to the participant that he could challenge and beat the application.

Another experimental dialogue system allowed the wizard to provide explicit word-level feedback with examples of the correct pronunciation of words alongside rejections. While not enough to correct all mistakes, participants were more willing to experiment with their utterances. Further, the wizard was more persistent in rejecting errors. If she initially gave participants a pass so as not to discourage him, she would return to rejections later. It is worth noting that

Figure 2. Experiments incorporated real-world props to immerse users.

participants did not enjoy this system as much as the other one described above, which highlights the difference between user improvement and user enjoyment.

Evaluating Efficacy

Evaluating user improvement is especially difficult when dealing with certain aspects of the speech signal. Verb conjugations are right or wrong and aberrant grammar (at least the most garish) can be detected by restricting a speech recognizer to a subset of phrases. Pronunciation and prosody arguably do not have the benefit of clear boundaries or discrete categories. Nonetheless, research suggesting the inability to perceive different target-language sounds depending on native tongue [3] and research suggesting teachers do not often teach pronunciation [5] make it a suitable candidate for CALL.

How to assess and train users in pronunciation and prosody is up for debate. A common approach to pronunciation error detection with convenient analogies to speech recognition is to make decisions based on some "distance" between the observed speech and a prototypical native speaker, e.g., the famous Goodness-of-Pronunciation algorithm [17]. This assessment is sometimes called "nativeness." Nativeness received backlash from the pedagogical community because it presupposes that the prototypical method of speaking is the correct one, insensitive to how intelligibly a student speaks [12].

While incorrect and unethical in pronunciation assessment, proximal assessments such as nativeness may be appropriate to CALL, assuming that users can learn from them. Evaluating the quality of learning may be even more difficult, given the sophistication of learning. An application should also promote long-term retention, unlike cramming. In addition, (post-)communicative second language researchers stress the importance of "deep" acquisition of meaning embedded into social context [14]. Furthermore, if a language concept is embedded amongst others, an application needs to manage the attention of the user across the concepts. The relative importance of prosody over, say, a comprehensive lexicon is not addressed by pronunciation researchers.

Choosing the correct balance of content solely from a pedagogical perspective is not guaranteed to be appreciated by the user. The Toronto experiments, for example, found a significant positive correlation between participant performance in post-treatment quizzes and good feelings about the application. The quizzes could only test vocabulary, grammar, and understanding. The wizard's post-treatment evaluation of pronunciation improvement, however, was not significantly correlated to any post-study survey response. This might relate to participants' inability to perceive their pronunciation errors. If a user cannot perceive her improvement, he may lose interest in the application.

Gaming the Game

Speech corpora used for training and testing pronunciation error detectors are often filled with "clean" data: they contain learners of intermediate ability reading prompts aloud [4,7,10]. Though it might be possible to automatically remove these utterances in deployment, it is far more difficult to enforce the implicit assumption that users are acting in good faith.

In the Toronto Wizard-of-Oz experiments, participants were remunerated regardless of their performance. This is both bad and good from a motivational standpoint: ideally they would arrive self-motivated to learn, evinced by their purchase of software, but at least they were not induced to complete scenarios. Those uninterested in learning were expected to laze about. There was further social disincentive to cheat due to the presence of an experiment administrator. Nonetheless, even during the piloting stage when feedback was completely controlled by the wizard (albeit incognito), participants would engineer utterances designed to "game the system," or force some parameter of the application to accept an invalid utterance. One common gaming behavior, described earlier, was to produce identical utterances to force a pass. Another was to produce a nonsense utterance to be rejected. Rather than thinking about what to say next, the participant could rely on the phrase provided with the rejection to be correct, and repeat it verbatim. Another was to replace target language words with native language words, especially when the participant's native language was in the same language family. Sometimes these words were cognates; at other times, rough homonyms. Another was to obscure the word by speaking very quickly, mumbling, or humming. Though some error detectors accommodate for a specific first language's phonemic inventory to combat non-native words [11], the last three techniques are anathema to techniques based on automatic speech recognizers

Over the course of piloting the experiment, cheating became a given: it was more prudent to detect and punish cheating than to try to distract participants with interesting, fun, or educational activities. Participants did and would continue to enjoy finding ways to game the system over deriving any long-term benefits. As mentioned, participants were not guaranteed to share motivations with those self-motivated users who would invest in similar software. Still, an awareness of the limitations of the underlying technologies, especially those involving the vagaries of speech, and attention to cheating during user studies are warranted by these findings.

Culture Clash

In an analysis of a post-study survey that measured participant sentiment towards the experimented software, participants were grouped broadly according to the first languages they reported in a pre-study survey: those that spoke only English fluently, those that spoke English and another language fluently, and those that spoke only a Chinese dialect fluently. One-way ANOVA yielded significant differences in the average phrase rejection rate, number of questions correct in post-treatment summative evaluations, how engaged participants felt, and how quickly they learned with the method. For all of these, English speakers outperformed dual language speakers who outperformed Chinese speakers; English speakers found the method faster and more engaging than dual language speakers than Chinese speakers. Bonferroni-adjusted pairwise analysis revealed only significant differences between the Chinese and English speakers, and only in engagement and perception of the speed at which they learned. Though it is only an assumption that the Chinese speakers share a common cultural background (their nationalities were not asked for), the results would corroborate the extensive findings of the dichotomy between Western and Chinese language teaching [6,9]. In brief, whereas "the West" has tended

towards a communicative language-teaching system full of dialogue and role-play, China has remained faithful to drill-and-test methods, teaching meta-linguistics and words directly.

These sorts of cultural findings are of utmost import when considering target demographics. Depending on the prevalence of certain cultures in that demographic, supplemental material (reference lists and meta-linguistic write-ups, for example) could help to allay the concerns of new users until they become comfortable with the application's method of teaching.

Conclusions

Over the course of the Toronto experiments, four prevalent issues emerged that highlight the challenges of a speech-based CALL dialogue system that are different from those that either speech engineers and second language researchers have been willing to face. The dialogue system had to earn its expertise, as opposed to a teacher that is granted respect by virtue of an institution. The dialogue system has to not only teach *well* but also be well delineated. The dialogue system cannot rely on the user to act in his best interest to learn, instead of acting to subvert. Finally, the dialogue system must be sensitive to the cultural background of its users in order to maintain their interests.

Each of these issues supports the notion that user experience and perception are paramount in CALL applications. Unlike a classroom, wherein a student accepts and adheres to the implicit agreement that if he follows instructions then he will learn, an application must continuously convince, bribe, cajole, and punish the user into believing that there even is an agreement. In other words, the user does not grant the same social and institutional affordances to the application that she would a teacher. The user's constant awareness that he is interacting with a computer makes designing speech-based CALL dialogue systems both very challenging and very relevant to HCI. For example, an application that is too strict on users' pronunciation combined with their inability to perceive errors may cause them to question the efficacy of the system, undermining its authority. Alternatively, a system too lenient on pronunciation will cause frustration later when speech systems designed to process "clean" speech are faced with poor pronunciation. This is just one of the many balances that must be found when developing CALL applications with speech.

Acknowledgements

XXX

References

[1] von Ahn, L. Duolingo: Learn a Language for Free While Helping to Translate the Web. *Proceedings of the 2013 International Conference on Intelligent User Interfaces*, ACM (2013), 1–2.

[2] Bernstein, J., Najmi, A., and Ehsani, F. Subarashii: Encounters in Japanese spoken language education. *CALICO journal 16*, 3 (1999), 361–384.

[3] Birdsong, D. Nativelike pronunciation among late learners of French as a second language. In *Language experience in second language speech learning: in honor of James Emil Flege*. John Benjamins Publishing, 2007, 406.

[4] Bratt, H., Neumeyer, L., Shriberg, E., and Franco, H. Collection and detailed transcription of a

speech database for development of language learning technologies. *ICSLP*, (1998), 1539–1542.

[5] Breitkreutz, J., Derwing, T.M., and Rossiter, M.J. Pronunciation teaching practices in Canada. *TESL Canada Journal 19*, 1 (2009), 51–61.

[6] Burnaby, B. and Sun, Y. Chinese Teachers' Views of Western Language Teaching: Context Informs Paradigms. *TESOL Quarterly 23*, 2 (1989), pp. 219–238.

[7] Cucchiarini, C., Strik, H., and Boves, L. Automatic evaluation of Dutch pronunciation by using speech recognition technology. *Automatic Speech Recognition and Understanding, 1997. Proceedings., 1997 IEEE Workshop on*, (1997), 622–629.

[8] Edge, D., Searle, E., Chiu, K., Zhao, J., and Landay, J.A. MicroMandarin: Mobile Language Learning in Context. *Conference on Human Factors in Computing Systems (CHI)*, ACM (2011), 3169–3178.

[9] Hu, G. Potential Cultural Resistance to Pedagogical Imports: The Case of Communicative Language Teaching in China. *Language, Culture and Curriculum 15*, 2 (2002), 93–105.

[10] Kawai, G. and Hirose, K. A method for measuring the intelligibility and nonnativeness of phone quality in foreign language pronunciation training. *ICSLP*, (1998).

[11] Moustroufas, N. and Digalakis, V. Automatic pronunciation evaluation of foreign speakers using unknown text. *Computer Speech & Language 21*, 1 (2007), 219–230.

[12] Munro, M.J. and Derwing, T.M. Foreign accent, comprehensibility, and intelligibility in the speech of second language learners. *Language Learning 45*, 1 (1995), 73–97.

[13] Pearson, J., Hu, J., Branigan, H.P., Pickering, M.J., and Nass, C.I. Adaptive language behavior in HCI: how expectations and beliefs about a system affect users' word choice. *Proceedings of the SIGCHI conference on Human Factors in computing systems*, ACM (2006), 1177–1180.

[14] Savignon, S.J. Communicative Language Teaching. *Theory into Practice 26*, 4 (1987), pp. 235–242.

[15] Strik, H., van Doremalen, J., Colpaert, J., and Cucchiarini, C. Development and Integration of Speech Technology into COurseware for Language Learning: The DISCO Project. In P. Spyns and J. Odijk, eds., *Essential Speech and Language Technology for Dutch*. Springer Berlin Heidelberg, 2013, 323–338.

[16] Wik, P. and Hjalmarsson, A. Embodied conversational agents in computer assisted language learning. *Speech Communication 51*, 10 (2009), 1024–1037.

[17] Witt, S.M. and Young, S.J. Phone-level pronunciation scoring and assessment for interactive language learning. *Speech communication 30*, 2 (2000), 95–108.

[18] *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge University Press, 2001.