Improving Literacy in Developing Countries Using Speech Recognition-Supported Games on Mobile Devices

Anuj Kumar[1], Pooja Reddy[2], Anuj Tewari[3], Rajat Agrawal[1], Matthew Kam[1]
[1]Human-Computer Interaction Institute, Carnegie Mellon University, Pittsburgh, PA, USA, 15213
[2]American Institutes for Research, Washington D.C., USA, 20007
[3]Computer Science Division and Berkeley Institute of Design, University of California, Berkeley, USA
anujk1@cs.cmu.edu

ABSTRACT

Learning to read in a second language is challenging, but highly rewarding. For low-income children in developing countries, this task can be significantly more challenging because of lack of access to high-quality schooling, but can potentially improve economic prospects at the same time. A synthesis of research findings suggests that practicing recalling and vocalizing words for expressing an intended meaning could improve word reading skills – including reading in a second language – more than silent recognition of what the given words mean. Unfortunately, many language learning software do not support this instructional approach, owing to the technical challenges of incorporating speech recognition support to check that the learner is vocalizing the correct word. In this paper, we present results from a usability test and two subsequent experiments that explore the use of two speech recognition-enabled mobile games to help rural children in India read words with understanding. Through a working speech recognition prototype, we discuss two major contributions of this work: first, we give empirical evidence that shows the extent to which productive training (i.e. vocalizing words) is superior to receptive vocabulary training, and discuss the use of scaffolding hints to "unpack" factors in the learner's linguistic knowledge that may impact reading. Second, we discuss what our results suggest for future research in HCI.

Author Keywords

Educational games; developing countries; information and communication technology and development (ICTD); literacy; mobile learning; speech recognition

ACM Classification Keywords

H.5.m [Information Interfaces and Presentation]: Miscellaneous;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

CHI'12, May 5–10, 2012, Austin, Texas, USA.
Copyright 2012 ACM 978-1-4503-1015-4/12/05...$10.00.

General Terms

Design; Human Factors

INTRODUCTION

More than half of the world speaks at least two languages [9], to the extent that "bilingualism and multilingualism are a normal and unremarkable necessity to everyday life for the majority of the world's population" [12]. Multilingualism arose because of reasons that include colonialism, diglossia, and efforts to promote national identity [12, 32]. Consequently, many children in both industrialized and developing nations grow up having been exposed to, and even learning to read in, at least two languages, neither of which may be spoken at home. However, most languages in multilingual societies do not share equal status; a "global" or "national" language (e.g., English, French, Mandarin, or Spanish) may co-exist with the vernacular languages, with the former privileged as the medium of instruction or official language of business.

Sadly, many children in the developing world stand to lose significant opportunities in life when their schools struggle to provide high-quality literacy instruction, sometimes for a vernacular language and more often in a second language (e.g., English), which teachers themselves lack proficiency in. The British Council estimates wage differences between salaried professionals with and without English skills to be between 20% and 30% in Bangladesh, Cameroon, Nigeria, Pakistan and Rwanda [32]. In postcolonial Morocco, where French is the official language of business, switching the medium of instruction from French to [the local language] Arabic is associated with a 50% reduction in the economic returns to schooling [2]. In India, from surveys with poor parents [27], English is one of the two most sought-after skills. In the city of Mumbai in India, for instance, "schooling in [the local language] Marathi channels the child into working class jobs, while more expensive English education significantly increases the likelihood of obtaining a coveted white-collar job" [24], such that English speakers experience returns on investment in schooling between 24% and 27%, while the returns for non- speakers are 10%.

Word reading skills, which includes vocabulary knowledge, is a significant predictor of the ability to read and comprehend longer passages of texts [30]. Studies have shown successes on the cellphone [18] and desktop computer [29] in developing countries with receptive vocabulary training (i.e., recognizing the meaning of a word when the reader sees it). Drawing on the research literature that we will discuss below, productive vocabulary practice (i.e., recalling the word for expressing a meaning and vocalizing aloud) is likely to bring about higher gains in vocabulary knowledge. We therefore hypothesize that productive vocabulary training, which language learning software applications can support via speech recognition (which checks that the learner is saying the correct word), can yield stronger literacy gains. While speech recognition remains a computationally difficult problem, speech in the niche domain of vocabulary learning is significantly more tractable when recognition is isolated to individual words (i.e., no need to locate boundary words, when recognizing phrases and longer sequences of words in the speech signal) and small vocabularies (which is usually feasible in practice because the words to be taught can be organized into short vocabulary lists in most situations).

This paper therefore proceeds as follows: first, based on the literature in second language learning and the psychology of reading, we motivate how speech recognition can be applied to improve word reading. We describe two games which we designed to support both receptive and productive practice (via voice commands), so that we can compare the learning gains across both conditions. We describe how we developed a speech recognizer that supports both games, and usability testing with rural children in India to ensure speech recognition accuracy. We report an experiment that compares receptive against productive training. Based on observations about learner difficulties in this experiment, we redesigned the games to incorporate two hints that aim to scaffold learners in the productive training. We report a second experiment that compares the learning benefits the hints confer. We conclude with implications for designing vocabulary training applications.

Our contributions are twofold. First, through a working prototype, we give empirical evidence that shows the extent to which productive training is superior to receptive vocabulary training. More importantly, we compare two scaffolding techniques to “unpack” factors in the learner’s linguistic knowledge subsystems that may impact word reading. Second, we discuss what our results suggest for future research directions in HCI.

MOTIVATION AND RELATED WORK

Theoretical Framework

“Word reading,” which refers to the reader’s ability to read *and* understand individual words [31], is fundamental for enabling her to comprehend the sentences and texts that she encounters [30]. Much of the research in the psychology of reading supports the position that word reading involves two sub-skills [31]: mapping orthographic units (e.g. individual or clusters of letters) onto its appropriate phonological unit, so that she can *decode* (or “sound out”) the written word. For instance, for the written word “cat”, the letter ‘c’ must be mapped to the sound /k/, ‘a’ to /a/, and ‘t’ to /t/.; and associating the phonological form (i.e., the sound /cat/) and its contextually-appropriate semantics (i.e., small, furry animal), so that *semantic extraction*, or the extraction of the word’s meaning from its sound, occurs.

According to the lexical quality hypothesis (LQH) [31], as illustrated in Figure 1, word reading involves the reader integrating her knowledge of three linguistic subsystems, namely: orthography (visual script), phonology (sound) and semantics (meaning). The LQH implies that the reader’s word reading skill can be enhanced if the linkage between her orthographic and phonological subsystem, or linkage between her phonological and semantics subsystem, or the quality of her orthographic or phonological subsystems themselves, are strengthened [31]. These improvements are especially challenging for the reader to achieve for her knowledge subsystems in the case of a second language.

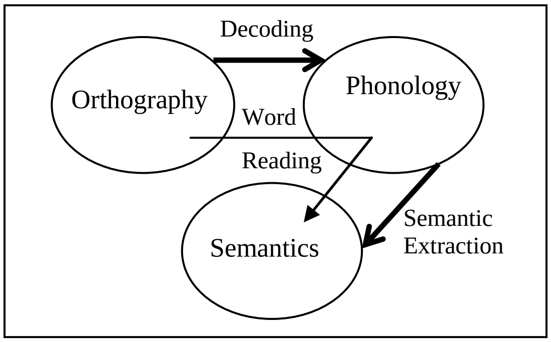

Figure 1: The three components of word reading knowledge (i.e., orthography, phonology and semantics).

To address the challenge of strengthening the subsystems and/or linkages between them in the learner’s knowledge, we turn to second language acquisition research. On top of recommendations to include vocabulary drills in language teaching [26], which can include both receptive vocabulary practice (i.e., the ability to recognize the meaning of a word when the reader sees it) and productive vocabulary practice (i.e., recalling the word for conveying a meaning and saying it aloud), the literature argues that speaking aloud provides the learner with phonological input back to the mind, which enables her to strengthen her word knowledge [9]. Similarly, vocalization helps her realize that she either lacks this knowledge, thus prompting her to acquire it, or that she has this knowledge, thus enabling her to further consolidate it [9, 36]. As such, we hypothesize that speech recognition can be used to facilitate productive vocabulary practice (i.e., check that she is speaking aloud the correct word), so as to strengthen her word subsystems, and thereby, ability to read with understanding.

Speech Technologies for Low Resource Contexts

Past efforts in human-computer interaction for development have explored speech as an input modality, especially as an alternative to graphical user interfaces. For instance, Patel et al. [28] investigates a voice forum for low-literate rural farmers to access information. Similarly, Agarwal et al. [1] describes "Spoken Web", a framework to help low-literate users access and create voice sites on the Web, similar to how literate users currently access the (text-based) World Wide Web. However, none of their user studies involves a working version of an automatic speech recognizer (ASR).

In attempts to incorporate ASRs into low-resource contexts, prior efforts have focused on accuracy improvements. By "low-resource," we mean contexts that lack the off-the-shelf databases for the acoustic model, pronunciation dictionary, grammar, etc. that are necessary for building accurate recognizers. The Meraka Institute [5] is currently exploring the technical feasibility of building an ASR for 11 South African languages, but to our knowledge, has yet to deploy a working ASR. Where working ASRs for low-resource contexts are concerned, one of the earliest successes came from UC Berkeley's TIER group for their Tamil Market project in rural India [33], which showed that it is feasible to build an accurate recognizer for a small number of users. But their approach does not scale beyond tiny vocabulary sizes. Healthline's "poor man's recognizer ++" [35] is one of the most generalizable approaches for building a speech recognizer for languages that lack off-the-shelf resources. However, besides accent, it does not consider other factors (e.g., background noise, pitch in children's voices [20], etc.) that can severely impact speech recognition accuracy.

Speech Technologies for Language Learning

So far, most language and literacy learning software in the developing and industrialized world do not employ speech recognition when teaching vocabulary and language skills, and as such, is restricted to receptive language knowledge. Research projects in this category that target learners in the developing world include MILLEE [15, 16, 18], Multiple Mice [29], Kane [10], and Same Language Subtitling [3].

Among the research projects that use speech recognition to improve language learning, most of them aim to improve pronunciation (vs. vocabulary), and none of them examines the impact that productive vocabulary training has on word reading. For instance, MIT's literacy tutor [21] targets pronunciation skills by providing interactive feedback for poorly articulated or mispronounced words. CMU's project LISTEN [23] is a reading tutor that has been adapted for children in developing countries. It presents a sentence at a time and provides feedback only at the end of a sentence unless the reader is stuck or clicks for help. The University of Colorado's Foundation to Literacy Program [8] started a reading program that was designed for beginning and poor readers. It targets foundational reading skills such as letter knowledge, phonological awareness and decoding skills in order to improve listening and reading comprehension. In addition, none of the above projects target mobile devices in developing regions, which are increasingly prevalent and less expensive than the desktop computers, which the above projects are based on. An exception is Tewari et al. [37], which is a pronunciation tutor on smartphones that aims to help Hispanic children in the USA acquire English skills.

There are also a number of commercial projects: Rosetta Stone [34] uses proprietary speech recognition technology to elicit oral responses from learners. They have tested and fine-tuned accuracy, both for native and non-native speakers. Similarly, Carnegie Speech [6] teaches speaking and listening skills using proprietary speech recognition technology and models. While both ventures have shown great potential for speech technologies in language learning, their components for building a speech recognizer (e.g., acoustic model, dictionary, grammar) were built from years of automated data collection and analysis. This approach requires financial resources that smaller ventures, research groups and even well-funded corporations trying to cover a wide spectrum of user groups, such as diverse contexts in the developing world, cannot attempt.

GAME DESIGNS

Building on the growing popularity of educational games [13] and successes with games to teach vocabulary in a second language in the developing world [15,18,29], we designed two educational games: Market Game and Farm Game. At the game design level, these games drew on a previous study on the characteristics of traditional village games that rural children in a developing country such as India find to be intuitive (compared to the features found in contemporary Western videogames), and the systematic differences between traditional and Western videogames [16]. These games incorporated actions such as the ability to either *catch* or *evade* a player, both of which were two popular game actions in traditional village games. As such, both games were designed to be culturally appropriate for children in rural regions in India. Each game followed a teaching, game play, and practice sequence. This sequence was informed and refined through earlier usability testing in the field [17]. Our prior research has not, however, involved educational games that support speech recognition.

Market Game

In the Market Game, the teaching phase (as shown in Figure 2A) entailed introducing the vocabulary words to the user. As in popular commercial software [34], introduction of a word meant mapping the image of the vocabulary word to its sound, as well as, its native language and English text. At a time, the software introduced five words (one after another), which was based on prior psychological studies that show that an average person can retain 7±2 new items in their short-term working memory at any time [22]. Once the word has been introduced, the user could replay the audio playback for its pronunciation at the teaching phase for any number of times, before going to the next screen.

Next, in the game play screen (Figure 2B) of the market game, the aim was to move the boy character from left (home) to right (shop in the market), while avoiding monkeys en-route. This activity was an adaptation of the daily routine of going to the market for children in rural India; and integrated actions from the popular physical games that they play, such as evading an opponent. Next, depending on the experimental condition that the player was assigned to, the player could purchase items from the shop by either *selecting* the correct item that corresponded to the said word (Figure 2C, Re condition), or by *saying the word aloud* that corresponded to the image displayed (Figure 2D, Pr condition).

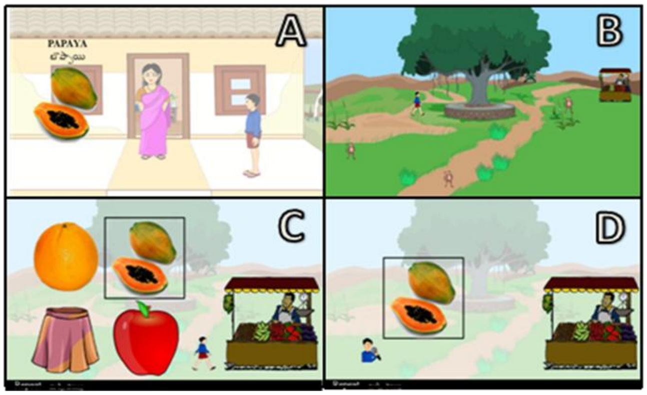

Figure 2: These screenshots describe the game sequence for the market game. (A) *Teaching phase*: Introduces the English word for the common nouns, which are the items to be bought in the market. (B) *Game play*: shows the boy attempting to move towards the shop, while the monkeys try to catch him. (C) *Receptive practice condition*: to purchase the item from the shop, the user has to map the word that the software plays aloud with its image from a list of four choices. (D) *Productive practice condition*: to purchase the correct item at the shop, the user has to say aloud the word that corresponds to the image.

Farm Game

The Farm Game (Figure 3) had the same pedagogical cycle of "teaching, game play, and practice", but used a different game action than Market game. The objective of the farm game was to save the farm by "catching" all the thieves and retrieving the items that they had stolen. Again, this activity was based on the common rural India scenario where children helped their parents keep vigil on the farm in the farming season, where a lot of animals or thieves tend to steal the crop. In order to recover an item from the thief, the user could do it in one of the two ways as described above for the Market Game in Figure 2C (receptive practice) or Figure 2D (productive practice), depending on the condition that the participant was randomly assigned to.

SYSTEM DEVELOPMENT AND USABILITY TESTING

Our system (consisting of the above games and the speech recognizer) was iteratively prototyped and refined on the Nokia N810 mobile phone. In May-June 2010, for a period of 4 weeks, we conducted a series of usability tests in the field with our working recognizer to improve both the game designs and speech recognition accuracy. In total, 10 children in grades 4-5 in Hyderabad play-tested successive versions of both games. In addition, we collected speech samples from 50 children to improve recognition accuracy.

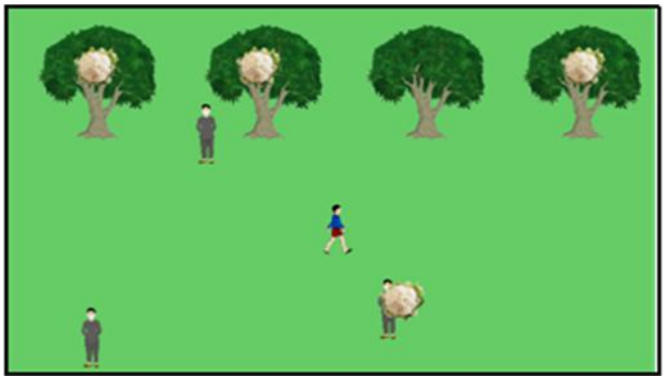

Figure 3: This screenshot shows the *game play* screen for the farm game, where the boy character attempts to catch the thief who is stealing the vegetable from the farm.

Interface Improvements

One of the important requirements for our speech user-interface (SUI) was to signal to the user when to speak. Typically, in SUIs indicate this by displaying a microphone icon when it is time to speak. However, from the initial usability tests, we realized that earlier icons for prompting the user to speak (Figure 4) were unintuitive to participants since most of them were never introduced to the notion of a microphone. Instead of showing the microphone on its own, they liked the idea of showing a boy holding a microphone, because they were better able to relate to instances of local elections or rallies where people would hold a microphone and speak into it.

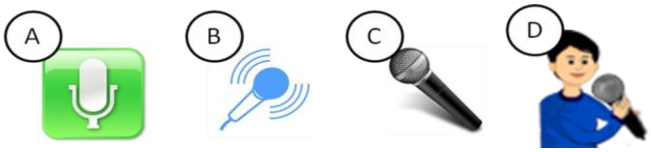

Figure 4: Microphone icons tested in the usability tests; A, B, C represent icons that are currently used in existing speech user-interfaces; D represents the icon that was understood by participants of our study.

Speech Recognition Improvements

To recognize user's speech in the conditions that required productive practice (with or without hints), we used pocketSphinx [14] – an open-source, small footprint, automatic speech recognizer (ASR) for mobile devices. Our choice of pocketSphinx was based on experimental evaluation of mobile speech recognizers [19], where pocketSphinx outperformed other mobile ASRs such as TinySphinx on small vocabulary tasks. However, due to the multiple challenges of speech recognition for non-typical contexts such as ours, off-the-shelf resources were not directly applicable. To account for numerous acoustic variabilities arising due to varying accents, background noise etc. we followed a three-step process to adapt and fine-tune the accuracy of the recognizer. First, we collected a speech corpus from representative speakers and trained our own baseline acoustic model. Second, we analyzed this corpus for factors that affected the speech recognition accuracy, and third, we adapted the baseline speech recognizer on these factors to improve its accuracy. Below we describe these three steps in more detail.

Our speech corpus was made up of 6250 utterances (approximately 6 hours) from 50 rural Indian children, equally divided across gender and grades 4-5. Each utterance in the dataset was labeled with the word that it represented at the time of recording, and these labeled inputs were used for training the speech recognizer. As in the original pocketSphinx paper [14], our baseline acoustic model used Hidden Markov Models with a 5-state Bakis topology. They were trained on the 6250 utterances, using 1000 tied Gaussian Mixture Models (senones) and 256 tied Gaussian densities. Since the application required the user to speak words in isolation (and there's no need to capture history), we used a unigram statistical language model with a vocabulary size of 25 words.

To improve recognition accuracy beyond the baseline performance of the above acoustic and language models, we next sought to adapt our recognizer to account for the variabilities in users' speech. Since adaptation can happen along several dimensions, we first quantitatively analyzed the speech utterances along seven voice metrics that collectively describe the characteristics of voice, namely articulation (a), speaking rate (r), sound quality (q), pitch (f0), and the first three formant frequencies (f1, f2, f3). Using univariate and multivariate analyses (the description of which is beyond the scope of this paper), the influence of each metric was calculated on the word error rate (WER), which appropriately measures the recognizer's accuracy in isolated word speech recognition tasks, such as ours. Such an analysis was necessary to identify the most critical factors and save time on performing all possible adaptation techniques.

As shown in Figure 5, speaking rate and articulation were the most significant factors that impeded recognition performance in the context of our users. Based on this analysis, to account for variations in the speaking rate, we adapted the baseline speech recognizer using a widely used continuous frame rate adaptation technique [7]. Similarly, to account for non-native articulations, using the technique described in "poor man's recognizer" [35], we added pronunciation variants to the dictionary that were a closer match to the accent of the users than the ones in the baseline dictionary.

In addition to the above quantitative analysis, we also qualitatively explored the recognition results of the incorrectly recognized words from our usability tests. We observed that in some cases of misrecognized words, the results had extra words – for instance, "papaya" was decoded as "peas papaya cup", perhaps, because background noise from cars or nearby mosques was also decoded in addition to the actual speech from the user. To overcome this, we considered the output of the ASR to be correct if the target vocabulary keyword was spotted anywhere in the decoded sentence. In addition to keyword spotting, we also used noise-canceling microphones to minimize misrecognitions due to external noise.

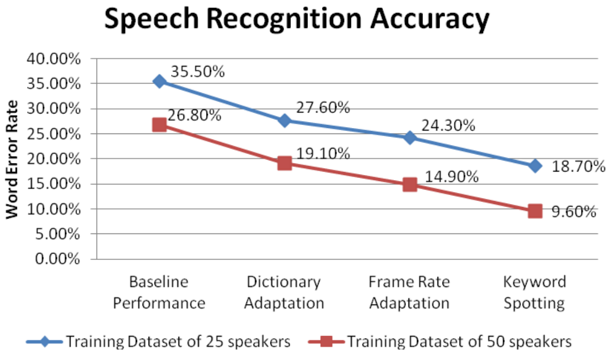

Figure 5: Summary of progressive performance improvements when the amount of training data was increased, and with the type of adaptation technique used.

After a series of context-based adaptations such as those arising due to non-native accent and noise (as outlined above), the speech recognizer's accuracy was 91.4% (Figure 5). In other words, approximately, 1 out of 10 times the speech recognizer misunderstood what the user spoke. In cases when the user happened to speak the correct answer (but the recognizer misunderstood it to be incorrect), the participants learned to repeat the word from the short, 15-minute training sessions prior to actual gameplay (see *Experiment* sub-section of Study 1). In general, these training sessions had a dual role: first, to acquaint the participants with the game dynamics using five curriculum words, and second, to help them realize that the system might make recognition mistakes. In order to proceed in the latter case, they could either change their response, or if they were confident, then repeat it. From our observations, most of the times when there was a false negative (i.e. user spoke correctly but speech recognizer misrecognized), learners repeated the vocabulary word rather than switching their response, and thus system's misrecognition is less likely to have impacted any learning gains.

STUDY 1: RECEPTIVE VS. PRODUCTIVE PRACTICE

Informed by the above theoretical framework, we have the following hypothesis for this study:

H1: Productive training[*] (Pr) is more beneficial for word reading than receptive training[*] (Re),

[*] With no hints (as will be introduced in the second study).

Participants

21 participants (11 boys and 10 girls) took part in our first study in June-July 2010 (5 weeks). They were 9 to 13 years old (mean=10.5 years) in 4th and 5th grades. We selected children in this age and grade level because word reading skills are increasingly important for reading comprehension and academic success at this level [4]. All participants were attending a public school in a rural area near Hyderabad, India. Telugu and English were taught as their first and second language respectively at this school. Most of their families owned at least one cellphone.

Experiment

The experiment involved a pre-post test block design, with the intervention comprising the above two market and farm games. To ensure relevance of the curriculum in the games, each game targets 10 unique words (all concrete nouns). The 20 words were chosen from government-issued English textbooks for 4th and 5th graders. To avoid introducing a confounding variable from our selection of these words, we consulted teachers at the school to verify that they do not plan to teach these words in their classes at the time that the experiment was planned. Each child played both games, but was randomly assigned to one of the two conditions: Re or Pr. With this assignment, when the 21 participants played the market game, 10 were in the Pr condition and 11 were in Re; and when they played the farm game, 11 were in Pr and 10 were in Re. Each participant played the games in an hour-long, after-school session comprising a word reading pre-test, training session, 30 minutes of gameplay, and a similar word reading post-test immediately after gameplay. Specifically, in the pre- and post-tests, the outcome variable (word reading score) was measured before and after each game using an adapted visual format from the Peabody Picture Vocabulary Test [11], which is used for measuring vocabulary knowledge. In our word reading test, every word was displayed on the screen with four images, and participants were asked to match the word with the picture that best represented its semantics. A correct match adds one to the participant's score, while an incorrect match or the lack of a response does not change the score. In a training session before the study, experimenters explained the games and allowed each participant to play the games with 5 practice words.

Results

The means and standard deviations for pre- and post- word reading scores, and post-test gains, are shown in Table 1 for both the Re and Pr conditions. As a sanity check, the t-tests did not reveal any significant differences in post-test gains between the market and farm games, for both the Re (p=.12) and Pr (p=.48) conditions; we expected both games to exhibit comparable post-test gains since they adopted similar designs and instructional principles. Likewise, between both games, there was no statistically significant difference on pre-test scores (p=.24) for both conditions, indicating that participants in both conditions started with the same baseline knowledge of word reading. This allowed us to focus on analyzing word reading gains without using pre-test scores as a covariate.

Variable	Condition	Market Game	Farm Game
Pre-Test	Re	7.8 (3.5)	6.9 (2.5)
	Pr	8.4 (2.8)	8.9 (3.0)
Post-Test	Re	8.5 (3.7)	8.3 (2.2)
	Pr	10.6 (1.5)	12.1 (3.0)
Gain	Re	.6 (1.1)	1.4 (2.3)
	Pr	2.3 (1.9)	3.2 (1.8)

Table 1: Means and standard deviations (in parentheses) for pre-test score, post-test score, and post-test gains for both games, for each of the two conditions (i.e. Re and Pr).

Given the above sanity checks and in order to perform a more robust statistical analysis, for each condition, we combined the scores across both games. On this combined dataset, post-hoc power analysis indicated a power of 0.62 (with cohen's d as 0.9 and significance level as 0.05), which was deemed adequate for inferential analysis for our purposes. A one-way 2-factor ANOVA showed a significant difference between the two conditions on post-test gains (F(1,40)=5.4, p=.001). More specifically, after 30 minutes of game play, gains were observed under both conditions: 1.0 word under Re, against 2.7 words under Pr. These results suggest that with even as little as 30 minutes of game play practicing words – receptively or productively – word reading scores increase significantly. Even more critically, there is evidence that productive training is significantly more beneficial for word reading development than simply receptively practicing words.

STUDY 2: SCAFFOLDING WORD PRODUCTION

Our first experiment has shown that productive vocabulary training, enabled by speech recognition that checks that the learner is recalling and articulating the correct words, leads to stronger gains on word reading scores (as measured for short-term retention), in comparison to receptive training. Given that word reading consists of both decoding and semantic extraction, however, it is not clear to what extent the benefit from productive training is shared between these two constituent sub-skills. Similarly, it is not clear if these gains would persist over a longer time after gameplay, i.e., long-term retention.

Most importantly, in the Pr condition, we observed at least 12 (out of 21) instances in which learners struggled to give their answers. For instance, they appeared to exert great effort to recall and vocalize a word. Therefore, to "unpack" the subsystems in the learner's linguistic knowledge which facilitates word reading via productive training, we added both an orthographic and phonological hint to both games (described in more detail below). We expected the hints to lower the cognitive load on selected subsystems in her word knowledge, so that she receives support to strengthen other subsystems in her word knowledge that are also responsible for word reading. Of course, it is not clear, and we need to learn, if the hints are counter-productive (e.g., she may not wean herself from being over-dependent on them).

We set out to address the above gaps in our second study. Informed by the lexical quality hypothesis, our hypotheses for study 2 were therefore:

H2: Productive training with an orthographic hint (Pr + Or) is more beneficial for decoding than productive training with no hints (Pr),

H3: Productive training with a phonological hint (Pr + Ph) is more beneficial for semantic extraction than productive training with no hints (Pr),

H4: Productive training with both a phonological and orthographic hint together (Pr + Or + Ph) is more beneficial for word reading than productive training with no hints (Pr).

Games Redesign

More specifically, we redesigned both games to support additional conditions: (A) *Orthographic hint* (in form of the first alphabet of the word and the length of the written form of the word) to support cognitive search for the vocabulary item using letter-to-sound mapping rules. (B) *Phonological hint*, in form of the first sound (or phoneme) to prompt a targeted search for the word by narrowing down the (semantic) search space to only those words that begin with that sound, and (C) *Orthographic and Phonological hint*, to support cognitive search of the word using either or both the letter-to-sound rule and narrowed-down semantic search space. The scaffolding design behind both the orthographic and phonological hints was informed by the lexical quality hypothesis in our theoretical framework [31] that illustrates how the hint features work in the games, for the given word “papaya”. When the phonological hint was available, based on experiences from prior usability tests [18], we gave the participant the option to repeat the hint, in case she missed it or wanted to hear it again.

Participants

40 participants (27 boys and 13 girls) took part in our second study in May-June 2011 (6 weeks). They were 9 to 13 years old (mean=11 years) and were 4th and 5th graders. All participants were attending a different public school than the one in the first study in a rural district near Hyderabad, India. This gave us the chance to test our prototypes with a different user group in the same area, without making major changes to the speech recognizer. Telugu and English were taught as the first and second language respectively at this school. Most of their families owned at least one cellphone.

Experimental Design

The second study had four conditions: Pr, Pr + Or, Pr + Ph, and Pr + Or + Ph. The experimental design for the second study was identical to that of the first study, except for two additions. First, to assess long-term retention of the vocabulary words covered in the games, we administered a delayed post-test between 7 and 10 days after gameplay.[1] Second, besides scoring for word reading on the pre-, post- and delayed post-tests, we also scored each participant on decoding and semantic extraction for all three conditions. For semantic extraction, the scoring was similar to that of word reading (as in Study 1), except that instead of a written word, participant heard a word and then selected matching image. For decoding, participant was shown a written word, which s/he read aloud, after which administrator clicked "Yes" if correct (irrespective of how well it was pronounced), and "No" otherwise. This allowed us to separately study the impacts of each hint condition on each reading subsystem. This also allowed us to enhance the validity of our assessment instruments, i.e., since decoding and semantic extraction should theoretically total up as word reading, we could check for measurement errors in our test instruments.

Results

Similar to study one, as a sanity check, we do not find any significant difference between the two games on any test score, for all four conditions (namely: Pr, Pr + Or, Pr + Ph, and Pr + Or + Ph). As such, in order to perform a statistical analysis with greater inferential power, for each of the four conditions, we combined the scores from both games so as to double our effective sample size.

Word Reading Gains

A one-way 4-factor ANOVA showed a significant difference between the four conditions on word reading gains as measured using the immediate post-test (*F*(3,76)=2.98, *p*=.03). Our independent sample t-tests with Bonferroni correction showed a significant difference between Pr and Pr + Or (*p*=.02), Pr and Pr + Ph (*p* =.04), Pr and Pr + Or + Ph (*p*=.05). There was no significant difference in word reading gains across other combinations of hints. [2]

[1] We aimed to conduct the delayed post-test on the 8th day after gameplay. Given relatively high absenteeism rate in some schools in the developing world (such as this school), however, for 12 students, we were only able to schedule the post-test on the 9th or 10th day. This minor variation (i.e., up to 2 days) after the 7-day lag is not expected to impact delayed post-test performance significantly. We were successful in conducting this test with all 40 participants.

[2] For brevity, when discussing results for the second experiment, we exclude combinations that were not significantly different i.e. *p* > 0.05 (except for a few cases in which significance changed from the immediate to the delayed post-test).

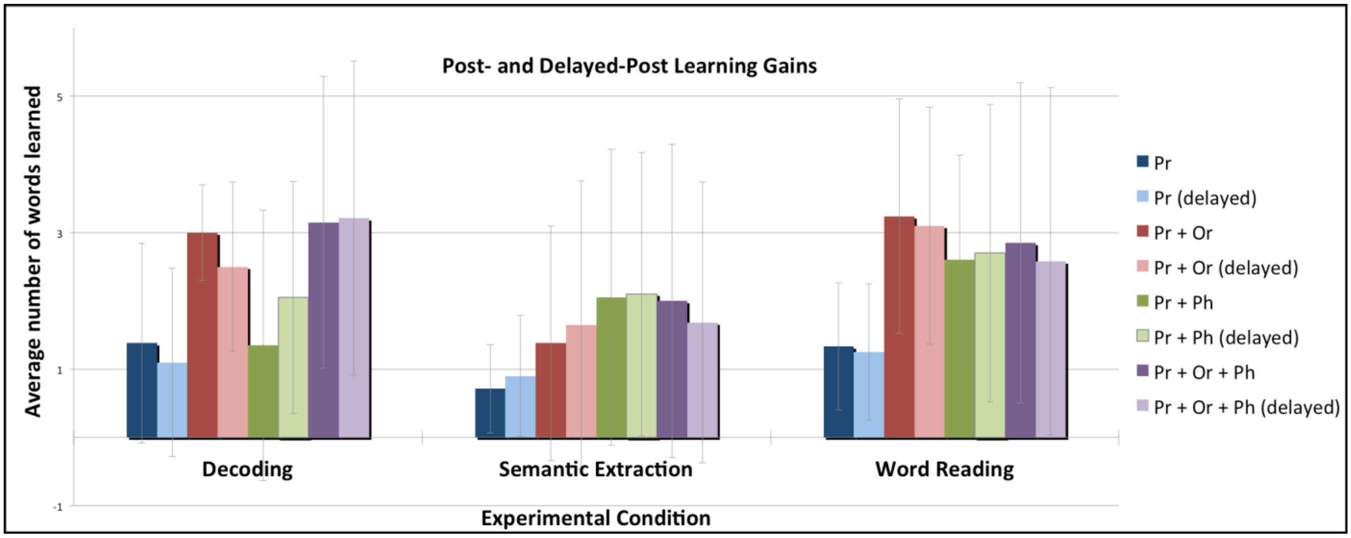

Figure 6: This figure shows the average post- and delayed-post test gains for the three dependent variables in the second study i.e. decoding, semantic extraction, and word reading. 'Pr' refers to productive practice (with no hint), 'Pr + Or' refers to productive practice with orthographic hint, 'Pr + Ph' refers to productive practice with phonological hint, and 'Pr + Or + Ph' refers to the productive practice with both orthographic and phonological hints.

For the delayed post-tests, a one-way 4-factor ANOVA revealed a significant difference between the four conditions for gains on word reading scores (*F* (3,76)=2.28, *p*=.05). Independent sample t-tests with Bonferroni correction showed a significant difference between Pr and Pr + Or (*p*=.03), and Pr and Pr + Ph (*p*=.05). But the difference between Pr and Pr + Or + Ph (*p*=.11) was not significant.

Decoding Gains

For the non-delayed post-test, a one-way 4-factor ANOVA revealed a significant difference on the decoding gains across the four conditions (*F*(3,76)=4.87, *p*=.003). Independent sample t-tests with Bonferroni correction showed significant differences between Pr and Pr + Or (*p*=.02), Pr and Pr + Or + Ph (*p*=.02), and Pr + Or and Pr + Ph (*p*<.01).

The delayed post-tests reflected similar results; the ANOVA revealed a significant difference on the decoding gains across the four conditions (*F*(3,76)=3.99, *p*=.01). Post-hoc analysis with Bonferroni corrected t-tests showed significant differences between Pr and Pr + Or (*p* = .04), Pr and Pr + Or + Ph (*p* = .01), and Pr + Or and Pr + Ph (*p* = .04)

Semantic Extraction Gains

For gains on the immediate post-test, a one-way 4-factor ANOVA for semantic extraction gains did not reveal any significant difference between the four conditions, F(3,76)=1.88, *p*=.13; however, post-hoc analysis with Bonferroni correction indicated a significant difference between (and only between) Pr and Pr + Ph (*p*=0.04). Similarly, where the delayed post-test was concerned, ANOVA analysis did not reveal a significant difference between the four conditions, F(3,76)=1.2, *p*=.31; but post-hoc analysis revealed significant difference between Pr and Pr + Ph (*p*=0.05) only.

DISCUSSION

Speech Recognition Support for Word Reading

The results from the second experiment support our three hypotheses, i.e. H2, H3 and H4. That is, when productive vocabulary practice is made possible by speech recognition, adding an orthographic hint facilitates significant decoding gains (but not semantic extraction), adding a phonological hint facilitates significant semantic extraction gains (but not decoding), and introducing both hints together (in the short-term only) and separately (in both the short- and long-term) facilitates significant gains on word reading. While these results are suggested by the lexical quality hypothesis, our work nevertheless contributes to our understanding of word reading in many ways. Firstly, the LQH has not been applied to inform productive vocabulary practice as an instructional approach. As such, we have provided stronger support for the LQH by showing that its implications continue to hold when applied to productive training.

Secondly, the results from our second experiment provides evidence that there are distinctive categories of productive vocabulary training, each of which is useful for a different knowledge subsystem within word reading. This implies that computing systems that employ speech recognition for productive vocabulary training need to be designed to allow for a more targeted approach to word reading training. Specifically, for the decoding subsystem to be strengthened, it is important to help learners connect their orthographic knowledge of a word with its meaning (e.g., through an orthographic hint). On the other hand, for semantic extraction, it is important instead to help learners link their phonological knowledge of the word with its meaning (e.g., through a phonological hint). More research in speech user-interface design could be done to experiment with more variations of the scaffolds (including hints) that could support each subsystem, so that we understand how to better improve overall word reading skill.

Implications for Design

Thirdly, thus far, no language learning software that contains speech recognition support has been studied in the context of improving literacy skills through productive vocabulary practice. While speech recognition systems have generally experienced limited adoption, our promising results suggest that this niche domain of speech recognition-based systems for productive vocabulary practice deserves more attention by speech technology experts, language learning specialists, as well as researchers and practitioners in human-computer interaction. At the least, this niche domain could accelerate the adoption of speech technologies in this specific context. We shall now outline some new areas for research in speech user-interfaces that this work opens.

A game that uses speech recognition to interpret the actions that a player issues through a verbal command is commonly known as a voice-command game [38]. Although we have not encountered such a game on the cellphone, they have been implemented on gaming devices such as the Nintendo DS and Playstation, and have appeared to be popular with players. For instance, the two most popular voice-command games on Nintendo DS called Nintendogs and Brain Age have sold 22.27 and 17.41 million copies, making them the third- and fifth-most popular game in the history of all Nintendo games by 2009 [25]. Existing voice-command games have focused purely on entertainment, not learning goals, and have not received much attention from academic researchers. As the first study to demonstrate the effects of voice-command games on literacy skills (or for that matter, learning), we hope to encourage more HCI and educational games researchers to experiment with more ways in which the interface and gameplay experience in voice-command games can be designed to better support educational goals.

Next, given the cost to manufacture visual displays, access to computing systems and services could be made more widespread through non-visual devices. Examples include IBM's Spoken Web [1] and speech-based Interactive Voice Response (IVR) systems. While IVRs have not received as much attention in the HCI community, and our games involve an automatic speech recognizer running locally on a mobile device, our results suggest that speech-based applications for productive vocabulary practice could be designed and delivered over an IVR system. More research is needed to understand how such vocabulary training can be delivered over a non-visual user-interface. Although it is possible to provide only phonological and not orthographic hints over a non-visual interface, we view this as an opportunity and not a limitation. As our results have shown, phonological hints can be used to improve semantic extraction skills, but not decoding skills. Literacy instructors and researchers believe that decoding skills are far easier to master than semantic extraction [26]. Through low-cost interfaces like IVRs, we can potentially increase the reach of literacy training among low-income learners in the developing world, where we have observed teachers and other instructional resources to be much better prepared to target decoding skills (but not semantic extraction).

Along the same line of affordability, educational platforms such as the Multiple Mice computer (that could be attached to and accept input from multiple mouse devices) have been designed to allow computing resources -- which are fairly expensive in low-resource communities – to be shared by more than one user at the same time. Speaker identification via speech recognition is potentially another way for many learners to share a computational device for productive vocabulary training. More research on designing such user-interfaces and digital learning experiences is needed.

ACKNOWLEDGMENTS

This research was funded by Nokia Research and CMU HCII. We thank Venkatesh Keri, Gaurav Arora and Yash Soni for their technical contributions; Indrani Vedula for her invaluable help with setting up both the studies, and Kaushik Shelat for his help in administering these studies.

REFERENCES

1. Agarwal, S., Jain, A., Kumar, A., Rajput, N. The World Wide Telecom Web Browser. In *Proc. ACM DEV 2010*.
2. Angrist, J., and Lavy, V. The Effect of a Change in Language of Instruction on the Returns to Schooling in Morocco. *Journal of Labor Economics, 15* (1), 1997.
3. Arora, P. Karaoke for Social and Cultural Change. *ICES (4)* 3, *Paper 1*. Troubador Publishing Ltd. 2006.
4. August, D., Carlo, M. S., Dressler, C., & Snow, C. (2005). The critical role of vocabulary development for English language learners. *Learning Disabilities Research & Practice, 20,* 50-57.
5. Barnard, E., Davel, M., and Huyssteen, G. Speech technology for information access: a South African case study. In *Proc. AAAI Symposium on Artificial Intelligence 2010*.
6. Carnegie Speech, http://www.carnegiespeech.com/
7. Chu, S., Povey, D. Speaking rate adaptation using continuous frame rate normalization. In *Proc. IEEE ICSSP*, March 2010.
8. Colorado Literacy Foundation – To Enhance and Enrich Literacy Throughout Colorado. http://www.coloradoliteracyfoundation.org
9. De Bot, K. The psycholinguistics of the Output Hypothesis. *Language Learning, 46,* 529-555, 1996.
10. Dias, M.B., Mills-Tettey, G.A., and Mertz, J. The TechBridgeWorld Initiative: Broadening Perspectives in

Computing Technology Education and Research. In *Proc. of the international symposium on Women and ICT: Creating Global Transformation*, ACM Press (2005).

11. Dunn, L. M., & Dunn, L. M. *Peabody Picture Vocabulary Test (3rd edition).* American Guidance Service, 1997.
12. Edwards, J. *Multilingualism*. Routeledge, 1994.
13. Gee, J.P. *What Video Games Have to Teach Us About Learning and Literacy*. Palgrave Macmillan, 2004.
14. Huggins-Daines, D., Kumar, M., Chan, A., Black, A., Ravishankar, M., and Rudnicky, A.I. PocketSphinx: a free, real-time continuous speech recognition system for handheld devices. In *Proc. ICASSP*, 185–188, May 2006.
15. Kam, M., Kumar, A., Jain, S., Mathur, A., and Canny, J. Improving Literacy in Rural India: Cellphone Games in an After-School Program. *In Proc. of IEEE/ACM Conference on Information and Communication Technology and Development* 2009.
16. Kam, M., Mathur, A., Kumar, A., and Canny, J. Designing Digital Games for Rural Children: A Study of Traditional Village Games in India. In *Proc. CHI 2009*.
17. Kam, M., Agarwal, A., Kumar, A., Lal, S., Mathur, A., Tewari, A., and Canny, J. Designing E-Learning Games for Rural Children in India: A Format for Balancing Learning with Fun. In *Proc. DIS 2008*.
18. Kam, M., Ramachandran, D., Devanathan, V., Tewari, A., and Canny, J. Localized Iterative Design for Language Learning in Underdeveloped Regions: The PACE Framework. In *Proc. CHI 2007*.
19. Kumar, A., Tewari, A., Horrigan, S., Kam, M., Metze, F., and Canny, J. Rethinking Speech Recognition on Mobile Devices. In *Proc. ACM Conf. on IUI 2011*.
20. Lee, S., Potamianos, A., Narayanan, S., Acoustics of children's speech: developmental changes of temporal and spectral parameters. *J. Acoust*. Soc. Amer. 105, 1999, 1455–1468.
21. McCandless, M. Word rejection for a literacy tutor. S.B. Thesis, MIT, May 1992.
22. Miller, G. A. (1956). The magical number seven, plus or minus two: some limits on our capacity for processing information. *Psychological Review* 63 (2): 81–97. doi:10.1037/h0043158.
23. Mills-Tettey, A., Mostow, J., Dias, M. B., Sweet, T. M., Belousov, S. M., Dias, M. F., & Gong, H. Improving Child Literacy in Africa: Experiments with an Automated Reading Tutor. *Prof. 3rd IEEE/ACM Conference on Information and Communication Technologies on Development*, 2009.
24. Munshi, K., and Rosenzweig, M. Traditional Institutions Meet the Modern World: Caste, Gender, and Schooling Choice in a Globalizing Economy. In *American Economic Review, 96* (4), 1225-1252, 2006.
25. Nintendo. Financial Results Briefing for the 69th Fiscal Term Ended March 2009. pp. 6, 2009. URL: http://www.nintendo.co.jp/ir/pdf/2009/090508e.pdf, last accessed: May 8, 2009
26. NLP. (2006). Developing Literacy in Second-Language Learners: Report of the National Literacy Panel on Language-Minority Children and Youth. In D. August,& T. Shanah (Eds.) Mahwah, NJ: Lawrence Erlbaum.
27. Pal, J., Lakshmanan, M., and Toyama, K. My Child Will Be Respected: Parental Perspectives on Computers in Rural India. In *Proc. of 2nd IEEE/ACM International Conference on Information and Communication Technologies and Development*, 2007.
28. Patel, N., Chittamuru, D., Jain, A., Dave, P., Parikh, T. Avaaj Otalo - A Field Study of an Interactive Voice Forum for Small Farmers in Rural India. In *Proc. CHI 2010*.
29. Pawar, U., Pal, J., Gupta, R., Toyama, K. Multiple Mice for Retention Tasks in Disadvantaged Schools. In *Proc. CHI 2007*.
30. Perfetti, C. A. Reading ability: Lexical quality to comprehension. *Scientific Studies of Reading, 11*, 357-383, 2007.
31. Perfetti, C. A., & Hart. L. The lexical quality hypothesis. In L. Verhoeven, C. Elbro & P. Reitsma (Eds.), *Precursors of Functional Literacy* vol. 11, 67–86. John Benjamins, 2001.
32. Pinon, R., and Haydon, J. The Benefits of the English Language for Individuals and Societies: Quantitative Indicators from Cameroon, Nigeria, Rwanda, Bangladesh and Pakistan. A custom report compiled by Euromonitor International for the British Council, 2010.
33. Plauche, M., Nallasamy, U., Pal, J., Wooters, C., and Ramachandran, D. Speech Recognition for Illiterate Access to Information and Technology. In *Proc. of International Conference on Information and Communications Technologies and Development*, 2006.
34. Rosetta Stone, Ltd. Rosetta Stone Language Learning. http://www.rosettastone.com/schools, accessed Sept. 22, 2011.
35. Sherwani, J., Ali, N., Mirza, S., Fatma, A., Memon, Y., Karim, M., Tongia, R., Rosenfeld, R. HealthLine: Speech-based Access to Health Information by Low-literate Users. In *Proc. IEEE/ACM Int'l Conference on Information and Communication Technologies and Development*, 2007.
36. Swain, M., & Lapkin, S. (1995). Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied Linguistics*.
37. Tewari, A., Goyal, N., Chan, M., Yau, T., Canny, J., Schroeder, U. SPRING: Speech and Pronunciation Improvement through Games, for Hispanic children. In *Proc. IEEE/ACM International Conference on Information and Communication Technologies and Development*, 2010.
38. Zyda, M., Thukral, D., Ferrans, J., Engelsma & Hans, M. Enabling a Voice Modality in Mobile Games through Voice XML. *Proc. of ACM SIGGRAPH Symposium on Video Games*, 143-147, 2008.