Facilitating Development of Pragmatic Competence through a Voice-driven Video Learning Interface

Gabriel Culbertson[1], **Solace Shen**[1], **Malte Jung**[1], **and Erik Andersen**[2]

[1]Department of Information Science, [2]Department of Computer Science, Cornell University
{grc74, solace.shen, mfj28, ela63 }@cornell.edu

ABSTRACT

Authentic foreign language videos are effective for developing pragmatic competence, or sensitivity to meanings expressed by tone and word choice, and the ability to effectively express these meanings. However, established methods for learning from foreign language videos are primarily text-based (e.g. captioning). Using text, learners do not practice aspects of oral performance (e.g. intonation, pausing, and pitch) that are important to pragmatic competence. In this paper we present a voice-driven system where learners practice and learn a foreign language by repeating phrases out loud from any video. Utterances are transcribed and translated and, if captions are available, the system indicates the correctness of the utterance. In an evaluation with 27 participants, we show that participants more frequently used the voice-driven system than a comparison text-based system. Furthermore, in a field study of 130 independent learners, we show potential for community-driven resource collection.

Author Keywords

language learning, pragmatic competence, video learning, speech-recognition

ACM Classification Keywords

K.3.0. Computers and Education: General

INTRODUCTION

Mastering a foreign language requires pragmatic competence, a sensitivity to meanings expressed by tone and word choice, and the ability to effectively express these meanings [18]. Textbooks rarely offer pragmatic input and gaining pragmatic competence through traditional classroom activities is difficult because "classroom discourse is highly conventionalized in ways that severely constrain both the quantity and the quality of learners' participation" [6].

Foreign language videos are an underutilized source of pragmatic input. Sites like YouTube make foreign language videos available that are timely and fit a large variety of topical interests. However, only limited pragmatic competence can be gained from passive video watching [31]. Active engagement with the videos that involve actual practice would lead to more optimal acquisition of pragmatic competence.

A simple, yet effective, form of active engagement with foreign language videos is repetition. Studies have shown that merely repeating a sentence requires a learner to be able to completely process a language [12]. Existing language learning tools like DuoLingo[1] and Rosetta Stone[2] do not prioritize speaking and existing video learning tools (e.g. [19]) use text rather than speech. Creating tools that focus on repetition of language, and integration of repetition into workflows involving native speaker materials could open new possibilities for language learning tools.

In this paper, we present Seiyuu-Seiyuu, an online video-based learning tool that takes a step towards these goals. Seiyuu means voice actor in Japanese. In Seiyuu-Seiyuu, users suggest videos to watch through a crowdsourced website by linking to YouTube videos. Then, Seiyuu-Seiyuu allows users to repeat utterances they hear in the video, and what's more, to take on the role of an voice actor and speak with paralinguistic cues such as intonation, pitch, etc. Seiyuu-Seiyuu takes advantage of Google Chrome's speech recognition technology to recognize what the user is saying. When using videos for which a transcript has been uploaded, the system allows users to see how much of the video they have correctly repeated.

We present results from an online evaluation study of 27 participants comparing our system to a text-based translation interface. In the study, learners used both the voice and text interface. We found that learners searched 53% more words using the voice interface than the text interface. Furthermore, learners who used voice first conducted more than twice as many total searches with voice and text, indicating an ordering effect from using the voice interface. Furthermore, our qualitative findings support previous research that shows the value of learning with foreign language videos, and suggests that the voice interface is better suited for learning practical conversational skills.

RELATED WORK

Cross-cultural communication

Even with linguistic competence, foreign language learners have difficulty communicating effectively with native speakers. Studies have shown that communication styles largely differ between native and non-native speakers [20, 23] and that these differences impair successful communication. For example, in a lab study, Wong [36] showed that native Mandarin speakers

[1]https://www.duolingo.com/
[2]www.rosettastone.com

CHI 2017, May 06 - 11, 2017, Denver, CO, USA

ISBN 978-1-4503-4655-9/17/05…$15.00

DOI: http://dx.doi.org/10.1145/3025453.3025805

had difficulty with repairing misunderstandings in face-to-face communication. Furthermore, in an interview study, Yuan et al. [39] showed that non-native speakers in universities found it difficult to interact with native speakers in part because of lack of common ground and unfamiliarity with informal and idiomatic English. These studies show that understanding the linguistic constructs of a language (e.g. vocabulary and grammar) is not sufficient to communicate effectively, and that communication skills are also important.

Importance of pragmatic competence

"Pragmatics is the study of language from the point of view of users, especially of the choices they make, the constraints they encounter in using language in social interaction and the effects their use of language has on other participants in the act of communication" [10]. Pragmatics is an essential part of communication. Kasper [18] provided the following example: "feed the cat", "can/could/would you feed the cat?", and "the cat's complaining" all communicate the same request, but may have very different impacts on the relationship between the requester and interlocutor. In a study of interpretation of dialogues, Bouton [8] showed that in 27% of cases, non-native speakers with high levels of linguistic competence interpreted the meaning of indirect statements differently from native speakers [18]. Furthermore, the inability to express requests effectively can disadvantage non-native speakers. Bardovi and Hartford [5] showed that in academic advising, non-native speakers were less likely to have their requests fulfilled because of a lack of pragmatic competence.

Some aspects of pragmatic competence learning can be transferred from a learner's first language. However, when norms differ across cultures, this transfer can hurt rather than help learners. In a study of Iranian learners of English, [1] showed that learners with higher proficiency made more pragmatic errors than low-proficiency learners in making refusals because higher proficiency learners tended to transfer more pragmatic knowledge. Because pragmatics differ from language to language, it is important to learn them. This transfer of pragmatic knowledge is nicely illustrated through a classroom exchange between two students that was observed by Hamid and Naeimi [1]. In this exchange, an upper-intermediate Iranian learner of English who plays a teacher responds to a classmate playing a student: "I have been teaching for many years, and I have experienced many paths. I think it's the best way" using formulaic Iranian refusals whereas a native English speaker responded with "Thanks for your suggestion, but we're following a very strict curriculum" [1]. In sum, if a learner wishes to communicate effectively in a second language, pragmatic competence cannot directly transferred from the learner's native language.

Pragmatic competence education

Because of the vast quantity and diversity of contexts that learners may find themselves in, traditional teaching methods that explicitly assign ideal responses to specific scenarios can fail to prepare learners for the real world. In a review of education literature, Kasper [18] writes that pragmatic competence cannot be taught and instead, "[t]he challenge for foreign or second language teaching is whether we can arrange learning opportunities in such a way that they benefit the development of pragmatic competence in L2". The key to teaching pragmatic competence is in giving learners as much experience as possible with as many different scenarios as possible. A later study by Matsumura [21] showed that in Japanese learners, the amount of exposure to English inside and outside of the classroom was more predictive of pragmatic development than English proficiency. Furthermore, other studies have shown that high linguistic competence is not necessary for pragmatic learning. For example, Takehashi and Beebe [28] showed that even early learners can acquire pragmatic competence given enough experience with real scenarios. Furthermore, constructivist views of education suggest that learners must use these scenarios to construct knowledge for themselves [7]. Together, these findings indicate that pragmatic competence is developed independently of linguistic competence and the key factor in pragmatic development is the amount and diversity of language experiences that learners engage with.

Existing language learning tools

There has been a lot of interest in HCI in creating tools for foreign language learning. Some of this work has focused on learning vocabulary, such as MicroMandarin, which uses the user's location to suggest relevant words [14] and Wait Learning, which teaches vocabulary during natural lulls in online conversations [9]. Other systems focused on teaching grammar, such as Ingenium, which provides virtual manipulatives for learning Latin grammar [40]. Finally, video games have been shown to provide a fruitful context for language learning, such as ToneWars, which teaches Chinese tones [17], and Crystallize, a 3D immersive game for learning Japanese vocabulary and grammar [11]. However, in order to learn how to understand speech effectively, learners need to go beyond vocabulary and grammar and listen to a large amount of speech from many different native speakers [15] and in multiple contexts [24].

Learning with foreign language videos

Given the importance of learning language in a variety of contexts, videos are a natural learning resource because they are available across the world and provide many diverse contexts for learning. According to Secules et al. [26], "[v]ideo permits second language learners to witness the dynamics of interactions as they observe native speakers in authentic settings...". Use of video in classrooms has also been shown to improve learner's fluency. For example, Weyers [33] found that the use of Spanish television programs in the classroom improved students' confidence and diversity of words used in communication. Some educators have identified video as an effective tool specifically for pragmatic learning (e.g. [32]).

However, learning directly from authentic foreign language videos can be very challenging for learners. In traditional learning settings, instructors reduce rates of speech and adjust their vocabulary to accommodate learners [3], but in authentic input (such as videos), learners must adapt to the increased rate of speech and more complex word choice and linguistic structure [4]. Because of this, many researchers have studied how to structure activities for learning from videos and developed tools to improve video learning. Foreign language captions are one of the most widely used and studied methods, and have

been shown to be more effective for learning than translated subtitles or audio-only video [35, 22]. Others have studied ways to improve over just displaying captions. For example, Kovacs and Miller [19] developed a system for simultaneously displaying the native-language caption along with the English translation and included word-by-word translations upon hovering over individual foreign language words.

Caption generation itself has been explored as a learning method. For example, Williams and Thorne [34] developed a course for subtitle generation and found it to be effective for learning, but it required extensive training in order for learners to participate. A system was developed to make the captioning process more approachable to learners by using a machine-generated transcript as a starting point and providing word-by-word translations [2].

While these systems enable learners to learn from videos, they are all text based, which may limit how deeply learners process speech from the video. For example, Swain and Lapkin [27] argue that learners notice problems with their output during the process of speaking, and then can go on to correct these problems. We therefore see an opportunity for novel language learning systems that draw from a voice-driven approach to learning from videos.

Voice-driven systems

Voice recognition has been explored in several different educational contexts. For example, Yoon et al. [38] developed a system to annotate texts with voice and found it to allow for better expression of complex ideas. Specifically in language learning, applications such as Rosetta Stone[3] incorporate automated pronunciation feedback. However, as noted by Gaur et al. [16], speech recognition systems are often inadequate for producing high-quality text transcripts. Yoon et al. [38] noted similar issues with voice transcription in their system: "participants preferred using the [audio] waveform over transcription because of the detrimental effect of transcription errors". In summary, previous research indicates that voice driven systems have promise in learning tools, but more work needs to be done to determine if current speech-recognition technology is adequate to provide good foreign language learning experiences.

UNDERSTANDING LEARNER USE OF MEDIA DESIGNED FOR NATIVE SPEAKERS

To explore the role of authentic media use in language learners, we conducted a survey with 36 university participants. Participants were between 18 and 25 years old (mean=20), and learned a combined 11 different languages. 55% had spent at least some time learning 2 or more languages. 63% used use native speaker media in their learning. Of the learners that reported using materials designed for native speakers for learning, when asked how they dealt with material they could not understand, many reported using online translation systems such as Google translate (45%). Other strategies included using subtitles (18%), native speaker friends or family (9%) and continuing to listen despite not understanding (9%).

[3]www.rosettastone.com

Motivation for learning	percent
Class	21%
Professional development	12%
Interest in language or culture	19%
Tourism	14%
Broaden world view	16%
Personal Challenge	11%
Communicate with family or friends	7%

Figure 1. Motivations for surveyed learners.

Learning method	percent
Class	59%
Media designed for native speakers	14%
Native speaker friends	10%
Applications	8%
Living abroad	9%

Figure 2. Learning sources for surveyed learners.

Learners were asked to report their biggest challenge in learning a language and results were coded and reported in Figure 3. When asked about the greatest challenge that learners faced, 41% of learners learners reported challenges related to oral competence: 14% said conversation (e.g. *Speaking in conversation as a native would*), 8% said limited interaction with native speakers (e.g. *it was very hard to find other speakers to practice with*), 8% said nuances in language (e.g. *Remembering the correct use of certain words even though they have the same meaning*), 8% said native-like pronunciation (e.g. *Accents to sound natural*), and 3% said listening (e.g. *having to extract what a speaker was saying*). Some learners also indicated difficulty finding resources for improving oral proficiency: one of the most significant challenges was accessing learning resources that allowed me to practice speaking and listening. The reported difficulties with oral proficiency match with the findings of Edge et al. [13]. These findings indicate a need for more effective oral learning tools, specifically those that focus on pragmatic competence.

DESIGN

The most common existing methods for learning a foreign language make use of carefully designed learning materials, which, we argue, fail to provide the breadth of experiences that are necessary to gain situational fluency. Use of authentic materials has been shown to enable pragmatic competence learning. However, using materials designed for native speakers can be incredibly challenging because there are many unfamiliar words and structures in these materials and finding out the meaning of these words and structures is difficult. Therefore, our primary design question was: how can we design to make authentic foreign language materials more accessible?

We observed three key opportunities which inspired the design: (1) the internet is home to countless free authentic foreign language videos, (2) repeating phrases or dialogues from videos is a natural activity and helps improve oral proficiency, and (3) speaking bypasses the need to type which is often very challenging and time consuming in a foreign language.

Challenge	percent
Motivation	17%
Grammar	17%
Conversation	14%
Limited interaction with native speakers	8%
Nuances in language	8%
Reaching native-like pronunciation	8%
Self-confidence	8%
Bad teachers	6%
Lack of time	6%
Listening	3%
Reading	3%
Writing	3%

Figure 3. Greatest challenges faced by learners as identified by our survey. Motivation, grammar and conversation appear to be the most critical.

(1) Freely available videos on YouTube include 76 languages and over 100 million videos[4]. Some countries also have their own streaming video sites (i.e. China - http://www.youku.com/) which can increase the number of videos even further.

(2) Copying words and phrases is how we learn our first language. Videos specifically have been shown to be an important source of learning for infants [25]. In adult learning, many educators use oral repetition as a central learning excercise (e.g. [30]). Furthermore, listening to foreign language before speaking (known as word priming) has been shown to improve recognition and pronunciation of those words [29].

(3) In our survey of authentic material use by foreign language learners, we found that of those that used materials designed for native speakers, 45% reported using a text based tool such as a dictionary or Google translate in order to learn from the material. Furthermore, many learners use videos with flashcard systems such as Anki[5] which requires transcribing text from videos.

Considering these opportunities, we developed an interface for watching any foreign language video while the learner can speak words or phrases to see them transcribed and translated below the video.

Learners use the system by first selecting a video to learn from. For example, a learner of English might choose the television program *Friends*. As the learner watches the program, they listen carefully for words and phrases they can pick out and then repeat them. For example, maybe the learner hears the phrase "Chandler, I sensed it was you." but is unsure of what "sensed" means. The learner would then hold the spacebar to pause the video and repeat the phrase aloud. The speech recognition system would recognize and display the text for the phrase. Below the transcribed text, a translation would appear in the learner's native language. After the learner reads the transcription and translation, they release the spacebar and continue watching the video.

[4]https://www.youtube.com/yt/press/statistics.html
[5]anki.net

The system was implemented using node.js as a backend. Speech recognition was realized using the Speech Recognition Webkit built into Google Chrome. Translation used the Google translate API.

Website

To improve long-term learning and engagement, the web version of the system includes additional features to support existing language learning practices and community building. To support long-term learning, spoken utterances can be saved to a history and edited. A plugin was also developed to allow learners to sync saved phrases with the spaced repetition system Anki[6].

To better engage learners with the system, game features were added. Learners can choose to upload a caption file, which allows the system to check for the accuracy of utterances. When captions are available, learners receive a score based on how much of the video they are able to repeat and learners can watch the same segment multiple times to increase overall progress. Furthermore, a transcript of the video is shown on the side to indicate which words have already been spoken and which words the learner still needs to say. The system is shown in Figure 4.

The website includes a *popular* page where recently viewed YouTube videos are displayed and learners can view their progress as shown in Figure 5. Learners can also choose to add their own video from YouTube or their hard drives. If learners choose a new YouTube video, the video gets displayed on the *popular* page. Videos from users' hard drives videos are not uploaded to the server or displayed on the *popular* page.

EVALUATION

Field study

The site was announced on the reddit LanguageLearning forum[7] as well as individual sub-reddits for Japanese[8] and Spanish[9]. Data was collected through usage logs. A total of 130 participants tried the system and 71 learners spoke 10 or more phrases. Learners that spoke 10 or more phrases spoke an average of 71 utterances. Users uploaded 22 new YouTube videos (in French, Spanish and Japanese) and used 6 unique media from their hard drives. This suggests that the tool can function in the wild.

Furthermore, since some videos may be more effective for pragmatic learning than others, we believe that identifying and sharing effective resources is an important task. Our findings about learner use of our system suggest that the tool could provide motivation for learner-sourced resource evaluation and sharing.

The results indicated that the system has potential to function as an independent learning tool, but we wanted to do a more systematic exploration in order to better understand how the tool compares to existing tools.

[6]anki.net
[7]reddit.com/r/languagelearning
[8]reddit.com/r/LearnJapanese
[9]reddit.com/r/learnspanish

Figure 4. The interface with game features provides feedback when learners sayid phrases correctly (1) over transcribed and translated text (2). A progress bar and text displays how much of the video the learner has correctly repeated (3). Learners can add utterances to their library or remove them using buttons (4), or upload transcripts and adjust how text is displayed through the settings (5). When available, a transcript is also displayed to show how much of the video a learner has repeated, and help learners find new words and phrases to listen to (6). Screenshot taken from *Ode to Joy* on YouTube (https://www.youtube.com/watch?v=4wGpu56WQGQ).

Formal Evaluation

To gain insight into the usability and effectiveness of our system compared to other video learning methods, we conducted an online study with foreign language learners. Originally, the study was available in six languages, but we only had participants use Spanish, French, and Chinese (Mandarin). In our survey of language learners, we found that learners most frequently used Google Translate to learn from native speaker materials. Therefore, to examine the effectiveness of Seiyuu-Seiyuu we developed an interface as a control that allowed learners to type into a textbox upon which translations appeared below using the Google translate API as shown in Figure 7. We used a within subjects design where approximately half of the participants first used the speech interface for video learning and the remaining participants first used the text interface. For all languages, realistic dramas or comedies were chosen for learners to watch. The video was different for each language. This is because culture is an essential element of pragmatic learning, so it was important to us to choose videos coming from cultures where each language was spoken. The videos used are shown in Figure 8.

Participants clicked on a link which redirected them to a webpage where they spoke a test phrase into their microphone in order to verify that speech recognition was working properly. Participants then chose a language and completed a short survey about their prior experience with the chosen language. Next, participants were randomly assigned to either use the speech interface first or the typing interface first. When using the speech interface, text was displayed to indicate that participants should hold the spacebar to pause the video and begin speaking. After pressing the spacebar, the interface would indicate that they should begin speaking as shown in Figure 6. As the participants spoke, their utterances were recognized and translated below the video. In the typing condition, the interface indicated that participants could pause the video by clicking on it, and participants could type words and phrases to see their translations. In each case, the video segment was 10 minutes long. Following the first video segment, participants were asked to rate difficulty, usefulness and enjoyment as well as recall words from the video with their surrounding contexts. Participants then watched a different 10 minute segment of the same video using the interface that they did not use in the first part of the study. Following the second task, participants were asked to report the same information as in the first part. Finally, participants were redirected to a survey where they provided demographic information and were asked to discuss their learning experience and compare the two interfaces.

Participants

Participants were recruited through a campus research system. Participants were compensated either $10 or research credit for participation. Three participants were excluded because they indicated that they were already native speakers of the chosen language or because they skipped parts of the experiment. A total of 27 participants (15 typing first and 12 speech first) were used in the final analysis. 67% of the participants did the study in Spanish, 26% of participants did the study in French and 7% of participants did the study in Chinese.

Figure 5. When visiting the website, learners choose a language (1) and can then view videos that other learners watched in that language (2). Learners can also choose their own video from YouTube or their computer (3). Links to Youtube become visible for all users, but personal videos are only visible to the user who uploaded them.

No significant differences were found in reported measures between languages.

Hypotheses

Given previous research on foreign language learning and voice-driven system design, we set up two hypotheses to explore possible differences between voice-driven and text-driven conditions. First, (H1) learners will try to look up more words in the speaking condition. The cognitive cost of speaking should be less than typing, so we expect learners will be more willing to look up words. Furthermore, previous research on voice-driven learning tools indicates a preference for generating speech over text [38]. (H2) Learners will find the speaking version more useful. Speaking is a goal of many learners, so we expect practicing speaking will be seen as more practical.

Measures

Usage

Usage was measured as the number of times a learner spoke a new utterance or typed a new phrase. This is a numerical score that counts the number of interactions.

Usefulness measure

After each video section, learners were asked to report how useful they found the system using a continuous slider (0 to 100).

Figure 6. In the voice interface used in our evaluation, learners were given instructions on how to use the system (1), and spoken phrases appeared below the video (2) with a translation below the utterance (3). In Japanese and Chinese, pronunciation was displayed beneath the characters. Since the speech recognition was not always accurate, the "more" button (4) could be clicked to show alternatives from the speech recognition system. Screenshot was taken from *Keikon Dekinai Otoko* on YouTube (https://www.youtube.com/watch?v=dX8vYhztrxM).

FINDINGS

Participants' usage frequencies were analyzed using linear mixed regression model. The independent variables were interface type (speech vs. typing) and condition (speech interface first vs. typing interface first). The interface type x condition interaction term was also entered into the model but was non-significant, $F(1,25) = 0.498$, $p = .487$, $\eta^2 = .015$. The within-subject effect of interface type was significant, $F(1,25) = 7.23$, $p = .013$, $\eta^2 = .221$, indicating that participants on average used the speech interface more ($M = 26.52$, $SD = 17.42$) than the typing interface ($M = 18.44$, $SD = 13.95$). The between-subject effect of condition was also significant, $F(1,25) = 29.45$, $p < .001$, $\eta^2 = .541$, indicating that when participants used the speech interface first, they interacted (speech and typing combined) with the system more ($M = 67.00$, $SD = 18.29$) than those participants that used the typing interface first ($M = 27.33$, $SD = 19.32$). Pairwise comparisons of simple effects also revealed the same finding. For the speech interface, participants in the speech interface first condition used it more often ($M = 38.75$, $SD = 12.08$) than participants in the typing interface first condition ($M = 16.73$, $SD = 14.77$), p < .001. Similarly, for the typing interface, participants in the speech interface first condition used it more often ($M = 28.25$, $SD = 12.75$) than participants in the typing interface first condition ($M = 10.60$, $SD = 9.23$), p = .001.

No significant differences were found in enjoyment, difficulty or reported number of phrases remembered.

Perceived usefulness and usability

A 2x2 mixed ANOVA, with interface type as the within-subject factor and condition as the between-subject factor was performed on participants' perceptions of usefulness. Participants on average found the speech interface more useful ($M = 27.67$, $SD = 22.96$) than the typing interface ($M = 21.41$, $SD = 20.90$). This difference was marginally

Figure 7. In the typing interface used in our evaluation, learners were given instructions below the video (1), could type word or phrases into a text field (2) and translations would appear below after the learner stopped typing (3). Screenshot taken with *Sur Le Fil* on YouTube (https://www.youtube.com/watch?v=bapP3JM3SZA&t=314s).

Language	video
Spanish	*Mi Corazón Es Tuyo*
French	*Sur Le Fil*
Chinese (Mandarin)	欢乐颂 *(huan'le'song), Ode to Joy*

Figure 8. Learning sources for surveyed learners.

significant, $F(1,25) = 3.51$, $p = .073$, $\eta^2 = .120$. The main effect of condition as well as interaction effect were non-significant.

However descriptively, when asked on a preference scale (1-7), with 1 being strongly prefer typing interface, 4 being neutral, and 7 being strongly prefer speech interface, most participants found the speech method more useful (63% of participants) than the typing method (15% of participants) and some were neutral (22%).

Although some participants had difficulty with the accuracy of the voice recognition engine (e.g. P6: *"sometimes it could not understand what I was trying to say"*), many participants found speaking was easier than typing (e.g. P18: *"...less cognitive overhead than typing"*, P13: *"...you could just say what you heard [instead of typing]"*). Furthermore, it eliminated the need to worry about spelling (P19: *"The dictionary method was hard because I didn't know how to spell some phrases so it was easier to repeat them."*, P12: *"I was struggling with how to spell the words so that distracted me."*). Other participants indicated that saying words aloud helped with memorization (e.g. P12: *"Saying the words out loud makes me remember them more."*, P15: *"You can pick up on words more quickly by actually saying them out loud."*).

However, some participants preferred the text method because it helped to train spelling (e.g. P5: *"[the voice method] did not help me with placing accents on letters as the program did that for me"*, P7: *"you may be able to hear the words being spoken in conversation but you may not know how to spell them when writing or reading."*). This finding indicates that learner type is important to consider when choosing between text- and voice-driven systems, and perhaps both methods are necessary for comprehensive learning.

While some participants wanted more feedback on pronunciation (e.g. *"I won't know if I pronounce or use these words correctly compared to the method of talking this language with the native speaker."*, P10: *"Saying it out loud doesn't give you a basis for pronunciation so I was saying them incorrectly."*) or blamed their pronunciation for trouble with the speech recognition engine (e.g. P21: *"[The voice interface is] harder if you have terrible pronunciation"*), many participants found the system to be helpful for improving pronunciation (e.g. P15: *"with the voice learning method you can practice speaking the words and sounding them out which is helpful for conversational Spanish."*, P23: *"Voice helped me understand accents more than typing"*).

Furthermore, some participants directly discussed learning pragmatic features with the system (e.g. P22: *"[M]uch better than doing it from a book. This way I know the right way to pronounce things and the context I might use the phrases in."*, P26: *"Voice learning is advantageous to other types of learning because you can hear the emotion in a person's voice. I find that Spanish speakers especially give a lot of clues to what they're saying in the way that they're saying it."*). Furthermore, the system could help learners overcome lack of confidence in pronunciation (e.g. P11: *"[practicing with the voice interface] will not be embarrassing if I pronounce the words badly."*).

Learning with native speaker materials

Using materials designed for native speakers is perceived as difficult by many learners. In the comments, many learners indicated that the speech was difficult to follow. When asked about the materials, 60% of learners indicated that the material was very challenging and learners reported an average of 7/10 points when asked how difficult the material was. For example, P3 wrote *"I found it very difficult to follow along because they were talking so fast"* and P13 wrote *"I may have overestimated my French-speaking ability; but I find it difficult to understand films in which the people are speaking fluidly because of how quickly they talk."*.

However, just because this activity is challenging does not mean it is not worth doing. Many participants indicated that the use of authentic materials made the learning more valuable (e.g. P10: *"I think the materials designed for native speakers is more challenging but it is more original and I believe that I can use it in related situations."*, P22: *"...made me feel like I was learning phrases I would actually use."*). Furthermore, some participants recognized that practice with native speech can can help with communicative goals (e.g. P13: *"[T]his is more real-world applicable. When one speaks to native French speakers; they are not going to enunciate every syllable or speak slowly like we learn in French class."*). We also speculate that the tool raises awareness in learners about the thoroughness of their understanding. It is easy for a learner to think that they understood the phrase adequately, but repeating phrases out loud can highlight the parts they missed.

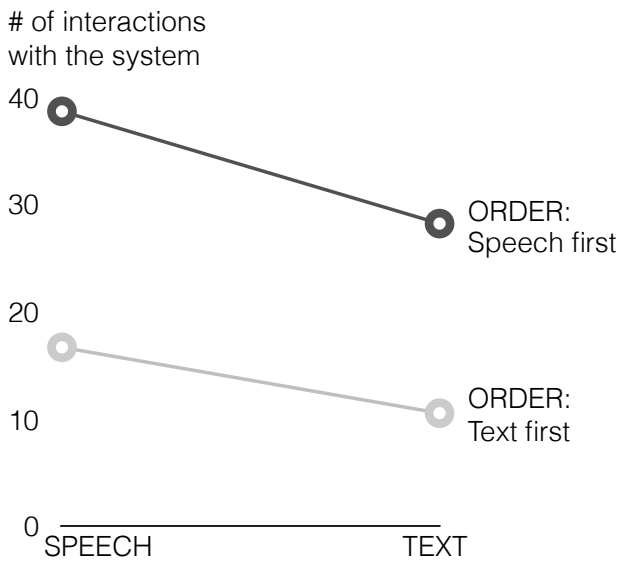

Figure 9. Number of interactions with the system for each interface in each ordering. Learners used the speech interface significantly more than the text interface, and using speech first resulted in more overall interactions.

Our findings also indicate that learners had diverging perspective on how to learn from native speaker materials. Learners were told that they did not need to understand everything, but some learners felt uncomfortable with this learning method (e.g. P18: *"Not as structured. I don't learn the exact grammar. I missed a lot of the dialogue."*, P6: *"It made it a lot harder because they were talking so fast and it was assumed that I could understand when I really had no idea what was going on"*). However, others found this method to be a refreshing change from classroom learning (e.g. P19: *"It felt less intimidating because I knew I wasn't supposed to understand everything. In a classroom; a lot of the material is designed for people at your level to there is more pressure to know exactly what everything means."*, P21: *"...It was also fun to try and repeat the words and trying to see if the translation was correct or made sense."*). Given different perspectives on learning through authentic materials, future work should more closely examine these perspectives and could explore designs to promote or support different learning styles.

Order effect

The linear mixed regression analysis reported above revealed that when participants used the speech interface first, they interacted with the system more through either speaking or typing than those participants that used the typing interface first as shown in Figure 9. We found this to be very interesting, but can only speculate on the reasons for this.

We speculate that the speech activity increased learner engagement or sharpened learners' listening ability. In open-ended comments, of participants that started with text, 40% used the word "fast", "rapid" or "quick" to describe the speech, whereas of those with that started with speech only 25% used one of those words to describe the speech. This could indicate that the speech activity prepared learners for better listening. However, more work needs to be done to explore this area. Regardless of the reason, we find the increased engagement with the video after using the voice interface to be an encouraging sign.

LIMITATIONS

Limitations with speech recognition technology and machine translation

Both speech recognition and machine translation are imperfect technologies. Many participants mentioned that they wished the speech recognition were more accurate. However, we chose to use these technologies over pre-annotated videos because our original design intent was to allow learners to tap into any video or audio resource rather than just those prepared by instructors. By choosing to use raw video and audio, we believe that our findings are more generalizable. Furthermore, despite learners' frustration with inaccuracies, we found that the majority of learners still preferred the speech interface over the text interface.

Furthermore, correctness feedback was not always available. Correctness feedback is beneficial, but there is a tradeoff, as it requires resources that are not available for every video. Our goal was to design a tool that enables users to use any video as a learning resource and we showed that this is possible. Previous literature has shown that even without feedback, learners notice and fix errors when generating second language writing [27]. In another study, novice learners utilized visuals, audio, and narrative understanding to identify errors in foreign language transcripts when provided Google translate [2]. Similarly, we expect users of this tool to be able to identify speech errors using speech recognition output and translation service.

In further exploration we found that 13% of utterances in our study were repeated, suggesting that learners were making corrections during use.

Limitations with videos

In the formal evaluation, participants did not have the option to choose a video, and were only shown specific segments from the video. In real use, we would expect that learners could choose their own videos and choose which parts to watch, as they did in the field study. We would expect that if the videos were better matched to learners' interests and skill level that results would only improve.

Learning limitations

Previous work has shown that interaction with videos can lead to pragmatic competence learning. Therefore our goal was to increase interaction with videos and we did not take quantitative measures of learning. However, we do have qualitative evidence that learners were gaining competence. For example, learners reported the system improving their listening skills, pronunciation, and contextual vocabulary. Furthermore, we asked learners to recall words and phrases along with their contexts and learners reported an average of 2.7 word-context pairs in both the text and voice conditions. This suggests that the video learning activity was leading to contextual learning, but more work will need to be done to improve our understanding of the quantity and quality of learning. Previous work on

pragmatic competence learning has used longitudinal studies with durations ranging from a semester to multiple years. In order to perform a qualitative evaluation of our system, future work should perform a longitudinal study of the learning method.

Finally, we wish to point out that language learning is a complex and nuanced activity and, as written by Xiao and Ishii, "traditional HCI methodologies with their focus on optimizing quantifiable metrics risk blinding researchers to the richness and nuance of artistic practices" [37]. Similarly, we fear that a focus on quantifiable metrics in language learning (e.g. vocabulary and grammar) blinds us to essential aspects of foreign language learning, such as pragmatic compotence.

CONCLUSION

The presence of rich context and authentic language makes videos invaluable resources for learning pragmatic competence, but learning with foreign language videos is very difficult. The challenge is to design tools that make the videos more accessible and allow learners to absorb as much as possible from the video materials.

Our results show that using voice is a natural and effective way for learners to engage with videos, and repeating words and phrases from videos can cause learners to engage more with text-based video activities. We found that the tool affords learning through videos that learners enjoy (shown by the variety of cartoons and dramas that learners uploaded during the field study), understanding where phrases might be used (e.g. one participant said: "*This way I know the right way to pronounce things and the context I might use the phrases in.*"), and practicing speech rich in emotion and subtlety (e.g. one participant reflected: "*Voice learning is advantageous to other types of learning because you can hear the emotion in a person's voice*"). To our knowledge, this is the first study to explore using automatic speech recognition to support video learning, and much work remains to be done in interface design to explore other methods for providing feedback, structuring learning within the system, and boosting learner confidence.

Learning with native speaker materials has the potential to be engaging and effective for learning deep language abilities such as pragmatic competence. This tool is a first step in exploring this space, but much work remains to be done to better understand the what can be learned through native speaker materials, and how best to support learners that wish to use these materials.

REFERENCES

1. Hamid Allami and Amin Naeimi. 2011. A cross-linguistic study of refusals: An analysis of pragmatic competence development in Iranian EFL learners. *Journal of Pragmatics* 43, 1 (2011), 385–406.
2. Anatomized. 2017. Have your cake and eat it too: Foreign language learning with a crowdsourced video captioning system. In *Proceedings of the 20th ACM Conference on Computer-Supported Cooperative Work & Social Computing*. ACM. In press.
3. Susan M Bacon. 1992a. Authentic listening in Spanish: How learners adjust their strategies to the difficulty of the input. *Hispania* 75, 2 (1992), 398–412.
4. Susan M Bacon. 1992b. The relationship between gender, comprehension, processing strategies, and cognitive and affective response in foreign language listening. *The modern language Journal* 76, 2 (1992), 160–178.
5. Kathleen Bardovi-Harlig and Beverly S Hartford. 1990. Congruence in native and nonnative conversations: Status balance in the academic advising session. *Language learning* 40, 4 (1990), 467–501.
6. Julie A Belz. 2007. The role of computer mediation in the instruction and development of L2 pragmatic competence. *Annual Review of Applied Linguistics* 27 (2007), 45.
7. Phil Benson. 2013. *Teaching and researching: Autonomy in language learning*. Routledge.
8. Lawrence F Bouton. 1988. A cross-cultural study of ability to interpret implicatures in English. *World Englishes* 7, 2 (1988), 183–196.
9. Carrie J Cai, Philip J Guo, James Glass, and Robert C Miller. 2014. Wait-learning: leveraging conversational dead time for second language education. In *CHI'14 Extended Abstracts on Human Factors in Computing Systems*. ACM, 2239–2244.
10. David Crystal. 2011. *Dictionary of linguistics and phonetics*. Vol. 30. John Wiley & Sons.
11. Gabriel Culbertson, Shiyu Wang, Malte Jung, and Erik Andersen. 2016. Social Situational Language Learning through an Online 3D Game. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 957–968.
12. P Duff. 2000. Repetition in foreign language classroom. *Second and foreign language learning through classroom interaction* (2000), 109.
13. Darren Edge, Kai-Yin Cheng, Michael Whitney, Yao Qian, Zhijie Yan, and Frank Soong. 2012. Tip tap tones: mobile microtraining of mandarin sounds. In *Proceedings of the 14th international conference on Human-computer interaction with mobile devices and services*. ACM, 427–430.
14. Darren Edge, Elly Searle, Kevin Chiu, Jing Zhao, and James A Landay. 2011. MicroMandarin: mobile language learning in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 3169–3178.
15. Maxine Eskenazi. 1999. Using a computer in foreign language pronunciation training: What advantages? *Calico Journal* (1999), 447–469.
16. Yashesh Gaur. 2015. The Effects of Automatic Speech Recognition Quality on Human Transcription Latency. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, 367–368.

17. Andrew Head, Yi Xu, and Jingtao Wang. 2014. Tonewars: Connecting language learners and native speakers through collaborative mobile games. In *Intelligent Tutoring Systems*. Springer, 368–377.
18. Gabriele Kasper. 1997. Can pragmatic competence be taught. *NetWork* 6 (1997), 105–119.
19. Geza Kovacs and Robert C Miller. 2014. Smart subtitles for vocabulary learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 853–862.
20. Han Z Li. 1999. Grounding and information communication in intercultural and intracultural dyadic discourse. *Discourse Processes* 28, 3 (1999), 195–215.
21. Shoichi Matsumura. 2003. Modelling the relationships among interlanguage pragmatic development, L2 proficiency, and exposure to L2. *Applied Linguistics* 24, 4 (2003), 465–491.
22. Holger Mitterer and James M McQueen. 2009. Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS one* 4, 11 (2009), e7785.
23. Duyen T Nguyen and Susan R Fussell. 2012. How did you feel during our conversation?: retrospective analysis of intercultural and same-culture instant messaging conversations. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work*. ACM, 117–126.
24. Martha C Pennington and Jack C Richards. 1986. Pronunciation revisited. *TESOL quarterly* 20, 2 (1986), 207–225.
25. Mabel Rice. 1983. The role of television in language acquisition. *Developmental Review* 3, 2 (1983), 211–224.
26. Teresa Secules, Carol Herron, and Michael Tomasello. 1992. The effect of video context on foreign language learning. *The Modern Language Journal* 76, 4 (1992), 480–490.
27. Merrill Swain and Sharon Lapkin. 1995. Problems in output and the cognitive processes they generate: A step towards second language learning. *Applied linguistics* 16, 3 (1995), 371–391.
28. Tomoko Takahashi, Leslie M Beebe, and others. 1993. Cross-linguistic influence in the speech act of correction. *Interlanguage pragmatics* 138 (1993), 158–169.
29. Pavel Trofimovich and Elizabeth Gatbonton. 2006. Repetition and focus on form in processing L2 Spanish words: Implications for pronunciation instruction. *The Modern Language Journal* 90, 4 (2006), 519–535.
30. Michio Tsutsui and Masashi Kato. 2001. Designing a Multimedia Feedback Tool for the Development of Oral Skills. *2001). CALL-The challenge of Change: Research & Practice* (2001), 81–88.
31. Robert Vanderplank. 2010. Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language teaching* 43, 01 (2010), 1–37.
32. Gay N Washburn. 2001. Using situation comedies for pragmatic language teaching and learning. *TESOL journal* 10, 4 (2001), 21–26.
33. Joseph R Weyers. 1999. The effect of authentic video on communicative competence. *The modern language journal* 83, 3 (1999), 339–349.
34. Helen Williams and David Thorne. 2000. The value of teletext subtitling as a medium for language learning. *System* 28, 2 (2000), 217–228.
35. Paula Winke, Susan Gass, and Tetyana Sydorenko. 2010. The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology* 14, 1 (2010), 65–86.
36. Jean Wong. 2000. Delayed next turn repair initiation in native/non-native speaker English conversation. *Applied Linguistics* 21, 2 (2000), 244–267.
37. Xiao Xiao and Hiroshi Ishii. 2016. Inspect, Embody, Invent: A Design Framework for Music Learning and Beyond. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 5397–5408.
38. Dongwook Yoon, Nicholas Chen, François Guimbretière, and Abigail Sellen. 2014. RichReview: blending ink, speech, and gesture to support collaborative document review. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. ACM, 481–490.
39. Chien Wen Yuan, Leslie D Setlock, Dan Cosley, and Susan R Fussell. 2013. Understanding informal communication in multilingual contexts. In *Proceedings of the 2013 conference on Computer supported cooperative work*. ACM, 909–922.
40. Sharon Zhou, Ivy J Livingston, Mark Schiefsky, Stuart M Shieber, and Krzysztof Z Gajos. 2016. Ingenium: Engaging Novice Students with Latin Grammar. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. ACM, 944–956.