Have your Cake and Eat it Too: Foreign Language Learning with a Crowdsourced Video Captioning System

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

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

ABSTRACT

Learning from captioned foreign language videos is highly effective, but the availability of such videos is limited. By using speech-to-text technology to generate partially correct transcripts as a starting point, we see an opportunity for learners to build accurate foreign language captions while learning at the same time. We present a system where learners correct captions using automatic transcription and machine-generated suggested alternative words for scaffolding. In a lab study of 49 participants, we found that compared to watching the video with accurate caption, learning and quality of experience were not significantly impaired by the secondary caption correction task using interface designs either with or without scaffolding from speech-to-text generated alternative words. Nevertheless, aggregating corrections reduced word error rate from 19% to 5.5% without scaffolding from suggested-alternatives, and 1.8% with scaffolding. Feedback from participants suggest that emphasizing the learning community contribution aspect is important for motivating learners and reducing frustration.

Author Keywords

crowdsourcing, language learning, video learning

ACM Classification Keywords

K.3.0. Computers and Education: General

INTRODUCTION

Many learners and educators see learning through foreign language video watching (i.e. video learning) as highly effective. Learning in this way allows learners to gain cultural knowledge alongside language mechanics such vocabulary and grammar [14, 30, 19]. Furthermore, learning through video can be highly engaging to learners if the learners find the video content interesting [32].

Studies of learning through foreign language video have shown that the most effective way to learn through foreign language video is by watching the video accompanied by foreign language captions rather than watching without captions or with translated subtitles [27, 25]. However, depending on the language and specific video that a learner wants to use for learning, these captions may be difficult to obtain or unavailable. Previous studies have shown that freedom of choice is essential to fostering intrinsic motivation [4, 7]. Furthermore, theories of the zone of proximal development [35] and flow [5] explain learning challenge difficulty as the essential element in effective learning and engagement. The theory of the zone of proximal development suggests that the key to learning is finding materials just outside of learners current skill level, and the theory of flow suggests that challenges at just the right difficulty level lead to increased motivation. Videos span a vast spectrum of difficulty, and videos at the right level of challenge for a learner may be missing captions. Given that having the right video is important for learner engagement and learning, it is important that learners have the ability to learn through any video.

Traditionally, if a learner wants to learn from a video without foreign language captions available, they would need to get those captions from an expert caption creator (either a native speaker or longtime learner). Furthermore, existing caption creation systems are tedious to use and require extensive training to be effective [37]. This means that language learning communities are forced to rely on the select few experts for caption generation. To overcome these restriction, we ask, what if learner communities could build their own captions while learning from and enjoying the video in the process?

We present a novel system where learners are given machine generated captions and caption editing tools. When learners notice mistakes in a caption, they can edit the caption on the video. We conducted a lab study to evaluate learning and engagement, and compared editing methods. The results suggest that there were similar amounts of language learning across conditions, despite participants making many corrections to the captions in the imperfect captions condition. This finding suggests that a caption correction task does not impair learning, so learners can improve their language skills while helping to build video caption learning resources. This opens up a broader space of possible applications that allow learners to improve their language skills while helping build shareable learning resources.

RELATED WORK

Many researchers have investigated adapting existing technology and developing new technology for use in foreign language learning [36, 21, 9, 12]. For example, in the human-computer interaction community, work has looked at using location-informed vocabulary learning [8], desktop wallpapers for passive vocabulary learning [6], mobile games for Chinese tone learning [18], and utilizing natural waiting periods, such as awaiting chat replies, for vocabulary learning

CSCW'2017, February 25-March 01, 2017, Portland, OR, USA

DOI: http://dx.doi.org/10.1145/2998181.2998268

[2]. These systems are helpful in learning many language mechanics such as vocabulary and grammar. However, in the long term, it is important for learners to "hear large quantities of speech" and "hear many different native speakers" [11] in order to understand speech effectively. Furthermore, hearing speech in multiple contexts is essential for learners to grasp the nuances of communication [28].

Video learning

Given the importance of listening to large quantities of foreign language audio, learning through watching foreign language videos has become a common practice for both independent learners and classroom learners. Video learning fits well with communicative language learning approaches as it emphasizes high level understanding and foreign videos offer windows into foreign cultures [10]. Ethnographic research in education has indicated that video learning "engage[s] students in language-based tasks and cultivate[s] collaboration and creativity" [16]. Furthermore, in classroom experiments, Secules et al. found video learning to be effective at improving students' high-level comprehension skills and inference skills [30].

Subtitles play a critical role in video learning. Winke et al. compared captions to no-captions in 4 different languages and found significantly more learning in captioned conditions [38]. The language that subtitles are in is especially important. For example, Ina [25] showed that incidental learning of vocabulary is more likely to occur when watching a foreign language program with subtitles in the learner's native language than watching with translated subtitles. Other studies in video learning have shown that watching videos with both audio and captions in the foreign language is an even more effective way to learn. For example, Mitterer and McQueen found that subtitles in the video's original language increased speech learning when compared to subtitles in the learner's native language [27]. These studies all suggest that foreign language videos with foreign language captions are an effective video learning method. Furthermore, the findings that subtitles in the original language of the video are better than those in the learners native language provide an opportunity for technology support as technology for speech transcription is currently far more robust than technology for speech translation.

The ways in which captions are displayed also affects language learning. For example, Perez et al. tried displaying only keywords to participants, but found that participants strongly preferred complete captions over keyword captions [29]. Kovacs and Miller developed system to display target language subtitles with hover definitions and found more vocabulary learning using this interface when compared to simultaneous native and foreign language subtitles [23]. A language learning website, FluentU[1], uses a similar setup for commercial language learning. These systems show that augmented subtitles can be an effective learning method, but they require perfect transcripts and translations to function, which limits the videos that these systems can be used on.

[1] http://www.fluentu.com/

Generating subtitles can be a language learning task in itself. This task has been shown to be helpful to students because students "feel they play an active role in their own learning process" and gives students a "familiar and motivating context" [39]. Williams and Thorne [37] developed a course for teaching subtitle generation as a learning task and found it to be highly effective, but learners required extensive training (2 days) before they were able to begin the task. Furthermore, learners involved in the study already had several years of language learning experience. To avoid the need for extensive training we designed our system with a low barrier for entry.

Crowdsourced and machine assisted caption generation

Many systems have been developed to assist humans in generating captions. For example the Re-speak system has professionals repeat the voice they are hearing to improve the machine generated output [20]. Lasecki et al. have also developed a system where listeners type part of a speech that they are listening to in real time and then aggregates the input of several transcribers to create a transcript [24].

Although imperfect, even raw machine generated captions can be helpful. Studies of deaf accessibility have also shown that captions, even those generated with speech-to-text systems, improve accessibility [33]. Other work by Gaur et al. examines use of machine transcripts as a starting point for generating correct transcripts [15]. They found that if the quality of the transcript is high enough, the machine transcript can reduce latency in correct transcript generation. This indicates that we should carefully consider the accuracy of the transcript when designing captioning systems. However, unlike most existing systems, what's important to us in the development of the current system is not the speed of captioning. Rather, it's important to us that learning is occurring during the captioning process and for that task to be engaging for learners.

Crowdsourcing and learning

The idea that learners can be used to crowdsource learning content has been shown to be promising approach in building new learning support systems.

Work by Kim et al. has shown learners can be prompted for information that can be used for improving learning conditions for other learners in how-to videos [22]. In their study, learners were prompted to generate summaries of segments of learning videos for the generation of step-by-step annotations. However, some of the challenges in foreign languages and caption generation require us to approach the problem differently. First, because learners have only partial language knowledge, we cannot expect that learners can understand all of the video, and therefore regular prompts are not an appropriate mechanism for learners to participate in the caption generation process. Furthermore, this work showed that the end product is useful, but has not shown that learners can learn through the content generation process.

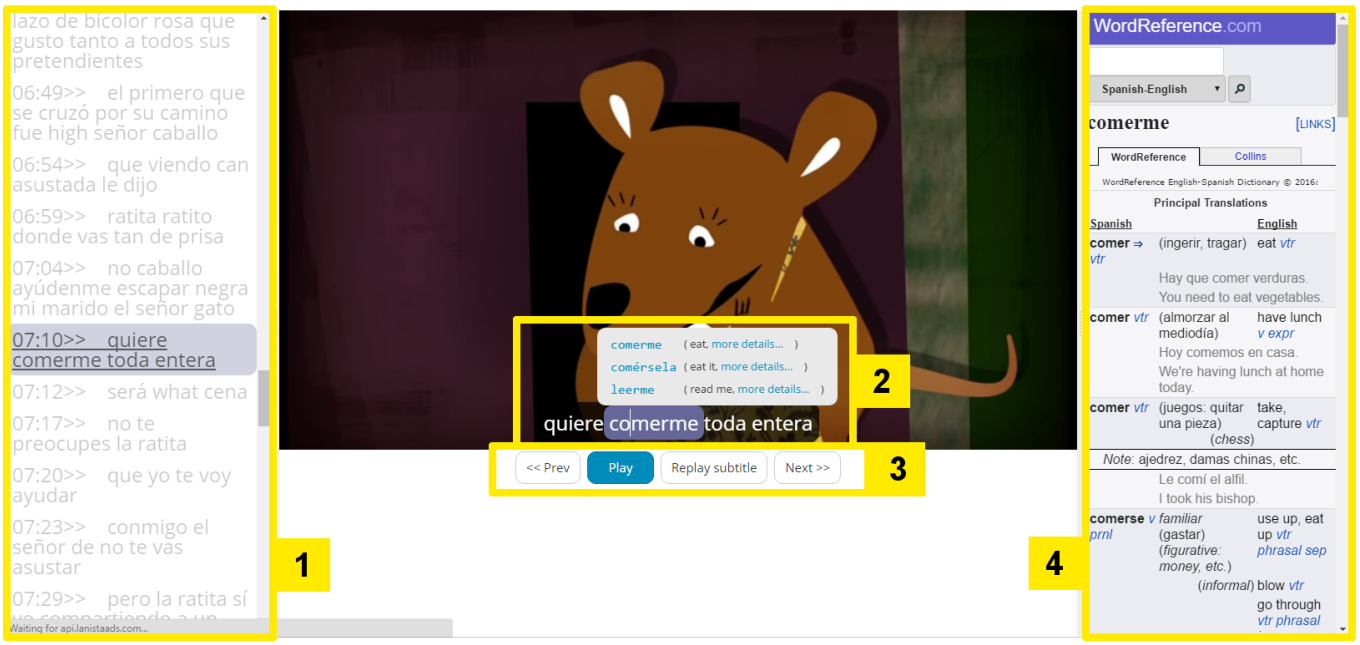

Figure 1. The caption video viewing and caption editing interface. Users can navigate through the video using the transcript (1), look at word definitions and edit the caption directly on the video (2), toggle video playback and jump one caption forward or back (3), and view expanded dictionary information using an external dictionary website (4).

The website DuoLingo[2] uses language learners to translate foreign language material on the web while learning in the process [34]. However, the lessons section of DuoLingo lacks context and culture around the material. By using videos as learning material, learners have rich context to situate their learning. Furthermore, DuoLingo lacks natural audio, which can be important for learning prosody of a target language. The captions that are generated in our system can also be used to improve learning for the rest of the language learning community.

In summary, our literature review shows that foreign language learning through videos with captions is an effective learning method, but working only with existing captioned video severely limits the effectiveness of video learning due to limited choice in finding videos that fit learners' personal preferences and difficulty level. Other work has explored different methodologies of generating captions, but all of these methods either require expert captioners or crowds of native speakers. Drawing inspiration from work that enables learners to contribute to building learning resources, we built a system to allow learners to build complete captions with and for other learners while learning during the process.

INTERFACE DESIGN

The high-level goal of our system is to enable learning while participating in the captioning process. However, captioning is a difficult and time-consuming task. For example, for a learner beginning a subtitling task from scratch, "a single 5-min excerpt may take many hours for a novice to complete" [37]. Novice learners are handicapped by having incomplete or no knowledge of the caption language. Thus the key design focus is on minimizing the effort required to edit captions and providing enough scaffolding to support novice learners.

The system was developed as a website to maximize accessibility. A screenshot of the system along with descriptions of each element are shown in Figure 1. On the site, learners select a video and captions are generated using the IBM Watson speech-to-text system[3]. The video is shown in the center with the captions overlaid on top of the bottom-center of the video. To edit the caption, learners simply click anywhere in the caption area to begin typing and changing words. If the speech-to-text system produces multiple possible words for an audio segment, these words will all be shown above the word containing the cursor (for example "comerme," "comersela," and "leerme" in area 2 of Figure 1). The word selected by the cursor will also show a translation generated by the Google Translate API, and the word can be searched in a dictionary by clicking a link "more details" next to the translation. Additional dictionary information is then displayed to the right of the video (area 5). If alternative words are displayed, each word will have a translation and dictionary link. Buttons (area 3) are displayed to allow users to move one caption forward or back and a transcript on the side (area 1) can be used to jump to a specific caption.

Producing words from scratch can be especially challenging for learners who may be unfamiliar with typing in that language. By beginning with the a machine generated transcript, we greatly reduce the number of words that a learner would need to produce from scratch. We further reduce this by providing machine generated alternative words. Instead of typ-

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

[3]http://www.ibm.com/

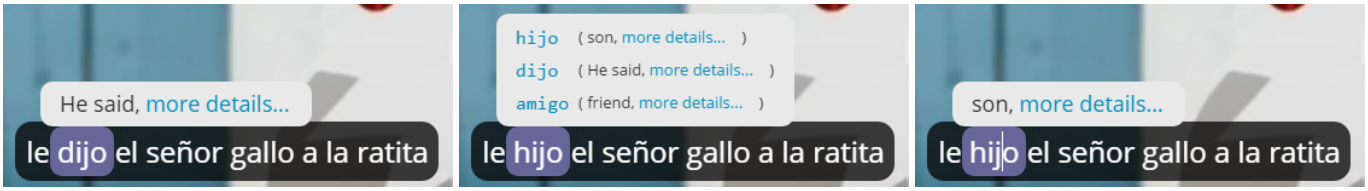

Figure 2. The 3 conditions used: (A) accurate captions, (B) imperfect captions with suggested alternative words correction task, and (C) imperfect captions with free response correction task. When participants clicked words in the subtitle, the word was highlighted and suggestion or translation information was displayed. In condition B, participants clicked the alternative Spanish word from the popup box to edit the caption. In condition C, participants typed words directly into the black area where the caption is displayed. In all cases, the displayed translation was generated with the Google translate API, and the 'more details...' link searched the selected word in the external dictionary.

ing the correct word, learner can instead just choose the correct word from a list.

Furthermore, word-by-word translations along with the video context enable learners to piece together the meaning of the caption. Work by Kovacs and Williams have shown that providing English translations for individual words can be helpful to learners to establish overall meaning [23]. However, because words can have slightly different meanings or nuances depending on the context, the translation will not always be entirely accurate. Therefore, we provide quick access to a full dictionary so that learners can view multiple possible translations and disambiguate the meaning.

LAB STUDY: ASSESSING LEARNING WITH DIFFERENT CAPTIONING WORKFLOWS

As stated earlier, one of primary objectives in developing this system was to ensure that learning is occurring during the captioning process and that the task is engaging for learners. Thus we conducted a lab study to evaluate the system. We had three key questions. First, (RQ1) are there significant differences in learning outcomes between using accurate vs. imperfect captions? If learning is significantly impaired by imperfect captions, then the system will not be a useful tool for language learners, and bringing captions from an external source may be more effective. Second, (RQ2) does the quality of the learning experience significantly differ between learning with accurate vs. imperfect captions? Because the system will be learner driven, it's important that quality of experience is not significantly impacted by imperfect subtitles or learners will drop out of the system. And finally, (RQ3) does the complexity of the correction interface affect learning outcomes and quality of learning experience? If complexity of the editing task significantly influences learning or quality of experience, we can pursue designs that simplify the editing task.

To establish our expectation for RQ1 (*are there significant differences in learning outcomes between using accurate vs. imperfect captions?*), we consider the difference between active and passive learning. In the accurate caption condition, learners only need to understand the video, so we consider it a passive learning task. On the other hand, learners in the imperfect caption condition need to edit the caption to resolve disparities between what is being said and the caption, so we consider this an active learning task. Work by Michel et al.[26] has shown that active learning in the classroom is more effective than passive learning. Furthermore, specifically in the context of language learning, Gu and Johnson [17] found that Chinese learners of English were more successful in learning vocabulary when using active learning strategies. Therefore we expected that there would be more learning in the imperfect caption condition than the accurate caption condition.

To establish our expectation for RQ2 (*does the quality of the learning experience significantly differ between learning with accurate vs. imperfect captions?*), we looked at the work of Perez et al. [29] which looked at the use of keyword-only captions (where only keywords were displayed) in comparison with complete captions for foreign language learning. They found that the keyword-only captions and complete captions result in similar learning gains, but participants reacted negatively to the keyword-only captions. Therefore, we expected that while participants would learn from watching video with imperfect captions, they will perceive the quality of their video learning experience as worse when presented with imperfect subtitles.

To establish our expectation for RQ3 (*does the complexity of the correction interface affect learning outcomes and quality of learning experience?*), we looked at work by Cades et al. [1] which has shown that the interruption task complexity affects primary task performance. Therefore, we expected that a less complex caption editing method (e.g., one in which selectable options are offered) will result in better performance than a more complex editing method (free open-ended response).

Experiment design

To assess the effects of different captioning systems, this study used a between-subjects design with 3 conditions. As shown in Figure 2, the three conditions used were: (A: accurate) accurate captions, (B: suggested-alternative) imperfect captions with suggested-alternative-word editing, and (C: free-response) imperfect captions with free-response editing. In all conditions, participants viewed a Spanish language video entitled *La Ratita Presumida* and the dictionary link searched the word in SpanishDict[4]. The same video was used in all conditions to control for differences in interest and difficulty that may come with different videos. In the accurate caption condition, participants were presented with error-free Spanish caption for the video. This served as the baseline condition. In the suggested-alternative imperfect

[4]http://www.spanishdict.com/

	Accurate		Suggested-alternative		Free-response			
	M	*SD*	*M*	*SD*	*M*	*SD*	*F*	*p*
Vocabulary change	3.44	2.48	1.88	1.93	2.40	2.23	2.15	.128
Reading speed ratio change	-0.27	0.30	-0.27	0.19	-0.20	0.25	0.44	.650
Effectiveness perception	4.33	1.20	3.77	1.46	3.75	1.42	1.00	.376
Video interesting and engaging	4.56	1.29	3.84	1.14	4.03	1.06	1.69	.197
Frustration	2.89	1.23	3.94	1.77	3.36	1.34	2.19	.089
Comprehension	10.72	6.01	4.81	5.90	6.39	4.75	4.98	.011
Difficulty perception	4.25	1.62	4.03	1.38	3.62	1.34	0.71	.497
Time on video	15.70	7.83	18.32	10.15	20.00	9.77	0.91	.409
Edits	-	-	53.31	58.12	42.13	39.91	1.18	.316

Figure 3. Mean Ratings/Score (Standard Deviations) of Quality of Experience Measures

{ una, a, al, ha }
{ errática, ratita, ratito, regatista, raquítica }
{ quiero, quiere, quédate }
{ pero, pino, pedro, pido, peer }

Figure 4. Sample errors generated by speech-to-text system

caption condition, participants were presented with caption that contained errors and offered a list of phonetically similar alternatives they could choose from to correct the caption. Word options for this condition were generated by the IBM Watson speech-to-text system[5]. In the free-response imperfect caption condition, participants were presented with the same imperfect caption, but instead of having selectable options, they had to type words to correct the caption. For both the suggested-alternative and free-response imperfect caption conditions, the imperfect caption was generated by beginning with a correct transcript and introducing a 19% word error rate, where the incorrect words were drawn from the machine generated word alternatives.

```
for each word in correct transcript:
 rand := random number between 0 and 1
 if rand < 0.2:
  add random incorrect from ASR at word
                        to bad transcript
 else:
  add word to bad transcript
```

A maximum of four alternatives were displayed, but as few as one alternative would be displayed if the speech-to-text system did not identify any alternative possibilities for a given word. Sample errors are shown in Figure 4.

In all conditions, participants were told that the captions were machine generated and may contain errors. We did not tell participants the purpose of the subtitle correction task because we wanted to control for framing effects in measuring engagement and understanding intrinsic motivation.

Experiment procedure

We used the following procedure. After giving assent, participants were asked to sit in front of a laptop with headphones on. Participants completed a brief survey about their demographic information and Spanish learning background, along with a Spanish vocabulary and reading speed test. Then an experimenter explained to participants how to use the caption editing interface. The experimenter told participants that they were free to edit the captions if they wished and could take as long as needed to watch the entire 11 minute video. While watching, participants could click the Spanish words in the captions to see their English translations and could edit mistakes in the captions. After completing the video, participants were automatically redirected to a post-test on vocabulary, reading speed, and comprehension. Finally, participants completed a survey about their experience.

[5]http://www.ibm.com/

Participants

Participants were recruited using a university recruitment system and email lists. Native speakers of Spanish were not permitted to participate. A total of 54 participants (55% female) were recruited and 49 participants were included in the final analysis. The 5 participants that were not included experienced technical difficulties which invalidated their data. The age of participants ranged from 18 to 39. Participants were randomly assigned to conditions (accurate [A] = 18, suggested-alternative [B] = 16, free-response [C] = 15).

RESULTS

Measures

Learning measures

Learning was measured using a vocabulary translation test and a reading speed test.

In the vocabulary test, 20 Spanish words were picked randomly from the video and learners were asked to type the English translation for each word. The same 20 words were used on the pretest and posttest, but the order was randomized. A pretest and posttest score were calculated by counting the number of words correctly translated. Number of words learned was calculated by subtracting the pretest vocabulary score from the posttest vocabulary score.

For the reading speed test, a standard sentence processing test was used [13]. Words appeared one word at a time, and in the end the learner was asked whether the sentence they just read was correct or not. This is a useful measure for testing language ability, because, like listening to speech, learners

Pearson Correlations

	Self-rated Spanish skill	Vocabulary (pre)	Comprehension	Reading speed ratio change	Effectiveness perception	Frustration	Video interesting and engaging	Vocabulary (change)	Edits
Self-rated Spanish skill	—	0.721***	0.482***	-0.360*	0.460***	-0.215	0.076	0.394**	0.174
Vocabulary (pre)		—	0.377**	-0.338*	0.408**	-0.088	0.167	0.433**	0.111
Comprehension			—	0.031	0.250	-0.259	0.276	0.473***	0.299*
Reading speed ratio change				—	-0.259	0.249	-0.116	0.053	0.079
Effectiveness perception					—	-0.323*	0.370**	0.263	0.134
Frustration						—	-0.253	-0.338*	-0.009
Video interesting and engaging							—	0.137	0.060
Vocabulary (change)								—	0.121
Edits									—

* p < .05, ** p < .01, *** p < .001

Figure 5. Table of correlations for measures of interest.

must process words as they read them or they will be unable to understand the entire sentence. We calculated reading speed on 30 sentences (15 from the video and 15 not in the video) before and after the video captioning task. Sentences not from the video were included because, unlike vocabulary learning, we expect that grammar learning from the video should generalize beyond just the sentences from the video. Because individuals have differences in general reading speed, we used the percent change in reading speed as our reading speed learning measure.

Quality of experience measures

Four measures were used to gauge the learner's quality of experience. For the first three measures on participants' perceived quality of experience, participants rated on seven-point scales (describes the task... "not well at" [1] to "extremely well" [7]) how effective for language learning they found the task to be, how engaging and interesting they found the video to be, and how frustrating they found the task to be. Finally, we measured comprehension using 20 true/false/unsure questions about the details of the video story. The comprehension score was calculated as the number of incorrect responses subtracted from the number of correct responses, and "unsure" answers were ignored. We used this scoring system because it gives us a more precise measure by discouraging guessing in cases where the learner is uncertain.

Editing behavior

In order to measure caption editing behavior, we measured the number of times the learner changed a word. This change could either be through clicking on a menu interface (suggested-alternative condition) or typing (free-response condition). Furthermore, we measured whether this change was correct or not.

Final caption accuracy measures

Two final sets of captions were generated by aggregating the changes made by participant groups in the suggested alternative condition (B) and the free response condition (C). In each condition, the final transcript was generated by beginning with the incorrect caption set. Then, for each word, if no participant made a change to that word, the word remained the same as the one in original incorrect caption set. If at least one participant made a change, the new word was determined by taking the change made by the majority of participants. In the case of a tie, a word from the tied-majority words was chosen pseudo-randomly. Finally, a word error rate (WER) was calculated for each final transcript. The average improvement made by each participant was also measured. Note that this is not equal to the final WER divided by the number of participants because many participants made the same changes.

Furthermore, in order to better understand the viability of the system for novice learners, we looked at the final accuracy of captions generated by just accounting for changes made by learners who reported little to no Spanish experience (1 on a 7-point scale).

Findings

A table of results are presented in Figure 3 and learning outcome graphs are shown in Figure 9. Participants in all conditions showed significant ($p < .001$) improvements in reading speed between the pre- and post-test. The mean percent change was -26.9% for the accurate condition, A, ($SD = .30$), -27.4% for the suggested-alternative condition, B, ($SD = .19$), and -19.6% for the free-response condition, C ($SD = .25$). A one-way ANOVA indicated that the differences between conditions were not statistically significant for percent change in reading speed, $F(2, 45) = 0.44$, $p = .650$, partial $\eta^2 = .019$. Participants in all conditions also showed significant ($p < .001$) improvements in vocabulary scores between the pre- and post-test. Thus we replicated earlier findings that

word	Acc.	Sug.-alt.	Free-res.	Correction
ayúdeme	22%	18%	20%	no
barrio	33%	12%	13%	yes
caballo	11%	18%	13%	no
cada	11%	6%	26%	yes
casarte	16%	6%	26%	yes
cerca	5%	0%	6%	yes
corriendo	16%	6%	20%	yes
cruzó	0%	6%	0%	no
encontró	11%	6%	6%	no
entera	38%	12%	20%	no
lazo	55%	37%	20%	yes
llegaron	0%	0%	6%	no
marido	22%	25%	33%	no
paseando	0%	0%	13%	yes
pendientes	0%	0%	0%	yes
presumida	66%	31%	13%	no
prisa	11%	18%	20%	yes
sacó	0%	6%	6%	yes
uñas	27%	6%	0%	yes
viendo	11%	6%	0%	no

Figure 6. Percent of participants that learned each of the words on the vocabulary test.

video learning leads to learning gains. Participants on average increased scores by 3.44 points in the accurate condition, A, (SD = 2.48), 1.88 points in the suggested-alternative condition, B, (SD = 1.93), and 2.40 points in the free-response condition, C (SD = 2.29). A one-way ANOVA indicated that the differences between conditions were again not statistically significant for change in vocabulary score, $F(2, 46) = 2.15$, $p = .128$, partial $\eta^2 = .085$. These results suggest that as expected, learning, in terms of reading speed and vocabulary improvements, was not significantly different with either accurate or imperfect captions. In the suggested-alternative condition, 17% of words learned were learned after the participant edited that word and in the free-response condition, 15% of words were were learned after the participant edited that word. Making a correction did not increase the probability of learning a word.

Participants also largely reported similar quality of the learning experience across conditions (see Figure 3). One-way ANOVAs indicated that the differences between conditions were not statistically significant for how effective participants found the task to be for learning, how engaging/interesting they found the video to be, and how frustrating they found the task to be. However, a statistically significant difference between conditions was found for video comprehension, $F(2, 44) = 4.98$, $p = .011$, partial $\eta^2 = .185$. Post hoc comparisons using the Tukey HSD test revealed that the mean video comprehension score was significantly higher with accurate caption in Condition A (M = 10.72, SD = 6.01) than with imperfect caption in Condition B (M = 4.81, SD = 5.90), p = .011. Descriptively, the mean video comprehension score was also higher with accurate caption in Condition A than with imperfect caption in Condition C (M = 6.39, SD = 4.75); this difference however was not statistically significant, p = .100. Quality of experience graphs are shown in Figure 10 These results suggest that contrary to expectation, participants did not perceive the quality of their learning experience to be significantly worse when presented with imperfect captions, although the imperfect captions seemed to have negatively affected their comprehension of the video.

No statistically significant differences were found for any of the learning or quality of experience measures between Condition B and C in the above ANOVAs and if applicable, post hoc comparisons. This suggests that the complexity of the caption editing method, whether with selectable options or requiring open-ended responses, did not seem to affect participants learning outcomes or experience in this study.

Results of the accuracy analysis are shown in Figure 7 for all learners and Figure 8 for novice learners. Aggregating learners' corrections overall improved final caption accuracy by 13.5% (from 19% to 5.5% WER) in the free response condition, and an even larger improvement of 17.2% was observed in the suggested-alternative condition (from 19% to 1.8% WER). While no improvement was observed for novice learners in the free response condition, a 10.1% improvement was observed in the suggested-alternative condition (from 19% to 8.9% WER). These results suggest that our caption correction system was effective in producing accurate captions from aggregated learner inputs, and that novice learners were able to benefit more from the scaffolding feature of suggested alternative words.

To gain insights for system improvements, we conducted correlation analyses to explore potentially important relationships (see Figure 5). One relationship of interest that emerged is a moderate correlation between measures of preexisting language ability and how effective participants found the the interface for learning (r = 0.460), as well as vocabulary learning (r = 0.394) and comprehension (r = 0.482). This indicates that, according to our measures, the learning task was generally more effective for learners with some experience.

DISCUSSION

Research questions

No significant differences in learning outcomes were found using accurate vs. imperfect captions (RQ1)

There were no significant differences between learning outcomes across conditions, but all conditions showed evidence of learning. Therefore, we conclude that while the caption editing task did not improve learning, it also did not impair it. Although this does not match our expectation, this finding aligns with the findings of Semke [31] that shows corrections to writing do not influence writing improvement, but rather the act of writing leads to improvement. Semke [31] found that having students correct their own foreign language writing rather than having teachers explicitly marking errors made no difference in writing improvement over a 10 week period. Similarly, learning by our participants was likely driven by the act of watching the foreign language video rather than the act of correction.

It is also possible that learners in all conditions were actively learning. Although we initially expected that learners in the imperfect caption conditions would be learning more actively,

Condition	N	Final WER	Avg. WER Improvement	Correct edits	Incorrect edits
Suggested-alternative	16	1.8%	4.1%	634	50
Free-response	15	5.5%	1.6%	241	31

Figure 7. Final caption accuracy for all learners

Condition	N	Final WER	Avg. WER Improvement	Correct edits	Incorrect edits
Suggested-alternative	8	8.9%	3.3%	251	23
Free-response	7	19%	0%	0	5

Figure 8. Final caption accuracy for novice learners

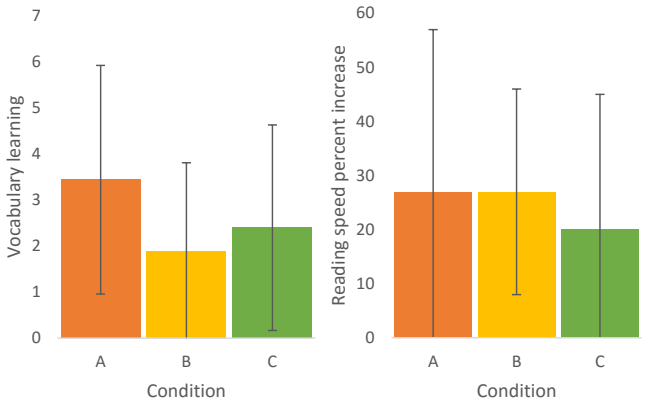

Figure 9. Learning measures are shown above for the three conditions (A: accurate captions, B: suggested-alternative imperfect captions, C: free-response imperfect captions). We found no significant difference between learning measures across conditions. Each error bar indicates +- 1 standard error.

it likely that although learners in the accurate caption conditions did not need to make edits, learners still needed to work actively to understand the story. This finding suggests that in distinguishing between active and passive learning tasks, the structure of the task is less important than the mental processes of the learner. Because the study took place in the lab, we expect that all of the learners would feel obligated to engage with the task, so it is possible that we would see differences in learning activity in less controlled conditions.

Despite having no significant difference across conditions, learning did occur in all conditions. Therefore, we believe that our system can enable learners gain language skill while contributing to the generation of accurate video captions.

The quality of the learning experience did in some ways significantly differ between learning with accurate vs. imperfect captions (RQ2)

We found that comprehension was lower in the suggested-alternative (B) condition and free-response (C) condition than accurate caption (A) condition, but the difference was only significant between the accurate caption (A) condition and suggested-alternative (B) condition. No significant differences were found across conditions for other quality of experience measures. This means that although learners with accurate captions understood more, they felt similarly engaged and frustrated. We suggest that the loss of comprehension with imperfect captions is due to the increased focus on low-level details in the editing conditions. Learners needed to focus on individual words rather than the high level story in order to complete the task.

In future work, it will be important to explore whether learners feel a loss of understanding is detrimental to the usefulness of the system. If the primary goal of the learner is to gain language knowledge, this may be acceptable, but in scenarios where complete comprehension is important, the system may not be appropriate. It may also be useful in the long-term to alternate between focusing on high-level comprehension and low-level attention to detail.

We also looked at comments written by frustrated participants describing why they felt frustrated. Because we only gave participants a limited picture of the overall system, many participants were confused by the presence of the errors or felt they could have learned more with accurate captions. For example, one participant wrote *I didn't really understand the whole "some subtitles aren't correct" thing.* In future work, we plan to explore ways to make the rationale and benefits of the system more clear to learners. However, it should be noted that while some participants indicated frustration with the editing task, we did not measure any statistically significant overall differences in frustration between correct and incorrect caption conditions.

The complexity of the correction interface did not affect learning outcomes and quality of learning experience (RQ3)

We found no significant difference between the two interface setups that we tested. More edits were made in the C condition (open ended editing) than the B condition (suggested alternative editing), but the difference was not significant. Originally we expected that the suggestion interface would offer a simpler and less taxing way to edit the caption, but our study suggests that there is no difference between this and open ended editing. We believe that this may have been because additional time needed to be spent reading the alternative words and their definitions which turned out to be roughly equivalent to the amount of time that was needed to type the word.

Beginner and intermediate learners

Our exploratory correlation analysis revealed that the learning task was generally more effective for learners with some experience. However, work in language acquisition has shown us that learners need to learn to parse and chunk sounds in a language before they can learn higher order processing such as vocabulary learning [3]. Although we did not measure

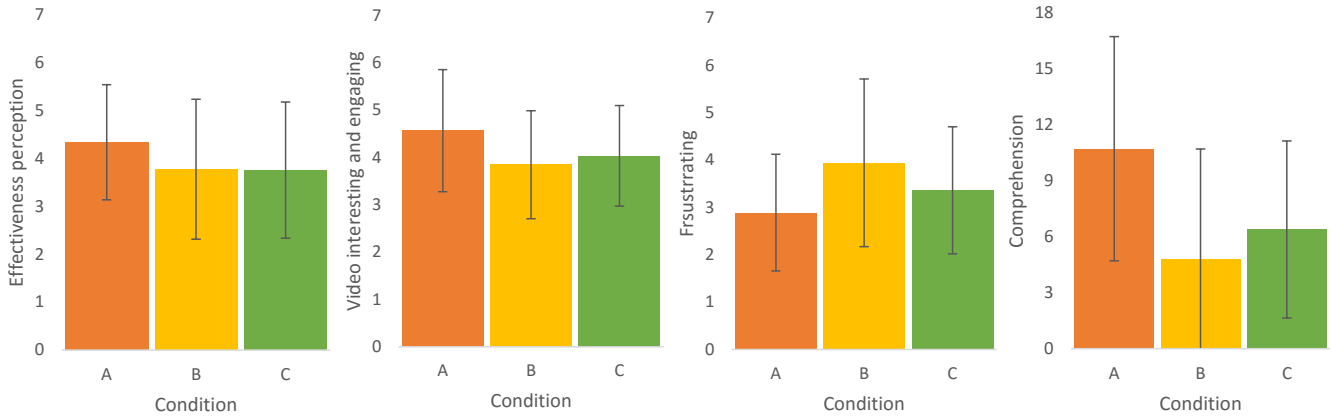

Figure 10. Quality of experience measures are shown above for the three conditions (A: accurate captions, B: suggested-alternative imperfect captions, C: free-response imperfect captions) with each error bar indicating +- 1 standard error. Significant differences were only found in the comprehension score.

this, learners were listening and trying to understand the foreign language audio, so naturally they would gain some skill in sound processing through this task. Furthermore, despite not remembering all of the vocabulary, beginning learners reported similar interest and engagement with watching the video. Although future work will need to tease out the nuances of video learning and skill level, we feel that our findings indicate that with adequate scaffolding video learning can be productive and engaging for learners of any level.

Our accuracy analysis showed that novice learners made correct edits in the suggested-alternative condition, but participants were unable or unwilling to make correct edits in the free-response condition. We believe that the additional scaffolding provided by the suggested-alternatives gave novice learners enough confidence to make changes to the captions.

Motivations for caption editing

In order to guide future design, we explored what motivated learners to edit captions using comments from our survey. In the survey we asked participants: "what motivated you to edit the subtitles?". While some participants indicated that the research study was the primary motivator, others noted that the errors in the captions prevented them from understanding the story. In order to better understand the video, they corrected the caption and read through it again. Others made edits because they were bothered by the mismatch of audio and text, or the story context and the caption meaning. This suggests that these participants found satisfaction in having a correct end-product. Both of these orientations suggest different design approaches. For example, if the primary motivator is understanding the story, additional scaffolding could be provided to learners to help them reach a complete understanding of the video. On the other hand if learners are motivated by perfecting the caption, parts of the caption that are likely to contain errors could be made more salient.

Corrections by learners

Our final caption accuracy results suggest that our caption correction system can produce accurate crowd-sourced captions, and that the scaffolding feature of suggested alternative words provided critical support for novice learners. Despite the fact that learners did not know the words and grammar beforehand and the errors were not highlighted in any way, even learners with no experience were able to make caption corrections in the suggested-alternative condition. The responses to the post-task survey indicated that participants used a combination of cues to establish the existence of an error and then to make a correction. Participants used word translations along with the audio context ("many words did not align with what the narrator was saying"), the visual context ("the sentence did not quite express what was happening in the visual scene"), and the narrative context ("get [the caption] back on track with the story"). These results indicate that with adequate scaffolding and the right video, even absolute novices can learn and contribute to caption generation.

LIMITATIONS

Although the system is intended for any video, it should be noted that the same video was used for all participants. We expect that the match between a learner's skill level and interests with a video will influence learning and quality of experience. The video that we did select was created for children and contained a lot of repetition. Some participants pointed out that this helped them with learning. While the use of just one video does not adequately reflect how we intend the system to be used in practice, our finding, that measurable learning gains occurred despite the use of just one video, is providing us with additional confidence about the effectiveness of the system for language learning. Future work should more carefully examine the effect of video difficulty and fit on learning outcomes.

Learning through caption correction was a novel method for all participants and, given the short duration of the study, learners may not have totally acclimated themselves to this learning method. Although participants did not indicate difficulty learning to use the interface, learning effectively through correction may require a shift in learning strategy. Future work should examine whether the system is effective over longer periods of time, and how utility perception and engagement change with time.

Learners may have been primed to learn the words from the pretest. However, this would not affect differences between conditions as all participants were primed in exactly the same way. Furthermore, this type of priming would not be unusual in real use scenarios. Often teachers prepare students with vocabulary lists before introducing new dialogues, and we could easily implement similar learning methods with our system by asking learners to pay special attention to words of interest in the video. We have added discussion for this to the limitations section.

Furthermore, the insignificant learning differences across conditions may have been in part due to our small sample size. The power of our study was too low to detect potential small differences. However, it should be noted that learning was measured to be significant (pre to post) in all conditions. Future work should further explore the difference between learning while editing captions and learning through other video engagement methods with a large sample.

Finally, measuring language learning is incredibly challenging given the complex, intertwined processes that are necessary for language comprehension. Williams and Thorne identified abilities to "listen attentively, recognize and fully absorb the content" [37] were essential for students to effectively produce subtitles. Our measures of vocabulary, reading speed and comprehension provide only limited windows into the the learning processes that are taking place. Future work should investigate other ways to evaluate language learning through captioning.

FUTURE WORK

This work indicates that a crowdsourced learning system for captioning is a promising direction, but there are still many facets of the system that need to be designed and explored. There is extensive theoretical and laboratory work investigating the effect of freedom of choice on motivation, but with this system we have the potential to explore this idea further in the field.

Furthermore, given the different low-level motivations that learners had for correcting captions, future designs should explore how and when to emphasize different aspects of the correction process. For example, we could draw from work on task-breakdown to give learners tasks that better fit with their learning style.

There is also potential to build in more community features such as simultaneous editing in order to emphasize the community-driven nature of the system and create opportunities for learners to transfer knowledge to other learners.

CONCLUSION

In this work, we presented a system to enable foreign language learners to learn while correcting video captions that could then be used by a wider learning community. Our findings suggest that although learners' comprehension was reduced by the editing task, it did not influence learning or engagement with the video. Comments from our usability indicated that contributing to building captions was motivating for some learners. Given this combination of findings, we envision a system where learners motivated by social contribution and learners motivated primarily by their own learning could both learn and contribute by collaborating on generating captions. Where previously learners would need to rely on external sources to generate captions for learning videos, we see this work as first step in building tools to allow foreign language learning communities to become self-sufficient.

REFERENCES

1. David M Cades, Nicole Werner, Deborah A Boehm-Davis, J Gregory Trafton, and Christopher A Monk. 2008. Dealing with interruptions can be complex, but does interruption complexity matter: A mental resources approach to quantifying disruptions. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, Vol. 52. Sage Publications, 398–402.
2. 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.
3. Morten H Christiansen and Nick Chater. 2015. The now-or-never bottleneck: A fundamental constraint on language. *Behavioral and Brain Sciences* (2015), 1–52.
4. John Condry. 1977. Enemies of exploration: Self-initiated versus other-initiated learning. *Journal of Personality and Social Psychology* 35, 7 (1977), 459.
5. Mihaly Csikszentmihalyi. 1996. Flow and the psychology of discovery and invention. *New Yprk: Harper Collins* (1996).
6. David Dearman and Khai Truong. 2012. Evaluating the implicit acquisition of second language vocabulary using a live wallpaper. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1391–1400.
7. Edward L Deci and Richard M Ryan. 1981. Curiosity and Self-Directed Learning: The Role of Motivation in Education. (1981).
8. 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.
9. Nick C Ellis and Pamela SH Bogart. 2007. Speech and language technology in education: the perspective from SLA research and practice.. In *SLaTE*. 1–8.
10. Rod Ellis. 2003. *Task-based language learning and teaching*. Oxford University Press.
11. Maxine Eskenazi. 1999. Using a computer in foreign language pronunciation training: What advantages? *Calico Journal* (1999), 447–469.
12. Maxine Eskenazi. 2009. An overview of spoken language technology for education. *Speech Communication* 51, 10 (2009), 832–844.

13. Thomas A Farmer, Jennifer B Misyak, and Morten H Christiansen. 2012. Individual differences in sentence processing. *Cambridge handbook of psycholinguistics* (2012), 353–364.

14. Thomas J Garza. 1991. Evaluating the use of captioned video materials in advanced foreign language learning. *Foreign Language Annals* 24, 3 (1991), 239–258.

15. 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.

16. Jason Goulah. 2007. Village voices, global visions: Digital video as a transformative foreign language learning tool. *Foreign Language Annals* 40, 1 (2007), 62–78.

17. Yongqi Gu and Robert Keith Johnson. 1996. Vocabulary learning strategies and language learning outcomes. *Language learning* 46, 4 (1996), 643–679.

18. 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.

19. Carol Herron, Bastien Dubreil, Cathleen Corrie, and Steven P Cole. 2002. A Classroom Investigation: Can Video Improve Intermediate-Level French Language Students' Ability to Learn about a Foreign Culture? *The Modern Language Journal* 86, 1 (2002), 36–53.

20. Toru Imai, Atsushi Matsui, Shinichi Homma, Takeshi Kobayakawa, Kazuo Onoe, Shoei Sato, and Akio Ando. 2002. Speech recognition with a re-speak method for subtitling live broadcasts.. In *INTERSPEECH*.

21. Hee-Jung Jung. 2003. Overview of computer assisted language learning research with second language acquisition perspectives. *Teaching English with Technology* 3, 3 (2003), 3–15.

22. Juho Kim, Phu Tran Nguyen, Sarah Weir, Philip J Guo, Robert C Miller, and Krzysztof Z Gajos. 2014. Crowdsourcing step-by-step information extraction to enhance existing how-to videos. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems*. ACM, 4017–4026.

23. 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.

24. Walter Lasecki, Christopher Miller, Adam Sadilek, Andrew Abumoussa, Donato Borrello, Raja Kushalnagar, and Jeffrey Bigham. 2012. Real-time captioning by groups of non-experts. In *Proceedings of the 25th annual ACM symposium on User interface software and technology*. ACM, 23–34.

25. Ina Lekkai. 2014. Incidental Foreign-Language Acquisition by Children Watching Subtitled Television Programs. *TOJET : The Turkish Online Journal of Educational Technology* 13, 4 (2014).

26. Norbert Michel, John James Cater, and Otmar Varela. 2009. Active versus passive teaching styles: An empirical study of student learning outcomes. *Human Resource Development Quarterly* 20, 4 (2009), 397–418.

27. Holger Mitterer and James M McQueen. 2009. Foreign subtitles help but native-language subtitles harm foreign speech perception. *PloS one* 4, 11 (2009), e7785.

28. Martha C Pennington and Jack C Richards. 1986. Pronunciation revisited. *TESOL quarterly* 20, 2 (1986), 207–225.

29. Maribel Montero Perez, Elke Peters, and Piet Desmet. 2014. Is less more? Effectiveness and perceived usefulness of keyword and full captioned video for L2 listening comprehension. *ReCALL* 26, 01 (2014), 21–43.

30. 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.

31. Harriet D Semke. 1984. Effects of the red pen. *Foreign language annals* 17, 3 (1984), 195–202.

32. Pamela Sherer and Timothy Shea. 2011. Using online video to support student learning and engagement. *College Teaching* 59, 2 (2011), 56–59.

33. Brent N Shiver and Rosalee J Wolfe. 2015. Evaluating Alternatives for Better Deaf Accessibility to Selected Web-Based Multimedia. In *Proceedings of the 17th International ACM SIGACCESS Conference on Computers & Accessibility*. ACM, 231–238.

34. Luis von Ahn. 2013. Duolingo: learn a language for free while helping to translate the web. In *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 1–2.

35. L Vygotsky. 1987. Zone of proximal development. *Mind in society: The development of higher psychological processes* 5291 (1987).

36. Mark Warschauer and Deborah Healey. 1998. Computers and language learning: An overview. *Language teaching* 31, 02 (1998), 57–71.

37. Helen Williams and David Thorne. 2000. The value of teletext subtitling as a medium for language learning. *System* 28, 2 (2000), 217–228.

38. 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.

39. Noa Talaván Zanón. 2006. Using subtitles to enhance foreign language learning. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras* 6 (2006), 4.