

# Exprgram: A Video-based Language Learning Interface Powered by Learnersourced Video Annotations

Kyung Je Jo John Joon Young Chung Juho Kim

School of Computing, KAIST  
Daejeon, Republic of Korea  
{kyungjejo,johnr0,juhokim}@kaist.ac.kr

## Abstract

Foreign language learners are challenged to master pragmatic competence, the ability to use language in a contextually appropriate way. While a large number of language learning materials are accessible, they are often optimized for developing linguistic components (e.g., vocabulary and grammar). In this research, we turn to videos in foreign language as an underutilized source of real-life situations with rich contexts. To efficiently learn from diverse situations through videos, learners should be able to access relevant videos that share a context or an expression. We introduce Exprgram, a language learning interface that utilizes videos at scale to enable context- and expression-based browsing. To enable such browsing, contexts or semantically related expressions in videos should be annotated at scale. Exprgram combines crowdsourcing and machine learning to acquire the needed annotations. Specifically, we introduce a *learnersourcing* workflow that harvests and organizes video annotations of highly contextual data and relevant expressions to improve our browsing system. Results of a pilot study show that Exprgram helps participants learn diverse expressions in a given context and generate reliable artifacts for future learners.

## 1 Introduction

Development of pragmatic competence, defined as the ability to use language in a contextually appropriate manner depending on factors such as relationship, feeling, or the context of the situation, is an essential part of foreign language learning (Washburn 2001). Firth and Wagner claims that language learning is not only learning and memorizing various linguistic components but also acquiring knowledge situated in physical and cultural context (Firth and Wagner 1998). However, research shows that non-native speakers, even with high proficiency, often lack pragmatic competence, thus fail to communicate as successfully as native speakers (Bardovi-Harlig and Hartford 1990). Kasper says the key to acquiring pragmatic competence is exposing learners to as many scenarios as possible (Kasper 1997). However, traditional language learning materials tend to provide conversations in limited and ideal settings, and fail to prepare learners for diverse real-life situations.

Videos in a foreign language are an underutilized source of authentic conversations situated in a large variety of

contexts. Large-scale authentic conversations from videos can effectively expose learners to diverse contexts and expressions. In our analysis of 2,100 subtitles of English movies and dramas, “How are you?” appeared in 238 utterances, yielding diverse responses from commonly-taught (e.g., “good”, “fine”) to contextual (e.g., “depends”, “cold”, “I’m not drunk”) to sarcastic responses (e.g., “That’s the ‘hey’ that means ‘I need something’.”) Furthermore, learners can also naturally access proper usage through rich context information such as settings, gestures, or emotions captured in videos. With a vast spectrum of contexts and situations, videos at scale present a pedagogical opportunity for teaching the usage of diverse expressions and facilitating the development of pragmatic competence.

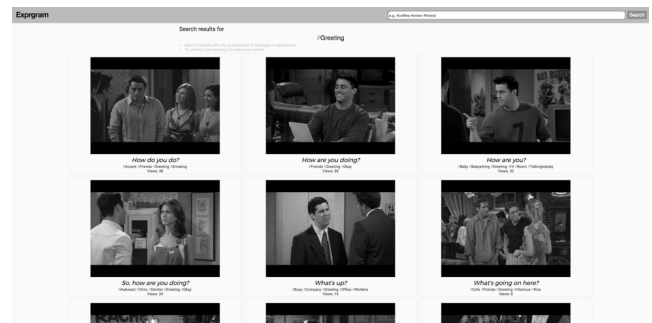


Figure 1: Search results for “#Greeting” in Exprgram. Learners can access a collection of videos that share a common situation (e.g., greeting) or related expressions (e.g., “how are you?” and “what’s up?”).

While learning through foreign language videos has become common and highly accessible through online platforms (e.g., Viki<sup>1</sup>, VoiceTube<sup>2</sup>, FluentU<sup>3</sup>), using videos to develop pragmatic competence presents challenges. First of all, existing language learning tools through video do not prioritize learning diverse expressions but rather focus on mastering vocabulary or improving pronunciation (Kovacs and Miller 2014; Culbertson et al. 2017b; 2017a). Secondly, mastering pragmatic competence not only requires

<sup>1</sup>www.viki.com

<sup>2</sup>www.voicetube.com

<sup>3</sup>www.fluentu.com

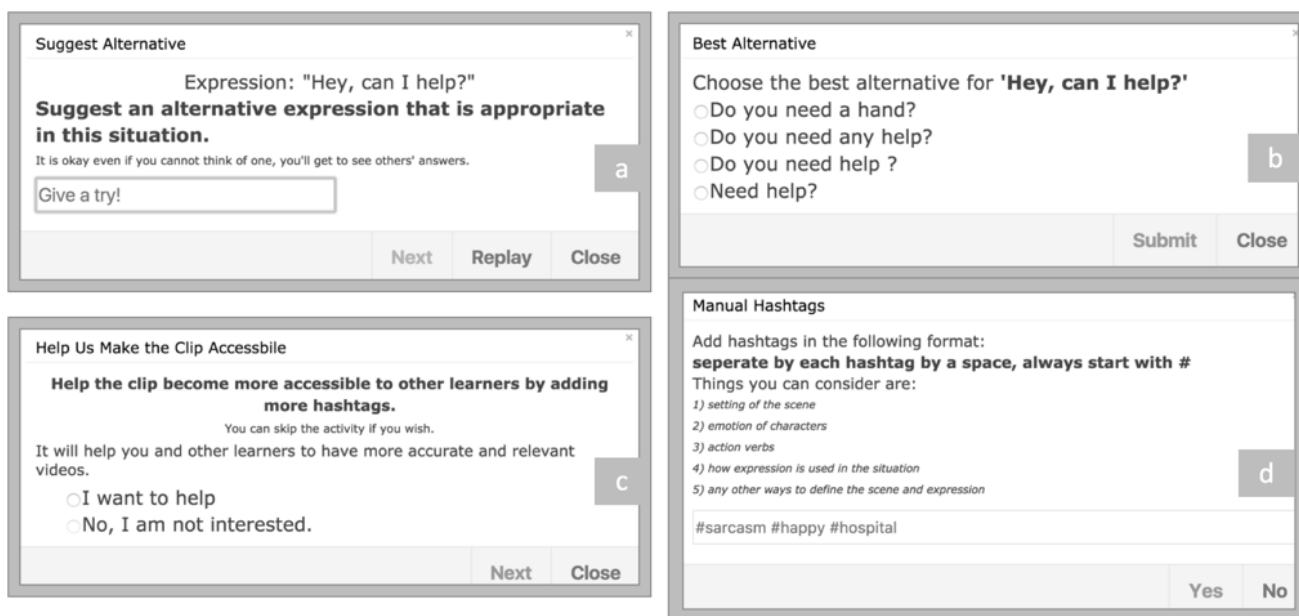


Figure 2: Four prompts that learners are given to annotate a video with relevant expressions and contexts - (a) providing an alternate expression in a given context, (b) protecting the system from malicious or incapable learners and learning diverse expressions from other learners’ artifacts, (c) motivating learners to annotate by explicitly stating the benefits to learners, and (d) adding context labels via hashtags.

proficiency in linguistic components but also mastery of diverse expressions in a contextually appropriate way. While existing language learning systems provide efficient tools for learning various linguistic components such as vocabulary, grammar, or punctuation, context- and expression-based browsing is often not supported for learners. Finally, foreign language learners are handicapped with a limited understanding of social and cultural context. Without an instructor, it may be challenging for learners to understand the subtlety or variations in nuance and learn the proper usage.

In this research, we attempt to combat the challenges and answer the following research question: “How can we use authentic conversations at scale from videos to expose language learners to various real life scenarios and teach diverse expressions?”

This paper introduces Exprgram, a web-based language learning interface for mastering diverse expressions under a given context. While machine learning techniques exist for computing semantic relatedness between sentences, the results are not always reliable due to their inherent limitations in capturing high-level concepts such as nuance, tone, or relationship between speakers. On the other hand, humans can generate video annotations that are possibly more accurate and understandable, but are inefficient compared to machines in terms of time and cost. Therefore, Exprgram takes a human-machine hybrid approach to annotating large-scale videos to enable context- and expression-based browsing (Figure 1). Exprgram motivates learners to answer prompts that involve meaningful learning activities (Figure 2), in which the resulting artifacts become video an-

notations that improve browsing. Our approach was inspired by *learnersourcing*, in which learners collectively generate useful material for future learners while engaging in a meaningful learning experience themselves (Kim et al. 2015). In Exprgram, the learnersourcing workflow powers a context- and expression-based browsing interface that is improved by learners’ voluntary participation.

## 2 Related Work

Learners play an essential role in Exprgram through annotating context labels, which directly affects the quality of the system. Prior research shows that learners can produce fine artifacts for the system while participating in meaningful activities themselves. In the Crowdy system, innately motivated learners generated video summary labels of comparable quality to those of experts (Weir et al. 2015). Culbertson et al. show that the learning is not impaired by human computation tasks that involved correcting video captions (Culbertson et al. 2017a). Building on these prior systems, Exprgram attempts to use learners for large-scale video annotations to identify relevant videos and expressions.

## 3 System Design

### 3.1 Design Considerations

We focus on teaching diverse expressions under diverse situations rather than other linguistic components such as vocabulary or grammar. Exprgram is implemented with two important design considerations to achieve the goal. Firstly, given the importance of exposing learners to vast real-life situations, learners should be able to easily browse videos

that are related in terms of context or expressions that are used. Secondly, learnersourcing activities should be pedagogically useful, specifically on learning the usage of diverse expressions.

### 3.2 Human-Machine Hybrid Approach

Exprgram allows learners to search for videos by expression or context. Our goal is to enable browsing similar contexts or relevant expressions, along with all excerpts of videos that involve relevant expressions to the learner’s query. For example, for learners interested in learning ways of greeting, they can browse “#greeting” to learn how people greet in various situations (Figure 1).

We take a human-machine hybrid approach to acquiring the labels required to support such browsing. Initially, we adopted Skip-thought (Kiros et al. 2015), an unsupervised learning method for a generic sentence encoder, to identify relevant expressions at scale. However, Skip-thought is limited in picking up small distinctions that drastically change the meaning of a sentence (Kiros et al. 2015); it scores “tricks on a motorcycle” similar to “tricking a person on a motorcycle”. In addition, due to inherent limitations in how machine learning works, it is difficult to capture rich contexts such as nuances, subjectivity, and highly contextual information such as the relationship between speakers. For instance, image captioning, a machine learning technique that automatically generates caption for an image, is optimized for capturing actions (e.g., ‘walking’, ‘standing’) rather than rich contexts. On the other hand, despite concerns for subjectivity, humans can easily capture diverse categories of context such as setting, emotion, or gesture.

### 3.3. Learnersourcing Video Annotations

Prior research introduced learnersourcing workflows to use learners as a crowd to generate meaningful artifacts through voluntary participation during their natural learning activity (Williams et al. 2016; Glassman et al. 2016; Weir et al. 2015). Exprgram consists of four prompts that are presented to learners in order at the end of each video (Figure 2). The purpose of the prompts is to expand our corpus of relevant expressions and to collect rich context labels.

**Suggesting an alternative** Firstly, learners are asked to give an alternate expression that is contextually appropriate in the given situation (Figure 2(a)). The educational purpose of this prompt is to verify learners’ understanding of context and ability to use language in a given context. Exprgram uses the collected expression in two ways: 1) to present them to other learners who are not capable of suggesting an alternative, and 2) to improve machine-assisted computation of semantically relevant expressions. For example, whereas “What’s up?” and “How are you?” had low similarity from the original machine learning technique, as learners suggest one another as the alternative, Exprgram scores two expressions as relevant. As a result, future learners are likely to benefit from the improved results.

**Choosing the best expression** Exprgram asks learners to pick the best expression out of a few alternatives from other

learners in the second prompt (Figure 2(b)). Learners see and learn several expressions that are appropriate for the context. Moreover, as Exprgram uses learners’ artifacts not only to teach learners but also to improve video browsing, quality control is in need to protect Exprgram from malicious or incapable learners. The voting mechanism in this stage serves as an initial filter.

**Motivation for further video Annotation** This prompt is an attempt to motivate learners to voluntarily participate in providing contextual labels (Figure 2(c)). The system explains why and how learners’ annotations help themselves as well as future learners. Unlike the two previous prompts that focus on mastering related expressions, this prompt is not designed to provide any direct learning benefit. Rather, it emphasizes the community-level benefit to promote learners’ participation.

**Context labeling** In the last prompt, learners are asked to describe the scene they just watched with context labels, in hashtags (Figure 2(d)). Learners can add various kinds of labels such as location, settings of the scene, relationship between characters, speakers’ emotion, etc. The hashtags are displayed without duplicates on the interface for future learners to help them easily grasp the context of the scene. Furthermore, extensive context annotations mean that the video can now be browsed through its context.



Figure 3: Two videos used in the pilot study with scene annotations collected.

Likert Scale Questions (Yes = 7, No = 1)	Average	Standard Deviation
When you were asked to select the best expression from the given set of options, did you find them reliable and appropriate?	5.47	1.25
Were you confident with the alternative answer you submitted?	5.07	1.39
Did you find the alternate expressions provided by the system helpful for learning more expressions?	4.53	1.55
Yes/No Questions	Percentage of people who responded yes	
Do you think hashtags described the context of the video you watched?	56%	
Did you find any conflicting hashtags that you wanted to add but have already been added by others?	25%	

Table 1: Summary of the questionnaire responses from the pilot study

## 4 Preliminary Evaluation

### 4.1 Pilot Study

We ran an in-lab pilot study with 16 participants to examine the potential effectiveness and limitations of Ex-prgram. Participants were graduate or undergraduate students who use English as second language. Eight participants watched a short snippet of video on expression “Hey, can I help?”(Video 1: Figure 3(a)), while others were given “I feel great and confident” (Video 2: Figure 3(b)). After the participants watched videos, they answered the four prompts. Participants answered a short survey after they completed the task, and answered three questions on a 7-point Likert scale and two questions in yes/no, presented in Table 1. They also shared their general experience using the interface.

A total of 16 and 12 unique hashtags were collected from Video 1 and Video 2, respectively. Participants answered that the collected expressions were reliable and helpful in acquiring foreign language skills (Table 1). Also, many felt confident about their own answer. Even though a further evaluation on the quality of collected expressions against expert-generated ones is necessary, results show initial promise in that at least students found the quality of learnersourced expressions to be high. However, some participants noted that expressions contained a limited diversity, varying only in one or two words. The results seemed positive towards the main idea of the interface: to clusture highly similar expressions to avoid duplicate learners’ effort.

We also asked questions about the effectiveness of hashtags, and more than a half of participants answered that hashtags accurately conveyed the context in the video. However, when asked in detail, participants thought that some hashtags were too general (e.g., #funny) or limited in explaining contexts (e.g., using #quit to explain ‘quit smoking’). Lastly, 75% of them replied that they were able to generate new hashtags even with existing hashtags.

## 5 Conclusion and Future Work

We envision a learnersourced language learning interface that iteratively optimizes itself for learning diverse expressions through vast real-life situations in videos.

There are limitations to the current work. Firstly, we could not see how combining machine learning and learnersourcing improves the browsing system with large-scale data, because of the small-scale data collected from the pilot study. We plan to collect large-scale data through a live deployment and evaluate the effectiveness of the hybrid technique. Secondly, though quality control is crucial in crowdsourcing (Oleson et al. 2011), we currently do not rigorously control the quality of learners’ input. Quality control in learnersourcing presents interesting challenges in that learners might be willing to provide quality work but lack the skills. We plan to add a screening test by asking them a question on the expression they should have learned from the video, and not accumulating their data when they give a wrong answer. Authoring a screening question for each video snippet might be a challenge to this method, and we plan to use learner-sourced artifacts in building such a screening question.

For future work, we will evaluate the effectiveness of the learnersourcing prompts by both measuring learners’ learning gains and evaluating the quality of learnersourced artifacts. After verifying the core workflow through a lab study, we will release the interface to the public.

## Acknowledgement

This work was supported by Institute for Information communications Technology Promotion(IITP) grant funded by the Korea government (MSIP) (No.2017-0-01217, Korean Language CALL Platform Using Automatic Speech-writing Evaluation and Chatbot).

## References

- Bardovi-Harlig, K., and Hartford, B. S. 1990. Congruence in native and nonnative conversations: Status balance in the academic advising session \*. *Language Learning* 40(4):467–501.
- Culbertson, G.; Shen, S.; Andersen, E.; and Jung, M. F. 2017a. Have your cake and eat it too: Foreign language learning with a crowdsourced video captioning system. In *CSCW*.
- Culbertson, G.; Shen, S.; Jung, M. F.; and Andersen, E. 2017b. Facilitating development of pragmatic competence through a voice-driven video learning interface. In *CHI*.
- Firth, A., and Wagner, J. 1998. Sla property: No trespassing! *The Modern Language Journal* 82(1):91–94.

- Glassman, E. L.; Lin, A.; Cai, C. J.; and Miller, R. C. 2016. Learnersourcing personalized hints. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing, CSCW '16*, 1626–1636. New York, NY, USA: ACM.
- Kasper, G. 1997. Can Pragmatic Competence be taught?
- Kim, J.; Miller, R. C.; Gajos, K. Z.; Klemmer, S. R.; Moran, T. P.; and Agrawala, M. 2015. Thesis: Learnersourcing: Improving learning with collective learner activity.
- Kiros, R.; Zhu, Y.; Salakhutdinov, R.; Zemel, R. S.; Torralba, A.; Urtasun, R.; and Fidler, S. 2015. Skip-thought vectors. *CoRR* abs/1506.06726.
- Kovacs, G., and Miller, R. C. 2014. Smart subtitles for vocabulary learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '14*, 853–862. New York, NY, USA: ACM.
- Oleson, D.; Sorokin, A.; Laughlin, G. P.; Hester, V.; Le, J.; and Biewald, L. 2011. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. *Human computation* 11(11).
- Washburn, G. N. 2001. Using situation comedies for pragmatic language teaching and learning. *TESOL Journal* 10(4):21–26.
- Weir, S.; Kim, J.; Gajos, K. Z.; and Miller, R. C. 2015. Learnersourcing subgoal labels for how-to videos. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing, CSCW '15*, 405–416. New York, NY, USA: ACM.
- Williams, J. J.; Kim, J.; Rafferty, A.; Maldonado, S.; Gajos, K. Z.; Lasecki, W. S.; and Heffernan, N. 2016. Axis: Generating explanations at scale with learnersourcing and machine learning. In *Proceedings of the Third (2016) ACM Conference on Learning @ Scale, L@S '16*, 379–388. New York, NY, USA: ACM.