ViVo: Video-Augmented Dictionary for Vocabulary Learning

Yeshuang Zhu[1], **Yuntao Wang**[1], **Chun Yu**[1]†, **Shaoyun Shi**[1], **Yankai Zhang**[1], **Shuang He**[1], **Peijun Zhao**[1], **Xiaojuan Ma**[2], **Yuanchun Shi**[1]
Key Laboratory of Pervasive Computing, Ministry of Education
Tsinghua National Laboratory for Information Science and Technology
[1]Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China
[2]Hong Kong University of Science and Technology, Hong Kong
zhu-ys13@mails.tsinghua.edu.cn, {yuntaowang, chunyu, shiyc}@tsinghua.edu.cn, mxj@cse.ust.hk

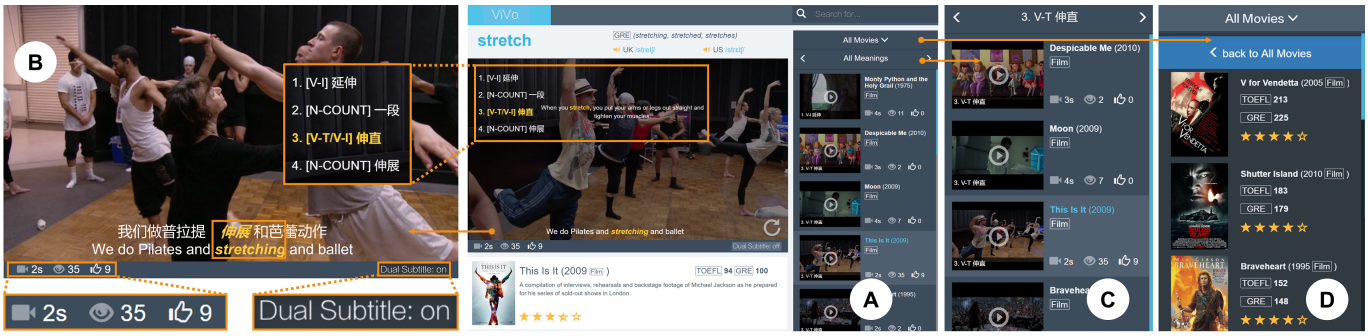

Figure 1. An overview of ViVo's interface. (A) The interface consists of the following main parts: a list of word definitions with the in-context sense in the video scene highlighted, a video area accompanied by the subtitles in the video's original language and optional dual subtitles, metadata of the current clip being played, and a clip navigation panel that supports filtering clips by meanings (C) or by movies (D). (B) As the clip is playing, bilingual keywords are automatically aligned and highlighted, and learners can interact with the clip by switching the subtitle's language and voting for a clip.

ABSTRACT

Research on Computer-Assisted Language Learning (CALL) has shown that the use of multimedia materials such as images and videos can facilitate interpretation and memorization of new words and phrases by providing richer cues than text alone. We present *ViVo*, a novel video-augmented dictionary that provides an inexpensive, convenient, and scalable way to exploit huge online video resources for vocabulary learning. ViVo automatically generates short video clips from existing movies with the target word highlighted in the subtitles. In particular, we apply a word sense disambiguation algorithm to identify the appropriate movie scenes with adequate contextual information for learning. We analyze the challenges and feasibility of this approach and describe our interaction design. A user study showed that learners were able to retain nearly 30% more new words with ViVo than with a standard bilingual dictionary days after learning. They preferred our video-augmented dictionary for its benefits in memorization and enjoyable learning experience.

†Denotes the corresponding author.

CHI 2017, May 06-11, 2017, Denver, CO, USA
 ISBN 978-1-4503-4655-9/17/05...$15.00
DOI: http://dx.doi.org/10.1145/3025453.3025779

Author Keywords

Subtitles; movie clips; dictionary; vocabulary learning

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

INTRODUCTION

Learner dictionaries, which rely on textual definitions and example sentences to illustrate words, are an important means for foreign language learning. However, learning with dictionaries is tedious and ineffective due to the monotonous textual illustration [5]. To improve the learning experience, researchers have explored and studied the use of multimedia to supplement dictionaries by coupling images, audios, and videos (e.g., [24]). Besides, modern online dictionaries such as *Bing Dictionary* already present word definitions together with images retrieved from the Internet. According to previous studies, multimedia contents provide abundant and multi-sensory contextual information through audible and visual data; they not only help to interpret the subtle usages of vocabulary but also provide rich cues for later recall [25, 29].

In this paper, we describe an inexpensive, convenient, and scalable video-augmented dictionary. The idea behind is to automatically retrieve short video clips from existing subtitled movies or TV genres to illustrate the keywords of interest, and incorporate them into a dictionary to facilitate vocabulary learning. The goal introduces major challenges such as to what extent the scenes and contexts in existing video

resources are rich and helpful to promote vocabulary learning, and how to extract appropriate video clips to illustrate individual words from large numbers of videos. Specifically, existing video resources contain contexts of vocabulary usage but without conveying the explicit definition of words; polysemous words render this problem even more challenging. Moreover, it is important to ask how one should model the user interface for video-augmented dictionaries in order to promote the learning effect. After carefully analyzing the feasibility and iteratively designing the user interface, we finally present *ViVo*, an online dictionary system augmented with movie videos. A user study shows that ViVo can improve learners' mid-term performance of vocabulary learning by 30% compared with a traditional dictionary.

This paper makes four main contributions:

- A quantitative analysis of vocabulary richness in existing subtitled movies and TV genres, which reveals an unrecognized opportunity for vocabulary learning.
- Accurately interpreting word meanings in video scenes to achieve dictionary-like illustrations of word senses by applying an effective word sense disambiguation (WSD) algorithm based on word embedding.
- A novel dictionary interface augmented with video clips to encourage learning from video contexts, highlight the bilingual contents for additional learning cues, and select the scenes with adequate contextual information.
- A user study that sheds light on the advantage of learning from movie videos as well as learners' interaction behavior with the video-augmented dictionary.

Captions and Subtitles

In this paper, two types of subtitles are examined: captions, a.k.a. transcripts in its original language (which means a foreign language for learners), and dual subtitles where both transcripts and its translations are shown. We will refer to both as subtitles in the following part of this paper, e.g., English subtitles and Chinese-English dual subtitles. In subtitles, a long dialog sentence is usually separated into short phrases called *lines* in order to display it properly. Besides, dual subtitles are often translated non-literally and only on phrase level rather than on individual words, as they are aimed at providing foreign audiences with a better comprehension of the movie plot. These facts pose challenges to constructing a video-augmented dictionary from subtitled videos.

RELATED WORK

A substantial body of work has been carried out to promote foreign language learning. Here, we review related work on contextual language learning, video-based language learning, and illustrating words with multimedia materials.

Contextual Language Learning

The core idea of contextual learning is to enable learners to acquire language knowledge with associated context. For example, Edge et al. proposed *Micromandarin* [10], which automatically provided contextually relevant content for language learning according to learner's location. Dearman and Truong [6] evaluated the learning effect of a live wallpaper on mobile phones and found that participants were much better at learning contextually relevant vocabulary than learning contextually independent ones. Edge et al. proposed *SpatialEase* [9], which contextualized the learning process of space- and motion-related vocabulary in body movements. Cai et al. [4] proposed the concept of wait-learning, which allowed learning conversation-related words during waiting for responses when chatting via instant messaging tools.

Although effective, one drawback of contextual language learning is that learning materials have to be heavily dependent on learner's context and thus are often limited and unpredictable. In contrast, our approach is based on another kind of context - the context in existing movies and TV genres - that can be analyzed and customized in advance to support dedicated and goal-directed learning. Learners can hence fully determine the vocabulary to learn according to their needs.

Language Learning via Subtitled Videos

Dual subtitles, which display subtitles in both the viewer's native language and the language of the video, are considered one of the best ways to present videos to language learners [22]. However, subtitles are usually comprehension oriented with non-literal, phrase-level translations, and interactions are poorly supported. Various kinds of enhancements have been attempted in order to make subtitles more suitable for language learning. For example, Zhou et al. [31] extended subtitles in a foreign language to contain an additional list of keywords contained in the current line together with definitions, memory skills, and so on. Similarly, *Gliflix* [23] added translations to monolingual subtitles for selected vocabulary words that appeared in the dialog. The purpose was to reduce the cognitive load on the viewers to discern and match specific vocabulary in two languages while comprehending the storyline. Kovacs et al. [15] designed interactive *Smart Subtitles* that showed word definitions and pronunciations on hover and supported dialog-based navigation by clicking on subtitle lines. Compared with dual subtitles, users could learn more vocabulary using Smart Subtitles with the same view time, understanding of the plot, and enjoyment.

However, current approaches to subtitle-aided language learning have focused on equipping the natural process of watching videos with more chances of incidental learning, i.e., they adopted a video to vocabulary workflow. In contrast to learning while watching videos, our research explores a vocabulary to video workflow: we gear our system design more toward dedicated learning by embedding video clips into a dictionary, with the purpose of providing rich contextual information for comprehending and remembering new words.

Illustrating Words with Multimedia Materials

In a dictionary, textual definitions and example sentences are usually insufficient for learners to grasp various and subtle differences in word meanings or emotions, especially for polysemous and synonymous vocabulary [5]. To overcome this, researchers have explored the use of multimedia contents (e.g. images, voices, and videos) as additional materials to illustrate words [2, 26, 20, 12]. Furthermore, other than manually collecting and compiling learning materials,

Savva et al. [24] proposed *TransPhoner*, which applied computational methods to transform a keyword in a foreign language into phonetically similar, highly imageable keywords in the learner's first language serving as memorable pronunciation guides, namely mnemonic keywords. A user experiment showed that TransPhoner could generate keywords with comparable quality to manually selected keywords, and was especially helpful for learners' memory of complex words.

To the best of our knowledge, existing work has relied on either static images or video clips generated from images to illustrate words. However, images have been found less capable than videos of creating learning materials, especially for abstract words [14]. In our research, we exploit the richness of existing movies and TV genres and propose computational approaches to generating appropriate video clips for learning, which are context-rich and scalable as we will show in the remainder of this paper.

The creation of a video-augmented dictionary for language learners requires three essential parts of work: 1) validating the feasibility of using existing subtitles as a resource for the dictionary in terms vocabulary richness; 2) filling the gap between the sense-based video scenes and the word-based entries in the dictionary; 3) properly designing the interface that optimizes the video-based learning effect. In the following three sections, we will address the three issues step by step.

ANALYSIS OF VOCABULARY IN SUBTITLES

While existing subtitles contain rich vocabularies with contextual information, it remains unknown to what extent they can be used as a resource for augmenting the dictionary. Specifically, what is the range, frequency, and distribution of the vocabulary in all subtitles and in each one alone? To answer these questions, we analyzed the vocabulary use in more than ten thousand subtitles of English movies and TV genres.

Data Resources

We randomly crawled a corpus of publicly available subtitles contributed by voluntary users[1]. For each subtitle, we also collected the metadata necessary for this study including genre, language, IMDb rating, and etc., utilizing the OMDb API[2]. Finally, we obtained 11,447 unique subtitle files in English. Each file corresponded to a movie or an episode of TV series. Among all subtitles most were for TV episodes, and only a tiny portion (about 5%) were for movies.

As for the words of interest, we considered the TOEFL and GRE word lists. These two lists represent everyday and professional language use respectively, and are targeted at the large number of English learners around the world who can basically understand the lines in the movies but still need to learn more new words. The TOEFL word list contained 2,859 essential vocabulary words [18] and the GRE list had 6,194 words [30]. These two lists shared 1,608 words in common, which meant there were 7,445 unique words in total. All words in the subtitles were lemmatized using Stanford CoreNLP [17] and counted against the two word lists.

[1] `http://www.zimuzu.tv/`

[2] `http://www.omdbapi.com/`

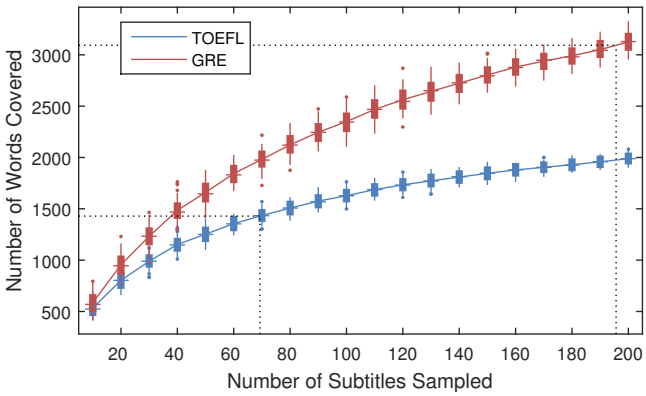

Figure 2. Curves and box plots for the coverages of TOEFL and GRE vocabularies in the subtitles. The dashed lines denote the estimated numbers of videos needed to cover half of the two word lists respectively.

Results and Findings

Existing videos are a rich resource of vocabulary use. In general, the subtitles covered almost all the TOEFL and GRE words (98% and 94% respectively), meaning that almost all these words could be illustrated by at least one video scene.

Furthermore, we estimated the growth of vocabulary coverage over the number of subtitles by randomly sampling N subtitles and counting the number of distinct vocabulary words in them. The sampling process was repeated 100 times for each N to obtain more reliable results. The growth curves with variances are shown in Figure 2. It can be seen that the word coverage grows unexpectedly rapidly with the number of subtitles. In particular, only 70 subtitles are sufficient to cover about 1,500 vocabulary words, which account for half of the TOEFL words or 1/3 of the GRE words. Moreover, vocabulary contained in a single subtitle can be quite rich. As an example, we found that the subtitle containing the most vocabulary words corresponded to the movie *Hamlet*[3], with 372 TOEFL words and 472 GRE words respectively in this 150-minute movie. It is encouraging news for language learners that hundreds of hard words can be illustrated with a live and coherent context of a single movie.

Moreover, the high vocabulary coverage was not a coincidental result of a small number of verbose subtitles. First, Figure 2 reveals the variance of the coverage at each point to be relatively small. Second, to further eliminate the possibility of the extreme case where a minority of subtitles would contribute to a main body of the vocabulary, we also analyzed the subtitle-vocabulary histograms and found that the distribution of the number of vocabulary words in each subtitle alone was nearly normal ($mean = 57$, $sd = 26$). These facts indicate that almost every movie or TV episode can be highly supportive of creating video-based learning materials.

It is worth noting that such a high vocabulary coverage in the subtitles is not an initially anticipated result, as the words used in speaking like movie dialogs tend to be shorter and less abstract than in writing [7, 1]. But in our analysis, for GRE words, which are usually considered to be formal and

[3] IMDb link: `http://www.imdb.com/title/tt0116477/`

academic, numerous occurrences have been found in the subtitles. The findings suggest that existing movies and TV genres are a viable resource to illustrate words, both in quantity and in diversity.

WORD SENSE DISAMBIGUATION

We have revealed the abundance of vocabulary in the subtitles. However, each word occurrence in one video scene can only illustrate one specific meaning of the word, as opposed to the fact that polysemy is common and is well presented in traditional dictionaries. It remains unknown to us whether the contexts in the subtitles are rich enough to illustrate the various polysemous words as a dictionary does. Hence in this section, we focus on identifying word senses by employing an effective, unsupervised word sense disambiguation (WSD) algorithm based on word embedding, with the ultimate goal of filling the gap between the sense-based video scenes and the word-based entries in our video-augmented dictionary.

WSD Based on Word Embedding

WSD is the task of determining the meaning of a word given its context in a sentence, which can be difficult as there are various cases of polysemy such as extended meanings (e.g. nominal and verbal usages) and nuanced meanings (e.g. analogy and sarcasm). To accurately calculate the likelihood of a word's meaning in a given sentence, we describe em-Lesk, an unsupervised WSD method based on the Lesk WSD algorithm [16] and word embedding [19].

The Lesk algorithm follows the distributional hypothesis in language semantics that words used in similar contexts tend to share similar meanings [13]. With the help of a standard dictionary, WSD can be achieved by finding the most closely matched meaning from the definition list of a given keyword.

To measure the semantic similarity between a keyword's context and each definition entry in the dictionary, we take advantage of word embedding. Word embedding is a way of word representation that also follows the distributional hypothesis and where each word is represented as a normalized vector of real numbers. By maximizing the dot product of a word vector and the vectors of its neighbors in a training corpus, the cosine similarity between two resulting word vectors is able to indicate the semantic similarity between the two corresponding words. This approach has previously been successfully applied to semantic analysis tasks such as in [11].

As for our em-Lesk algorithm for WSD, we first retrieve the word vectors pre-trained on a large text corpus. After averaging these vectors weighted by TF-IDF, we obtain the vector representing the keyword's whole context. The same process is also applied to each definition entry of the keyword to obtain the vector of each word meaning. The final similarity score between the keyword and each definition entry is calculated as the cosine similarity between the two corresponding vectors. The above approach is not language-specific but applicable to a variety of languages where a dictionary and a training corpus can be obtained.

For the implementation of em-Lesk, we used the word2vec tool to train word vectors on a 100-billion-word news corpus based on the skip-gram model with negative sampling [19], in coordination with a Collins Dictionary with bilingual definitions and example sentences for Chinese English learners[4]. Besides, as each word instance in a sentence and each meaning entry in the dictionary have a part-of-speech (POS) tag, em-Lesk also utilized the POS information to eliminate irrelevant senses so that the candidate list could be narrowed down. POS tags were obtained via the Stanford POS Tagger [27].

	TOEFL		GRE	
	Strict	General	Strict	General
MFS	.306	.434	.326	.521
POS-MFS	.367	.518	.390	.609
Em-Lesk	**.600**	**.670**	**.648**	**.745**

Table 1. Accuracy of em-Lesk compared with the baselines.

Evaluation of Em-Lesk

We compared our em-Lesk algorithm with two baselines to show its effectiveness: the most frequent sense (*MFS*) that always chose the first and also the most frequent sense of the keyword in the dictionary, and *POS-MFS* that always chose the most frequent sense that shared the same POS tag with the keyword to be disambiguated. The MFS baseline was considered to be quite competitive in the field of WSD [3].

Example sentences given for each word sense in the Collins Dictionary provided us with a handy dataset for automatically evaluating our em-Lesk algorithm. The examples were all selected from a corpus of real-world text comprised of text in fictions, broadcasts, recorded conversations, etc. As a large part of the examples were also dialog sentences, they were representative of the sentences in the subtitles for our purpose of an initial evaluation. For the words in the TOEFL and the GRE lists respectively, we constructed two test sets with a total of 14,345 pairs of mapping from an example sentence to the corresponding word sense. We also considered two test conditions, i.e., *strict* and *general*. In the *general* condition, we took all the test data into account, whereas in the *strict* condition, we considered only the proportion of words with ambiguous senses under the same POS, i.e., the words that had at least two definitions with the same POS. We ran the three WSD methods on the test data and computed the WSD accuracy.

As shown in Table 1, em-Lesk outperformed the two baselines and could correctly disambiguate up to 74% ambiguous GRE words in the *general* condition. Note that the evaluation took only the mapping of example sentence to its belonging definition as correct. In practice, the performance would be even higher as many different senses of a word were no more than the same meaning used in different contexts and were too trivial for learners to distinguish, such as *abandon* (somebody) versus *abandon* (a thing or place). Hence, em-Lesk was sufficient for our goal of analyzing senses in the subtitles and building a video-augmented dictionary.

There are several directions to improve our current WSD method. First, we trained word embedding on an external corpus due to the lack of sufficient amount of text in existing subtitles. It will be better to begin with the embedding trained

[4] `http://www.collinsdictionary.com/`

on the external corpus and then adapt the vectors to the context of movies according to the text from subtitles. Second, we found that in the evaluation, a large part of WSD errors made by em-Lesk could be attributed to POS tagging: if a word instance received an incorrect POS tag, then it was impossible for em-Lesk to tell its correct meaning. This can be alleviated by jointly considering the POS score and the context similarity score. Third, we can utilize the sentence-sense pairs in existing dictionaries as labeled data to train a supervised WSD model. However, these improvements are beyond the scope of this paper. Besides, in the automatic evaluation, we used example sentences from a dictionary as a ground-truth, which could potentially lead to performance bias for em-Lesk if the context around a keyword contained the same words as its associated definition. Since the evaluation was an initial one, a manual examination in the context of subtitles was desired in the future.

Sense Richness in Subtitles

We applied our em-Lesk algorithm described above to testing the sense diversity of our subtitle corpus. Among all the 15,518 distinct senses of the TOEFL and GRE words in the dictionary, 93.5% were covered by the subtitles. Hence, we draw the conclusion that there are not only abundant word occurrences in movie subtitles but also a high diversity of word meanings due to different contexts of use.

INTERACTION DESIGN

An Overview

The richness of vocabulary and contextual information in existing videos motivates us to design ViVo, a novel video-augmented dictionary for language learners. ViVo enhances traditional dictionary by illustrating words with short video clips automatically excerpted from movies. As shown in Figure 1, ViVo presents textual information of the looked-up word like in a dictionary, which includes the vocabulary level, pronunciation, and bilingual definitions. At the same time, ViVo fetches the associated video clips as example sentences where the keyword is enunciated and the subtitles are given.

However, the interface design involves several usability challenges, e.g., how to encourage learning from the video contexts other than the textual contents, how the learning contents should be properly highlighted in both subtitles and videos, and how to segment and filter the clips to identify the best ones for learning. In this section, we describe how our design of ViVo can address these considerations. The techniques supporting the design will be later elaborated in the Implementation section.

Encouraging Video-Based Learning

In order to emphasize on the video-based learning feature and encourage vocabulary learning from the context of use, two key design decisions are made.

First, textual definitions are delayed to show up until one of the video clips has finished. Once a clip is selected, it will automatically start playing with subtitles (Figure 1B). Once finished, it will jump back to the start and wait for learners to repeat it. Then a text layer is shown to teach learners the exact word meanings in the dictionary, along with the whole

Figure 3. An example of ViVo's keyword highlighting. Learners can acquire word meanings from both the rigorous definitions in the dictionary and the in-context uses in the video scene.

dialog sentence just being played (Figure 1A). This design encourages a process of learning from guessing where learners attempt to guess or recall the word meaning by means of the videos before they are exposed to the exact definitions [26, 20].

Second, Chinese subtitles are initially hidden from the interface and are only shown on user's demand. ViVo presents the English-only subtitles by default, although users can switch between the monolingual subtitle and the bilingual ones by clicking the subtitle button. This design also facilitates the process of learning from guessing and more importantly, it avoids learners' tendency to shift their attention from the word itself and other information in the foreign language when bilingual information is given [15].

With the above design of learning flow, learners are guided to concentrate on both the textual information and the scenes in the movies.

Sense and Keyword Highlighting

As shown in Figure 3, given a keyword in a video clip, ViVo identifies and highlights its best-matched sense in the dictionary. Other definitions in Chinese are also presented as supplements for learners to gain an overview of the various different meanings of the word. When a user hovers on each meaning, a detailed definition in English is shown.

Moreover, as has been discussed before, the translation in the subtitles is usually non-literal and dependent on the context, which in turn offers an alternative explanation for the word in addition to the rigorous definitions in the dictionary. In order to exploit such an additional opportunity for learning, the English keyword and its translated Chinese counterpart are both highlighted to draw learners' attention (Figure 1B and Figure 3). The highlighted translation in the subtitles, in combination with the accurate dictionary definition, provides two translations for the same keyword, which has been proved to help users notice the differences and infer the meaning from uncertain translations [28].

Clip Segmentation, Selection, and Navigation

Previous research has revealed the benefits of reviewing a word in different contexts on learner's retention of the word [5]. Moreover, viewing the same word across multiple video clips can exploit the phenomenon of dual coding

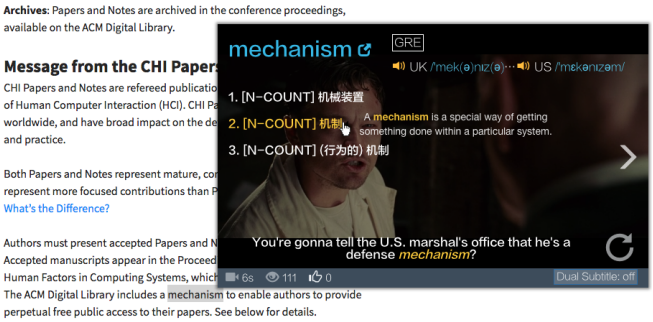

Figure 4. ViVo's pop-up search interface. While reading online, the user selects a word to immediately view its meaning in the context, illustrated by one of the best-matched movie clips.

which helps guard against the negative effects of encoding specificity [21]. Guaranteed by the richness of context in the subtitles, ViVo can retrieve an adequate number of clips to illustrate a single word. However, the issue of longer time cost arises when learners are faced with multiple, context-rich clips. We alleviate this issue by excerpting the clips at the length of a complete sentence, like the example sentences in the dictionary. Clips that last too long (more than 20 seconds) are also filtered out. The design is a trade-off between learning efficiency and contextual richness. With sentence-length clips, learners can still understand a large part of the video context while at the same time exploring multiple clips with a relatively high efficiency.

Furthermore, ViVo provides navigation features that help learners quickly scan through the clips without actually playing them one by one. All clips associated with the keyword are listed on the right panel of ViVo. Learners decide which clips to view according to the metadata including image previews, length in seconds, view times, and votes by other learners. In particular, votes and view times enable social interaction among users. If a learner finds a clip to be helpful for remembering a word, she/he can also vote it up. Learners can further explore the clips by meanings and by movies (Figure 1C and Figure 1D). To facilitate browsing by meanings, a concise definition of the keyword is provided in the learner's native language over the preview image of each clip. For browsing by movies, metadata is shown for each movie and TV episode including title, year, poster, as well as the quality of both the video content (IMDb rating) and its richness in vocabulary for learning (the numbers of TOEFL and GRE words covered).

Pop-up Search Interface

Apart from the use for dedicated vocabulary learning, ViVo can also be tailored as a pop-up search interface to facilitate vocabulary lookup and review at reading, as shown in Figure 4. While reading, users select a word in the document to inspect its meanings in a pop-up window. The interface analyzes the context of the word to determine its in-context definition to highlight, i.e. the definition in the dictionary that best matches the usage in current reading context. If no contextual information is detected, the most common sense of the word will be presented. As in ViVo's main interface, the textual definition is presented along with an associated video clip. However, in the pop-up interface, definitions are shown prior to the videos as understanding is the primary goal for readers. Whether to further review the word in the video's context is optional to the user. If the user clicks the play button or stays in the definition page for a time longer than a preset limit (five seconds in our case), the recommended clip will start playing automatically. Learners can replay it or jump to the next clip by pressing a corresponding button, or else they can simply click outside the pop-up window to close it and continue the reading process. Besides, a previous study has shown that when an unfamiliar word is re-looked up, it will enhance vocabulary learning by reminding learners of its contextual information learned in previous encounters [5]. The pop-up interface can yield this benefit by prioritizing those clips that the user has previously viewed.

IMPLEMENTATION

We implemented ViVo as an online application based on the HTML5 video elements, and implemented ViVo's pop-up search interface as a Google Chrome extension that monitors user's word selection activity. ViVo required all the videos to have no subtitles embedded in the frames. Each video was accompanied by two standalone subtitles: an English version and a Chinese-English dual version. We converted all the videos into H.264 format and all the subtitles into WebVTT format[5] so that they could be natively supported by HTML5 video players. In this section, we describe how we addressed the technical challenges in the implementation of ViVo.

Clip Generation

After all vocabulary words in the subtitles are indexed, each subtitle line containing the target word can be retrieved along with the timestamps indicating the start and the end of the clip in the whole video. However, in subtitles, a longer sentence of dialog usually spans multiple lines. To restore a full sentence from the lines, we design regular expression rules to traverse forward and backward starting from the central line to determine the actual sentence boundaries. Then a grammatically legitimate sentence is obtained by concatenating the lines within the boundaries. Furthermore, in order to avoid cutting off an ongoing speech at length, we opt to truncate the clip at the middle point of the interval between two consecutive sentences. The spared time offers extra benefits for learners to get prepared to focus on the target word and its context. Eventually, each clip is associated with a complete dialog sentence in both languages along with length in time, metadata of movie, a screenshot, the corresponding sense in the dictionary, etc. The sense is determined by performing our em-Lesk WSD algorithm.

Highlighting Bilingual Keywords

Since all English words in the subtitles have been lemmatized before being indexed, it is straightforward to locate and highlight an English keyword even in the sentences where the word is morphed. A particular issue arises when locating the Chinese counterpart of an English keyword in the translated subtitles, especially when the translation is non-literal and paraphrased. Therefore, given a pair of bilingual sentences, we first perform word segmentation. In particular, for

[5] `https://w3c.github.io/webvtt/`

languages without explicit boundaries between words, such as Chinese in our case, the segmentation is done based on a dictionary and word frequency[6]. Then we employ a statistical bilingual word aligner[7] to align the two lists of tokens, while the salience of the highlight is bound to the confidence score of the aligner in order to accommodate the uncertainty. The following snippet illustrates how bilingual keyword alignment can help learners to grasp word senses:

- *Transcription*: `I will` **`avert`** `my eyes.`
- *Translation*: 我会**非礼勿视**的

In the translation above, the word *avert (eyes)* is liberally translated as the Chinese idiom *to see no evil* with a little bit of humor. Our procedure successfully aligns the keyword to its Chinese counterpart and creates an impressive mnemonic of one specific use of the keyword.

We manually examined the quality of keyword highlighting on a set of 60 keyword-line pairs. For 37 of the pairs, our method correctly located the Chinese counterpart of the given English keywords. For 19 pairs, no Chinese counterpart was found. Only in four cases, the translation was highlighted in a wrong way. Missing or erroneous alignment was mainly caused by the limitation of the aligner we used: the aligner was not specifically trained on movie dialogs. However, as such cases are relatively rare and the missing highlighting of translated keywords has limited impact on learners, the above highlighting method is sufficient for our system.

Clip Ranking

ViVo tries to pick out those clips from the matched candidates that best highlight the keyword to learn. We adopt heuristics to rank the clips based on the observation that a keyword is better enunciated in a scene if the subtitle line containing the keyword stays longer on the screen while the remaining lines taking up less of the total time. On the one hand, as the number of words in a subtitle line is fairly limited by the width of the frame, a longer presence time of a subtitle line indicates a slower and clearer speech. On the other hand, given a fixed presence time of the keyword, a shorter clip lowers learner's cognitive load and wait time to be exposed to the target word. Therefore, we round the time span of all clips into intervals of every two seconds, and then sort them by the longest time of the matched line, and then by the shortest total length.

In the future, we will take into consideration more factors to further identify the clips that are better suited for learning, such as to reward the lines with no other difficult words around the keyword, or to incorporate a grammar checker to filter out those incomplete, informal sentences. Besides, we also take social interaction into consideration. The clips that have received more votes or views are ranked higher.

EVALUATION

We conducted a user study with 22 Chinese-native speakers who learned and used English as a foreign language. The goal of the study was to examine the effectiveness of ViVo for memorizing new word meanings, compared with a full-featured online dictionary - Youdao[8]. The Youdao Dictionary was widely used by English learners in China and was already equipped with a few multimedia features such as related images retrieved from the web and synthesized speech of example sentence. Given a list of words to be learned, we measured participants' short-term and mid-term retention performance, as well as learning time and learning strategy.

Video and Subtitle Resources

From the IMDb Top 250 list[9], we sampled 61 movies that were produced in English and later than the year 1960. These movies covered a wide range of genres, including science fiction, drama, history, and etc. Altogether, they contained 1,666 TOEFL and 2,448 GRE vocabulary words. Note that the coverage was higher than the results shown in Figure 2, as all the subtitles here were for movies, which were usually twice in length compared with the subtitles for TV genres.

Vocabulary

As the vocabulary size of each participant varied, we took a two-step process to determine the appropriate learning task for each participant. First, we created a long list of 140 vocabulary words from the TOEFL and GRE vocabulary sets; all the words were chosen by two non-native English speakers in a way that both of them were unfamiliar with the words. For each word, at least one occurrence could be found in the movie collection (4 on average). Then a day before the study, we asked each participant to filter the long list to obtain 36 completely unfamiliar words for the learning task. By "completely unfamiliar", the words that were guessable or seemingly seen before by the participant were excluded. We did not use a pretest here as participants might forget the exact meanings of the words they had seen before but would quickly pick them up if they encountered the words again.

Participants

Twenty-two undergraduate and graduate students with an engineering background were recruited for our study. The subjects aged from 20 to 34 ($mean = 23.8$, $sd = 3.1$) and seven of them were female. They were all Chinese-native with self-evaluated English proficiency as intermediate and had been learning and using English for years. Specifically, they all had previous experience with reading and writing academic papers in English. Seven of them had taken TOEFL or GRE exams before. Besides, they were unfamiliar with a majority of the selected movies prior to the study.

Experimental Design and Procedure

The 36 words were equally and randomly divided into three sessions, and in each session, half were learned with ViVo and the other half with the dictionary. Participants learned the words once and were tested twice for short-term and mid-term retention respectively. Word orders were all randomized for both learning and testing. Note that the interweaving of the two learning methods in each session prevented participants from establishing a global context of the movies and hence the result could more closely reflect the actual learning effect of local contexts provided by the short video clips.

[6] **http://www.xunsearch.com/scws/api.php**

[7] **http://nlp.csai.tsinghua.edu.cn/~ly/systems/TsinghuaAligner/TsinghuaAligner.html**

[8] **http://dict.youdao.com/**

[9] **http://www.imdb.com/chart/top**

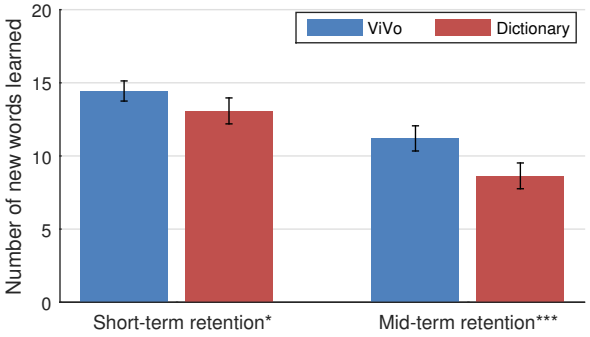

Figure 5. The performance of vocabulary learning, with standard error bars. (*: p <.01, ***: p <.0001)

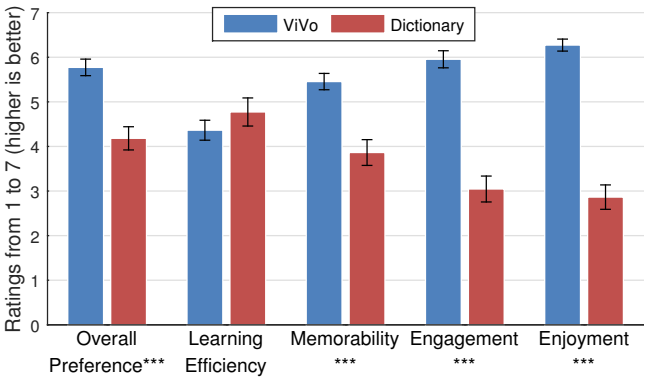

Figure 6. Subjective ratings of interface preferences, with standard error bars. (*: p <.0001)**

Upon beginning the study, participants first did a warm-up exercise to get familiar with the interfaces of ViVo and Youdao Dictionary. Then came the three sessions of learning and post-learning test. The words came out one after another. For each word, participants could learn as much time as they liked until they felt satisfied and proceeded by clicking the next button. The next word would show up after a five-second intervention of a distractor image. After one learning session was finished, participants took a test to recall Chinese definitions for the 12 word they had just learned. The questions were free-response and had no time limit. Participants were told not to guess the meanings. The above procedure of learning and testing was repeated twice. Participants took a five-minute rest between each two consecutive sessions. Finally, the first day of the study ended with a questionnaire. Participants were told to refrain from either reviewing the 36 words or actively learning new words by other means in the subsequent days. Two days later, they did a second test of all the 36 words learned and took part in an informal interview.

RESULTS

Vocabulary Gain

Participants' learning outcome was measured by the number of words correctly defined after learning. Any answer that was semantically relevant was counted as correct. For polysemous words, participants only had to recall one of the meanings they had encountered. Two raters independently marked the answers according to the meanings in the dictionary and then discussed with each other until a consensus was reached.

Figure 5 shows participants' vocabulary gain right after learning and two days later. On the whole, participants could remember more new words after learning with movie clips than using a standard dictionary. A paired t-test showed that with ViVo there were significantly more new words correctly defined ($t_{21} = 3.14$, $p < .01$) right after learning. From a longer-term perspective, the advantage of learning with ViVo enlarged by 30% more words successfully recalled than learning with a dictionary ($t_{21} = 6.95$, $p < .0001$).

Furthermore, note that the mid-term result is more reliable than the short-term one: as the post-learning test happened right after the learning process, it was not difficult for the participants with certain English proficiency to temporarily remember all the 12 words in each session, which led to a smaller difference in the test scores. Hence, the improvement in learning with ViVo was pronounced enough to conclude that vocabulary learning based on movie clips is more effective compared with traditional dictionaries. Although we did not conduct a long-term study, according to the forgetting curve [8], without any revision after learning, the retention decays exponentially and remains at a relatively stable level after two days. Hence, the mid-term result could speak for the long-term learning effect of ViVo.

Learning Time and User Interaction

The mean time for learning each word was 50 seconds ($sd = 23$) in ViVo versus 37 seconds ($sd = 23$) in the dictionary respectively. Time spent on ViVo was 35% higher and a significant difference was observed in the amount of learning time ($t_{21} = 7.39$, $p < .0001$). Note that in the learning stage of our study, we allowed participants to freely browse the materials for each word and did not impose any time limit, which represented a natural way of learning. It was therefore reasonable to expect that participants had taken equally self-initiative to devote adequate time to both learning conditions. We confirmed the hypothesis by running a Wilcoxon signed-rank test on participants' self-evaluated learning efficiency for both interfaces (see Learning Efficiency in Figure 6) and no significant difference was found ($Z = -22.5$, $p = .43$). Namely, the increase in learning time with ViVo is not evident enough to be perceived by learners.

Furthermore, different learning strategies adopted in both interfaces can be revealed by our system's interaction log and participants' self-assessed time allocation. In ViVo, learners intensively interacted with the video clips both in depth and in breadth: in order to learn each word, they actively viewed more than four clips averagely (the mean clip duration was about six seconds); for each distinct clips the learners had viewed, they would repeat it once again on average, with more than half of the clips were played at least twice. Besides, in ViVo, learners spent 2/3 of the time on video-related contents such as viewing the clips, reading the subtitles, and browsing the movies. In the dictionary condition, however, they paid little attention to the example sentences but spent more than half of the time reading the word definitions. The above evidence indicates that ViVo's design attracts learners to engage themselves in the video-based learning process. Such initiative in repetitive practice enhances

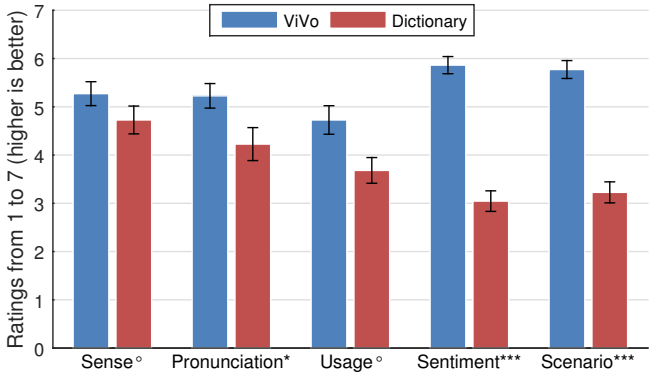

Figure 7. Subjective ratings of learning outcomes, with standard error bars. (°: <.05, *: p <.01, ***: p <.0001)

learners' long-term retention [8] and is exclusive to video-based learning [31], even if more time may be needed.

Subjective Feedback

Figure 6 shows participants' preference on the whole and on several aspects. We ran Wilcoxon signed-rank tests on the 7-point scale ratings. Results confirmed that learners preferred ViVo to the dictionary for vocabulary learning. Specifically, they felt significantly more engaged and joyful in the learning experience with ViVo, and considered the memorization of new words learned from video contexts would last longer.

Subjective feedback on the learning outcomes is shown in Figure 7. The learning outcomes of ViVo learners went beyond only the vocabulary gain. Statistical tests showed that users of ViVo significantly learned more about a word than that of the dictionary. In particular, the sentiment and use scenarios of the word were considered to be exclusive advantages of vocabulary learning based on video scenes, which benefited from the rich pragmatic context in movies as we will discuss below.

Furthermore, in the free-form feedback, learners agreed that learning with ViVo "*made remembering words much more interesting*" and "*created impressive cues by associating the words with the scenes*". Subjects stated that they were attracted to repeat viewing or explore multiple clips: "*I would like to repeat a clip until I could understand the word meanings all by myself*" (P5); "*I wanted to explore more scenarios where a word is used as different meanings*" (P16). The repetition, which might potentially cost longer time, also indicates learners' willingness and engagement in learning with ViVo. More importantly, learning from multiple clips provided richer cues for later recall and was especially beneficial to longer-term memorization: "*The video-based learning was more like the way I learned my native language where new words were gained in practical use*" (P11); "*There were many times during the posttests when I could clearly remember the scenes associated with the words*" (P9). P21 even correctly contextualized most of his answers in the scenes although not required in the tests. In contrast, dictionary learners tend to get weary with the textual information and are unwilling to explore more of a word even given enough learning time: "*There were too many example sentences but I was just tired to read them*" (P1); "*I thought the dictionary to be more efficient for word lookup and understanding, but poor information was provided for vocabulary learning*" (P5).

DISCUSSION

Proper Scenes for Learning

In the interview after the study, we asked participants to identify impressive words learned from movie scenes and provide reasons. From their feedback, we summarized several types of impressive scenes that are particularly helpful for illustrating and learning. Four typical categories are shown in Figure 8 and discussed below.

- **Repeated**: A word repeated in the same movie indicates its importance or relevance to the movie context and can inculcate a strong memory of it. P6 even expressed that he would *"never forget the word **sergeant** for the rest of life as it is repeated in too many scenes in Forrest Gump"*.
- **Self-explanatory**: Self-explanatory scenes are easy for learners to interpret without much text or plot. A thing, an activity, or a feeling can all be intuitively illustrated by the items or characters' performance in movies. For example, *tuck* reminded P3 of how an officer ordered a soldier and *hassle* reminded P18 of how a prisoner was regarded.
- **Humorous**: The dialogs containing a joke, a metaphor, or an exaggerated action can draw learners' interest, even if only a conversation setting is depicted visually. For example, a joking dialog impressed P14 and P15 with how the word *raccoon* was pronounced, even though the scene *"has nothing to do with raccoons"*.
- **Emotive**: Dialogs with an intense emotion, e.g., provocative and irritating, can convey an emotional impact on learners. P22 said, *"I can contextualize several words learned in the scenes depicting a debate"*. And P10 was particularly instilled by the line *"We can not **falter**!"*.

Impressive scenes usually combine multiple features above. In the future, we can further refine the clips for learning by incorporating semantic features of the subtitles and the scenes, other than emphasizing the keyword based on dwell time in our current stage. Moreover, superficial features like sentence terminators and video metadata might be helpful enough to indicate the visual content and semantic in a scene. For example, lines ending with exclamations or questions can serve as indicators of emotive scenes, and scenes excerpted from comedies are likely to contain humorous performance.

Effect of Familiarity

For each word learned via ViVo, we asked participants to label whether they were familiar with the associated clips before and then calculated the mean of vocabulary gain for each participant under both familiarity conditions. However, a t-test revealed no significant difference. The result can be partly attributed to the movie collection used in our study: the movies were not popular and 2/3 of them were unfamiliar to participants. A further study is needed in order to draw conclusions about the effects of familiarity on learning outcomes. From another perspective, the advantage of video-augmented learning with ViVo might not be limited by the familiarity with the movie. The choice of presenting clips at a sentence

Figure 8. Examples types of impressive scenes for learning, including the ones that are repeated (A and D), self-explanatory (B and D), humorous (A), and emotive (A and C).

level eliminates the need for learners' prior knowledge about the whole movie.

However, according to the subjective feedback, participants preferred to view the clips from the movies that were familiar to them. Five participants explicitly showed great interest in learning with clips from movies they were familiar with. In the future, it will be better to allow learners to choose videos of interest or upload their own video collections in order to create a personalized learning environment. Moreover, the number and level of vocabulary words contained can serve as a new criterion for recommending movies to learners. By tracking the viewing history of each individual learner, a personalized watchlist can be automatically generated to maximize the learner's exposure to new words.

The Vocabulary-Media Workflow

Many previous video-based learning approaches (e.g., [23] and [15]) used a media-vocabulary workflow. However, accurately mastering the usage of vocabulary in a foreign language needs much more dedicated learning effort. Dedicated learners usually have learning tasks and schedule in advance and wish to gain anticipated learning outcomes, which cannot be met in a learning process directed primarily by video watching. In contrast, ViVo adopted a vocabulary-media workflow which can bring multiple video clips together to illustrate a single word for learning while pertaining the advantages of multimedia materials. Moreover, as it is expensive to manually create multimedia materials for learning that meets the learner's level and goal, our vocabulary-media approach exploits rich language knowledge and context in existing online videos which are inexpensive, convenient, and scalable.

Both the vocabulary-media workflow and its reverse have their use scenarios and are hence complementary to each other. In the future, a system combining both with blended reference, consumption, and review modes might ultimately be a very promising and engaging way for language learning.

Limitations and Future Work

Our current lab study only focused on learners' vocabulary gain after one round of learning and hence suffers several limitations. First, the longer learning time could be due to a potential novelty effect in our lab study. A long-term, in-situ study is needed to explore how ViVo can become integrated into the learning process in practice, including the word look-up function at reading. Second, the vocabulary tests were not sufficient to reflect the learning outcomes in depth. As the outcomes of video-based learning tend to be compound, further studies are needed to quantify the learning effects in term of word pronunciation, usage, mood, and etc. Third, the movies used in our study were not customized for each subject. It is an interesting direction to let the user specify a collection of preferred movies and explore the effect of familiarity on video-based vocabulary learning.

Moreover, our video-based approach can be generalized, both to different levels of learners and to learning other languages. On the one hand, learning difficulty can be manipulated by picking out the lines with only words below a specific level. On the other hand, the proposed approach of providing rich movie contexts for language learning is not restricted to specific language. Although currently we did not analyze the vocabulary coverage in movies in other languages, we expect a similar result as movies reflect human life where language is used. However, several challenges need to be addressed for scaling-up, including accommodating the noises of online subtitles, obtaining subtitles in less-spoken languages, and video license issues.

CONCLUSION

We have presented ViVo, a novel video-augmented dictionary that makes the abundant language knowledge and usage context in movies easy-to-access by learners. In order to automatically generate illustrative video clips for learning at scale, we first verified the feasibility of utilizing existing movies with respect to the coverage range and spread of words in subtitles. Furthermore, we identified the scenes with adequate contextual information for learning the diverse word senses using a WSD algorithm. Based on the findings, we described the design and implementation of ViVo, which focused on encouraging learning from understanding the context and choosing clips that best highlighted the keywords and senses through both visual effects and proper length of time. Finally, the results of a lab study confirmed ViVo's advantage over the dictionary by a 30% increase in the mid-term vocabulary retention. Learners gained better interpretation and memorization of new words from the rich cues in the videos and they strongly preferred ViVo for its efficient and enjoyable learning experience. ViVo offers an inexpensive, convenient, and scalable way to exploit huge online multimedia resources for language learning.

ACKNOWLEDGEMENT

This work is supported by the National Key Research and Development Plan under Grant No. 2016YFB1001200, the Natural Science Foundation of China under Grant No. 61272230, Tsinghua University Research Funding No. 20151080408. This work is also funded by the NExT Search Centre (Grant R-252-300-001-490) supported by the Singapore National Research Foundation under its International Research Centre at Singapore Funding Initiative and administered by the IDM Programme Office.

REFERENCES

1. Akinnaso, F. N. On the differences between spoken and written language. *Language and speech 25*, 2 (1982), 97–125.
2. Amemiya, S., Hasegawa, K., Kaneko, K., Miyakoda, H., and Tsukahara, W. Long-term memory of foreign-word learning by short movies for ipods. In *Advanced Learning Technologies, 2007. ICALT 2007. Seventh IEEE International Conference on*, IEEE (2007), 561–563.
3. Bhingardive, S., Singh, D., V, R., Redkar, H., and Bhattacharyya, P. Unsupervised most frequent sense detection using word embeddings. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, ACL (2015), 1238–1243.
4. Cai, C. J., Guo, P. J., Glass, J. R., and Miller, R. C. Wait-learning: Leveraging wait time for second language education. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, ACM (2015), 3701–3710.
5. Dang, T.-D., Chen, G.-D., Dang, G., Li, L.-Y., et al. Rolo: A dictionary interface that minimizes extraneous cognitive load of lookup and supports incidental and incremental learning of vocabulary. *Computers & Education 61* (2013), 251–260.
6. Dearman, D., and Truong, K. 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 (2012), 1391–1400.
7. DeVito, J. A. Levels of abstraction in spoken and written language. *Journal of Communication 17*, 4 (1967), 354–361.
8. Ebbinghaus, H. *Memory: A contribution to experimental psychology*. No. 3. University Microfilms, 1913.
9. Edge, D., Cheng, K.-Y., and Whitney, M. Spatialease: Learning language through body motion. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, ACM (2013), 469–472.
10. Edge, D., Searle, E., Chiu, K., Zhao, J., and Landay, J. A. Micromandarin: Mobile language learning in context. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, ACM (2011), 3169–3178.
11. Fast, E., Chen, B., and Bernstein, M. S. Empath: Understanding topic signals in large-scale text. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM (2016), 4647–4657.
12. Han, C.-H., Yang, C.-L., and Wang, H.-C. Supporting second language reading with picture note-taking. In *CHI '14 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '14, ACM (2014), 2245–2250.
13. Harris, Z. S. Distributional structure. *Word 10*, 2-3 (1954), 146–162.
14. Kayama, T., Kaneko, K., Miyakoda, H., and Ishikawa, M. Effective materials for abstract words in foreign vocabulary learning. In *Wireless, Mobile and Ubiquitous Technologies in Education (WMUTE), 2010 6th IEEE International Conference on*, IEEE (2010), 207–209.
15. Kovacs, G., and Miller, R. C. Smart subtitles for vocabulary learning. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, ACM (2014), 853–862.
16. Lesk, M. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, ACM (1986), 24–26.
17. Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S. J., and McClosky, D. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations* (2014), 55–60.
18. Matthiesen, S. J. *Essential Words for the TOEFL*. Barron's Educational Series, 2014.
19. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems* (2013), 3111–3119.
20. Miyakoda, H., Kaneko, K.-i., and Ishikawa, M. Effective learning material for mobile devices. *JLCL 26*, 1 (2011), 39–51.
21. Paivio, A. Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology/Revue canadienne de psychologie 45*, 3 (1991), 255.
22. Raine, P. Incidental learning of vocabulary through authentic subtitled videos. *JALT-The Japan Association for Language Teaching* (2012).
23. Sakunkoo, N., and Sakunkoo, P. Gliflix: Using movie subtitles for language learning. In *UIST 2013 Adjunct*, ACM (2013).
24. Savva, M., Chang, A. X., Manning, C. D., and Hanrahan, P. Transphoner: Automated mnemonic keyword generation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, ACM (2014), 3725–3734.
25. Secules, T., Herron, C., and Tomasello, M. The effect of video context on foreign language learning. *The Modern Language Journal 76*, 4 (1992), 480–490.
26. Takigiku, Y., Kaneko, K., Ishikawa, M., and Miyakoda, H. Short movie materials based on tessellation for foreign vocabulary learning. In *Information Technology Based Higher Education and Training (ITHET), 2010 9th International Conference on*, IEEE (2010), 284–291.

27. Toutanova, K., Klein, D., Manning, C. D., and Singer, Y. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology-Volume 1*, Association for Computational Linguistics (2003), 173–180.

28. Xu, B., Gao, G., Fussell, S. R., and Cosley, D. Improving machine translation by showing two outputs. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2014), 3743–3746.

29. Yeh, Y., and Wang, C.-w. Effects of multimedia vocabulary annotations and learning styles on vocabulary learning. *Calico Journal* (2003), 131–144.

30. Yu, M. *New Oriental GRE Vocabulary Selection*. Qunyan Press, 2005.

31. Zhou, J., Dai, X., and Wang, P. Foreign language learning based on video scenes. In *Education Technology and Computer (ICETC), 2010 2nd International Conference on*, vol. 2, IEEE (2010), V2–350.