Effectiveness of textually-enhanced captions on Chinese High-school EFL learners' incidental vocabulary learning

Huizhen Wu
Tongji University/Shanghai Business School
Xiaohu Yang[1]
Tongji University

Received: 17 December 2021/ Accepted: 21 June 2022
DOI: 10.30827/portalin.vi38.23511
ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

ABSTRACT: This study employed a mixed-methods approach to investigate the impact of textually-enhanced captions on EFL learners' incidental vocabulary gains and learners' perceptions of the captioning usefulness in a multi-modal learning environment. A total of 133 Chinese EFL high school learners of the low-intermediate level were randomly assigned to English captions with highlighted target words and L1 gloss (ECL1), Chinese and English captions (CEC), Chinese and English captions with highlighted target words (CECGW), and no captions (NC). Our quasi-experimental findings did not detect any significant differences among the caption types on vocabulary form recognition while ECL1 was found the most effective in meaning recall and recognition. Captioning types and learners' language proficiency exerted medium-to-large effects on meaning recall and meaning recognition. Our qualitative data suggested the participants generally viewed captioned videos positively, with variability in perceptions of concurrent presentation of information. The saliency of L1 gloss could direct the viewers' attention to the semantic features of a word and reinforce sound-form-meaning connections. Videos lacking L1 glosses of target words had relatively little effect on learners' vocabulary learning while more textual inputs might not necessarily result in vocabulary gains. Pedagogical implications are proposed for teachers' adoption of L1 in captioned videos to enhance learners' learning effectiveness.
Key words: Captioning types, video viewing, foreign language learning, vocabulary learning, L1 gloss

Eficacia de los subtítulos mejorados textualmente en el aprendizaje de vocabulario de estudiantes chinos de inglés como lengua extranjera

RESUMEN: Este estudio investigó el impacto y las percepciones de los estudiantes de los subtítulos mejorados textualmente en las ganancias incidentales de vocabulario de los estudiantes de inglés como lengua extranjera en un entorno de aprendizaje multimodal.

[1] **Corresponding author:** Xiaohu Yang, School of Foreign Languages, Tongji University, 1239 Si Ping Road, Shanghai, 200092, P.R. China. Email: 19026@tongji.edu.cn.

133 estudiantes chinos de inglés como lengua extranjera de nivel intermedio bajo fueron asignados aleatoriamente a subtítulos en inglés con palabras objetivo resaltadas y brillo L1 (ECL1), subtítulos en chino e inglés (CEC), subtítulos en chino e inglés con palabras objetivo resaltadas (CECGW), y sin subtítulos (NC). Nuestros hallazgos cuasi-experimentales no detectaron diferencias significativas entre los tipos de subtítulos en el reconocimiento de formas de vocabulario, mientras que ECL1 resultó ser el más efectivo para recordar y reconocer significados. Nuestros datos cualitativos sugirieron que los participantes generalmente veían los videos subtitulados de manera positiva, con variabilidad en las percepciones de la presentación simultánea de información. La prominencia del brillo L1 podría dirigir la atención de los espectadores a las palabras objetivo y reforzar las conexiones de la forma del sonido y el significado.
Palabras clave: Tipos de subtítulos, visualización de videos, aprendizaje de idiomas extranjeros, aprendizaje de vocabulario, brillo L1

1. Introduction

Today's language learners have ever-increasing access to culturally rich and enjoyable online resources (Teng, 2019) such as videos, which have 'become an integral part of youth culture' (Leonhardt, 2020:229). In the Chinese context, learners are often exposed to bilingual input outside the school setting through TV, movies and the Internet, since information via bilingual presentation can reach a wider population (Li, 2016). Despite the easy accessibility and availability of bilingual captions and their potential value to English as a Foreign Language (EFL) learners, there is a lack of research into their effects on L2 development (see Yang 2013, Li, 2016, Liao et al. 2020).

Nowadays, technology affordances have resulted in more advanced captioning availability, e.g., glossed target words or keywords, and bilingual captioning with highlighted target words. Glosses refer to 'short definitions or explanations with nonlinearly linked data associated with graphics, audios, and videos in computerized texts' (Yun, 2011). The use of some form of scaffolding, like captioning or glossed keywords, could allow language teachers to tap into a readily available and wide range of genres likely to captivate learners (Yeldham, 2018; Costales, 2021), and help learners to visualize what they hear to enhance audio-visual input, especially when the material is slightly beyond their level of proficiency (Hsieh, 2020). Vocabulary acquisition can be arduous for English as a Second Language (ESL) learners (Webb & Nation, 2017). The effectiveness of the use of glosses in captioned video in relation to learners' vocabulary learning has been widely reported (e.g. Montero Perez et al., 2018; Hsieh, 2020). However, very little empirical research has investigated the impact of access to highlighted target words in bilingual captioning on lexical L2 development. Furthermore, there is scant research into how learners themselves perceive the usefulness of different types of captioned video.

To address this gap, this study intended to fulfil dual objectives: firstly, to examine if the differential effectiveness exists under different captioning conditions for high-school EFL learners' vocabulary learning; secondly, to document learners' perceptions of the usefulness of different captioning to language learning.

2. Theoretical framework and literature review

Captions provide learners with "polysemiotic texts" instead of "traditional monsemiotic texts" (Lertola, 2019:488). The polysemiotic nature of audio, visual products—verbal and visual channels—involves a complex mechanism for information processing. Captions not only provide learners with a synchronous written verbal representation of the oral text during the audio-visual input but also enables learners to build referential connections while learners are receiving the input aurally and visually. Intake is what learners consciously notice, and noticing is the necessary and sufficient condition for converting input to intake (Schmidt, 1990). To enable learning from multimedia materials, information needs to be attended to in order to be available for processing in working memory (van Gog, 2014:255). Given the dynamic and transient nature of information presentation during system-paced video viewing, signalling or cueing may help learners attend to the relevant information. Cues come in many forms and can be incorporated into text, a picture, or both (van Gog, 2014: 256). Text-based cues can be more subtle, drawing attention to certain key terms or ideas by using colour in written text (Moreno & Abercrombie, 2010), e. g. glosses. In view of the complex information processing in the multimedia context, this study will consider Mayer's (2014) Cognitive Theory of Multimedia Learning and van Gog's Cuing Principle (2014) in that they have the implications for language learning of the synchronous processing of information through acoustic and visual channels.

According to the Cognitive Theory of Multimedia Learning (Mayer, 2014:59), humans process visual/pictorial and auditory/verbal information via separate but limited-capacity channels. Humans engage in *active learning* by selecting relevant incoming information, organizing selected information into coherent mental representations, and integrating mental representations with existing knowledge. As van Gog (2014: 255) put it, since both processing less relevant information and not processing relevant information will be deleterious for learning, it is necessary to guide learners' attention to the essential material, which can be achieved by means of cueing. As the visual/pictorial and auditory/verbal channels in working memory are extremely limited, only a few items can be held or manipulated in each channel at any one time (Mayer, 2014). We may argue that with textually-enhanced captions, i.e. glossing, "cueing facilitates learning by reducing the amount of visual search required, because the cues guide attention to the right location or to the important information", which should foster active learning (van Gog, 2014:264).

The existing captioning research has found that exposure to captioned videos may help L2 learners gain more access to online streamed texts, comprehend unknown words and hence "enhance incidental vocabulary learning, that is, vocabulary learning as a by-product of listening for meaning" (Gass, 1999); though "defining which particular type of subtitling is most effective seems difficult to determine and results in this area are somewhat inconclusive" (Kanellopoulou, 2019: 145). Zanón (2006) posits that a captioned video provides a triple connection between image, sound and text, and this type of connection encourages strong associations for retention and language use. The potential of audio-visual input as a source of initial vocabulary acquisition has been suggested to be comparable to written input in terms of learning gains (Webb, 2015). However, the facilitative captioning effects with glosses in very few studies were differential depending on these factors (but not limited

to): test designs, aspects of vocabulary measures, EFL/ESL contexts, L1-L2 orthographic differences, learners' proficiency level and selection of the video content.

Montero Perez et al. (2014) investigated 133 Dutch-speaking French undergraduates under 4 conditions: full caption, keyword caption, and full caption with highlighted keywords and no caption. The findings revealed that the captioning groups scored equally well on form recognition and clip association and significantly outperformed the control group. Contrary to their hypothesis, only the keyword captioning and full captioning with highlighted keywords groups outperformed the control group on meaning recognition, and captioning did not affect comprehension or meaning recall. Montero Perez et al. (2018) later expanded their research by delving into the effects of adding L1 gloss access to the target words on vocabulary learning, and found the glossed keyword caption resulted in a better outcome in form recognition and recall than full captioning, keyword captioning, and no captioning. However, Hsieh (2020) revealed that FCL1 (full caption with highlighted target-word and L1 gloss) was most effective for learning both word form and meaning relative to the other caption types, namely, full caption with highlighted target-word (FCHTW), full caption with no audio (FCNA) and no caption. But to the best of our knowledge, captioning effects on vocabulary learning under the bilingual captioning condition as well as the glosses embedded in the bilingual condition remain under-explored.

In sum, no clear conclusion can be drawn from the existing research regarding what type of captions (e. g. full, bilingual, textually enhanced) can facilitate L2 vocabulary acquisition most effectively. Firstly, the existing research primarily focused on captioning effects among undergraduates when L1 and L2 were mostly Indo-European languages. Thus, further research is needed to know what types of captioned videos would facilitate high school learners' vocabulary gains in an EFL context in which there exist distinct L1-L2 orthographic, phonologic and semantic differences. Secondly, given the value of subtitles as a scaffold for lower-proficiency learners to access multimodal authentic input (Pujadas and Muñoz, 2020), it would be worth investigating the effect of the concurrent presence of L1 and L2 (i.e. bilingual captions) on learners' comprehension and vocabulary acquisition. Thirdly, learners' perceptions of the usefulness of captions to vocabulary learning remain under-explored. As Guillot (2020) noted in her research of the audiovisual translation of cinematic contexts, audiences' experience is normally holistic, and individual: what viewers make of what they see, hear and/or read is dependent on what is to be seen, heard and/or read. Since perceptions could refine how we understand learners' use of captions and help course designers provide opportune scaffolding to EFL learners, more research is needed to investigate why and under what conditions learners considered captions as useful, not useful, or even harmful. It is widely accepted that learner perceptions of instruction and instructional design play an important role as they direct learning (Frick et al., 2009). They are often weighted heavily by instructors and may sway their decisions regarding the addition or removal of learning supports (Leveridge & Yang, 2014). In light of these issues, the current study was designed to explore the effects of textually-enhanced captions on vocabulary learning and learners' perceptions of captioning usefulness in Chinese students, whose L1-L2 manifests orthographic, phonologic and semantic differences. The specific research questions to be addressed are posited:

1. Does the type of captioning: full caption with highlighted target words and L1 gloss (ECL1), bilingual captions (CEC), bilingual captions with highlighted target

words (CECGW), and no captioning (NC)) have a differential effect on EFL learners' vocabulary learning, as measured by form recognition, meaning recall and meaning recognition tests?

2. Do EFL learners perceive captions as a help or hindrance to their vocabulary learning?

3. METHODS

3.1. Participants

One hundred and thirty-three learners from a high school in rural China participated in the study. This school was chosen because we have long-term collaboration and we constantly provide language teachers of the school with research support. All the participants were native speakers of Mandarin Chinese. Students' average grade provided by their English instructor was used to measure their English proficiency (comparable to the B1 level based on CEFR). Classes were organized in groups of approximately forty students. To test the consistency of students' language proficiency scores, students' final exam was used to compare with the average performance provided by the teachers (Cronbach's alpha = 0.77). An ANOVA analysis was performed to confirm that their English proficiency did not significantly differ (F (3,129) =1.98, p =0.12) across the four groups. Students' age ranged from 17-18 and learners had the same weekly exposure time (3.75 hours per week, 45 minutes per day). Their average English learning length from the onset age was 8.97 years (SD=1.2). None of them reported study abroad experience. All participants had normal (or corrected to normal) vision and reported sound hearing. Written and oral consent were obtained from all participants' parents before conducting the experiment. The study was also approved formally by the school and the teachers.

3.2. Materials

The video material was chosen from a TED talk titled *Every kid needs a champion*. The talk was selected by two researchers (the first author and an independent researcher) and evaluated by three EFL teachers with a teaching experience of more than 16 years. Topic familiarity, length, accent and difficulty level were considered in selection of the video. We also consulted the instructors to ensure the video was not viewed in class before the study. It lasted 7 minutes and 32 seconds (1055 words in length). The Flesch Reading Ease[2] score of the videos' text readability was 87.6, indicative of standard and average level for high school learners. The talk was delivered by a passionate educator, addressing the importance of relationship between teachers and kids. The moderate speaking speed, standard American English accent and strong storytelling nature of the video made it suitable for

[2] https://readabilityformulas.com/free-readability-formula-tests.php. The Flesch Reading Ease Readability Formula is considered one of the oldest and most accurate readability formulas, developed by Rudolph Flesch in 1948. The formula is: RE (Readability Ease) = 206.835 – (1.015 x ASL) – (84.6 x ASW). ASL = Average Sentence Length, ASW = Average number of syllables per word. The total score ranges from 0 to 100. The higher the number, the easier the text is to read.

students' viewing. The videos with full caption, Chinese subtitles, and without captions are all accessible from the official website[3]. The video with bilingual captions was created by the software *Arctime*[4], without any alteration to the original Chinese and English captions available at the official website. The examples of three captioning conditions were displayed in **Figure 1**. In the CEC mode, a full Chinese translation was added to the English caption. In the ECL1 condition, target words were highlighted in red, and an equivalent Chinese translation was displayed below. The CECGW condition was created based on the CEC mode by adding L1 and L2 glosses of the target words.

ECL1

CECGW

CEC

Figure 1. Screenshots of the three caption types (ECL1, CEC and CECGW)

We adopted the following procedure to select the target words, which were deemed important for understanding the meaning of a sentence (Teng, 2019). First, 20 possible target words from the transcript were selected by 2 researchers. Considering the frequency of word occurrence might impact word learning, we chose the target words which occur only once in the video. Then, we consulted the three experienced EFL high-school instructors to determine the suitability of the selected target words. Furthermore, 10 pilot participants

[3] The video is accessed via https://www.ted.com/talks/rita_pierson_every_kid_needs_a_champion. "Every kid needs a champion" by Rita Pierson is licensed under CC BY–NC–ND 4.0 International.
[4] https://arctime.org/

(similar to our participants' language proficiency but not included in our study) were invited to view the video and take the trial tests to ensure that the difficulty level of the selected video was appropriate and that the target words chosen were not familiar to them. Afterwards, a semi-structured interview was conducted among the pilot participants to elicit their opinions on the selected words and open questions. We incorporated their suggestions to modify the wording of the open question and delete the words known to them. The target words not chosen by both the instructors and the pilot participants were deleted. Finally, 10 target words were selected, including nouns, verbs, adjectives, and adverbs.

3.3. Instruments

3.3.1. Vocabulary tests

Participants took the vocabulary tests after viewing the video twice. To prevent them from intentionally noting down the target words during viewing, they were not informed of the specific forms of test items. We followed Hsieh's (2020) scheme to design the vocabulary tests containing True/False, translation and multiple-choice questions, which respectively correspond to form recognition, meaning recall and meaning recognition (see examples in **Table 1**).

Table 1. Design of the comprehension questions with examples

	TEST ITEMS	EXAMPLES
Vocabulary Learning	Form Recognition	Please judge if the word has occurred in the video. e.g., *arduous, legacy, academically, strut, reserve*
	Meaning Recall	Please give the Chinese translation of the word.
	Meaning Recognition	Please choose the correct answer. *Well, your year is going to be long and **arduous**, dear.* a) 难受的 b) 糟糕的 c) 辛苦的

The vocabulary tests contained 10 target words and 5 filler items in a written form. The three vocabulary tests required different types of cognitive processing. In form recognition, participants needed to judge whether the given words had appeared in the video. Meaning recall and meaning recognition measured vocabulary meaning. In both tests, the target words and the full sentence that appeared in the video were shown to give contextual clues. Meaning recall was more difficult than the meaning recognition, because it required producing meaning from memory without reliance on contextual clues. Thus we changed the order of Hsieh's (2020) test administration for the latter two tests. Namely, students completed meaning recall prior to the meaning recognition test to avoid possible interference from previous test performance, as indicated in Montero Perez's study (2020). To avoid guessing, students could not move to the next question until they had finished the previous question.

For all three sub-tests, instructions in L1 were written on top of the paper as well as given orally by the researchers to ensure full understanding.

3.3.2. Questionnaire

The first part asked students to complete a short demographic questionnaire regarding age, length and onset age of English learning, vision and hearing condition (normal, corrected to normal or not). The second part included open-ended questions: *How do you like this captioning video? Please specify the reasons.* The control group did not need to answer this question since captions were not available to them.

3.4. Procedure

The research was administered during regular classes, where students were taught English course using a multimedia computer system with a projector on a communal screen. Learners from four classes were applied one of the ECL1, CEC, CECGW and NC conditions. Under each condition, one researcher was responsible for playing the video and one teacher monitored the classroom to ensure that each student was following the instructions correctly. The sound and visual quality were checked to ensure all participants could hear and see the videos clearly. During the experiment, pausing and replaying the video were not allowed, but note taking was permitted. After viewing the video twice, the students proceeded to complete content comprehension task (by recalling the video gist in either English or Chinese) and three paper-and-pencil vocabulary tests sequentially. Finally participants completed an open-ended questionnaire. The whole experiment took around 45 minutes (See **Figure 2**). Students were informed in advance that their performance on the comprehension exercises would not affect their course grades, and were assured that all the data would be analyzed and used only for research purposes in an anonymous way. This study only focused on investigating learners' vocabulary learning under different captioning modes, so results of the content comprehension were not reported here.

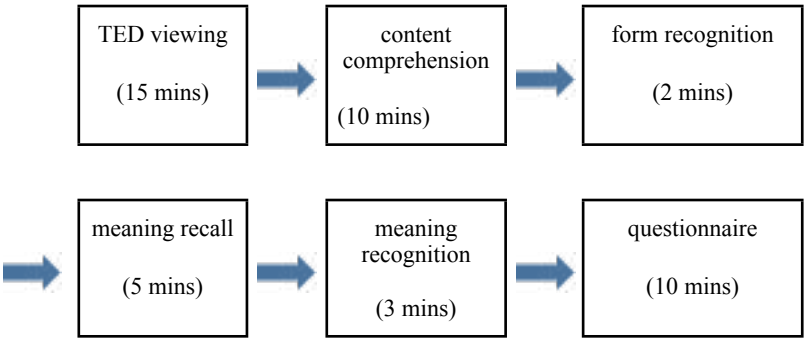

Figure 2. Experiment procedure.

3.5. Scoring

All questions were scored dichotomously by two independent raters. One point was given for a correct response and zero for an incorrect one. For meaning recall test, we adopted a more lenient attitude towards the answers. If the Chinese meanings were deemed proper in the context, we would count them as correct. The inter-rater reliability reached 99.67%. The two raters discussed to address any disagreements in scoring.

3.6. Data analysis

For the first research question, the three vocabulary test results were analyzed using one-way multivariate analysis of covariance (MANCOVA), with video caption type as the independent variable and the scores on the three vocabulary tests as the dependent variables. English proficiency scores and the onset age of English learning were included as covariates in the model for vocabulary test scores. We set the significance level in the statistical analysis at 0.05, and reported partial eta squared (η_p^2) as effect size, by interpreting η^2 values of 0.01, 0.06 & 0.14 as small, medium and large effect sizes, respectively (Cohen, 1988). For the second research question, we followed a content analysis approach (Hsieh and Shannon, 2005) using MAXQDA Analytics 2020 to analyze the responses to the open-ended question.

4. RESULTS

4.1. Effectiveness of captioned videos for vocabulary learning

Table 2 presents the descriptive statistics of vocabulary learning tests under four conditions.

Table 2. Descriptive statistics of vocabulary tests

		FORM RECOGNITION		MEANING RECALL		MEANING RECOGNITION	
Conditions	*N*	*M* (SD)	*CR*	*M (SD)*	*CR*	*M (SD)*	*CR*
ECL1	34	9.53 (2.48)	71.5%	7.74 (2.18)	77.4%	8.56 (0.86)	85.7%
CEC	32	8.75 (3.02)	65.6%	4.97 (1.79)	49.7%	8.05 (1.13)	80.5%
CECGW	34	8.92(2.74)	66.9%	4.97 (2.30)	49.7%	8.43 (0.79)	84.1%
NC	33	9.15 (1.74)	68.7%	5.60 (1.69)	56.1%	7.28 (0.85)	72.8%
Total	133	9.09 (2.54)	68.2%	5.83 (2.30)	58.2%	8.09 (0.99)	80.8%

As presented in **Table 3**, the one-way MANCOVA yielded a significant main effect of English proficiency on meaning recall and meaning recognition after controlling for English proficiency and onset age of English learning. The effect size statistics revealed a

medium-to-large effect of caption type on the meaning recall and meaning recognition test outcomes (F=6.49, p<0.001, η_p^2 = 0.135), but not on form recognition. However, the onset age of English learning as the covariate showed negligible effect (F=1.244, η_p^2 = 0.029, p=0.297).

The Bonferroni post-hoc test was performed to further investigate the differences across captioning conditions, as shown in **Table 4**. No pairwise comparisons achieved statistical significance in form recognition test (p>0.05). For meaning recall, post-hoc comparisons showed that ECL1 group scored higher than the CEC, CECGW and the NC groups with statistical significance (p≤0.05). For meaning recognition, the ECL1, CEC and CECGW scored significantly higher than the no caption group; the ECL1 scored higher than the CEC and the CECGW group without statistical significance. The CECGW was slightly higher than the CEC group without statistical significance.

Table 3. MANCOVA results of captioning effects on vocabulary learning (with English proficiency and onset age of English learning as covariates)

WILK (P) (η_p^2)	SOURCE	VOCABULARY TESTS	F	P	η_p^2
0.971(0.297)(0.029)	Onset age	Form recognition	0.039	0.843	0.000
		Meaning recall	0.247	0.620	0.002
		Meaning recognition	3.713	0.056	0.028
0.865(<0.001)(0.135)	English	Form recognition	2.446	0.120	0.019
	proficiency	Meaning recall	11.724	0.001*	0.085
		Meaning recognition	10.325	0.002*	0.075
0.601(<0.001)(0.156)	Caption	Form recognition	1.104	0.350	0.025
	type	Meaning recall	14.331	0.000*	0.253
		Meaning recognition	10.722	0.000*	0.202

*Notes. *Indicates a statistically significant difference between groups.*

Table 4. Bonferroni post-hoc test for meaning recall and meaning recognition scores

VOCABULARY SCORES	TREATMENT PAIRS	MD	SE	P
Meaning recall	ECL1-CECGW	2.680*	0.475	0.000
	ECL1-CEC	2.606*	0.483	0.000
	ECL1-NC	2.313*	0.479	0.000
	NC-CECGW	0.367	0.481	1.000
	NC-CEC	0.293	0.492	1.000
	CEC-CECGW	0.074	0.478	1.000

Meaning recognition	ECL1-CECGW	0.043	0.214	1.000
	ECL1-CEC	0.309	0.218	0.950
	ECL1-NC	1.088*	0.216	0.000
	CECGW-CEC	0.266	0.216	1.000
	CECGW-NC	1.044*	0.217	0.000
	CEC-NC	.778*	0.222	0.004

No pairwise comparisons achieved statistical significance in form recognition test ($p>0.05$). For meaning recall, post-hoc comparisons showed that ECL1 group scored higher than the CEC, CECGW and the NC groups with statistical significance ($p<0.05$). For meaning recognition, the ECL1, CEC and CECGW scored significantly higher than the no caption group; the ECL1 scored higher than the CEC and the CECGW group without statistical significance. The CECGW was slightly higher than the CEC group without statistical significance.

4.2. Perceptions of Captioned Videos for Vocabulary Learning

The open-ended question in the questionnaire concerned learners' perceptions of the given captioning video and what were the specific reasons underlying their perceptions. We performed a content-analysis on students' responses based on several consecutive reading and coding cycles. We analyzed and coded all valid responses, which were entered to the MAXQDA software prior to coding. After three rounds of careful reading, we first categorized each learner's responses to the usefulness of the captions into *useful*, *partially useful, not useful*, and *harmful* for each caption condition (See **Figure 3**). Following several rounds of reading one researcher coded all responses and the second researcher cross-checked them to ensure consistency. We then grouped all themes into different coding categories: language acquisition, form-meaning connection, perceptual processing, learner comfort and video impact. Reasons were specified per coding category and frequency for each coding category was presented (See **Table 5**).

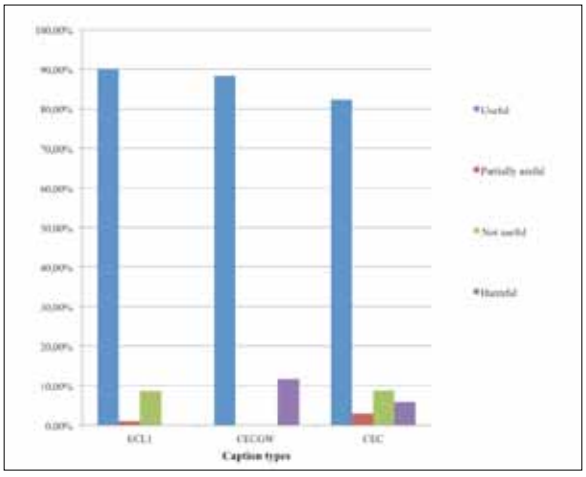

Figure 3. Learners' perceptions of caption usefulness

The majority of students regarded their assigned captions as being useful, with the ECL1 group taking the lead. Very few students in the CEC and the CECGW group held

Table 5. Examples of learners' perceptions of caption usefulness

	CATEGORY	DESCRIPTIONS	ECL1	CEC	CE-CGW
PERCEPTIONS OF USEFULNESS	Language acquisition	To foster vocabulary learning, accumulate new words and phrases, improve listening, know the spelling and pronunciation of new words, and better my language intuition	44%	25%	31%
	Sound-form-meaning connection	To better understand video content To build connection between form and meaning very easily	50%	10%	40%
	Perceptual processing	To recognize quickly and remember easily because of salience of glosses To reduce visual search because of glossed target words To optimize the bimodal effect	50%	4%	46%
	Learner comfort	To alleviate anxiety and mental effort during viewing because of availability of captions To help deal with fast rate of speech	80%	/	20%
	Video impact	To arouse learners' interest because of the speaker's engaging tone and speech content	14%	36%	50%
		Total	**93%**	**58%**	**74%**

			60%	20%	20%
PERCEPTIONS OF USELESSNESS/ HARMFULNESS	Language acquisition	To retain some new words in a short period of time, but they fade away in a longer period of time To forget very easily after exposure to the video only twice			
	Form-meaning connection	To direct attention to Chinese subtitles To give learners cognitive overload via bilingual captioning Hard to find the equivalent expressions between native language and target language, because of different sentence constructions	/	43%	57%
	Perceptual processing	To increase mental effort because of information overload To fail to attend to target words because of transient information	/	50%	50%
	Learner comfort	Speech rate is fast, presentation modes of bilingual captions and glossed target words are distracting learners' focus on the speech.	/	33%	67%
		Total	**7%**	**40%**	**26%**

negative attitudes towards the use of captions. Students reported that they found captions useful for reasons related to language acquisition such as "*vocabulary acquisition, listening comprehension, oral production*" and building sound-form-meaning connection like "*It can help me disambiguate the unclear spelling, pronunciation and meaning of the audio input*." (CEC-9) Some students made particular mention of advantages of bimodal effects by saying "*I wouldn't have known some of new words if I solely listened to the audio input*." (ECL1-23)

Many students in the ECL1 and CECGW groups noted the gloss effect in facilitating form-meaning connection: "*The glossed target words were highlighted and provided Chinese meanings, helping me not only better comprehend the video but also accumulate new words and phrases*." (ECL1-13)

In addition, some students perceived captions as providing learner comfort:

> For students who have not reached the level of proficiency in English communication like me, the speed at which I understand the 'aural cues' will not be on a par with the speaker's thinking and speaking rate. Slangs coupled with accents in foreign language videos complicated the difficulty of video comprehension. And videoed captions can alleviate this situation effectively. (ECL1-32)

Several students made particular mention of the video content impact on learning experience by saying "*The video is very engaging. I like the speaker's impressive tone.*" (CECGW-53).

5. Discussion

5.1. Captioning effects on form recognition

Our results revealed that there was no statistically significant difference across the four conditions in form recognition test, following the pattern of descending effect: ECL1>NC>CECGW>CEC. This corroborated Kang et al.'s (2020) finding that L1and L2 glosses failed to significantly enhance the acquisition of visual word forms. But our result contradicted other previous research (Montero Perez et al., 2018; Hsieh, 2020), which showed that glossed keyword caption facilitated learners' form recognition of vocabulary learning compared to no caption. In Neuman and Koskinen's (1992) and Sydorenko's (2010) studies, captions significantly helped learners in written form recognition. One of the possible reasons for the discrepancy of the findings might be due to the selection of video materials. Hsieh (2020) used animations as materials, which had a stronger correlation of the audio-visual content than the Ted talk chosen in our study. The other possible reason for the inconsistency might relate to how the glossed target words were presented. In Montero Perez et al.'s learner-paced video viewing experiment, participants were able to tap the spacebar to access the meaning of glossed target words. By pausing and playing the video, learners could spend as much time as they wanted studying the word. Hence learners can better optimize their attention allocation to the dynamic and transient information presented in a multimedia learning environment. Because presenting glosses along with captions would minimize the interruption of the video or lead to cognitive overload (Hsieh, 2020), our study did not allow participants to pause or replay the video. The third possible cause might relate to the subjects' different language proficiency. Our study focused on high school learners, whose English proficiency might be lower than undergraduate participants in Hsieh's and Montero Perez et al.'s studies. In addition, ECL1 specifically provided Chinese translations below the glossed target words, which drew participants' attention to the salient non-Roman Chinese characters instead of the English words. This might aid word meaning, but would not be very useful in form recognition.

Furthermore, for the EFL learners with lower linguistic proficiency, bilingual captioning (CEC) and bilingual captioning with glossed target words (CECGW) failed to exert a facilitative effect on learners' vocabulary learning. With limited cognitive capacity to simultaneously store and process ongoing transient information during video viewing, multimodal input may have excessively loaded EFL learners with lower linguistic proficiency, hence hampering learners' active processing. Learners noted "*Because the messages presented are very transient, I cannot attend to any specific words when my focus is on comprehension of the auditory input*." (CECGW-68) It would be very difficult for learners at a low proficiency level to effectively allocate attention to take advantage of the "enriched multimodal input" (Kam et al., 2020) while viewing the fast-paced video. Consequently, the presence of textually enhanced captions made little difference to learners' vocabulary form recognition.

5.2. Captioning effects on meaning recall and recognition

The meaning recall test results followed this order: ECL1>NC>CEC>CECGW groups. The finding that ECL1 was significantly effective for vocabulary meaning learning echoed the previous findings (Montero Perez et al.,2018; Hsieh,2020; Kang et al.,2020) regarding the facility of glossed target word caption. The English caption with L1 gloss helped draw learners' attention to the language in the video, which seemed to help isolate what the learners perceived to be important and helped them determine what to pay attention to in subsequent viewings. For foreign language learners, lack of English vocabulary often poses difficulty for comprehending, so they especially need the assistance of L1 target words when watching videos to obtain novel vocabulary (Hsu et al.,2013). Most students in the ECL1 group reported that because of the salience of glosses they could quickly recognize the target words' Chinese meanings, thus fostering sound-form-meaning linkage. This also echoed Kang's finding that L1 and L2 glosses were significantly effective in consolidating form–meaning associations (Kang et al., 2020). For low-intermediate learners, because of the salience and orthographic Chinese characters, L1 glosses of the target words were especially helpful for meaning recall and facilitative in building form-meaning connections. It can be inferred from this study that enabling students to have access to meaning through glossed captions led to a significantly higher uptake rate from video in terms of initial form-meaning connections. Making words more salient through glossing but with moderate information presented would "stimulate learners' noticing and initial form-meaning connection" (Montero Perez et al.,2018).

The meaning recognition test results followed this order: ECL1>CECGW>CEC>NC. This finding suggested that captioning as a source of input, whatever type it is, facilitated learners' vocabulary learning compared to the non-caption condition. Our qualitative data also confirmed this pattern that learners reported more usefulness (93%) than harmfulness or uselessness (7%) under captioning conditions. The availability of captions "helped learners more readily segment words from the stream of speech" (Yeldham, 2018). When information via audio-visual input is beyond learners' capability, captioning would facilitate L2 learners' meaning making process, though with differential effects. Similar to the meaning recall test, the ECL1 group scored the highest among all conditions. This finding may be due to the fact that the English caption with L1 glossed target words was concise and clear, and the amount of information was moderate enough to provide learners with enough time to attend to each target word and hence foster meaningful learning.

In contrast, under the CEC and the CECGW conditions, textual information was much longer and denser. Given the same time limit for students to allocate their attention to the transient and dynamic information, greater cognitive load would occur and might pose challenge to low-intermediate high-school L2 learners' meaning-making process. Due to the limited capacity in working memory, attention to both visual channel via bilingual captioning (native language & target language) and auditory channel (target language) may overload learners' cognitive processing. Although captions could provide textual aids for learners regarding what they heard or watched in a video, the extra information could also distract learners' attention (Hsieh, 2020). Our findings about the differential effects of captioned videos on vocabulary learning may enrich Mayer's (2014) cognitive theory of multimedia learning and van Gog's (2014) cueing effect. In our study, learners' extraneous

cognitive load seemed to have increased under the CEC and CECGW modes. As Sweller et al. (2011) argued, processing information that is not relevant to learning induces extraneous cognitive load that is ineffective for learning or may even hamper learning. What's more, the orthographic differences between Chinese and English may be another contributing factor to complicating the difficulty of information processing and hence lead to an increase in learner's cognitive overload. Learners also noted "*It's difficult for me to build connection between unknown words and Chinese meaning, because sentence constructions in Chinese and English are very different*." (CECGW-76) The Chinese subtitles "interpreted the original utterance semantically instead of a word-for-word literal translation" (Dai, 2005) because of the differences underlying sentence constructions in Chinese and English. Learners might feel it hard to find the equivalent expressions between subtitles (in Chinese) and captions (in English). Since CEC mode doesn't contain features that draw learners' attention to novel items in order to activate learning, learners might lose direction in the absence of salient features of glossed target words. In our qualitative data, learners reported that they experienced difficulties understanding new words under the bilingual caption (CEC) mode, because the phonological, semantic and writing systems of English language are very different from those of the Chinese language. Presence of bilingual captions seemed to distract learners from attending to the transient auditory cues. *Failing to attend to essential message* or *attending to extraneous message* may hamper the subsequent process of integrating and synthesizing information.

In the current study the multimedia message was delivered via both channels at a rapid pace while learners need adequate time to process this information beyond their linguistic proficiency. Therefore, it would become very challenging for low-intermediate high school EFL learners to select, organize and integrate the essential multimedia instructional message given the limited time during system-paced video viewing. Carrying out cognitive processing takes time, but a fast-paced presentation that requires a lot of mental model building may not allow enough time (Mayer & Pilegard, 2014). In this regard, effective cueing can be implemented to help learners use their limited working memory capacity in an optimal manner through selecting, organizing, and integrating the information presented in the text and pictures (van Gog, 2014). The use of suitable cognitive load-reducing techniques via effective cueing, like creating visual saliency under ECL1 condition in our study, may reduce learners' visual search and then foster learning.

5.2. Pedagogical implications

Given the wide variety in cues used, it is hard to distill one-size-fits-all instructional guidelines for EFL teachers regarding when cueing is needed, what elements of the text should be cued, and what type of cue is most useful. As Gerbier et al. (2018) claimed, highlighting words should trigger a visual-attentional capture toward the word being highlighted (and simultaneously heard), and with respect to literacy skills, this audio-visual synchrony may boost phonological, orthographic and semantic learning thanks to the temporal congruency between visual and audio inputs.

Based on findings in our study, instructional designers could utilize technological affordances to create captioned videos with glossed target words such as ECL1 to facilitate

the low-intermediate high-school EFL learners' vocabulary learning. This may very well, as Webb & Chang (2012) suggested, strengthen the form (phonological and orthographic) and meaning connections that contribute to vocabulary development. Bilingual captioning (CEC), though commonly practiced among EFL learners after class in the Chinese context, seemed not to be so effective and did not yield satisfactory vocabulary learning gains for learners at low proficiency. Learners are not exploiting different sources of information under this mode, probably because the redundant information may increase the cognitive load of the task. Therefore, having an effective visual cue to auditory word segmentation might contribute to learners' enjoyment and engagement with audio cues.

In addition, as learners reported engaging videos could optimize learning experience and teachers are advised to carefully select and provide opportune multimedia learning resources for learners to utilize after class. Students also reported they tend to use the audio-visual input as a "model" to improve their spoken English by emulating the speaker's intonation and pronunciation. Namely, they were shadow-reading and were consciously vocalizing to themselves language that they have been exposed to during video viewing. Shadow-reading offered repeated opportunities for listening, reading, understanding, and vocalizing L2 segments (De Guerrero & Commander, 2013). Although imitation may result in a close copy of the original, it has a potential for creativity, transformation, and transcendence. The saliency of the highlighted target word increased the likelihood that it will be noticed and acquired and that connections between forms and meanings will be built.

When using videos to facilitate vocabulary acquisition, EFL teachers should be encouraged to use glossed L1 target word captioning because 'it might facilitate students' recognition of unknown words and their making initial form-meaning connections' (Montero Perez et al., 2014). However, to avoid over-reliance on captions in video viewing, teachers can adopt a 'staged video approach' (Danan, 1992). Teachers could show the video first with ECL1, then transitioned to the no caption condition. By gradually decreasing the amount of text, teachers might help learners progressively minimize the amount of support while optimizing chances for word learning. In so doing, affective benefits can flow from caption use through providing learners with the confidence and self-efficacy in knowing that such comprehension of these authentic materials is within their grasp.

6. Conclusion

This study focused on the effectiveness and perceived usefulness of textually-enhanced captions for Chinese high-school EFL learners' vocabulary learning. Our findings revealed that: (1) ECL1 captioning videos positively affected vocabulary learning in terms of meaning recall and meaning recognition. The salient glosses of the target words and L1 translation reduced the possibility of information overload and enhanced the sound-form-meaning association. (2) Videos with bilingual captions (CEC) did not seem to foster vocabulary growth and learners reported that presentation modes of bilingual captions are distracting learners' focus on the speech. (3) Videos without any caption did not seem to help vocabulary learning, indicating that when the difficulty of the video exceeded the learners' proficiency level, captioned videos lacking L1 translation to the essential message (e.g. target words) had

little effect on low-intermediate learners' vocabulary acquisition. Therefore, effective and opportune scaffolding, such as glossed target words captioning (ECL1) adopted in this study, can alleviate the cognitive burden on the students and help nurture their English learning.

Though evidence was found for different captioning effects on vocabulary learning, this study has several limitations. First, the present experiment was limited in scope involving high school EFL learners at a low-intermediate level. Second, we used questionnaires to elicit learners' reports about their perceptions of caption usefulness, their viewing behavior and attention allocation under different viewing conditions. Future study needs to employ other data collection methods, like eye-tracking technology to gather a more objective picture of learners' viewing behavior. Third, though we conducted a pilot study to elicit suggestions from some participants and experienced teachers when selecting the target words, a baseline vocabulary test can be administered in the future study to measure participants' vocabulary size more objectively.

The abundant possibilities outlined above illustrate the vast and yet uncharted potential of the use of glossed target words captioning (ECL1) in the practice of teaching and learning in the EFL classroom.

7. References

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Academic Press.

Costales, A. F. (2021). Subtitling and dubbing as teaching resources in CLIL in Primary Education: The Teachers' perspective, *Porta Linguarum*, *36*, 175-192. https://doi.org/10.30827/portalin.v0i36.16228.

Dai, J. (2005). Captioned video and foreign language learning. *Technology Enhanced Foreign Language Education*, *103*, 18-22.

Danan, M. (1992). Reversed subtitling and dual coding theory: New directions for foreign language instruction. *Language Learning*, *42*(4), 497–527. https://doi.org/10.1111/j.1467-1770.1992.tb01042.x.

De Guerrero, M., & Commander, M. (2013). Shadow-reading: Affordances for imitation in the language classroom, *Language Teaching Research*, 2013, *17*(4), 433-453. https://doi.org/10.1177/1362168813494125.

Frick, T.W., Chadha, R., Watson, C., Wang, Y., & Green, P. (2009). College student perceptions of teaching and learning quality. *Educational Technology Research and Development*, *57*(5), 705–720. https://doi.org/10.1007/s11423-007-9079-9.

Gass, S. M. (1999). Discussion: Incidental vocabulary learning. *Studies in Second Language Acquisition*, *21*(2), 319–333.

Gerbier, E., Bailly, G., & Bosse, M. (2018). Audio-visual synchronization in reading while listening to texts: Effects on visual behavior and verbal learning. *Computer Speech and Language*, *47* (1), 79-92. https://doi.org/ff10.1016/j.csl.2017.07.003f.

Guillot, M. N. (2020). Ocean's eleven stand-alone scene 12 with subtitles–a gift for teaching, what lessons for research? *Perspectives: Studies in Translation Theory and Practice*. *28*(6), 822-836. https://doi.org/10.1080/0907676X.2019.1701053.

Hsieh, H.F., & Shannon, S. E. (2005) Three approaches to qualitative content analysis. *Qualitative Health Research*, *15*(9), 1277–1288. https://doi.org/10.1177/1049732305276687.

Hsieh, Y. (2020). Effects of video captioning on EFL vocabulary learning and listening compre-

hension. *Computer Assisted Language Learning, 33*(5-6), 567-589. https://doi.org/10.1080/09588221.2019.1577898.

Hsu, C.K., Hwang, G.J., Chang, Y.T., & Chang, C.K. (2013). Effects of video caption modes on English listening comprehension and vocabulary acquisition using handheld devices. *Educational Technology & Society*, *16*(1), 403-414.

Kam, E.F., Liu, Y.T., & Tseng, W.T. (2020). Effects of modality preference and working memory capacity on captioned videos in enhancing L2 listening outcomes. *ReCALL*, *32*(2), 213–230. https://doi.org/10.1017/S0958344020000014.

Kanellopoulou, C. (2019). Film subtitles as a successful vocabulary learning tool. *Open Journal of Modern Linguistics*, *09*(02), 145–152. https://doi.org/10.4236/ojml.2019.92014.

Kang, H., Kweon, S. & Choi, S. (2020). Using eye-tracking to examine the role of first and second language glosses. *Language Teaching Research, 6*, 1-22. https://doi.org/10.1177/1362168820928567.

Leonhardt, J. (2020). Using film and media in the language classroom: reflections on research-led teaching. *ELT Journal, 74*(2), 229–231. https://doi.org/10.1093/elt/ccaa007.

Lertola, J. (2019). Second language vocabulary learning through subtitling. Revista *Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics*, *32*(2), 486–514. https://doi.org/10.1075/resla.17009.ler.

Leveridge, A.N., & Yang, J.C. (2014). Learner perceptions of reliance on captions in EFL multimedia listening comprehension, *Computer Assisted Language Learning*, *27*(6), 545-559. https://doi.org/10.1080/09588221.2013.776968.

Li, M. (2016). An Investigation into the differential effects of subtitles (first language, second language, and bilingual) on second language vocabulary acquisition. PhD thesis. Edinburgh: University of Edinburgh.

Liao, S., Kruger, J., & Doherty, S. (2020). The impact of monolingual and bilingual subtitles on visual attention, cognitive load, and comprehension. *The Journal of Specialised Translation*, 33, 70-98. https://www.jostrans.org/issue33/art_liao.pdf.

Mayer, R.E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp.55-81). New York: Cambridge.

Mayer, R.E., & Pilegard, C. (2014). Principles for managing essential processing in multimedia learning: segmenting, pre-training, and modality principles. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp.307-337). New York: Cambridge.

Montero Perez, M. (2020). Incidental vocabulary learning through viewing video: the role of vocabulary knowledge and working memory. *Studies in Second Language Acquisition*, *42*(4), 1-25. https://doi.org/10.1017/S0272263120000145.

Montero Perez, M., Peters, E., & Desmet, P. (2018). Vocabulary learning through viewing video: The effect of two enhancement techniques. *Computer Assisted Language Learning, 31*(1), 1-26. https://doi.org/10.1080/09588221.2017.1375960.

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning and Technology*, *18*(1), 118-141. http://llt.msu.edu/issues/february2014/monteroperezetal.pdf.

Moreno, R., & Abercrombie, S. (2010). Promoting awareness of learner diversity in prospective teachers: Signaling individual and group differences within classroom cases. *Journal of Technology and Teacher Education*, *18*(1), 111–130. https://www.learntechlib.org/primary/p/29271/.

Neuman, S.B., & Koskinen, P. (1992). Captioned television as comprehensible input: Effects of incidental word learning from context for language minority students. *Reading Research Quarterly*, *27*(1), 95–106.

Schmidt, R.W. (1990). The role of consciousness in second language learning. *Applied Linguistics*, *11*(2), 129-158. https://doi.org/10.1093/applin/11.2.129.

Sweller, J., Ayres, P., & Kalyuga, S. (2011). Cognitive load theory. New York: Springer.

Sydorenko, T. (2010). Modality of input and vocabulary acquisition. Language Learning & Technology, *14*(2), 50–73. http://llt.msu.edu/vol14num2/sydorenko.pdf.

Teng, F. (2019). Maximizing the potential of captions for primary school ESL students' comprehension of English-language videos. *Computer Assisted Language Learning*, *32*(7), 665-691. https://doi.org/10.1080/09588221.2018.1532912.

van Gog, T. (2014). The signaling (or cueing) principle in multimedia learning. In R. E. Mayer (Ed.), *Cambridge handbook of multimedia learning* (pp. 255-269). Cambridge.

Webb, S., & Chang, A. (2012). Vocabulary learning through assisted and unassisted repeated reading. *The Canadian Modern Language Review*, *68*(3), 267-290.

Webb, S., & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.

Yeldham, M. (2018). Viewing L2 captioned videos: What's in it for the listener? *Computer Assisted Language Learning*, *31*(4), 367-389. https://doi.org/10.1080/09588221.2017.1406956.

Yun, J. (2011). The effects of hypertext glosses on L2 vocabulary acquisition: A meta-analysis. *Computer Assisted Language Learning*, *24*(1), 39–58. https://doi.org/10.1080/09588221.2010.523285.

Zanón, N. T. (2006). Using subtitles to enhance foreign language learning. *Porta Linguarum: Revista Internacional de Didáctica de Las Lenguas Extranjeras*, *6*, 41–52. http://www.ugr.es/~portalin/articulos/PL_numero6/talavan.pdf.

Acknowledgements This research was supported by a Grant from the National Education Sciences Planning of Chinese Ministry of Education (No. DBA210298). We would like to express our thanks to Ms. Li Wang for helping collect the data and all students for voluntary participation in the research, and to the anonymous reviewers and the editors for their constructive feedback.

Note: All research instruments will be provided by the authors upon request. Please contact the Corresponding author Xiaohu Yang via Email: 19026@tongji.edu.cn.