Transforming our World through Universal Design for Human Development
I. Garofolo et al. (Eds.)

doi:10.3233/SHTI220884

Towards a More Inclusive Learning Environment: The Importance of Providing Captions That Are Suited to Learners' Language Proficiency in the UDL Classroom

Shamira VENTURINI[a], Michaela Mae VANN[a], Martina PUCCI[a], and Giulia M. L. BENCINI[a, 1]
[a] *Ca' Foscari University of Venice*

Abstract. Captions have been found to benefit diverse learners, supporting comprehension, memory for content, vocabulary acquisition, and literacy. Captions may, thus, be one feature of universally designed learning (UDL) environments [1, 4]. The primary aim of this study was to examine whether captions are always useful, or whether their utility depends on individual differences, specifically proficiency in the language of the audio. To study this, we presented non-native speakers of English with an audio-visual recording of an unscripted seminar-style lesson in English retrieved from a University website. We assessed English language proficiency with an objective test. To test comprehension, we administered a ten-item comprehension test on the content of the lecture. Our secondary aim was to compare the effects of different types of captions on viewer comprehension. We, therefore, created three viewing conditions: video with no captions (NC), video with premade captions (downloaded from the university website) (UC) and video with automatically generated captions (AC). Our results showed an overall strong effect of proficiency on lecture comprehension, as expected. Interestingly, we also found that whether captions helped or not depended on proficiency and caption type. The captions provided by the University website benefited our learners only if their English language proficiency was high enough. When their proficiency was lower, however, the captions provided by the university were detrimental and performance was worse than having no captions. For the lower proficiency levels, automatic captions (AC) provided the best advantage. We attribute this finding to pre-existing characteristics of the captions provided by the university website. Taken together, these findings caution institutions with a commitment to UDL against thinking that one type of caption suits all. The study highlights the need for testing captioning systems with diverse learners, under different conditions, to better understand what factors are beneficial for whom and when.

Keywords. automatic speech recognition, captions, English language learners, foreign language instruction, multi-modality, universal design, universal design for learning, universal design and individual differences.

[1] Giulia Bencini, Department of Linguistics and Comparative Cultural Studies, Ca' Foscari University of Venice, Ca' Bembo, Fondamenta Toffetti – Dorsoduro 1075, Venice, Italy; E-mail: giulia.bencini@unive.it

1. Introduction

Captions (same-language subtitles) have been shown to be one of the tools for enhancing language comprehension and learning of content in diverse populations. Gernsbacher [1] summarizes the positive findings of over 100 published empirical studies that reported benefits of providing captions for children learning how to read, adults, learners with and without hearing impairments, and learners of a second language (L2). The benefits across different studies include improvements in listening comprehension, vocabulary acquisition, memory for content and literacy development. In some countries the use of captions in TV and multimedia products is regulated by law (see - for example - the 21st Century Communications and Video accessibility Act of 2010 in the US [2]) and implemented by national broadcasting channels (such as the BBC in the UK [3]). The use of captions in video content is also recommended by some Universal Design for Learning (UDL) guidelines [4]. Since diversity can be along many dimensions, including sensory, cognitive, and linguistic, it is suggested that providing captions is one of the options instructors have to turn unimodal (spoken) or bimodal content (video+spoken) into multi-modal content, thus increasing accessibility to lecture content [4, 5, 6]. Universities and institutions involved in Higher Education worldwide are increasingly adopting UDL guidelines to promote inclusion and to meet the needs of diverse student populations [7, 8], for example adopting the use of captions during live lectures [5].

In the field of second language learning (L2) and instruction, researchers have studied the effects of captioned video-content on second language learning for more than 30 years, generally finding positive effects [9]. Many studies point out how L2 learners benefit from captions, reporting positive effects on listening comprehension, vocabulary learning, and pronunciation [6, 9, 10]. Other researchers, however, caution against generalizing these results to all types of L2 learners, because of individual differences in second language proficiency. Currently, the existing data on the relationship between language proficiency and caption use (whether beneficial, neutral or detrimental) is mixed [11]. This motivated the current study. The aims of the study are twofold. First, we wished to assess if captions are always useful, or if learner differences - namely their proficiency in the language of the audio - play a role in the efficacy of the captions. To do so, participants were asked to watch an audio-visual recording of an unscripted seminar-style lesson in English retrieved from a university website [12]. Second, we wished to compare the effects of different caption types on viewer comprehension. We tested participants under three viewing conditions: premade captions provided by the host University (UC), automatic captions (AC), and no captions (NC).

2. Methods

2.1. Participants

80 non-native speakers of English participated in the study. They were Italian university students enrolled in an English as a Foreign Language course (L2 English) (age $M = 22$, $SD = 5$). Prior to participating, their English proficiency was assessed using the grammar portion of the Michigan Test of Language Proficiency (MTELP). This test consists of a set of 45 multiple choice questions that are presented aurally.

2.2. Experimental Task

To investigate whether learners with different proficiency levels benefit from captions, participants were asked to watch and listen to a 10-minute video of a seminar style lecture in English, under different viewing conditions (see 2.3). The material was an authentic video-lecture downloaded from MIT Courseware [12]. The video-lecture came with captions that could optionally be added. In the video, an instructor discussed a topic in linguistics (Creoles and Pidgins) and interacted with students. This type of audiovisual content is commonly used in English as a foreign language university programs, so the task was familiar to the participants.

2.3. Design and Procedure

The experimental design was a one-way factorial between-subjects design with three levels, corresponding to three viewing conditions for the audio-visual lecture: video-lecture with human corrected captions, available on the university's website (UC), video-lecture with automatically generated captions re-generated using YouTube (AC), and video-lecture with no captions (NC). Participants were randomly assigned to one of the three viewing conditions: AC, $N = 26$; UC, $N = 27$; NC $N = 27$. The language assessment test (MTELP) and the experimental task were embedded in Qualtrics XM and administered remotely. MTELP was administered first. After viewing the video-lecture, participants completed a 10-question multiple choice comprehension test on the content of the lecture.

3. Data Analysis

Our dependent variable was comprehension of the content of the video-lecture. Comprehension scores were computed for each participant by dividing the number of correct answers on the comprehension test out of the total number of questions (percent correct). To analyze the data we used a generalized linear model predicting comprehension. English language proficiency scores were numerical: participants received one point for each correct answer on the MTELP (maximum score: 45). MTELP scores were used as continuous covariate in our data analyses. The three-level viewing condition (AC, UC, NC), proficiency, and their interactions were entered in the model. Caption conditions (AC and UC) were contrasted with the no captions condition (NC), and follow up pairwise comparisons contrasted AC with UC.

4. Results

4.1. Proficiency

Participants' English language proficiency on the MTELP ranged from 15 to 45 (maximum score on the test). The mean score was 40.3 ($SD = 6$) and the median was 42, indicating that our sample contained a large proportion of higher proficiency speakers, corresponding to advanced B2-C1 on the Common European Framework of Reference for Languages (CEFR).

4.2. Comprehension performance

The graph in Figure 1 plots comprehension scores (percent correct) as a function of English language proficiency (MTELP score, 15-45) and viewing condition: captions provided by the university (UC, blue line), automatically generated captions (AC, red line) and no captions (NC, green line). Learners' comprehension of the content of the video lecture in English was modulated by proficiency. Learners who had higher proficiency scores in English, on average, performed better on the comprehension test, as is to be expected. The effects of viewing condition and caption type differed greatly between lower proficiency and higher proficiency speakers, though. This is best seen by inspecting the graph, where there is an indication of a cross-over interaction between caption type and proficiency (graphically: red line above the blue line at lower proficiency levels, crossing over to blue line above the red line at higher proficiency levels). Because the pattern of results is complex, we will go over it step by step, dividing the presentation by English language proficiency range.

4.2.1. Low-to-middle proficiency range (MTELP between 15-30, corresponding to lower B2 on the CEFR)

Those learners who viewed the video with the captions provided by the University website (UC, blue line) performed worse than viewers at the same proficiency level who were not given captions. Automatic captions, however, did provide benefits to these viewers (red line above both green and blue lines).

4.2.2. Mid-to-high end of the proficiency scale (MTELP between 35-38, corresponding to B2 on the CEFR)

The difference between viewing conditions disappears, and comprehension performance is not affected by viewing condition.

4.2.3. High range (MTELP > 38, corresponding to B2+ to C1 range on the CEFR)

For learners who had higher levels of proficiency, especially those in the MTELP 40-45 range, there is a numerical trend suggesting that they benefited more from the university captions (UC condition) than the automatic captions (AC) or no captions (blue line above the red and the green lines). This difference, however, is not significant.

Pairwise comparisons of estimated marginal means of linear trends confirmed that there is a significant difference between the AC (red) and UC (blue) lines' slopes ($p < 0.05$). The positive estimate indicates that learners who were less proficient in English performed significantly worse with the captions provided by the university (UC) than with automatic captions (AC). In the same fashion, the negative estimate resulting from comparing UC and NC suggests that participants in the latter group performed better with no captions than with the UC, even though this was not statistically significant. The results of the pairwise comparisons are reported in Table 1.

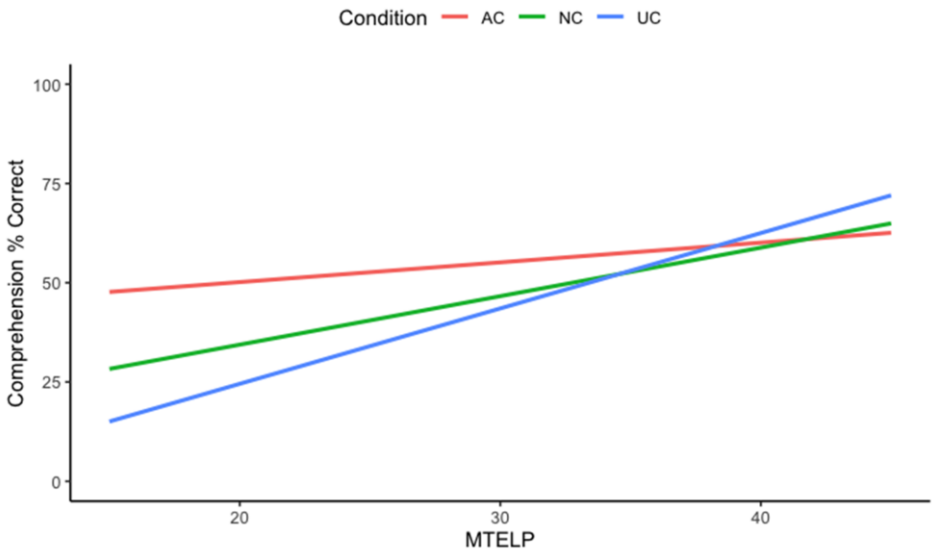

Figure 1. Learner comprehension scores when watching video-lecture under three viewing conditions (AC = automatic captions; UC = captions provided by the University; NC = no captions) as a function of English language proficiency scores (MTELP).

Table 1. Pairwise estimated marginal means of linear trends.

Comparison	Estimates	SE	DF	z ratio	p value
AC - UC	0.007	0.003	Inf	2.42	<0.05
AC - NC	0.003	0.002	Inf	1.44	n.s.
UC - NC	-0.003	0.003	Inf	-10.7	n.s.

5. Discussion and conclusion

Many studies have shown that captions (same-language subtitles) support listening comprehension, and learning of content in diverse populations [1]. Some of the UDL guidelines encourage instructors to adopt captions, promoting inclusion of diverse learners [4] as providing captions for lectures may help students with different cultural and linguistic backgrounds as well as diverse needs, and may support the learning process [1, 4, 5, 7, 8].

In this study we assessed the usefulness of captions in L2 English learners with different levels of proficiency, as learners' language proficiency is an understudied variable in the UDL literature, and results that consider this variable are mixed [11]. We also wished to compare the effects of different caption types (UC, AC and NC) on viewers' comprehension. We tested L2 learners' comprehension of a seminar-style lecture under different viewing conditions to examine how proficiency may interact with the type of captioning system provided.

Our comprehension results showed that whether students benefited from captions depended *both* on their language proficiency level and on the type of caption provided. In general, participants' language proficiency predicted content comprehension with a significantly positive correlation ($p < 0.001$). Captions had a larger impact on

comprehension at lower language proficiency levels, but their effect (whether positive or negative) depended on the type of caption. At lower proficiency levels, the captions provided by the University (UC condition) turned out to have a detrimental effect, whereas automatic captions (AC condition) significantly supported comprehension in comparison to the UC condition ($p < 0.05$). At mid and high proficiency levels, viewing condition did not affect comprehension.

We now turn to discuss why lower proficiency speakers did better in the AC condition than in the UC condition, and why the UC condition appeared to be so detrimental for low proficiency learners. In fact, perhaps surprisingly, the captions provided by the university (UC condition) turned out to make things worse for low proficiency speakers than not providing any captions at all. We propose that the reason for these opposite effects is to be attributed to differences in how the different captioning systems displayed content, which interacted with viewer proficiency. The one salient feature that differed substantially between systems was the amount and distribution of text on the screen relative to speech onset in the audio. In the AC condition text was presented in a word by word (incremental) format, closely synchronized with the speech signal (this is characteristic of automatic speech recognition systems) (see Figure 2a). The captions presented in the UC condition, on the other hand, presented a greater amount of text that appeared all at once, distributed over two lines (see Figure 2b). So, in addition to containing more text, UC captions were not synchronized with speech.

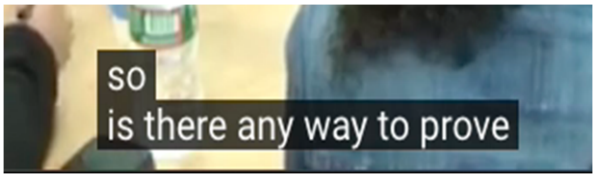

Figure 2a. Display format for the AC condition (speech-synchronised incremental format)

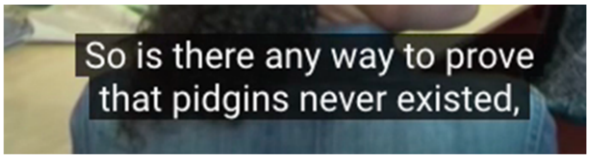

Figure 2b. Display format for the UC condition (two-line format)

We suggest that automatic captions facilitated speech segmentation and word identification because this system provided a better temporal alignment with the actual timing of spoken input [5, 6, 9, 13]. Conversely, captions in the UC condition were not aligned, and this may have resulted in cognitive over-load and hindered comprehension. Studies have begun to investigate the impact of different caption types on cognitive load, which is believed to play a role when it comes to processing multimodal input [14, 15, 16, 17, 18]. CL is defined as "the load imposed on the learner's cognitive system while performing a particular task" [14, p. 241].

An open question is whether students with higher proficiency in English had higher speech decoding and spoken English comprehension abilities, thus managing to integrate the information provided in the captions with the audio, or whether they simply ignored these captions all together, relying on the spoken input only. This question can only be addressed in studies that examine whether and how viewers are processing captions, such as eye-tracking studies. This is a topic for future research.

In summary, in this study we found that students' language proficiency in the language of the audio and the type of captioning system, are both important variables to take into account when choosing to provide learners with multi-modal material in the form of captioned video-lectures.

Two general recommendations emerge from this work: first, instructors need to be aware of the fact that variability in language proficiency in the language of the lecture will impact whether or not students find captions useful. Second, for low proficiency speakers captioning systems that are closely synchronized with speech may work better than ones that are not.

Author contributions

Shamira Venturini - conceptualization, formal analysis, investigation, methodology, writing
Michaela Mae Vann - conceptualization, formal analysis, methodology, writing
Martina Pucci - writing
Giulia Bencini - conceptualization, methodology, supervision, writing

Research Ethics Statement

The study was conducted following the principles of the 1964 Declaration of Helsinki. The experimental procedures were approved by the University Ethics Board at Ca' Foscari University of Venice (protocol approval number 1/2020). All participants provided their informed consent prior to participating.

Funding

M. M. Vann was supported by a doctoral student scholarship under the "Department of Excellence" award to the Department of Linguistics and Comparative Cultural Studies, funded by the Italian Ministry of Universities and Research (MUR).
M. Pucci was supported by a doctoral student Research and Innovation scholarship (PON Ricerca & Innovazione 2014-2020 – CUP: H75F21002110005) funded by the Italian Ministry of Universities and Research (MUR) and the European Union (FSE REACT EU).

References

[1] Gernsbacher, MA. Video Captions Benefit Everyone. PIBBS. 2015;2(1):195-202.
[2] 21st Century Communications and Video accessibility Act, 2010, H. R. 3101, 111th Cong. [cited 2022 Apr]. Available from: https://www.govinfo.gov/content/pkg/BILLS-111hr3101pcs/pdf/BILLS-111hr3101pcs.pdf.
[3] BBC.github.io, BBC Subtitle Guidelines [Internet]. 2021 [cited 2022 Apr]. Available from: https://bbc.github.io/subtitle-guidelines/.
[4] CAST The UDL Guidelines, Universal Design for Learning Guidelines version 2.2 - checkpoint 1.2. 2018 [cited 2022 Apr]. Available from: https://udlguidelines.cast.org/representation/perception/alternatives-auditory.

[5] Wald M, Bain K. Universal access to communication and learning: the role of automatic speech recognition. UAIS. 2008;6:435-47.
[6] Garza TJ. Evaluating the use of captioned video materials in advanced foreign language learning. Foreign Language Annals. 1991;24:239-58.
[7] Bencini G, Garofolo I, Arenghi A. Implementing Universal Design and the ICF in Higher Education: Towards a Model That Achieved Quality Higher Education For All. Proceedings of Universal Design and Higher Education in Transformation Congress (UDHEIT2018): learning from the past, designing for the future. Studies in Health Technologies and Informatics; 2018 Oct 30-Nov 2; Amsterdam, Berlin. Washington DC: IOS Press; 2018. p. 464-72.
[8] Bencini G, Arenghi A, Garofolo I. Is my University Inclusive? Towards a multi-domain instrument for Sustainable Environments in Higher Education. Proceedings of Universal Design 2021: From Special to Mainstream Solutions. IOS Press; 2021. p. 137-43.
[9] Montero Perez, M. Second or foreign learning through watching audio-visual input and the role of on-screen text. Language Teaching. 2022;55:163-92.
[10] Montero Perez M, Van Den Noortgate W, Desmet P. Captioned video for L2 listening and vocabulary learning: A meta-analysis. System. 2013;41:720-39.
[11] Gass S, Winke P, Isbell DR, Ahn J. How captions help people learn languages: A working-memory, eye-tracking study. LLT. 2013; 23(2):84-104.
[12] Michel DeGraff , ocw.mit.edu, MIT Opencourseware, Creole Languages and Caribbean Identities, course 24.908, lesson 1. Do "Pidgins" exist? Do creoles come from pidgin? Spring 2017 [cited 2022 Apr]. Available from: https://ocw.mit.edu/courses/24-908-creole-languages-and-caribbean-identities-spring-2017/.
[13] Shimogori N, Ikeda T & Tsuboi S. Automatically generated captions: Will they help non-native speakers communicate in English? Proceedings of the 3rd ACM International Conference on Intercultural Collaboration, ICIC '10; 2010 Aug 19-20; New York, NY: Association for Computing Machinery; 2010. p. 79–85.
[14] Chan WS, Kruger J, Doherty S. Comparing the impact of automatically generated and corrected subtitles on cognitive load and learning in a first- and second-language educational context. LANS – TTS. 2019;18:237-72.
[15] Diao Y, Chandler P & Sweller J. The effect of written text on comprehension of spoken English as a foreign language. AJP. 2007;120(2):237–61.
[16] Kalyuga S, Chandler P & Sweller J. Managing split-attention and redundancy in multimedia instruction. ACP. 1999;13(4):351–71.
[17] Kruger J, Hefer E, Matthew G. Measuring the Impact of Subtitles on Cognitive Load: Eye Tracking and Dynamic Audiovisual Texts. Proceedings of Eye Tracking South Africa; 2013 Aug 29-31; Cape Town, South Africa. p. 62-66.
[18] Winke P, Sydorenko T, Gass S. Factors Influencing the Use of Captions by Foreign Language Learners: An Eye-Tracking Study. MLJ. 2013;97(1):254-75.