THE EFFECTS OF BI-MODAL INPUT ON FOSTERING L2 JAPANESE SPEECH SEGMENTATION SKILLS

by

Natsumi Suzuki

A Dissertation

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Doctor of Philosophy

School of Languages and Cultures
West Lafayette, Indiana
May 2019

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Atsushi Fukada, Chair

School of Languages and Cultures

Dr. April Ginther

Department of English

Dr. Mariko Moroishi Wei

School of Languages and Cultures

Dr. Jessica Sturm

School of Languages and Cultures

Approved by:

Dr. Jennifer William

Head of the Graduate Program

For my parents.

ACKNOWLEDGMENTS

Many people have assisted me in completing this dissertation, both directly and indirectly. I cannot thank them all here, but I would like to express my very great appreciation to the faculty and staff at the School of Languages and Cultures for giving me an opportunity to continuously grow, both as a researcher and as a language instructor. I would like to offer my deepest appreciation to the chair of my committee member and my advisor, Professor Atsushi Fukada. He has given me continuous help, support, and feedback ever since I arrived at Purdue University six years ago. Without his guidance, this Ph.D. would not have been achievable. I would like to extend my appreciation to my committee members, Professor April Ginther, Mariko Moroishi Wei, and Jessica Sturm, whose expertise was invaluable in the formulating of the research topic and methodology in particular. I also thank Mayu Miyamoto and Naoko Takano for helping me rate my data, and all my peers with whom I have had the pleasure to work during this and other related projects. I also must acknowledge the hard work of the participants in this study, as they are the motivation of my research. I was inspired by their dedication in learning Japanese, and it was rewarding working with them. Finally, I would like to thank my parents for giving me the opportunity to start my educational journey in the United States, and for always supporting me with every decision I take. And last but not least, to my husband and my biggest supporter, Lukas Brenner, who has been by my side throughout my graduate career. Thank you for your unconditional support and encouragement, and always being there for me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Author: Suzuki, Natsumi. PhD
Institution: Purdue University
Degree Received: May 2019
Title: The Effects of Bi-Modal Input on Fostering L2 Japanese Speech Segmentation Skills
Committee Chair: Atsushi Fukada

The purpose of this study was to investigate to what extent bi-modal input improves the word segmentation ability of L2 learners of Japanese. Accurately identifying words in continuous speech is a fundamental process for comprehending the overall message, but studies show that second language (L2) learners often find this task difficult, even when all individual words are familiar to them (e.g. Field, 2003; Goh, 2000). This is where the combination of written and audio input (bi-modal input), like when providing captions in the target language, could be helpful because it can provide orthographical image of the sound they hear, which in turn makes the input more intelligible (Charles & Trenkic, 2015). This study was implemented through a single-case design (SCD), where 12 third-year Japanese learners at a public university in the Midwestern United States underwent a semester-long pre-post design experiment. Participants watched a series of Japanese documentary with sound and captions (bi-modal input) throughout the semester. Before and after viewing each video, participants took Elicited Imitation Tasks (EIT) as the pre-post-tests, as well as at the beginning and at the end of the semester. The result showed that most participants improved their EIT scores throughout the semester, even to utterances from videos and speakers to which they had not been exposed. This study provided evidence that bi-modal input has the potential to help learners' internal phonological representations of lexical items to become more stable and sophisticated, which would in turn contribute to L2 Japanese learners' speech processing efficiency.

CHAPTER 1. INTRODUCTION

Background

The ultimate goal of language learning is to be able to communicate meaning, and this means to share and receive information. When receiving information through auditory input, the listener must relate the sound to meaning by mapping phonological forms to intended words in their memory (Magnuson, 2016). While many languages have clear word boundaries such as spaces in written text, continuous speech has no boundaries (for instance, pauses) between words, and it needs to be divided into discrete units such as phrases, words, and morphemes. This process of detecting word boundaries in continuous speech is called speech segmentation, and is a complex process that requires multiple simultaneous activation of phonetic, semantic, syntactic, and contextual information. Also, in written text, readers can always go backwards in the text in order to decode the overall message, but this cannot be done in listening since auditory information is timebound and transient, which requires listeners to quickly process the words online.

Despite its complexity, it does not usually pose a problem for native language listeners in understanding speech because they are highly efficient in processing information and can rely on their abundant linguistic knowledge while referring to other information such as contextual cues and their prior knowledge (Mattys & Bortfeld, 2016). Second language (L2) learners, on the other hand, find this task difficult because their lexical and syntactic knowledge is underdeveloped compared to native language users, and they also must learn the phonology of the L2. In general, languages with comparable phonotactic rules help facilitate the transfer of segmentation, but languages with different regularities may create interference (Mattys &

Bortfeld, 2016). As Mattys and Bortfeld state, “the process of becoming a proficient L2 listener involves not only vocabulary growth and syntactic learning but also increased control over potentially contradictory sets of non-lexical segmentation strategies” (p. 68).

Studies have shown that vocabulary growth does not always lead to better segmentation skills (Bonk, 2000; Ching-Shyang Chang, 2007; Goh, 2000; Milton & Hopkins, 2006; Van Zeeland & Schmitt, 2013). Learners often learn new words by memorizing the spelling and the meaning of the word, but neglect to remember how the words sound (Goh, 2000) and this results in a mismatch of their word orthography and word phonology knowledge. Milton and Hopkins' (2006) study showed that even advanced level learners of English had a phonological recognition vocabulary much smaller than their written vocabulary recognition, and other studies have shown that even when learners were familiar with the lexical items presented aurally, they were not able to recognize them in connected speech (Bonk, 2000; Ching-Shyang Chang, 2007). In language textbooks, supplementary vocabulary words are often provided when learners listen to a passage or a dialogue, but simply presenting them in writing does not necessary mean that they would comprehend them when they listen to the passage; they may not even notice them in connected speech. Although listening is one of the most complex skills in language learning (Montero Perez, Peters, & Desmet, 2015), listening practice is often subordinated to grammar and vocabulary learning in the classroom (Graham & Santos, 2015), and language teachers often find teaching of listening to be particularly challenging (Chambers, 1996; Field, 2008).

Rost (2011) states that misunderstanding or non-understanding of words in speech, whether through faulty identification of word boundaries or inadequate knowledge of word meanings, is a major source of confusion in language comprehension, particularly L2 comprehension. Field (2003) supports this view by stating that the most common cause of breakdown in L2 understanding is the identification of words in connected speech, or lexical segmentation. While speech segmentation is difficult to perform when listening to L2, it is essential for successful L2 listening comprehension and more generally for L2 learning. Then how can we assist L2 learners to foster speech segmentation skills to accurately identify words in connected speech? One way to do so is to use bi-modal input. Bi-modal input here means that people hear something and also see it written at the same time (simultaneous presentation of matching aural and orthographic stimuli), as it is used in target-language subtitles/closed captions provided for the hearing-impaired. These will be referred to as “captions” as in Vanderplank's article (2010), in which he uses “captions” to refer to the same-language subtitles (either L1 aural input with L1 subtitles or L2 aural input with L2 subtitles) and “subtitles” for translation or native language subtitles (L2 aural input with L1 subtitles). The combination of written and audio input, like when providing captions in the target language, has the potential to help develop speech segmentation skills by making the input more intelligible.

This idea of the use of bi-modal input started to receive considerable attention in the early 1980s through Allan Paivio’s Dual Coding Theory (Paivio, 1986), in which he promoted the idea that the brain processes verbal and non-verbal stimuli via two different cognitive systems, and when these two systems interact with one another, it results in improved information processing and better information recall because both visual and verbal codes can be used when recalling information. A more recent theory developed on the basis of Paivio’s Dual Coding Theory is Richard Mayer’s Multimedia Principle. Mayer postulated that learners learn better when they are presented with both visual and verbal form of material, because presenting materials in these two forms can utilize the full capacity of human processing system (Mayer, 2014). Although Mayer’s theory was developed based on L1 learners, it could be applied to the case of L2 learners as well.

The auditory input in L2 captioned video includes the speech in the video, and the visual input includes the image of the video and the text of the speech. When L2 learners are watching captioned video, they must integrate/match these different sources of information for efficient language processing. Figure 1 shows an image of this multimodal integration. Suppose a L2 English learner is watching a show in English with captions, and he/she hears and sees the word "pirate." When only looking at the word "pirate", the learner's L2 phonological representation of the word may be something like /ˈpɪrat/. However, once hearing and seeing the word at the same time, learners could modify their phonological representation of this word to the correct form, /ˈpaɪrət/. In addition to that, if the video shows an image of a pirate at the same time, this can help the learner to understand or confirm what the word "pirate" means. By integrating these multimodal information, learners could 1) improve their speech of auditory word recognition, 2) modify their phono-lexical representation if necessary, and 3) map phonetic form to meaning by both looking at the text and the corresponding visual image if it is available.

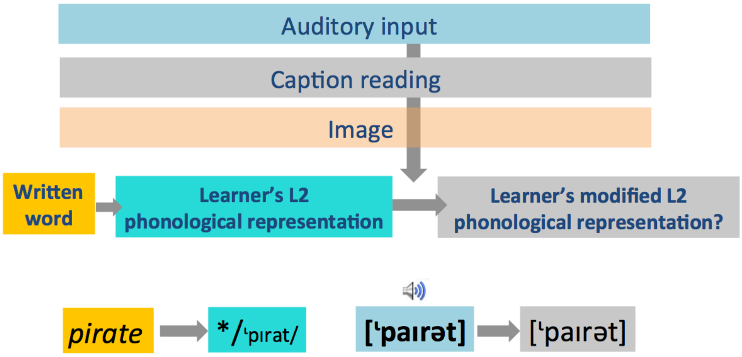

Auditory input
Caption reading
Image
Written word
Learner's L2 phonological representation
Learner's modified L2 phonological representation?
pirate
*/ˈpɪrat/
[ˈpaɪrət]
[ˈpaɪrət]

Figure 1 Multimodel Integration (Wisniewska & Mora, 2018)

The idea that captioned-video can be an effective tool for L2 learners is also partly based on Stephen Krashen's Input Hypothesis (1985), because captioned-video offers ideal comprehensible input (CI), input that is slightly more advanced than the learners' current stage of linguistic competence (i+1). In addition to the Input Hypothesis, Krashen's Affective Filter Hypothesis also plays a role here. This hypothesis explains that learners' will have "affective filter" when they learn the target language in a stressful environment, and this filter may impede language acquisition. Affective filter can be often times present when L2 learners are engaging in L2 listening, because they might feel that they lack a sense of control over the listening process (Graham & Santos, 2015). Watching videos with caption may put the learners at ease because it can provide aid to better understand the video, which in turn could increase learners' motivation in learning the target language.

The Noticing Hypothesis, proposed by Richard Schmidt (1990) has also been an influential theoretical framework for caption research. This hypothesis proposes that language acquisition takes place when learners give conscious attention to the form of input. This is, however, often difficult for L2 learners to do when listening to connected speech because of the fast and transient nature of spoken language. This is where caption can become useful, because caption provides orthographical image of the sound they hear, and learners may be able to notice and modify their phono-lexical representation if necessary. With increasing exposure to the target language through captioned-videos, more words will have greater familiarity, and at the same time, learners' phonological representation will become more stable and sophisticated. Both of these factors contribute to L2 learners' speech processing efficiency.

One of the most attractive factor for using captioned-video as a pedagogical tool, is that there are abundant resources available today for both teachers and learners. For example in the U.S., it was mandated that all video programs aired by television video programming providers be captioned by the year 2006 so that viewers including those with hearing impairment and elderly who have difficultly hearing have better access to the content being aired (Ellcessor, 2012). Although it is quite later compared to the progress of the U.S., report has shown that five major commercial broadcasting organization in Japan had made 100% of their contents available with captions by the year 2018, and *NHK*, Japan's national public broadcasting organization had made 88.5% of its broadcasted programs available with captions in the same year (Ministry of Internal Affairs and Communications, 2018). TV programs with captions could be accessed easily for learners studying the target language in a *second* language context (i.e. L2 learners of Japanese residing in Japan), but it may not be as easy for learners studying in a *foreign* language context (i.e. L2 learners of Japanese residing in the U.S.).

One way that learners in a foreign language context can easily access captioned-videos is through video streaming. Video streaming is a type of media streaming in which data from a video file is continuously transferred through the internet to a remote user. L2 learners can access global content videos from all around the world from the accessibility of video streaming, and in fact, report has shown that over half of U.S. households are now subscribing to at least one video streaming service, a 450% increase since 2009 (Westcott, Downs, Loucks, & Watson, 2018). Netflix, for example, is one of the biggest video streaming providers, which was founded in the U.S. in 1997 with over 139 million paid subscribers worldwide (Jarvey, 2019). With having this many subscribers, Netflix found the need to incorporate closed captions for viewers with hearing impairment so that all subscribers can have better television accessibility. As a result, Netflix promised the National Association of the Deaf (NAD) that they will have closed captions in 100% of Netflix streaming content by 2014 (Ellis, 2015). The good news is that this is not limited to English language contents; nearly all shows have captions in that target language (e.g. Japanese shows on Netflix have Japanese closed captions in addition to having subtitles in many other languages). Netflix is currently producing more and more of their own non-English language content for their global audience, and in fact, they have announced that 36% of their upcoming original contents will be non-English (“Netflix eye Europe,” 2019). This means that L2 learners in a foreign language context, for example, L2 learners of Spanish or French or Japanese learning the target language in the U.S., can watch foreign language contents with captions in that language. This is, of course not limited to Netflix; for example NHK has many videos online that can be watched with captions. There are also many online tools today that can easily create captions on videos, so instructors can create course materials with a show of their choice, even if the original source does not contain captions. As we have just seen above, there has been a large increase in resources available over the past couple of years, and as a result, a body of research on captions as a pedagogical language tool grew, most prominently in English language learning.

Statement of the Problem

Inspired by the theories mentioned above, many researchers started to investigate whether captions can be beneficial for improving L2 learners’ listening comprehension skills, and most have shown that caption groups scored higher on a comprehension test when compared to no caption control groups or to L2 subtitle groups (e.g. Chung, 1999; Guichon & McLornan, 2008; Hayati & Mohmedi, 2010; Markham, 1989, 1999). Brown (1977), however, explains that any comprehension test is largely a memory test, and it does not directly test what participants actually heard, but what they can remember from what they have learned, or what they can reconstruct from what they have heard based on their background knowledge. Many of the previous studies suffer from this very problem, and the results from the studies do not show that captions improved participants’ listening ability per se; it is unknown how much of the improved performance was actually due to participants’ ability to read the captions.

Buck (2001) argues that what is measured on most listening tests is largely general language proficiency or general comprehension rather than a specific measure of listening ability. Buck further states that if we are trying to assess learners' listening ability, we need to focus on aspects of proficiency and comprehension that are unique to listening. One of the aspects that are unique to listening is the ability for learners to match sounds to words and to identify word boundaries in a stream of speech. This is a fundamental aspect of listening comprehension, because listeners must first accurately match sounds to words so that they can move on to the trial-and-error process of speech segmentation. In this process, learners access their mental lexicon and their knowledge of the current context to determine where the speech is segmented, and to understand what that word means as intended by the speaker. As previous studies show, many L2 learners have difficulty in segmenting speech produced by native speakers, even when all words are familiar to them in text (Bonk, 2000; Ching-Shyang Chang, 2007; Goh, 2000; Milton & Hopkins, 2006; Van Zeeland & Schmitt, 2013). While previous research shows a beneficial effect of captions on general L2 listening comprehension, it remains unclear to what extent it can help improve L2 learners' ability to match sounds to words and to segment continuous speech into individual words.

In order to investigate whether bi-modal input (caption) supports perceptual learning of L2 speech and whether it can help improve learners' speech segmentation skills, Charles and Trenkic (2015) used a shadowing task to collect data on the influence of bi-modal input on L2 English learners' speech segmentation skills. In their four-week pre-post design experiment, 12 international students at a UK university watched a series of British documentaries either with sound and captions, with sound but no captions, or with captions but without sound. The result revealed that participants in the bi-modal group improved more than the other two groups on their ability to segment speech. The results were exciting not only for researchers, but also for teachers, who can easily incorporate L2 listening activities by utilizing videos with captions. Learners can also easily watch these videos at home on their own in front of a TV or their computer.

While much research on the effect of captions on L2 listening has been conducted on L2 English learners (especially ESL learners), none to the researcher's knowledge has been done on L2 Japanese learners. The use of bi-modal input of Japanese is worth investigating because Japanese has many different characteristics compared to English. First, Japanese has a different orthographic system than the English alphabet. Second, written Japanese does not have clear visible word boundaries in the form of white spaces positioned between words as in English, and third, since Japanese is a mora-timed language, speech is phonologically segmented differently compared to stressed-timed languages such as English.

Research Question and Study Design

The overarching research question for this study is whether bi-modal input assist L2 Japanese learners to improve their speech segmentation skills. This dissertation, inspired by Charles and Trenkic's (2015) study, was designed as a focused study of some of the core aspects of second language speech segmentation and word recognition skills. The data was collected from 12 third-year Japanese learners at a large state university in the Spring 2018 semester who participated in a semester-long pre-post design experiment. This study adopted a single-case design (SCD), a variation of time-series designs in which progress of a single participant or a single group of participant is examined through interventions over a period of time to determine whether a causal relationship exists between the variables (Kennedy, 2005). Since no empirical evidence of the effectiveness of captioned-videos on fostering L2 Japanese learners' speech segmentation skills has been found up to this point, it is important to first investigate whether a causal relation exists between the independent variable (i.e. the bi-modal input) and the dependent variable (i.e. L2 Japanese learners' speech segmentation skills) before conducting a larger comparison study. It is hoped that this study will serve as a springboard to further research projects on caption research in Japanese, and that the data produced in this study will be useful to both learners and educators.

CHAPTER 2. LITERATURE REVIEW

Listening Skill

Listening has often been defined as one of the most complex skills in language learning, because learners cannot control the speed of the input due to its real-time nature (Montero Perez, Van Den Noortgate, & Desmet, 2013). Researchers suggest that learners must listen extensively in order to become better listeners (Graham & Santos, 2015), and Ellis (2005) emphasizes the importance of exposure to the target language in L2 acquisition, stating "successful instructed language learning requires extensive L2 input" (p. 217). Ellis supports his claim on the basis of how the amount and quality of input children receive accounts for their L1 acquisition, and he strongly believes that the same can be said for L2 acquisition. In a *second* language learning context, learners can be expected to receive plenty of input even outside of the classroom, but in a *foreign* language learning context (as when learning Japanese in schools in the United States), there are usually far fewer opportunities to receive extensive input outside of the classroom. Then, it becomes even more important to practice listening in the classroom in foreign language learning context, but in reality, listening practice is often subordinated to grammar learning and vocabulary development (Graham & Santos, 2015). Previous studies also suggest that language teachers often find teaching of listening to be particularly challenging (Chambers, 1996; Field, 2008). If listening is "the most difficult skill to learn" (Vandergrift, 2004, p. 4) but nevertheless learners are not provided with much opportunity to practice improving this skill, we can naturally assume that learners' levels of anxiety and self-efficacy (one's belief in one's ability to succeed in specific situations or accomplish a task) are negatively affected (Graham & Santos, 2015), especially since listening is often an interactive process where pre-planning is difficult.

Good listeners must be ‘tuned in’ at all times and actively use their various knowledge — linguistic or otherwise — in order to make sense of what they hear.

Rost (2011) states that listening as a skill shares traits with other language skills, most notably the comprehension skill. Anderson (2015) explains that listening comprehension can be broken down into three stages. The first stage involves the perceptual processes that decode the spoken (acoustic) message. The second stage is the parsing stage, where the words in the message are transformed into mental representation of the combined meaning of the words. The final stage is the utilization stage, in which listeners use the mental representation of the sentence’s meaning. These three stages are by necessity ordered in time, but they also partly overlap. For example, listeners can make inferences from the first part of the sentence while they are perceiving a later part. Listeners also must use both bottom-up and top-down processes in comprehension. Bottom-up process is used when listeners construct meaning from the phoneme-level up to discourse-level features, and top-down process is used when learners use context information and their prior knowledge to build a conceptual framework for comprehension. Native language listeners can efficiently process what is heard because they have automated word recognition and need little conscious attention to individual words, but this task is much more difficult for L2 listeners (Vandergrift, 2004). While Field (2008) agrees that both bottom-up and top-down processes are necessary for successful L2 listening comprehension, he believes successful bottom-up process is the long-term key to skilled listening, because the ability to decode what is in the input accurately and automatically allows listeners to focus on the speaker’s intended meaning. As we can see, listening is a very complex process that is far from a passive activity.

As Field (2008) states, it is important to support learners bottom-up process to increase efficiency, so that they can focus more on utilizing their top-down process when comprehending speech. This is, however, not as easily done as simply having learners improve their linguistic knowledge. One may assume that as long as learners know the words that are being used in the speech, they can better comprehend the message, but studies have shown that this is actually not the case. Milton and Hopkins (2006) investigated whether the knowledge of word orthography and of phonology are closely linked. They deemed problematic that vocabulary tests normally rely testing the written form of words only, but even if learners know the written form of the words, they may not recognize them when they are spoken. For example, function words in English may vary in pronunciation according to their sentence position and the other words around them, and may also disappear completely in natural speech. Their results indeed showed that learners' knowledge of word orthography is not exactly the same as their knowledge of word phonology, and that even advanced level learners of English had a phonological recognition vocabulary much smaller than their written vocabulary recognition.

Ching-Shyang Chang's (2007) study also showed that knowing a word in its written form is not the same as knowing or recognizing it when it is spoken, especially when it occurs in connected speech. Her study investigated whether giving Taiwanese ESL learners more time to learn the vocabulary that is contained in the listening passage would make any difference in their listening comprehension. The result showed that the impact was minimal regardless of whether participants learned the vocabulary 30 minutes, one day, or one week before listening to the passage, concluding that listening comprehension is not enhanced by simply knowing the words. In fact, some learners said they did not recognize the vocabulary they learned when it was presented aurally, suggesting a lack of knowledge of how words sound in connected speech.

Goh (2000) investigated the cognitive difficulties Chinese university-level students experience when listening to English by eliciting data through learners' self-reports. The self-reports showed that more than half of the students faced the same problem during listening, which was related to a fundamental aspect of comprehension: their perceptual processing. Many students reported their inability to recognize words when they were spoken, even when these were words that they actually knew, showing that for some students, sound-to-script relationships have not been fully automatized. Therefore, although they knew certain words by sight, they could not recognize them by sound. Goh (2000) further explains that their listening vocabulary was underdeveloped. Their ability to understand spoken words was greatly handicapped because they had not stored the sounds of lexical items efficiently in long-term memory. This underdeveloped listening vocabulary could have been directly related to the way the students learned new words. Many of them said they learned by memorizing the spelling of words and often neglected to remember how the words sounded.

Bonk (2000) investigated whether there is a correlation between comprehension rating score after listening to a text once and the amount of familiar lexis in that text. Participants were 59 L2 English learners who had Japanese as their L1, enrolled as English majors at a university in Japan. In general, the result showed a moderate positive correlation (r (59) = .446, $p < .05$) between lexical recognition and comprehension ratings, but also found some variation in comprehension. While some listeners had quite good comprehension knowing only 75% of the words, Bonk also found that nearly a quarter of the participants never attained a rating of "Good" comprehension (as opposed to "Inferior" comprehension rating), despite showing evidence of familiarity with up to 100% of the lexical words in the text. Bonk believes that this result is attributed to these participants' inability to comprehend connected speech in the L2.

Most research on L2 reading comprehension has indicated that 98% lexical coverage provides adequate comprehension of written text, and many research studies have applied this result to listening comprehension as well (Van Zeeland & Schmitt, 2013). However, this is problematic since listening comprehension involves different process than reading comprehension and vocabulary seems to play a different role in the two modalities. This is why Van Zeeland and Schmitt (2013) researched the relationship between lexical coverage and listening comprehension among both native and non-native speakers, to find out whether the same degree of lexical coverage is necessary for L2 listening comprehension as that of L2 reading comprehension. In their study, the coverage of four spoken informal narrative passages was manipulated, and the participants' (36 L1 English speakers and 40 L2 learners of English) listening comprehension of the four passages was measured. The result showed that most native and non-native participants could adequately comprehend the spoken input with only 90% coverage, but similarly to Bonk's (2000) study, the L2 learners showed wider variation in their comprehension than those of native speakers'; that is, despite having the same vocabulary level, learners still achieved different levels of listening comprehension. Van Zeeland and Schmidt (2013) state that there was "likely some variation in their general L2 proficiency and their skillfulness in online word segmentation and automaticity, which may have influenced their listening ability" (p. 473). They also explained why the role of vocabulary knowledge seems to be smaller in L2 listening than in L2 reading by pointing out the fleeting nature of the spoken discourse. While readers can always go backwards in the text in order to decode the overall message, this cannot be done in listening. If listeners missed some materials, they may need to refer to other information (contextual cues, prior knowledge, etc) in order to understand the intended message by the speaker. Van Zeeland and Schmitt also state that this requirement of quick, online processing is what makes listening challenging to many learners.

As studies show, learners often have a difficult time recognizing words even if they are familiar to them in written form, and simply knowing the words used in the listening passage does not guarantee that they will have good listening comprehension. This seems to be even more the case in continuous speech, because learners cannot rely on clear word boundaries as in written text and must process online word recognition. This ability to match sounds to words and to identify word boundaries in continuous speech is a fundamental aspect of listening comprehension as it goes hand in hand with word recognition.

Speech Segmentation

Speech segmentation is the process of detecting word boundaries in continuous speech. It is a by-product of the word recognition process, as word recognition occurs when speech is segmented into individual words and activates all words with which it is compatible in the given context (Cutler, 2000). Speech segmentation is a major part of learning a different language, as one must accurately distinguish between the individual words spoken by speakers in order to make sense of what they said. It is crucial for listeners to be able to accurately recognize a word, because if not, they cannot access their mental lexicon and knowledge of the current context to understand what that word means as intended by the speaker. Furthermore, Rost (2011) states that misunderstanding or non-understanding of words in speech, whether through faulty identification of word boundaries or inadequate knowledge of word meanings, is a major source of confusion in language comprehension, particularly second language comprehension.

This is, however, not an easy task for L2 learners because spoken language is a continuous stream of words which contains very few salient cues to indicate word boundaries. In

English printed text, there are visible word boundaries in the form of white spaces positioned between words, and this allows readers to see where a word starts and ends. However, as we can see from the spectrogram in Figure 2, there are no word boundaries or indicators showing where a word starts or ends in a continuous stream of speech. The utterance made in Figure 2 is “Why, what weekend were you guys gonna be there?” spoken by a native speaker of American English (Warner, 2005). This sentence contains nine words in the printed version, but when spoken, we can only see roughly four chunks of sounds.

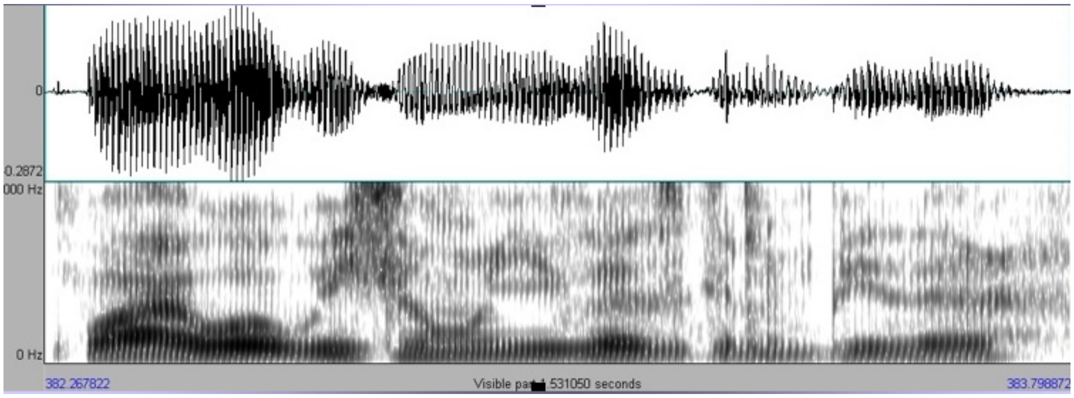


Figure 2 Spectrogram of a connected speech (Warner, 2005)

The absence of word boundary markers does not usually pose a problem for native language listeners because they can correctly map sound strings onto word representations by relying on their abundant lexical and syntactic knowledge while fine-tuning it via statistical regularities, such as lexical stress and phonotactics (Mattys & Bortfeld, 2016). Their efficient speech processing system in turn allows them to focus their attention more on contextual information. When listeners hear speech, multiple simultaneous activation and competition occur since auditory information is time-bound and transient. For example, the sentence “ice cream is so tasty” consists of word with fully overlapping segmentation alternatives (“ice cream” vs, “I scream”). However, it gets disambiguated by syntactic and semantic context, because “scream” is a verb and another verb ("is") cannot follow it, and the sentence "I scream is so tasty" is syntactically and semantically unfit. Eventually, "I scream" gets terminated through direct inhibition by the better fitting competitor ("ice scream"). Mattys and Bortfled (2017) explains that "all words in running speech are temporary activated, with relative activation levels fluctuating as a function of fit with the unfolding input" (p.56).

In contrast, L2 learners' lexical and syntactic knowledge is underdeveloped compared to native language users and therefore, they must depend more on non-lexical segmentation strategies. Cutler (2000) states that non-lexical segmentation of speech relies on language-specific phonology, namely rhythmic structure and phonotactic sequencing. In general, languages with comparable phonotactic rules will facilitate the transfer of segmentation strategies, whereas those with different regularities may create interference. Thus, "the process of becoming a proficient L2 listener involves not only vocabulary growth and syntactic learning but also increased control over potentially contradictory sets of non-lexical segmentation strategies" (Mattys & Bortfeld, 2017; p. 68).

Previous studies suggest that language rhythm determines the segmentation unit most natural to native listeners (Cutler, Mehler, Norris, & Segui, 1992), and this method of identifying word boundaries is known as the Metrical Segmentation Strategy (Cutler & Butterfield, 1992; Cutler & Norris, 1988). For example, French has a syllabic rhythm and Segui (1984) summarized a number of studies indicating that polysyllabic words, whether they are heard in isolation or in connected speech, are analyzed syllable by syllable by French listeners. Evidence even shows that French listeners use the syllable in segmenting a foreign language (English). On the other hand, English listeners segment based on stress since its speech rhythm has a characteristic pattern which is expressed as an alternation of a strong syllable (has primary or secondary stress and contains a full vowel) and a weak syllable (unstressed and contains a short, central vowel such as schwa). In fact, native speakers of English are found to rely heavily on strong syllables to identify the start of a new word because approximately 90% of content words are stressed on the first syllable in English (Cutler et al., 1992). While native speakers of English can use this strategy to identify word boundaries, it is not as easy for L2 English learners to employ the same strategy. Many of the Chinese university-level students in Goh's (2000) study reported their inability to recognize English words when they were spoken. The participants in Goh's study were L1 Chinese speakers, and Chinese is a syllable-timed language, which can explain why participants in this study often had a difficult time segmenting spoken English, even if the individual words were familiar to them.

Metrical Segmentation Strategy is employed differently in Japanese, because Japanese belongs to a different rhythmic class than, for example, English and French. While English is classified as a stressed-timed language and French as a syllable-timed language, Japanese is classified as a mora-timed language. By this, it means that the Japanese speech consists of a series of morae (a Japanese timing unit that cannot be further divided into subunits of onset, nucleus, and coda), and each mora bears roughly the same length of time (Port, Dalby, & O'Dell, 1987). Otake, Hatano, Cutler, and Mehler (1993) examined the segmentation strategy used by native Japanese listeners on spoken Japanese, and found that Japanese listeners' response patterns were consistent with moraic segmentation, while French listeners' response patterns were consistent with syllabic segmentation with the same material. The study had participants listen to short sequences of Japanese words preceded by two occurrences of a target specification (either CV or CVN; N is considered the only true syllabic coda in Japanese). They were instructed to listen for a word beginning with the sounds specified as a target for each sequence, and to press a response key as soon as they had detected an occurrence of this target. For example, words *tanishi* and *tanshi* have three morae: *ta-ni-shi* and *ta-n-shi.* But they differ in number of syllables: *ta-ni-shi* has three syllables but *tan-shi* has two syllables. Likewise, they differ in syllabic structure: the first syllable of *tanishi* is *ta,* but the first syllable of *tanshi* is *tan.* In this example, the target participants have to identify is either *ta* or *tan*. The syllabic hypothesis predicts that CV targets (*ta*) should be detected easily in CVCVCV words (*tanishi*) among French listeners and CVN targets (*tan*) should be detected easily in CVNCV words (*tanshi*), because in each case the target corresponds to the initial syllable of the stimulus word. The mora hypothesis, on the other hand, predicts that there should be no difference in responses to CV targets for CVCVCV or CVNCV words because the initial CV is also the initial mora; hence, the CV target should have identical status in both words. The result showed that French listeners exhibited a pattern of responses consistent with the syllabic hypothesis, and that Japanese listeners' response patterns were consistent with moraic segmentation. Based on their findings, Otake et al. (1993) states that listeners' speech segmentation is influenced by the way their native language organizes its language, and why this is so can be explained by how infants learn their native language. Since infants have no existing store of meaningful units that they can use to segment speech, they first need to build a lexicon. And for the lexicon-building task, infants must identify the characteristic rhythmic structure of the speech input in order to form the framework for the lexicon-building process. The authors claim that new-born child is likely equipped with "procedures which enable them to focus efficiently upon aspects of speech which assist in segmenting the continuous speech stream into meaningful units" (p.276).

Aspects of language-specific phonology other than rhythmic structure has also shown to play a role in speech segmentation. For example, vowel harmony can be used to detect word boundaries with languages which have this phonological feature. Finnish is one of those languages where certain vowels cannot co-occur in a word, and listeners can use this fact to insert word boundary to two successive syllables containing incompatible vowels. Suomi, Mcqueen, and Cutler (1997) demonstrated this in their study using word spotting task and found that a word such as *palo* ‘fire’ was easier for them to detect in *kypalo* than in *kupalo,* because the Finnish vowel /y/ cannot occur in the same word as /a/ or /o/, whereas the vowel /u/ can.

Another aspect of phonological structure that is clearly language-specific is phonotactic sequencing. Studies including McQueen's (1998) show that listeners appear to use phonotactic constraints (restrictions on permissible phoneme sequences within syllables) to segment continuous speech. McQueen (1998) showed that Dutch listeners were able to spot *rok* ‘skirt’ in *fiemrok* (/mr/ cannot be syllable-initial or final, meaning that a boundary must occur within the sequence) than in *fidrok* (/dr/ must be syllable-initial and /d/ cannot be syllable-final, meaning that a boundary cannot occur within /dr/). Weber (2000) investigated the influence of phonotactic constraints on L1 German and English listeners. L1 German listeners with very high competence in English and L1 English listeners who had no knowledge of German detected embedded English words such as *luck* in nonsense sequences such as *moyschluck* and *moysluck* in a word-spotting task. The result showed that L1 English listeners found it easier to detect it in *moyschluck* than in *moysluck* because no native English words begin with the sequence /schl/*,* while many can with /sl/. On the other hand, German listeners found it easier to detect *luck* in *moysluck* than in *moyschluck* because */sl/* cannot begin a word in German while /sch/ can. Although German listeners were most strongly influenced by information about their L1, they were still able to use information about English in order to segment English speech. They found it most difficult to detect words that were not aligned with a clear syllable boundary according to either German or English phonotactics, somewhat easier with words which were aligned according to English but not to German, even easier with words which were aligned according to German but not with English, and the easiest with words aligned according to both languages. This shows that listeners can make use of not only phonotactic information available in their own language, but also that in a language they are highly proficient in.

These studies show that process of listening is language specific, but what about the case of bilingual listeners? Cutler, Mehler, Norris, and Segui (1989, 1992) investigated early English and French bilinguals' segmentation procedures, and found that despite their exceptional mastery of both English and French, they tended to call upon one rhythmic segmentation procedure (either the procedure typical of French using syllable boundaries, or English using stress-based boundaries) and did not use the segmentation procedures interchangeably. However, listening to speech is all about efficiency; listeners exploit whatever aspects of linguistic structure they can to make listening as efficient as it can be (Cutler, 2000). The bilinguals in their study did not apply the rhythmic structure to segment the syllables when its use was inefficient in the language with a different rhythmic structure. Cutler et al. claims that sufficient experience with a language can lead to avoidance of such inefficiency, and that listeners can learn to abandon its use when necessary. The finding here is similar to Weber's (2000) as they suggest that listening may be constrained by one dominant language, but it is nonetheless possible to use the information from other languages to choose the most appropriate and efficient listening strategy.

Because native listeners can use strategies which are specifically tailored to the native phonology, they are highly efficient in segmenting speech. In addition to that, since native listeners have greater lexicon than L2 learners, this allows them for more automatic activation of word forms. As a result, they can use more top-down process in listening comprehension because they can focus their attention on using context information with less dependency on bottom-up cues to word boundaries. L2 learners on the other hand, have fewer words which they can recognize quickly enough to use for top-down segmentation compared to native listeners, which initially make learners be more dependent on bottom-up cues. This is why it is important to support L2 learners' bottom-up process and to make their phonological representation more sophisticated and stable, so that their efficiency of speech processing will increase, which in turn will allow them to focus more on the top-down process of listening comprehension. It is important to build learners' lexicon so that learners can more easily activate word forms that are available in their mental lexicon like native listeners, but as we saw from previous studies earlier, knowing words does not necessary lead to better listening comprehension. What is important for L2 learners is to have the sound-to-script relationship automatize. How, then can we assist learners to improve their ability to match sounds to words so that they can better identify word boundaries in continuous speech? One way to do so is to use bi-modal input. Bi-modal input here means that people hear something and also see it written at the same time (simultaneous presentation of matching aural and orthographic stimuli), as it is used in closed captioning media. The combination of written and audio input, like when providing captions in the target language, has the potential to help develop speech segmentation skills by making the input more intelligible.

Theories Underlining Caption Research

Bi-modal input is a term to describe the simultaneous presentation of matching aural and orthographic stimuli (Charles & Trenkic, 2015). The most common way to present these two forms simultaneously is by using captions. The use of captions to foster second language acquisition started to receive considerable attention in the 1980s inspired by Allan Paivio's Dual

Coding Theory (Paivio, 1986). In his theory, Paivio postulated that the brain processes verbal and non-verbal (visual) stimuli via two different cognitive systems, creating separate representations for information processed in each channel. When these two systems interact with one another and are both activated, it is expected to improve information processing, subsequently leading to greater depth of processing and better recall, because both visual and verbal codes can be used when recalling information.

A more recent theory studied in depth by Richard Mayer is the Multimedia Principle. Paivio's Dual Coding Theory was the basis of this theory, which was developed in the context of L1 learning. The basis of its idea is that learners learn better when they are presented with both visual and verbal form of materials, because presenting materials in these two forms can take advantage of the full capacity of human processing system (one for verbal and one for visual material), and that human understanding is enhanced when learners are able to mentally integrate visual and verbal representations. Mayer (2014) states "[i]n the process of trying to build connections between words and pictures, learners are able to create a deeper understanding than from words or pictures alone. This idea is at the heart of the theories of multimedia learning" (p.7).

Another influential theory for caption research is Stephen Krashen's Monitor Theory (1985). It consists of five main hypotheses, which are 1) the Acquisition-Learning Hypothesis, 2) The Monitor Hypothesis, 3) the Natural Order Hypothesis, 4) the Input Hypothesis, and 5) the Affective Filter Hypothesis. Within these five hypotheses, the last two have some relevance to L2 learning that takes place using captions. The Input Hypothesis states that language acquisition can only occur when learners receive a sufficient amount of comprehensible input (CI), and CI is defined as the input that is slightly more advanced than his/her current stage of linguistic competence (i+1). Researchers thought that captioned TV is an ideal medium for providing CI because it provides learners with the opportunity to view the video, hear the spoken words, and see the printed text, all of which assist learners to better comprehend the presented material (e.g. Huang & Eskey, 1999; Neuman & Koskinen, 1992; Vanderplank, 1988, 1990; Winke, Gass, & Sydorenko, 2010).

The Affective Filter Hypothesis explains that in a stressful environment where learners have ‘affective variables’ (e.g. anxiety, low motivation, low self-esteem), the “affective filter” will rise high and will block the input from coming in, impeding language acquisition. Affective filter can be present when learners are engaged in L2 listening, because learners may feel that they lack a sense of control over the listening process (Graham & Santos, 2015). Researchers state captions can help lower learner’s affective filter when watching videos in the L2 because captions provide aid to better understand the video, which could relieve some of the anxiety experienced by learners (Borrás & Lafayette, 1994; Danan, 2004). Captions can allow them to relax, grow more confident in their ability to understand, and increase their motivation in learning the target language

Richard Schmidt’s (1990) Noticing Hypothesis has also been an influential theoretical framework for caption research. Schmidt claims noticing, or conscious attention to the form of input, is necessary for second language acquisition to take place (Schmidt, 1990, 2001; Schmidt & Frota, 1986). The role of attention and awareness must also be discussed in the light of noticing, because attention is responsible for allocating the cognitive resources that lead to noticing, and there is no learning without awareness at the level of noticing (Robinson, 1995). According to Robinson (1995), the concept of attention has three uses; it can be used to 1) describe the processes involved in “selecting” the information to be processed and stored in memory, 2) describe our “capacity” for processing information, and 3) describe the mental “effort” involved in processing information. He defines attention as “the process that encodes language input, keeps it active in working and short-term memory, and retrieves it from long-term memory (Robinson, 2003; p.631). Schmidt (1995) suggests two level of awareness, which is at the level of noticing or conscious registration of the occurrence of some event and at the level of understanding, or recognition of a general principle, rule, or pattern. To put all these concepts together, Robinson (1995; p.318) defines noticing as “detection with awareness and rehearsal in short-term memory”. The rehearsal in short-term memory is important in noticing because short-term memory “serves as the interface between everything we know and everything we can see or do” (Cowan, 1993; p.166) and is conceived as where skill development begins (Anderson, 1983). Furthermore, short-term memory is later encoded into long-term memory (Robinson, 1995), and what is stored in learners’ long-term memory leads to language acquisition. Schmidt and Frota (1986) suggested that “a second language learner will begin to acquire the target like form if and only if it is present in comprehended input and ‘noticed’ in the normal sense of the word, that is, consciously” (p. 311). It is often the case that authentic language input (as in videos) maybe so complex that learners need some sort of aid to notice the linguistic forms. In these cases, caption can help learners notice L2 forms when they are faced with a string of incomprehensible input, because captioned-video provides orthographical image of the sound they hear. Since caption visualizes word boundaries (Bird & Williams, 2002), learners need less time decoding the speech, which in turn lets them pay more attention to the language being used in the video (Vanderplank, 1990).

Beginning of Caption Research

In the 1980s and 1990s, researchers started to investigate whether closed captions intended for the deaf and hearing-impaired could also be a valuable resource for language learning and teaching (e.g. Bean & Wilson, 1989; Markham, 1989, 1999; Vanderplank, 1988, 1990), and Price's work (1983) conducted more than 30 years ago has been said to be the pioneering work in caption research. Five-hundred L2 learners of English from 20 language backgrounds at Harvard University participated in this study, and they were randomly assigned to two groups after controlling for their English proficiency and their length of stay in the United States. One group saw the video excerpts with captions, and the other half without, and half of each group had one viewing, while the other half had two viewings. Price found that all those who saw the captioned film benefited significantly in their comprehension of the film even with only one viewing.

Despite this exciting finding, captioned materials were not often used in language classrooms because the school had to either purchase a closed captioned TV adaptor, or buy captioned videotapes that were created by the National Captioning Institute, which were both costly options. This, however, started to gradually change with the United States taking the lead in the late 1990s to early 2000s, when captions started to receive official support through legislation which required broadcasters to produce a certain percentage of their programming with captions for the deaf and hearing-impaired (Vanderplank, 2014). The Telecommunications Act of 1996 mandated that all new video programs aired by television video programming providers (networks, cable operators) be captioned by the year 2006 (Ellcessor, 2012).

As the video streaming industry, of which Netflix is a prominent example, started to grow, customer satisfaction became even more important. Netflix, founded in the U.S. in 1997, has accumulated tens of millions of streaming subscribers worldwide through its development and expansion of their streaming service. Netflix Inc. and the National Association of the Deaf (NAD) submitted a joint Consent Decree to the federal court in 2012 to ensure captions in 100% of Netflix streaming content by 2014 so that viewers with disability can have better television accessibility ("Netflix and the National Association of the Deaf," 2012), and other video streaming providers started to follow their lead. As a result of the increased resources, a body of research on captions grew most prominently in English language learning and most studies generally supported Price's findings that caption can help learners to better comprehend the audio-visual materials.

The Effects of Bi-Modal Input (Caption) on L2 Listening Comprehension

A number of studies have examined whether captioning is effective for improving L2 learners' comprehension of audiovisual material, and most have shown better comprehension test scores for captioning group when compared to no captioning control groups or to L1 subtitling groups (e.g. Chung, 1999; Guichon & McLornan, 2008; Hayati & Mohmedi, 2010; Markham, 1989, 1999). Chung (1999) compared listening comprehension rates for videos using variety of techniques including captions. Participants were 170 Taiwanese L2 English learners, and they viewed four different English video segments each attended with 1) advance organizers (information that is presented prior to learning and that can be used by the learner to organize and interpret new incoming information), 2) captions, 3) a combination of both 1 and 2, and 4) none of the foregoing. After each viewing, a comprehension test of multiple-choice items in the participants' L1 (Chinese) were given to each group. The result showed that the combined group and the caption group scored significantly higher than the other two groups, while there was no significant difference between the caption group and the combined group. In fact, the results revealed that captions on videos best helped bridge the competence gap between reading and listening and enhanced language learning.

Markham (1989) conducted an experiment where 76 university-level ESL learners watched two English TV programs. The participants were divided into two groups, where each intact group of students viewed one program with captions and one without captions in order to achieve balanced exposure to the treatment variable. After each viewing, participants completed multiple-choice comprehension tasks. As a result, participants demonstrated better performance on the test after being exposed to the bi-modal input (captions) than after they had watched the TV program without captions. Based on this finding, Markham claimed that bi-modal input is beneficial in improving L2 listening comprehension.

More recently, Hayati and Mohmedi (2010) conducted a similar study to Markham's where they investigated the effects of using films with or without captions on intermediate EFL students' listening comprehension. Participants in their study were L2 English learners from a university in Iran, and the participants viewed six episodes of an English show entitled "Wild Weather" in one of the following three treatment conditions: English captions, Persian subtitles, or only video. After each viewing session, multiple-choice comprehension tests were administered to examine the participants' listening comprehension. Each question contained language that actually occurred somewhere in the episode. The results revealed that the English caption group performed at a considerably higher level than the Persian subtitle group, who in turn performed at a substantially higher level than the video-only group on the comprehension test. This result suggests that bi-modal L2 input (English audio + English captions) strengthens the verbal message of the video, because viewers' attention can alternate from the auditory to the visual format or be directed along the visual and auditory routes simultaneously. Participants who watched the video with L1 subtitles commented that Persian subtitles distracted their attention and prevented them from concentrating on the spoken language.

Similarly to Hayati and Mohmedi's (2010) study, Guichon & McLornan (2008) examined the impact of different types of input upon L2 learners' comprehension of spoken English. Forty French intermediate-level L2 English learners were divided into either one of the following groups: 1) sound alone, 2) image and sound, 3) image, sound, and L1 (French) subtitles and 4) image, sound, and L2 (English) captions. In their respective groups, learners watched a three-minute BBC news report. Following the video viewing, participants were asked to produce a detailed written summary of the video in English with the help of their own notes. A written summary was used rather than multiple-choice question because their study aimed at identifying which of the four modalities facilitates comprehension the best and how students' written productions are influenced by the nature of each modality. The researcher identified 35 semantic units which were deemed central to understanding the video, and the results were obtained by calculating the number of semantic units that appeared in the student's written summary. Their study provided evidence that comprehension improves when learners are exposed to a text in several modalities, but caption is more beneficial than L1 subtitles because it causes less lexical interference. As in Hayati and Mohmedi's (2010) study, the findings from this study also contribute to the idea that L1 subtitles could inhibit the effective processing of L2 audio input, and learners best comprehend L2 video when they receive L2 bi-modal input.

The results from these studies showed that the presence of captions improved participants' general comprehension of the material, but it can be argued that they do not show that captions improved the participants' listening ability. Brown (1977) explains that any comprehension test (whether it is in reading, writing, listening, or spoken form) is largely a memory test. It does not directly test what participants actually heard, but what they can remember from what they have heard, or what they can reconstruct from what they have heard based on their background knowledge. These studies suffer from this very problem. In spite of the claim that bi-modal input leads to better listening performance, they do not show whether bi-modal input was beneficial in improving learners' listening skills per se. It could be the case that the improved performance was due to the participants' ability to read the captions.

The Effects of Bi-Modal Input (Caption) on Word Recognition and Speech Segmentation

Although studies show that captions may help make the audio input more intelligible resulting in improved comprehension of the audiovisual material, it remains a question whether captions can truly train learners to develop their listening skills and will allow them to eventually comprehend new speech without captions. It is valuable to find out whether bi-modal input can qualitatively change the phonological representation of the word in the learners' mind. Bird and Williams (2002) investigated this very issue, focusing on two types of learning; learning relating to auditory word recognition (implicit learning) and learning relating to word retention (explicit learning). In their first experiment, 16 native and 16 advanced nonnative English speakers either saw, heard, or simultaneously saw and heard English words, and had to decide as quickly as possible whether they knew the meaning of the target word. The target word consisted of 40 familiar and unfamiliar English words. The result showed that participants had the fastest reaction times in recognizing unfamiliar words when they both saw and heard the word simultaneously, suggesting that both groups tended to rely on the visual stimulus to make their decision. However, the familiar words were reliably recognized even in the absence of textual support. In a later phase, participants were presented with a list of words only in the sound modality, and had to judge whether each item had been presented in the earlier phase of the experiment. The result showed again that both native and nonnative participants showed superior recognition memory scores for the words with bi-modal presentation, which aligns with the dual-coding theory that presentation of simultaneous audio and text results in better information recall.

In order to see whether participants can also better recognize and retain nonwords when they are presented in the bi-modal condition, 24 advanced ESL students completed a rhyme monitoring task in their second experiment. In the first phase, participants were told that they would encounter pairs of nonwords, and had to decide upon encountering the second item whether it rhymed with the first item. The first item of each pair was presented auditorily, and the second would either be presented as sound only, as text only, or simultaneously as text and sound. The result showed that the bi-modal presentation of the second item improved their accuracy in rhyme judgment the most, showing that phonological information generated from the text was helping to establish phonological representations of nonwords, probably because the participants were less able to form stable representation of nonwords from sound input alone. This result supports Seidenberg and Tanenhaus's (1979) study where they studied the orthographic influences on auditory rhyme detection. In their study, they found that word pairs that were orthographically similar (e.g., stroke and joke) were judged correctly as rhymes more quickly than rhyme pairs that were orthographically dissimilar (e.g., soak and joke). This shows that orthographic inconsistencies between the rhyme and target items appeared to interfere with their judgments, even when participants should have been able to make rhyme judgements solely on the basis of phonological information.

Bird and Williams also tested whether bi-modal input can improve participants' recognition memory of the nonwords, and it indeed showed that their recognition memory score was the highest for words that were presented with sound and text. From testing nonwords, they were able to show that bi-modal presentation helped the recognition memory of auditory information, independent of any semantic context. They state "[b]y removing the sematic context that was available in the subtitling studies, the present experiments show more clearly that providing subjects with text and sound versions of known and unknown words can facilitate recognition memory relative to sound alone" (p.528).

With regard to implicit learning, they found the bi-modal input to be most effective for word recognition when the target words were either unfamiliar to the participants, or when they were nonwords. The familiar words in Experiment 1 was easily recognized even without the support of the text, and thus, they stated that bi-modal input beneficially affected word recognition only when new phonological forms needed to be encoded. This, however, could be overestimating listeners' ability to recognize words when they are presented in connected speech. Bird and Williams only investigated words in isolation, but familiar words could sound unfamiliar when they are presented in continuous speech, as earlier studies have shown (Bonk, 2000; Chung, 1999; Goh, 2000; Van Zeeland, 2014).

In order to investigate to what extent bi-modal input can help improve L2 learners' ability to recognize words in continuous speech, Charles and Trenkic (2015) conducted research employing a shadowing task as their test instrument to examine whether bi-modal input can help learners with L2 segmentation of speech. Their study focused on speech segmentation because they wanted to focus on learners' perceptual ability rather than their comprehension of the overall message in the speech. Bi-modal input in the form of sound and captions can provide an additional source of information about the words being spoken, and hence about the sounds being heard. Therefore, they hypothesized that the use of captions can help learners to identify words accurately in continuous speech. In their four-week pre-post design experiment, 12 international students at a UK university were randomly assigned to watch two 30-minute long videos taken from two documentaries in one of the following three conditions: 1) bi-modal (sound and captions), 2) no caption (sound but no captions), or 3) no sound (with captions but without sound). The no sound group was introduced to control for whether a potential advantage in the bi-modal condition stems from the captions alone or from the combination of captions and sound. This study aimed to explore low-level listening processes at the lexical segmentation level; hence, a shadowing task was employed as their test instrument in which participants listened to short utterances and immediately repeated what they heard. This task, therefore, simply tested their ability to identify words in a continuous speech stream (i.e. lexical segmentation). Participants completed a shadowing task in week one as a pre-test. In weeks two and three they watched 30 minutes of two different documentaries in their assigned condition. In each week, an immediate post-test was conducted to examine the cumulative effect of watching programs with captions on listening. In week four, participants took the post-test. All participants were trained and tested individually.

Test materials consisted of short pause-bound utterances extracted from five documentaries in total, two of which were used for the training. All excerpts were only two to nine words long. The pretest was composed of 20 excerpts from the five documentaries, totaling 100 excerpts. Immediate post-tests consisted of 120 excerpts each, and 40 of those were ‘old items’ (i.e. excerpts that the participants were exposed to during the training). Forty items were ‘new’ (i.e. excerpts from the same documentary, but from parts to which participants were not exposed.) These items were tested to check whether learning gets generalized to new utterances produced by the same speaker. Finally, there were 40 ‘unrelated items’ (i.e. excerpts from another documentary which the participants did not watch). This condition was employed to examine whether learning generalizes to different speakers of a broadly similar accent (standard British English in their study).

The post-test consisted of 160 excerpts, and 40 of those were ‘old items’, 40 were ‘new items’, 40 were ‘unrelated items’, and lastly 40 were ‘final unrelated items.’ The ‘final unrelated items’ were excerpts that participants did not encounter in the previous tests. The number of correctly repeated words (the word was counted as correct if it could be recognized as the target word) was counted for each excerpt for each participant, and these numbers were turned into proportion scores.

A mixed design ANOVA with time/test as the within-subject factor and group as the between-subject factor was performed. The result showed no main effect of group (*F*=.17, *p* > .05) suggesting that the overall performance of the three groups across four weeks was not significantly different from each other. There was, however, a main effect of time/test (*F* = 15.22, *p* < .01), confirming an overall improvement overtime. Crucially, there was also an interaction between group and time (*F* = 5.71, *p* < .05), suggesting that the groups did not improve overtime at the same rate. Specifically, the bi-modal group showed the most pronounced improvement overtime. The bi-modal group was lagging behind the two groups in their ability to correctly repeat words on the pre-test, but they consistently outperformed the other two groups on all the post-tests (see Figure 3).

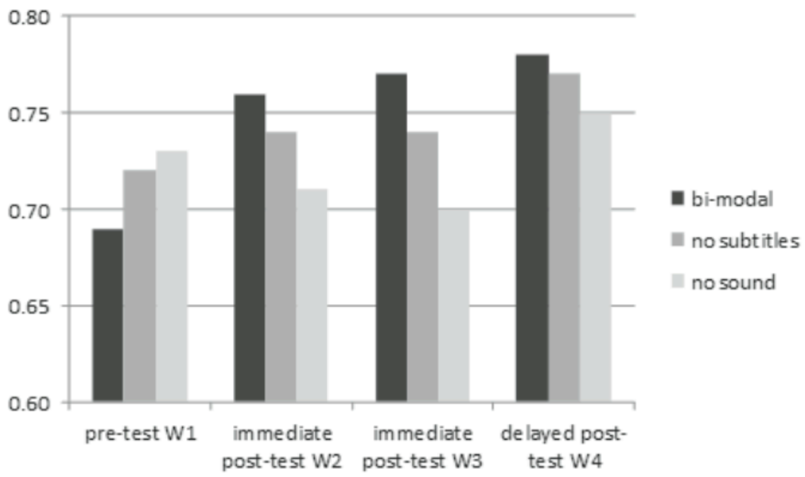


Figure 3 Overall scores across groups. Charles & Trenkic (2015).

When looking at the excerpt types, there was a trend for the bi-modal group to perform more strongly than the other two groups on both ‘old’ and ‘new’ items, but the difference did not reach statistical significance. However, the group by test/time interaction was significant for ‘unrelated’ and ‘final unrelated’ items, suggesting that watching programs with sound and captions may result in learning that generalizes beyond the programs watched.

The results from this study were obtained from a small sample of participants, and some of the predicted effects, while in the right direction, did not reach statistical significance. However, the overall findings from this study suggest that bi-modal presentation of input could benefit learners in developing speech segmentation skills, which is indeed an essential listening skill. Moreover, results from this study suggest that watching programs with captions may be helpful not only for segmenting the spoken input accompanied by captions, but may have a more far-reaching effect on the development of segmentation abilities in L2.

Charles and Trenkic’s study has shown that the improvements were in fact due to better segmentation skills, and not just to the fact that learners were provided with written input. Their study employed a good assessment task (the shadowing task which assesses lexical segmentation, because it requires listeners to identify the words in continuous speech), good controls (introducing the 'no sound' group), and employing excerpts that the participants did not hear during their treatment, which showed generalization to new sentences and new speakers. It is, however, too early to generalize this result, and there is a need to replicate and extend this study further. An entirely different result could be found if a language other than English is tested. While a great deal of research on the usage of captions on L2 listening has been conducted on L2 English learners, none to the researcher's knowledge has been done on L2 Japanese learners. The use of bi-modal input of Japanese will be an interesting case to investigate because 1) Japanese has a different orthographic system from the English alphabet, 2) written Japanese does not have clear visible word boundaries in the form of white spaces positioned between words as in English, and 3) since Japanese is a mora-timed language, speech is phonologically segmented differently from stressed-time languages such as English.

Written Japanese and Japanese Speech Segmentation

Unlike English, Japanese does not utilize the Roman alphabet as its orthographic system. The modern Japanese writing system is a combination of two character types, which are logographic *kanji* (characters originated in Chinese) and syllabic *kana*. There are two kana orthographies, *hiragana* and *katakana.* Both hiragana and katakana are mora-based, and each mora is represented by one character (or one digraph) in each system. This could be a vowel such as /a/ (hiragana あ, katakana ア); a consonant followed by a vowel such as /ka/ (hiragana か, katakana カ) or the moraic nasal /N/ (hiragana ん, katakana ン). Because kana characters do not represent single consonants (except in the case of /N/) but rather represents (C)V syllable, they are referred to as syllabaries rather than alphabets, and the mapping of Japanese phonology to kana orthography is nearly one-to-one. Function words and inflectional affixes are written in hiragana, and foreign words are written in katakana. Hiragana is also used to write content words, but many are written in kanji.

Studies have shown that there are effects of L1 orthography on L2 visual word recognition (Akamatsu, 1999; Chikamatsu, 1996; Koda, 1987, 1989, 1990; Mori, 1998; Wang & Geva, 2003). Visual word recognition is defined here as "a process by which a *reader* identifies a string of printed letters as a meaningful unit" (Chikamatsu, 2006; p.68). According to the literature on visual word recognition, there are different types of *orthographic depth*; there are "shallow" orthographies (alphabetic languages such as Serbo-Croatian, Italian, Spanish, or possibly English) which have highly or relatively consistent sound-spelling correspondence. There are also "deep" orthographies (logographic languages such as Chinese) which possess sound-spelling correspondence characteristics that are not consistent and more opaque. According to the Orthographic Depth Hypothesis, languages with shallow orthography with regular grapheme-phoneme correspondence (GPC) involve greater reliance on phonological information and activates it before lexical access, while languages with deep orthography without systematic GPC involve greater reliance on visual coding where one processes the visual representation of a word as a whole unit directly related to its meaning without phonological mediation (Chikamatsu, 2006). Studies have shown this hypothesis to be true, that the deeper L1 logographic groups are more impaired by the unavailability of visual information than shallower L1 alphabetic groups, showing that there is a heavy visual coding reliance among L1 deeper groups (Akamatsu, 1999; Koda, 1987, 1989, 1990; Wang & Geva, 2003). Similar results were obtained for the case of nonalphabetic L2s. Chikamatsu (1996) found that among L2 learners of Japanese, L1 Chinese learners relied more on visual information than L1 English learners when completing Japanese syllabic kana word lexical judgement tests. In Mori's (1998) study, she had

L1 English and Chinese learners take a Japanese logographic kanji memory test which used pseudo kanji characters. She found that the memory of L1 English participants deteriorated significantly due to phonologically inaccessible characters, compared with the memory of L1 Chinese participants, whose performance showed strong visual processing strategies in memorizing the pseudo kanji characters. Based on these cross-linguistic studies, Chikamatsu (2006) states that there seems to be different degrees of visual (or phonological) coding involvement in L2 word recognition as a function of a learner's L1 orthographic depth (p.69). What is useful about using captioned-videos in assisting learners' word recognition is that captioned-videos provide both visual cue in the form of captions, and phonological cues in the form of auditory information, which could assist word recognition for all learners regardless of their L1 orthographic depth. If this is the case, captions should be beneficial in helping all Japanese learners improve their speech segmentation skill, because speech segmentation is a by-product of the word recognition process (Cutler, 2000).

Another notable difference between English and Japanese text is that unlike English, Japanese words are not separated by spaces. This is why the mixed usage of kanji, hiragana, and katakana is important because it can help identify word and phrase boundaries. As explained above, kanji is always used for content words such as nouns, verb stems and adjectives, katakana for foreign-origin content words, and hiragana is always used for particles and for the endings of verbs and adjectives. Although Japanese writing system does not have spaces between words, the orthographic system provides clues as to where a word starts or ends in a sentence. This is because the three orthographic systems are visually distinct from one another (i.e. hiragana is curvy-shaped while katakana tends to be angular in shape, while kanji looks more complex than hiragana and katakana). A sentence that includes all three orthographic systems looks like this:

私はパデュー大学で勉強します (“I study at Purdue University”).

Table 1
The Breakdown of the Sentence: 私はパデュー大学で勉強します

Word/Morpheme	Orthography	Part of speech	Meaning
私	kanji	pronoun	‘I’
は	hiragana	topic particle	
パデュー	katakana	noun	‘Purdue’
大学	kanji	noun	‘university’
で	hiragana	locative particle	
勉強	kanji	verb stem	‘study’
し	hiragana	verb	‘do’
ます	hiragana	polite auxiliary	

The third notable difference is, as already discussed, that Japanese has a different rhythm than English. While English is a stress-timed language, Japanese is a mora-timed language, and as a result, they are segmented differently. Previous studies suggest that language rhythm determines the segmentation unit most natural to native listeners; English has a stress rhythm, and segmentation by English listeners is based on stress. On the other hand, the rhythm of Japanese is based on morae, and Japanese listeners’ response patterns are consistent with moraic segmentation (Otake et al., 1993). As we saw earlier from previous studies, L2 learners do not segment speech the way native listeners do; having this in mind, we can assume that many L2 Japanese learners experience difficulty in segmenting spoken Japanese because there are considerably fewer languages that are categorized into the mora-timed language compared to syllable-timed and stress-timed languages (Nespor, Shukla, & Mehler, 2011).

Having these differences compared to English, it is worthwhile to investigate whether bi-modal presentation of input (Japanese sound with Japanese captions) would foster L2 Japanese learners’ speech segmentation skills. If bi-modal input can lead to the improved L2 speech segmentation ability among L2 learners of Japanese, it will have obvious pedagogical implications. Captioned educational audio-visual materials are available in large quantities in Japanese as well.

Present Study

The present study was inspired by Charles and Trenkic's (2015) study, and examined the effectiveness of bi-modal input on fostering L2 Japanese learners' speech segmentation skills. The data was collected from 12 third-year Japanese learners at a large state university in the Spring 2018 semester. These levels of participants were chosen because the bi-modal L2 input procedure requires a certain level of skill in the L2 in order to comprehend authentic materials to be used.

In Charles and Trenkic's study, they divided their participants into three groups: 1) bi-modal (sound and captions), 2) no caption (sound but no captions), or 3) no sound (with captions but without sound). In this study, however, we only had a single experimental group since we adopted a single-case design (SCD). SCD is a variation of time-series designs in which progress of a single participant or a single group of participant is examined through interventions over a period time (Kennedy, 2005). This design was chosen for three reasons. First, it was due to time constraint. In Charles and Trenkic's (2015) study, their experimental session only lasted for four weeks, and as they mentioned themselves, this time period could have been too short to reach statistical significance. For this study, we preferred to have a longer treatment phase, but we could not allocate an extensive amount of class time to the experimental sessions, since each third-year level class was 50 minutes long and only met three times per week. Second, it was due to the restricted environment. Since there were only a total of 30 Japanese learners in the third-year level classroom, we believed that we would not have enough statistical power if the learners were divided into multiple comparison groups. SCD is useful when group-comparison designs are not feasible for the context. Third, since no empirical evidence of the effectiveness of captioned-videos on fostering L2 Japanese learners' speech segmentation skills has been found up to this point, it is important to first investigate whether a causal relation exists between the independent variable (i.e. the bi-modal input) and the dependent variable (i.e. L2 Japanese learners' speech segmentation skills). SCD lets us determine whether a causal relationship exists between the variables, which is an important information to have before conducting a larger comparison study.

As a result of these modifications, participants in this study (all belonging to a single bi-modal group), underwent a semester-long SCD experiment in which their speech segmentation abilities were assessed. Experiment details will be discussed in the next chapter. This research was designed to answer the following seven research questions.

Research Questions

Research Question 1

Does repeated exposure to bi-modal input (audio + visual) help L2 Japanese learners improve their ability to segment utterances into words in the video that they watched?

Research Question 2

Does repeated exposure to bi-modal input help L2 Japanese learners retain their speech segmentation skills?

Research Question 3

Does L2 Japanese learners' learning get generalized to utterances produced by the same speaker in the video they have not watched?

Research Question 4

Does L2 Japanese learners' learning get generalized to utterances produced by a different speaker of a broadly similar accent (i.e. standard Japanese/Tokyo dialect) in the video they have not watched?

Research Question 5

What are the L2 Japanese learners' reactions towards the bi-modal input?

Research Question 6

Does the learners' L1 affect the effectiveness of the bi-modal input?

Research Question 7

What are some possible reasons why some learners did not respond to the intervention as well as others?

Hypothesis

Previous studies have shown that captions can aid "with the phonological visualization of aural cues in the minds of listeners, who become more certain of ambiguous input" (Danan, 2004; p. 70). If this is the case, we hypothesize that participants in this study will improve their ability to segment utterances into words in the video they watched. In addition to that, captions can help learners to "accurately form a memory trace of the words, and can later more easily identify identical sounds without textual support" (Danan, 2004; p. 70). If this is the case, learners should be able to segment speech in the video they watched even if they are tested few weeks after watching the video (delayed post-test) just as well as when they are tested immediately after watching the video (immediate post-test).

Charles and Trenkic's (2015) study showed that the participants in the bi-modal group also made improvements in segmenting utterances beyond the programs watched, regardless of whether the speaker was the same or not. If bi-modal input can qualitatively change the phonological representation of the word in the learners' mind, which in turn helps them in making long-term changes in their speech perception, learners should eventually become able to recognize words in connected speech in which they were not exposed to. That is why we hypothesize that after receiving the bi-modal input treatment for over the course of one semester, participants will also improve their speech segmentation skills beyond the programs watched in the treatment phases. It is, however, natural to assume that participants will first become better able to segment speech and recognize words spoken by the same speaker they were exposed to during the treatment, and then gradually become better at segmenting utterances spoken by a different speaker. If this is true, we hypothesize that a greater improvement will be observed on utterances produced by the same speaker than from a different speaker in the video they have not watched.

The complexity and transient nature of spoken language often make it difficult for language learners to truly comprehend L2 utterances. Since captions can visually support learners comprehend the speech they hear, bi-modal input may help lower learners' affective filter by relieving some of the anxiety experienced by them. If this deems to be true, participants should experience an overall positive reaction towards the bi-modal input intervention.

Regarding the difference in the learners' L1, we hypothesize that it will not affect the effectiveness of the bi-modal input if the Orthographic Depth Hypothesis holds true. The Orthographic Depth Hypothesis states that alphabetic languages (such as English) tend to rely more on phonological information when recognizing words, whereas logographic languages (such as Chinese) tend to rely more on the visual/lexical information. Since captioned videos provide both phonological and visual cues, it should be able to help Japanese learners recognize words in connected speech regardless of their L1.

Lastly, the choice of a SCD as the design for this study allows the researcher to look more carefully into the individual differences of the participants in their response to the bi-modal intervention. Some participants may not respond to the intervention as well as other, due to the strategy they are using while watching the captioned videos. Studies have shown that the written input helps viewers understand the audio rather than distracting them (Charles & Trenkic, 2015; Guichon & McLornan, 2008; Hayati & Mohmedi, 2010; Markham, 1989), but if viewers are not familiar with watching videos with captions, it could lead to a sense of distraction and poor language gains (Danan, 2004). We will investigate whether there seems to be a relationship between participant's familiarity with viewing captioned-videos, how they interact with the videos (how much they pay attention to the captions), and their improvement on their speech segmentation skills.

CHAPTER 3. METHODOLOGY

Overview

This study examined the effectiveness of bi-modal input on fostering L2 Japanese speech segmentation skills through a Single-Case Design (SCD). Participants were 12 third-year Japanese learners at a public university in the Midwestern United States, and they participated in a semester-long pre-post design experiment in the Spring 2018 semester. They took Elicited Imitation Task (EIT) for their pre-post-tests, and watched 15 episodes of Japanese documentary series with captions over the course of 11 weeks. Table 2 is an overview of the study.

Table 2
Overview of the Study

Spring 2018	**Experimental Session**	**Session with the researcher**	**Individual video viewing & post-test**
2/5 - 2/9	Week 1	Background survey Take J-CAT	
2/12 - 2/16	Week 2	Grand pre-test (new/different-speaker items)	
2/19 - 2/23	Week 3 (Cycle 1)	Video 1, 2, 3 pre-tests	View video 1 → Video 1 post-test View video 2 → Video 2 post-test View video 3 → Video 3 post-test
2/26 - 3/2	Week 4 (Cycle 2)	Video 4, 5, 6 pre-tests	View video 4 → Video 4 post-test View video 5 → Video 5 post-test View video 6 → Video 6 post-test
3/5 - 3/9	Week 5 (Cycle 3)	Video 7, 8, 9 pre-tests	View video 7 → Video 7 post-test View video 8 → Video 8 post-test View video 9 → Video 9 post-test
Spring Break			
3/19 - 3/23	Week 7 (Cycle 4)	Video 10, 11, 12 pre-tests	View video 10 → Video 10 post-test View video 11 → Video 12 post-test View video 13 → Video 13 post-test

Table 2 (continued)

3/26 – 3/30	Week 8 (Cycle 5)	Video 13, 14, 15 pre-tests	View video 14 → Video 14 post-test View video 15 → Video 15 post-test View video 16 → Video 16 post-test
4/2 – 4/6	Week 9	Grand post-test (new/different-speaker items)	
No session			
4/16 – 4/20	Week 11	Delayed post-test video 1, 2, 3 Post-experiment survey	

Participants

Twelve third-year Japanese learners at a large state university volunteered to participate in this semester-long experiment in the Spring 2018 semester. This level of participants were chosen because they must have a certain level of proficiency in Japanese in order to comprehend authentic materials to be used. The present researcher recruited participants at the end of Fall 2017 semester who were then taking JPNS 301, a fifth-semester course at this university. During recruitment, participants received a detailed explanation of the study procedures and timeline, including an IRB-approved information form, and they were given the opportunity to ask questions.

A background survey was administered during Week 1 of the experimental session. Participants answered the survey through an online platform called *Qualtrics*, and the survey included 25 questions. The survey asked for participants' general background information, as well as what kind of Japanese input they regularly receive outside of the classroom. It also asked how used to/comfortable participants were with watching TV shows/movies with subtitles. All 12 participants completed the survey, and the complete survey is provided in Appendix A. The following information about the participants were obtained through the survey.

Five male and seven female students volunteered to participate, whose ages ranged from 20 to 24 with the average being 21.08 years old. Participants' L1s were English (4), Vietnamese (1), and Chinese (7).

Participants' majors were various, and only one majored in Japanese. Seven of them, however, were minoring in Japanese (See Table 3 for a complete list of majors). Most participants started learning Japanese at this university from 101 (the very first semester course at this university), except for three participants. One participant started learning Japanese in high school and tested into JPNS 201 (third-semester course) upon entering college, and the other two took an online Japanese course and tested into JPNS 201 and 202, respectively. The average length of their Japanese study was 2.8 years. Nine participants were taking the JPNS 302 course during the semester this experiment took place, while three were not, but have taken the JPNS 301 course in the previous semester. None of them had lived in Japan before.

In order to find out how much Japanese input they regularly received outside of class, the researcher asked 1) whether they regularly watch Japanese shows (such as Japanese TV dramas, movies, animes, and YouTube videos), 2) whether they regularly listen to Japanese music, and 3) whether they regularly read Japanese.

When asked whether they regularly watch Japanese shows, 11 of the participants answered *yes* while only one answered *no*. The time they spent watching Japanese shows ranged from three hours to 21 hours per week, with the average being eight hours. Nine out of these 11 participants said they watch Japanese shows in Japanese with subtitles in L1, one said they watch the shows in Japanese without any subtitles, and one said they watch it dubbed in his L1. All 12 participants answered that they were comfortable watching TV shows/movies with subtitles.

Nine participants said they regularly listen to Japanese music, ranging from 1.5 hours to 35.8 hours every week (the average being 16 hours). Only five participants answered that they regularly read Japanese other than their course materials, which included Japanese articles about video games, manga, anime news, twitter, newspaper articles, Japanese light novels, game instructions, and Japanese websites. All in all, all participants were highly motivated to learn Japanese as we can see from the amount of various input they were voluntarily receiving outside of class. This response though, was collected at the beginning of the semester (Week 1 of the experimental session), and when the researcher asked the participants again in Week 11 in an informal interview, the average hours watching Japanese shows slightly dropped from eight hours per week to six and a half hours per week, as well as time spent listening to Japanese music (Week 1 average was 16 hours per week but Week 11 was 15 hours per week). Participants answered either they watch/listen the same amount or somewhat less, because their schedule got busier as the semester progressed.

Table 3
Participants' Background Information

Participants	12 (5 Male, 7 Female)
Age	20-24
Major	Linguistics (3), Computer Graphics Technology (2), Asian Studies (2), Computer Engineering (1), Computer Science (1), Pharmacy (1), Japanese (1), Management (1), Actuarial Science (1), Industrial Engineering (1)
Average Length of Study	2.8 yrs ~ JPNS 301 (3) ~ JPNS 302 (9)
L1	English (4), Chinese (7), Vietnamese (1)
Do you regularly watch Japanese shows?	Yes (11) (Week 1 average 8 hrs/week; Week 11 average 6.5 hrs/week) No (1)

Table 3 (continued)

How do you watch the shows?	In Japanese with subtitles in L1 (9) In Japanese without subtitles (1) Dubbed in L1 (1)
Do you regularly listen to Japanese music?	Yes (9) (Week 1 average 16 hrs/week; Week 10 average 15 hrs/week) No (3)
Do you regularly read Japanese materials? If so, what do you read?	Yes (5) → Japanese articles about video games, manga, anime news, twitter, newspaper articles, Japanese light novels, game instructions, and Japanese websites No (7)

In order to assess participants' general Japanese language proficiency, the researcher administered the Japanese Computerized Adaptive Test (J-CAT) before the experimental session started. J-CAT is commonly used as a proficiency test, and was developed by the J-CAT Project team at the University of Tsukuba. It can be taken free of charge online (as of March, 2019).

Participants' four skills, which included listening, vocabulary, grammar, and reading skills were assessed using this test, all of which are skills needed to comprehend the captioned videos. Each section is scored out of 100 points, so the perfect score for the entire test is 400 points. Table 4 shows how to interpret J-CAT scores, taken from their website. According to their definition, "Basic" learners can exchange basic ideas, "Intermediate" learners can manage daily communication, and "Advanced" learners can manage academic and professional communication (J-CAT Project team, n.d.). The 12 participants' proficiency ranged from the lowest overall score of 130 points to the highest of 257 points, with the average being 197 points. All participants' scores are shown in Table 35 on p.97.

Table 4
Interpretation of J-CAT Scores

J-CAT	Proficiency Level
-100	Basic
100-150	Pre-Intermediate
150-200	Intermediate
200-250	Intermediate-High
250-300	Pre-Advanced
300-350	Advanced
350-	Near Native

Single-Case Design (SCD)

This study utilized a Single-Case Design (SCD). SCD is a variation of time-series designs in which progress of a single participant or a single group of participants is examined through repeated interventions over a period of time (Kennedy, 2005). The central goal of SCD is to determine whether a causal relation exists between the introduction of intervention and change in a dependent variable (Levin & Kratochwill, 2003). SCD often involves repeated, systematic measurement of a dependent variable before, during, and after the active manipulation of an independent variable (e.g. applying an intervention) (Kratochwill et al., 2010). The replication of intervention can be done by introducing and withdrawing the independent variable, known as the reversal/withdrawal (ABAB) design. Hence, the participant or group serves as its own control for purposes of comparisons. SCD also benefits from collecting data at multiple time points as opposed to a single pre-post intervention, as it can decrease the random variability within data (Greaves, Camic, Maltby, Richardson, & Mylläri, 2012).

In SCD, each outcome variable must be measured systematically over time by more than one rater, and the interrater agreement must be documented on the basis of a statistical measure of interrater reliability. An effect is shown when ABAB design have a minimum of four phases per case with at least three data points per phase. Any phases based on fewer than three data points cannot be used to demonstrate existence or lack of an effect (Kratochwill et al., 2010).

SCD has often been used in applied and clinical disciplines in psychology and education, such as school psychology and the field of special education (Kratochwill et al., 2010). This design is useful when group-comparison designs are either not suitable for the research question or not feasible for the context, as with very small samples or heterogeneous populations (Kennedy, 2005). Furthermore, SCD has the advantage of providing detailed documentation of the characteristics of those cases that *did* respond to an intervention and those that *did not* (Kratochwill et al., 2010).

As stated earlier, learners must have a certain level of Japanese proficiency to participate in this study since they would watch authentic Japanese TV shows with Japanese captions, not with subtitles in their L1. This meant that participants must be at least third or fourth year Japanese learners, and there were only a total of 30 learners taking the third-year Japanese course when this study took place, and eight fourth-year learners. The number of participants here were assumed not high enough to achieve statistical power with a group-comparison design study. In addition to that, watching 15 videos in class plus taking pre-post tests were expected to be too time consuming, especially since each class was only 50 minutes long and they only met two to three times per week. It was also difficult to integrate video viewing as a take-home assignment, because there were many other assignments to be done in the course. For these reasons, group-comparison designs were not feasible for the context, thus SCD was deemed as an appropriate methodology for this study. The advantages of using SCD is that it offers valuable data before conducting a larger-scale comparison study, as it is important to first investigate whether a causal relation exists between the intervention (i.e. the bi-modal input) and the dependent variable (i.e.

L2 Japanese learners' speech segmentation skills). Since no empirical evidence of the effectiveness of captioned video on fostering L2 Japanese learners' speech segmentation skill has been found, this study could serve as a basis to conduct a larger between-groups design study in the future if positive effects of the bi-modal input is found. Also, by using SCD, researchers are able to look more carefully into individual differences in interactions with the bi-modal input. In addition to that, working with a small number of participants can let researchers ensure that participants are following the instruction correctly, which is sometimes difficult to manage in a larger group-comparison design.

Materials

Survey.

A survey was administered twice during the course of the 11 weeks, once at the beginning and once at the end of the eleven-week experimental session. The first survey collected participants' background information, and the second survey collected participants' reactions to the experiment they participated in. All items were created and distributed through an online platform called *Qualtrics.* All survey items can be found in Appendix A and B.

Videos.

The intervention of this study was video viewing with captions. Treatment materials consisted of 15 short videos from a documentary series called *Mieruzo! Nippon* (見えるぞ！ニッポン) "I can see you! Japan" offered through *NHK*, Japan's national public broadcasting organization. This is a social studies program for Japanese 3rd and 4th graders, and each episode focuses on a specific prefecture in Japan, introducing its industries, people, livelihoods, and traditions. There are three characters in the show, which are two animated characters and one female narrator speaking in standard Japanese (i.e. Tokyo dialect). This show has an option to watch either with or without captions, and can be accessed without cost on the *NHK for School* website (website available on the reference). The captions use all three Japanese writing systems, kanji, hiragana, and katakana. However, most kanji have *furigana* on top, which is a Japanese reading aid, consisting of smaller hiragana printed normally either on top or next to the kanji to indicate its pronunciation (see Figure 4 for example). In modern Japanese, it is often used in children's or learners' materials but it is also used to assist native Japanese adults with rare, nonstandard, or ambiguous kanji readings.

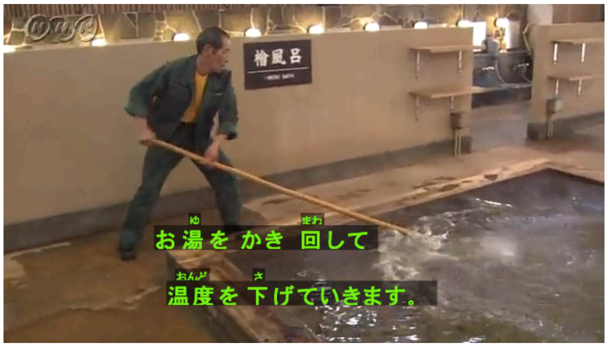


Figure 4 Example of a caption with the use of *furigana*

This show was chosen because 1) it can be watched with captions, and can be accessed without cost on the *NHK for School* website, 2) each video is only 15-minutes long, 3) teachers who were teaching the third-year Japanese course at this university at this time said this show is appropriate in terms of difficulty for learners at those levels 4) the narrator is always the same female-speaker speaking in standard Japanese (i.e. Tokyo dialect), 5) it was necessary to choose a show whose content will not influence the test results, and it was thought that most students will not be familiar with obscure prefectures in Japan (i.e. prefectures other than the well-known ones like Tokyo, Kyoto, Osaka, Hiroshima, etc.) and 6) the content of the video was educational and hence, has the potential to be integrated into Japanese course curricula.

Testing Materials.

Elicited Imitation Task. Participants took the pre-post-tests in the form of Elicited Imitation Task (EIT). EIT requires participants to listen to short sentences and to repeat them following a short pause. This task, therefore, simply tests their ability to identify words in a continuous speech stream rather than testing their general listening comprehension skills. The logic of the EIT is that if participants do not understand the string of sounds they are presented with, they cannot arrange it in chunks and are not able to hold it in their short term memory for the time necessary to decode it (Vinther, 2002). One may think that participants could "parrot[1]" what they have heard without understanding the sentences, but studies have shown that that is only possible when the sentences are short enough to be retained in immediate memory (Lee, 1970; Lust, Chien, & Flynn, 1987; Munnich, Flynn, & Martohardjono, 1994). For example, Vinther (2002) tested six adult native speakers of Spanish in imitating ten sentences, which consisted 16 and 26-syllable-long normal Spanish sentences and nonsense sentences that were linked into Spanish-sounding words and Spanish syllable structure. The result showed that the participants were able to repeat all the normal sentences and the 16-syllable nonsense sentences without any problem, but they had no success with the 26-syllable nonsense sentences. Miller and Chapman (1975) studied the factors of total number of words, number of content words only, and lexical density in their search for a counting index to predict the difficulty of sentence repetition. They concluded that morpheme length (defined as the number of meaningful units) rather than word length was the most influential factor.

[1] Parroting: senseless repetition of the acoustic image of a sentence which is so short that its sounds can be phonologically processed in working memory without their meaning being decoded (Vinther, 2002)

It may be difficult to determine whether a failure in imitating a sentence derives from insufficient listening skills or from a lack of speaking skills. However, Vinther (2002) states the following, which explains why EIT could be used to assess listening skills:

> If subjects exposed to listening comprehension training and tested with EIT improve pretest scores, i.e. imitate a greater number of sentences correctly in the test than in the pretest, it is most natural to ascribe the improvement to the training and consider that what has improved is their decoding and comprehension skills. Their productive ability may have improved too, or, in case it has existed in concealed form during the pretest, it may have been allowed to emerge in the test. Poor productive ability may result in low test scores even if the subject has been able to understand the model sentences. But the opposite situation, that subjects should be able to produce well-formed imitations of sentences they have not understood and are not able to remember, is highly improbable. (p.63)

Excerpts. EIT test materials consisted of short excerpts taken out of the videos. A total of 219 excerpts were taken out from the 15 videos that the participants have watched, where the same narrator is speaking ('old' items) and 15 excerpts with the same narrator from the same show ("I can see you! Japan") but an episode they did not watch ('new' items). These items were introduced to check whether learning gets generalized to new utterances produced by the same speaker. There were additional 15 excerpts from a different *NHK for School* show that the participants did not watch ('different-speaker' items). These excerpts were taken out from another social studies program called *Shittoku Chizucho* (知っトク地図帳) "Good to know the maps", which is for Japanese 3rd and 4th graders, alike. The narrator is different from the other program, but was also a female who spoke standard Japanese. The 'different-speaker' items were introduced to investigate whether learning generalizes to a different speaker of a broadly similar accent (i.e. Tokyo dialect).

All the vocabulary and grammar used in the excerpts have been learned previously, and the difficulty of the vocabulary was controlled across all excerpts using an online vocabulary level checker called *Reading Tutor* (リーディングチュウ太) (Kawamura, Kitamura, & Hobara, n.d.). All excerpts are chunks of meaningful phrases that are acceptable to a native Japanese speaker, and all excerpts had background noise as it was directly taken out of the videos (all videos constantly had background noise). Since previous studies show that morpheme length is the most crucial factor in determining the difficulty of EIT (Miller & Chapman, 1975; Vinther, 2002), the researcher piloted multiple times to find out the morpheme length that appeared to be longer than the participants' immediate memory span and yet not so long that they were too difficult for them to process. All excerpts contained six to 12 morphemes, with the average being eight morphemes per sentence. A 26-year-old female native Japanese speaker took the EIT for all excerpts and only the items where she was able to get 100% on her first try were used for the experiment.

EIT was carried out using an online oral practice/assessment platform called *Speak Everywhere* (Fukada, 2013), and the presentation of the excerpts were randomized for each participant. Participants' utterances were saved on *Speak Everywhere* where the raters later accessed for scoring.

Procedure

Before the experimental session started, participants took the J-CAT in a computer lab all at once, and they also answered the background survey.

In Week 2 and Week 9 of the experimental session, participants took a grand pre-post-test, in the form of EIT. The excerpts in the grand pre-post-test were the 'new' and 'different-speaker' items (i.e. excerpts participants were not exposed to during the intervention). Each participant met individually with the researcher in her office to take the grand pre-post-test. Participants were explained about the EIT, and practiced a few times before they took the test.

Between the grand pre-post-test, there were five cycles of pre-post-tests, which were all EITs. In cycle 1, participants individually met with the researcher to take three pre-tests, corresponding to three videos to be assigned for the cycle. Within the cycle, each time they viewed a video, they took a post-test for that video. They were instructed to watch the assigned videos only once on their own and to complete the post-test immediately afterwards, and to complete this task before they met with the researcher the following week. Four more cycles of the same format followed, which resulted in them having watched 15 videos in total.

The following are the reasons why the participants viewed the videos and took the post-test on their own time. First, it was due to a lack of time. The three pre-tests took about 15 minutes to complete, as well as the post-tests, and each video was 15 minutes long. Hence, participants would have had to meet with the researcher 75 minutes each week, and this would have been too time-consuming for the participants as they were all unpaid volunteers. Second, we wanted to give them flexibility to work with their schedule by having them view the videos and take the post-test at home so that they could complete the semester-long experiment.

Participants were instructed to watch three videos each week, and the presentation of videos were counterbalanced across participants. To prevent the presentation order of the videos from affecting the results, we randomly assigned six different orders to the participants. The videos can be accessed through the *NHK for School* website, but in order to make sure that the participants were watching the correct videos and were watching it with captions, the researcher captured each video with captions and made it into a YouTube video. Each week, participants received the links to the three YouTube videos after they took the pre-test. Participants were strictly instructed to watch each video once and to take the post-test only once immediately after viewing the video in order to assess their improvement accurately. It was emphasized that the experiment was not a competition amongst the participants and had no impact on their grades (if they were taking a Japanese course at the time).

In Week 11, participants met individually with the researcher and took the delayed-post-test, which was the same EIT they had taken in cycle 1 in Week 2. This was done to investigate how much learners were able to retain their word segmentation skills from the 'old' items after nine weeks. Participants also answered the post-experiment survey at this time using *Qualtrics* again.

Scoring and Analysis

Scoring.

Because Japanese is an agglutinative language, one "word" can contain several morphemes that may be expressed as separate words in other languages. For example*,* *yomaserareta* is read-CAUSATIVE-PASSIVE-PAST 'was made to read'. Since native speakers are not able to analyze such strings of morphemes reliably, a morphological analyzer developed within a computer program called *Chakoshi* was used to produce correct analysis for grading.

Two raters who were instructors of Japanese at the same university listened to all excerpts that were saved into *Speak Everywhere* and counted the number of correctly repeated morphemes from each excerpt for each participant, and the result were turned into proportion scores. Each word was counted as correct if it was recognized as the actual morphemes used in the video. Synonyms were also counted as correct if it meant the same thing as the actual word used in the excerpt. For example, the word for "car" could be said both as *kuruma* or *jidōsha*, and both of them were accepted since saying either word shows their ability to identify and render the meaning of the word for "car." The interrater reliability was assessed using *IBM SPSS Statistics*, and the intraclass correlation coefficient for all excerpts were >.99, showing a very strong interrater reliability. The average of the two raters' scores were used as the final score for analysis.

Analysis.

Old Items.

To determine whether a causal relation exists between the introduction of intervention (bi-modal input) and change in the dependent variable (EIT scores), an analysis for the old items (items taken out from the videos participants watched) was made following the steps provided by *What Works Clearinghouse* (Kratochwill et al., 2010). In this technical documentation report which was compiled to provide standards for SCDs, a detailed explanation on data analysis is provided:

> The rationale underlying visual analysis in SCDs is that predicted and replicated changes in a dependent variable are associated with active manipulation of an independent variable. The process of visual analysis is analogous to the efforts in group-design research to document changes that are causally related to introduction of the independent variable. In group-design inferential statistical analysis, a statistically significant effect is claimed when the observed outcomes are sufficiently different from the expected outcomes that they are deemed unlikely to have occurred by chance. In single-case

research, a claimed effect is made when three demonstrations of an effect are documented at different points in time. (p.21)

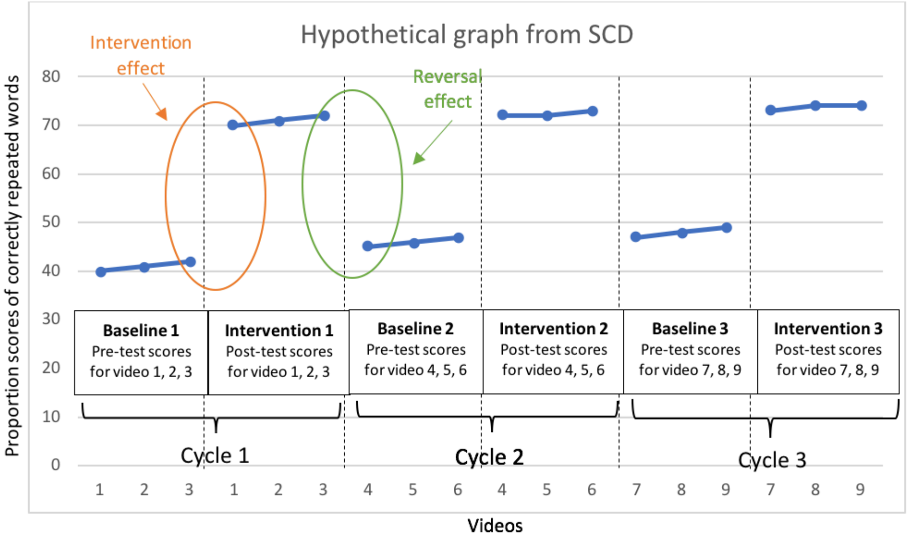


Figure 5 Hypothetical graph from SCD

Figure 5 is a hypothetical graph that shows the single-case reversal design (i.e. ABAB design) used in this study. The vertical axis is the proportion scores of correctly repeated words (EIT scores) and the horizontal axis is the number of captioned-videos participants watched. In each cycle, there are three pre-test scores, which are called the baseline scores. There are also three post-test scores, indicating that the participants watched three captioned-videos (i.e. received intervention of bi-modal input) with which the post-tests were paired. The participants' EIT scores increased from baseline phases to the intervention phases, showing an intervention effect and suggesting a possible functional relationship between the IV (the bi-modal input intervention) and the DV (EIT scores). We can also see a reversal effect, where there is a decrease in the scores from the intervention 1 in cycle 1 to the baseline 2 scores in cycle 2. This is because the scores from baseline 2 is another set of three pre-test scores before they watch the corresponding captioned-videos. This relationship of the intervention effect and the reversal effect is the main focus of the data analysis step for the old items, with the primary question being whether a causal relationship between the IV and DV can be inferred. According to Kratochwill et al. (2010), at least three attempts are needed to demonstrate an intervention effect at three different points in time, as we can see from Figure 5. In this graph, the experiment is replicated two times (where a single experiment is a baseline + intervention). The current study replicated four times.

Kratochwill et al. (2010) provide standard rules for conducting visual analysis in SCDs in *What Works Clearinghouse* (Kratochwill et al., 2010). They have synthesized the methodology literature in SCD and established best practices when conducting the visual analysis, in order to improve the inferences that can be made within a given study. Their rules for conducting visual analysis involve six variables and four steps. The following six variables are used to examine within- and between-phase data patterns:

1. **Level**: The mean score for the data within a phase. This will be the mean EIT score in this study.
2. **Trend**: The slope of the best-fitting straight line for the data within a phase. A relatively flat trend is expected in this study since the difficulty of the EIT items were controlled across all excerpts.
3. **Variability**: The range or standard deviation of data about the best-fitting straight line. This refers to how scattered the data points are around the trend line; the more scattered they are, the lower the ability to establish a stable level and trend.
4. **Immediacy of the effect**: The change in level between the last three data points in one phase and the first three data points of the next. Since this study contains only three data points within each phase, all data points will be used for analysis. The more rapid

(or immediate) the effect, the more convincing the inference that change in the outcome measure was due to manipulation of the independent variable.

5. **Overlap**: The proportion of data from one phase that overlaps with data from the previous phase. In Figure 5 there are no overlaps between any points of baseline phases and intervention phases. The larger the separation (as in Figure 5), the more compelling the demonstration of an effect.
6. **Consistency of data patterns across similar phases**: Looking at data from all phases within the same condition (i.e. all "baseline" phases; all "post-intervention" phases) and examining the extent to which there is consistency in the data patterns. The greater the consistency, the more likely the data represent a causal relation.

The following four steps were taken when assessing visible patterns within and between phases and to test whether the data support the inference:

Step 1: Document predictable baseline pattern.

The two purposes of a baseline are to "(a) document a pattern of behavior in need of change, and (b) document a pattern that has sufficiently consistent level and variability with little or no trend, to allow comparison with a new pattern" (Kratochwill et al., 2010, p.19). The effect of the independent variable cannot be assessed unless the baseline demonstrates the proposed concern/problem. The baselines in Figure 5 illustrates such a pattern of the speech segmentation skills. The baseline in each cycle consists of three EIT scores which follow a flat trend across the phase and demonstrate limited variation. The scores are rather low, all below 50% accuracy.

Step 2: Examine the data within each phase of the study.

This step is taken in order to assess within-phase pattern(s). The level, trend, and variability of the data within each phase should be assessed.

Step 3: Compare the data from each phase with the data in the adjacent (and similar phases).

This step assesses "whether manipulation of the independent variable was associated with an "effect." An effect is demonstrated if manipulation of the independent variable is associated with predicted change in the pattern of the dependent variable" (Kratochwill et al., 2010, p.18). In Figure 5, the baseline phase 1 is compared to intervention phase 1 and intervention phase 1 is compared to baseline phase 2. Similar phases (i.e. baseline phase 1 to baseline phase 2, intervention phase 1 to intervention phase 2) should also be compared.

Step 4: Integrate all information from all phases of the study.

The final step in visual analysis is to integrate all the information gleaned from steps earlier "to determine whether there are at least three demonstrations of an effect at different points in time (i.e., documentation of a causal or functional relation)" (Kratochwill et al., 2010, p.18).

According to Kratochwill et al. (2010), if the criteria of the six features are met following the four steps for visual analysis, the data are deemed to document a causal relation, and the inference may be made that change in the outcome variable (EIT scores) is causally related to manipulation of the independent variable (bi-modal input).

New Items and Different-Speaker Items.

New items consisted of 15 excerpts with the same narrator from the same show ("I can see you! Japan") but an episode they did not watch. Different-speaker items consisted of 15 excerpts from a different show that the participants did not watch, but were spoken by a speaker with a broadly similar accent (i.e. Tokyo dialect) as the speaker from the old and new items. The new and different-speaker items were introduced to investigate whether learning generalizes to new utterances produced by the same and different speaker. A paired sample *t*-Test was run using *IBM SPSS Statistics 25* to compare the 12 participants' mean scores between the grand pre-test and post-test for the new items, as well as for the different-speaker items. This was because unlike the old items, there were only two points to compare, and the same SCD analysis procedure could not be done for the new and different-speaker items. Instead, a simple statistical analysis was carried out to see whether participants had made a significant improvement in their EIT scores after the intervention, even on items they had not watched. Before running the analysis, it was made sure that 1) there were no significant outliers in the pre-post test scores, and 2) that the scores were approximately normally distributed.

Delayed-post-test Scores.

For the same reason explained above, a repeated measures ANOVA was run using SPSS to compare the means from the pre-test and post-test for the three videos participants viewed in Week 3 (cycle 1) to the means from the delayed-post-test they took in Week 11 to see whether repeated exposure to bi-modal input help L2 Japanese learners retain their speech segmentation skills on items they were exposed to after eight weeks. Before running the analysis, it was made sure that the scores were approximately normally distributed, there were no significant outliers, and the sphericity was tested as well.

Level of Significance.

The level of significance was set to .05 for all statistical analysis. Because this study contains three instances of statistical hypothesis testing (two *t*-Tests and one ANOVA), α is set to .016 (Bonferroni corrected) for detecting 5% level of significance.

Pre-post-experiment Survey.

A qualitative analysis was done to see whether a relationship exists between the participants' background information (such as their L1, amount of Japanese input, how much they paid attention to the captions in the video, etc.) and the pattern of their improvement. The data from the pre-post-experiment survey were intended to further elucidate the quantitative results.

CHAPTER 4. RESULTS AND DISCUSSIONS

Overview

This chapter presents the results of the data analysis to investigate the primary research question in this study: does bi-modal input assist L2 Japanese learners to improve their speech segmentation skills?

The data analysis will first show the results of all 12 participants whose scores were compiled into a single group. Then, participants will be grouped based on their L1 and whether or not the L1 groups behaved differently will be investigated. Finally, data from two participants who *did* respond particularly well to the intervention will be compared to the two who *did not* to investigate individual differences in response to the intervention.

Results Pertaining to Research Question 1

Results regarding Research Question 1, "Does repeated exposure to bi-modal input (audio + visual) help L2 Japanese learners improve their ability to segment utterances into words in the video that they watched?" will first be analyzed. Table 5 shows a summary of means (in proportion scores) of all 15 pre-test (baseline) and post-test (post-intervention) scores and the pre-post-test means for the five cycles. Figure 6 displays a bar graph of the average pre-test scores and the corresponding videos' average post-test scores. The dotted bars are the pre-test scores and the solid bars are the post-test scores.

Table 5

Summary of Pre-Post-Test Scores for RQ1

		Pre-test %	Post-test %	Pre-test mean %	Post-test mean %
Cycle 1	Video 1	56.15%	63.11%	58.82%	66.38%
	Video 2	61.56%	69.36%		
	Video 3	58.76%	66.67%		
Cycle 2	Video 4	62.37%	73.21%	62.32%	74.46%
	Video 5	57.99%	69.78%		
	Video 6	66.60%	80.38%		
Cycle 3	Video 7	72.45%	80.85%	66.35%	73.40%
	Video 8	62.69%	67.93%		
	Video 9	63.90%	71.41%		
Cycle 4	Video 10	61.88%	71.85%	60.14%	69.72%
	Video 11	57.03%	67.37%		
	Video 12	61.52%	69.95%		
Cycle 5	Video 13	59.85%	62.83%	54.45%	61.30%
	Video 14	57.22%	65.37%		
	Video 15	46.28%	55.69%		
Average				60.42%	69.05%

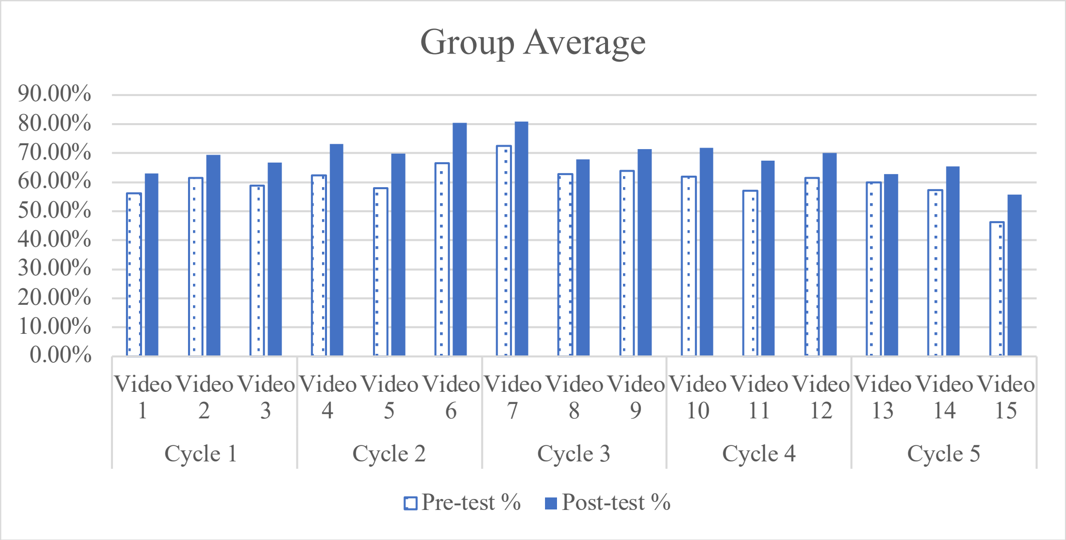


Figure 6 Bar graph of the pre-post-test EIT scores

While a native Japanese speaker was able to obtain 100%, the average pre-test score for third-year learners was only 60.42%. Figure 7 shows the line graph with both trend (linear) lines for the pre-test and post-test phases. We can see a relatively stable trend for the baseline (pre-test), except for the scores from Video 7 and 15. The score from "Pre Video 7" is considerably higher than the trend line and "Pre Video 15" lower. Although it was made sure that all the grammar in the excerpts were previously learned and that the vocabulary level and the morpheme length was controlled, excerpts from Video 7 could have been easier than average and excerpts from Video 15 could have been more difficult. More of this will be discussed in the limitation section in Chapter 5. All in all however, the overall graph in Figure 7 shows small variability in the baseline (pre-test) phases, allowing us to make comparison with the post intervention (post-test) phases.

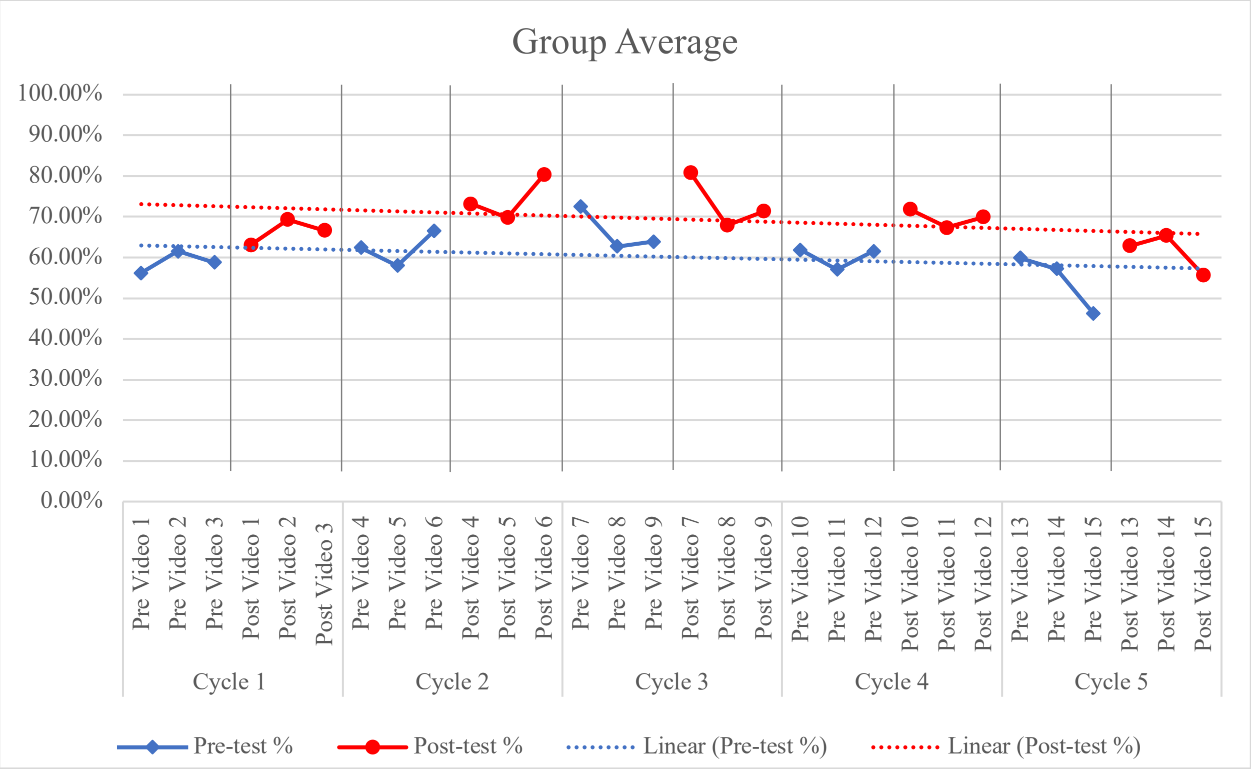


Figure 7 Line graph of EIT scores with trend lines

From the bar graph in Figure 6, we can see that all post-test scores are higher than the corresponding pre-test scores, and Figure 7 shows that the trend line for the post intervention phase is higher than the baseline trend line. The average post-test score was 69.05%, and within each cycle, participants made steady gain of 6.85% to 12.14%. These consistent gains across five weeks of watching Japanese documentary series with captions are strong evidence for improvements made in the participants' speech segmentation skills for utterances they were exposed to.

The other question to address is whether there was a reversal effect between the intervention phase and the baseline phase. Table 6 shows the mean EIT scores with intervention and reversal effect sizes and Figure 8 shows the visual analysis of the five cycles with the two phases alternating. We can see from Figure 8 that all scores from the baseline phases are constantly lower than the intervention phases, except for a small overlap we can see on Video 7 pre-test score to the post-test scores in cycle 2. The average gains participants made were relatively constant throughout all cycles, with the largest intervention effect being 12.14% in cycle 2, while the smallest intervention effect was a gain of 6.85% in cycle 5. The largest reversal effect was a decrease of 15.27% from cycle 4 to cycle 5, while the smallest was a decrease of 4.06% from cycle 1 to cycle 2. An overall general pattern of a reversal to baseline was obtained, and we see an intervention effect at all five different points in time, in which three points are needed to demonstrate an intervention effect (Kratochwill et al., 2010). Hence, we can conclude that this SCD study shows an intervention effect, which serves as evidence for a causal relation between the IV (bi-modal input) and the DV (participants' EIT scores). The foregoing analysis gives an affirmative answer to Research Questions 1: repeated exposure to bi-modal input indeed helps L2 Japanese learners improve their ability to segment utterances into words in the video that they watched.

Table 6
Mean EIT Scores with Intervention and Reversal Effect Sizes

	Pre-test Mean Score	Post-test Mean Score	Intervention Effect	Reversal Effect
Cycle 1	58.82%	66.38%	+ 7.55%	
				- 4.06%
Cycle 2	62.32%	74.46%	+ 12.14%	
				- 8.11%
Cycle 3	66.35%	73.40%	+ 7.05%	
				- 13.25%
Cycle 4	60.14%	69.72%	+ 9.58%	
				- 15.27%
Cycle 5	54.45%	61.30%	+ 6.85%	
Overall Mean	60.42%	69.05%	+ 8.63%	- 10.17%

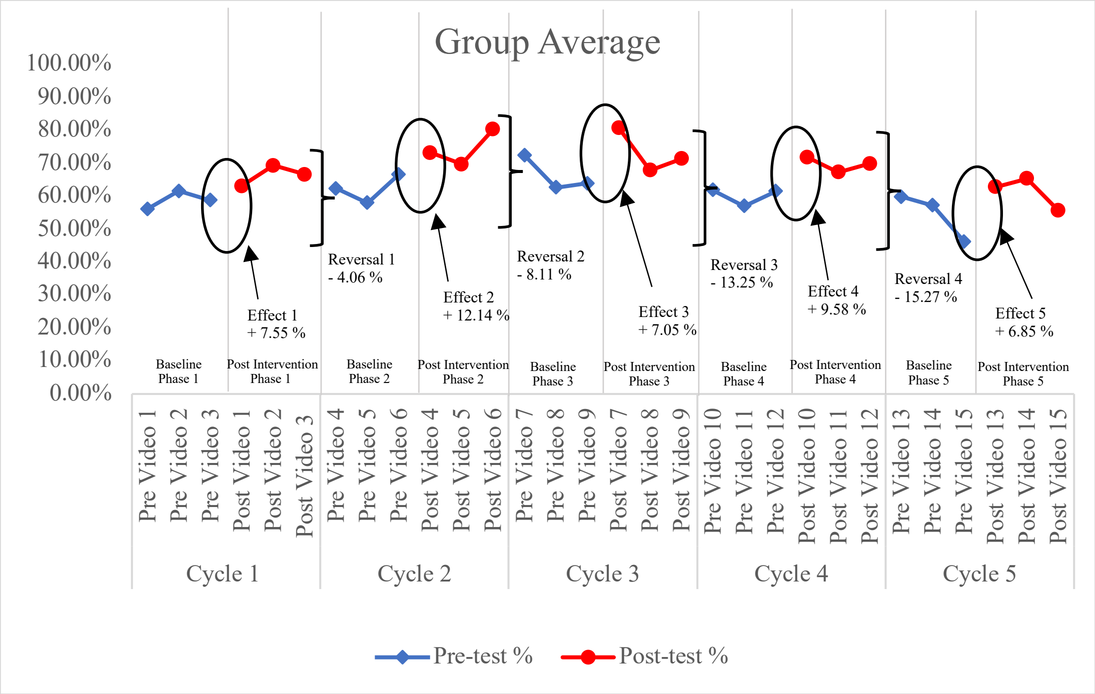


Figure 8 Reversal graph of EIT scores with intervention effects

Results Pertaining to Research Question 2

Next, results regarding Research Question 2, "Does repeated exposure to bi-modal input help L2 Japanese learners retain their speech segmentation skills?" will be analyzed. A repeated-measures ANOVA was run using *SPSS version 25* to compare the means of the pre-post-test scores for the three videos in Week 3 to the means of the delayed post-test scores in Week 11. The result of each video will be presented in order below.

Video 1

Table 7 shows the descriptive statistics for Video 1. There were 14 excerpts with a total number of 130 morphemes. There were on average 9.29 morphemes per item (sentence). The mean proportion score for the pre-test was 56.15%, post-test score was 63.11%, and the delayed post-test score was 64.94%.

Table 7
Descriptive Statistics for Video 1 Pre, Post, Delayed Post-Tests

		Pre-test		Post-test		Delayed Post-test	
		Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Mean		.561538	.0522233	.631090	.0544185	.649359	.0588745
95% CI	Lower	.446596		.511315		.519777	
Mean	Upper	.676481		.750864		.778941	
5% Trimmed Mean		.560684		.630057		.652066	
Median		.559615		.615385		.698077	
Variance		.033		.036		.042	
Std. Deviation		.1809068		.1885112		.2039473	
Minimum		.2923		.3808		.3115	
Maximum		.8462		.9000		.9385	
Range		.5538		.5192		.6269	
Interquartile Range		.3221		.3827		.3913	
Skewness		.066	.637	.087	.637	-.407	.637
Kurtosis		-1.083	1.232	-1.774	1.232	-1.187	1.232

Note. CI = confidence interval.

Before carrying out the repeated-measures ANOVA, we made sure that the following five assumptions were met:

Assumption 1: The dependent variable should be measured on a continuous scale (i.e. interval or ratio level).

Assumption 2: The independent variable should consist of two categorical, "related groups" or "matched pairs". "Related groups" indicates that the same subjects are present in both groups.

Assumption 3: There should be no significant outliers in the differences between the related groups.

Assumption 4: The distribution of the differences in the dependent variable between the two related groups should be approximately normal.

Assumption 5: The variances of the differences, known as sphericity, between all combinations of related groups must be equal.

Assumption 1 and 2 were met for all three videos analyzed for Research Question 2 because 1) the dependent variable (EIT scores) was measured on a ratio level and 2) the same 12 participants took all the pre, post, and delayed post-tests. Assumption 3 was also met because there were no outliers: i.e. scores three standard deviations or more from the mean. A test of normality was conducted to see whether Assumption 4 was met. The significance level was set to .05, and the null hypothesis was that the data is normally distributed, while the alternative hypothesis was that the data is not normally distributed. Since this dataset has only n=12, the Shapiro-Wilk test was used. The p-value for the pre, post, and the delayed post-test scores were p = .748, .176, and .299, respectively (see Table 8). We therefore conclude that the data for all three videos are normally distributed.

Table 8
Tests of Normality (Shapiro-Wilk Test) for Video 1 Pre, Post, Delayed Post-Tests

	Statistic	df	Sig.
Video 1 Pre-test	.958	12	.748
Video 1 Post-test	.904	12	.176
Video 1 Delayed	.922	12	.299

Mauchly's Test of Sphericity was used to test whether Assumption 5 was met. Mauchly's Test of Sphericity tests the null hypothesis that the variances of the differences among the pre, post, and delayed post-test scores are equal, and Table 9 shows the result. The result of this test was $x^2(2) = 2.356$, $p = .308$, showing that the assumption of sphericity is not violated. Since all five assumptions were met for Video 1, we moved on to running the repeated-measures ANOVA.

Table 9
Video 1 Mauchly's Test of Sphericity

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Epsilon Huynh-Feldt	Lower-bound
time	.790	2.356	2	.308	.827	.954	.500

Table 10 shows the result of the within-subject effects generated by repeated-measures ANOVA. We see that the mean scores for EIT are significantly different ($F(2, 22) = 14.412$, $p = .000$), indicating that there is an overall significant difference between the means at different time points. The result here, however, does not tell us where those differences occurred. That is why Bonferroni post hoc test was conducted, which allows us to find out which specific means differed.

Table 11 shows the result of the Bonferroni pairwise comparisons, with $p = .05$. We can see that there was a significant difference in the mean EIT scores between both pre-test and post-test and pre-test and delayed post-test ($p = .001$). This means that participants scored significantly better both at the immediate and delayed post-test compared to the pre-test. There was no significant difference between the post-test and delayed post-test scores (*p* = 1.000). This means that participants retained their immediate post-test scores after eight weeks. If we look at the descriptive statistics in Table 7, we can see that not only did participants retain their EIT scores from immediate post-test in Week 3 to delayed post-test in Week 10, but actually scored even better.

Table 10
Video 1 Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
time	Sphericity Assumed	.052	2	.026	14.412	.000	.567
Error (time)	Sphericity Assumed	.039	22	.002			

Table 11
Video 1 Bonferroni Pairwise Comparisons

					95% CI for Difference	
(I) time	(J) time	MD (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Pre	Post	-.070	.014	.001	-.110	-.029
	Delayed	-.088	.016	.001	-.133	-.042
Post	Pre	.070	.014	.001	.029	.110
	Delayed	-.018	.021	1.000	-.077	.040
Delayed	Pre	.088	.016	.001	.042	.133
	Post	.018	.021	1.000	-.040	.077

Note. MD = Mean Difference.

Video 2

Next, we will examine the results from Video 2. Table 12 is the descriptive statistics of the Video 2 pre, post, and delayed post-test mean scores. There were 17 excerpts totaling 133 morphemes. There were on average 7.82 morphemes per item (sentence). The mean proportion score for the pre-test was 61.56%, post-test score was 69.36%, and the delayed post-test score was 66.38%.

The same five assumptions as above were met, with the data having no outliers and being normally distributed (Table 13 pre, post, delayed post-tests tests of normality results: *p* > .05 (*p* = .182, .440, and .743, respectively)). The result of the Mauchly's Test of Sphericity was $x^2(2) = 2.821$, *p* = .244, showing that the assumption of sphericity had not been violated.

Table 12
Descriptive Statistics for Video 2 Pre, Post, Delayed Post-Tests

	Pre-test		Post-test		Delayed Post-test	
	Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Mean	.615602	.0530250	.693609	.0489433	.663847	.0508635
95% CI	.498894		.585885		.551897	
Mean	.732309		.801333		.775797	
5% Trimmed Mean	.618212		.694862		.666806	
Median	.671053		.751880		.716165	
Variance	.034		.029		.031	
Std. Deviation	.1836841		.1695447		.1761963	
Minimum	.3271		.4098		.3383	
Maximum	.8571		.9549		.9361	
Range	.5301		.5451		.5977	
Interquartile Range	.3571		.2914		.2641	
Skewness	-.381	.637	-.434	.637	-.511	.637
Kurtosis	-1.495	1.232	-.828	1.232	-.408	1.232

Table 13
Tests of Normality (Shapiro-Wilk Test) for Video 2 Pre, Post, Delayed Post-Tests

	Statistic	df	Sig.
Video 2 Pre-test	.905	12	.182
Video 2 Post-test	.935	12	.440
Video 2 Delayed	.957	12	.743

Table 14
Video 2 Mauchly's Test of Sphericity

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Epsilon Huynh-Feldt	Lower-bound
time	.754	2.821	2	.244	.803	.919	.500

Since all assumptions were met, we ran the repeated-measures ANOVA on items from Video 2. Table 15 shows the results from the tests of within-subjects effects, and we found significant differences in the three EIT scores (*F*(2, 22) = 14.273, *p* = .000). A Bonferroni post hoc test was run to see where the difference(s) lies. Table 16 shows the pairwise comparisons results. We can see that there is a significant difference between pre and post-test scores (*p* = .001). The *p*-value for scores between pre and delayed post-test was almost significant (*p* = .05). Like Video 1, there was no significant difference between the post and delayed post-test scores (*p* = .052). Although participants did not continue to make gains in the delayed post-test as in Video 1, they still scored better than the pre-test and were able to retain the improvement they made from post-test to delayed post-test even after eight weeks.

Table 15
Video 2 Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
time	Sphericity Assumed	.037	2	.019	14.273	.000	.565
Error (time)	Sphericity Assumed	.029	22	.001			

Table 16

Video 2 Bonferroni Pairwise Comparisons

					95% CI for Difference	
(I) time	(J) time	MD (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Pre	Post	-.078	.016	.001	-.122	-.034
	Delayed	-.048	.017	.050	-.096	7.060E-8
Post	Pre	.078	.016	.001	.034	.122
	Delayed	.030	.011	.052	.000	.060
Delayed	Pre	.048	.017	.050	-7.060E-8	.096
	Post	-.030	.011	.052	-.060	.000

Video 3

The same procedure as in Video 1 and 2 was taken in order to analyze the results from Video 3. Items from Video 3 consisted of 14 excerpts with a total number of 127 morphemes. There were on average 8.73 morphemes per item (sentence). Table 17 is the descriptive statistics of the Video 3 pre, post, and delayed post-test scores. The mean score for the pre-test was 58.76%, post-test was 66.67%, and 66.57% for the delayed post-test. No outliers were detected from all three test scores and the results from the test of normality are shown in Table 18: all pre-post-delayed scores were normally distributed with $p > .05$ ($p = .470$, .073, and .120, respectively). Mauchly's Test of Sphericity shows $x^2(2) = .524$, $p = .770$, confirming that sphericity is assumed in all three data (see Table 19).

Table 17
Descriptive Statistics for Video 3 Pre, Post, Delayed Post-Tests

	Pre-test		Post-test		Delayed Post-test	
	Statistic	Std. Error	Statistic	Std. Error	Statistic	Std. Error
Mean	.587598	.0489629	.666667	.0537250	.665682	.0535545
95% CI	.479832		.548419		.547810	
Mean	.695365		.784915		.783555	
5% Trimmed Mean	.589895		.671843		.668781	
Median	.624016		.740157		.732283	
Variance	.029		.035		.034	
Std. Deviation	.1696125		.1861087		.1855182	
Minimum	.3031		.3465		.3740	
Maximum	.8307		.8937		.9016	
Range	.5276		.5472		.5276	
Interquartile Range	.3219		.3465		.3396	
Skewness	-.282	.637	-.650	.637	-.587	.637
Kurtosis	-1.035	1.232	-1.203	1.232	-1.143	1.232

Table 18
Tests of Normality (Shapiro-Wilk Test) for Video 3 Pre, Post, Delayed Post Tests

	Statistic	df	Sig.
Video 3 Pre-test	.938	12	.470
Video 3 Post-test	.874	12	.073
Video 3 Delayed	.891	12	.120

Table 19
Video 3 Mauchly's Test of Sphericity

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	df	Sig.	Greenhouse-Geisser	Epsilon Huynh-Feldt	Lower-bound
time	.949	.524	2	.770	.951	1.000	.500

Table 20 shows the result of the tests of within-subjects effects. We see that *F*(2, 22) = 20.659, *p* = .000, showing that there is a significant difference between the means. The result of the Bonferroni post hoc in Table 21 shows that there is a significant difference between pre and post-test (*p* = .000) and between pre and delayed-post-test (*p* < .005). There was, however, no significant difference between the post and delayed post-test (*p* = 1.000). As we can see from the descriptive statistics in Table 17, the score from delayed post-test was only slightly lower than the immediate post-test (less than .001%).

Table 20
Video 3 Tests of Within-Subjects Effects

Source		Type III Sum of Squares	df	Mean Square	F	Sig.	Partial Eta Squared
time	Sphericity Assumed	.049	2	.025	20.659	.000	.653
Error (time)	Sphericity Assumed	.026	22	.001			

Table 21
Video 3 Bonferroni Pairwise Comparisons

					95% CI for Difference	
(I) time	(J) time	MD (I-J)	Std. Error	Sig.	Lower Bound	Upper Bound
Pre	Post	-.079	.013	.000	-.116	-.042
	Delayed	-.078	.016	.001	-.122	-.034
Post	Pre	.079	.013	.000	.042	.116
	Delayed	.001	.014	1.000	-.038	.039
Delayed	Pre	.078	.016	.001	.034	.122
	Post	-.001	.014	1.000	-.039	.038

Research Question 2 Summary

Although the degree of retention was different for the three videos, we were able to observe retention of the EIT scores from the immediate post-test participants took in Week 3 to the delayed post-test they took in Week 11 for all three videos. Figure 9 shows the estimated marginal means of all three videos in cycle 1. The vertical axis is the estimated marginal means and the horizontal axis is the time. We have assumed that the delayed post-test score will decrease somewhat after the immediate post-test score (which we can see in Video 2 notably and very slightly in Video 3 in Figure 9), because participants would have had better context information in the immediate post-test from seeing the videos before they took the test, as well as hearing the excerpts immediately before. However, participants scored higher in the delayed post-test than in the immediate post-test in Video 1. This could be because their speech segmentation skills improved from watching 15 videos with captions during the intervention phases, which led to the improvement from the immediate post-test to the delayed post-tests scores in Video 1 and the retention on Video 2 and 3 delayed post-test scores. This suggests that watching videos with captions may have a long-term effect for segmenting the spoken input accompanied by captions.

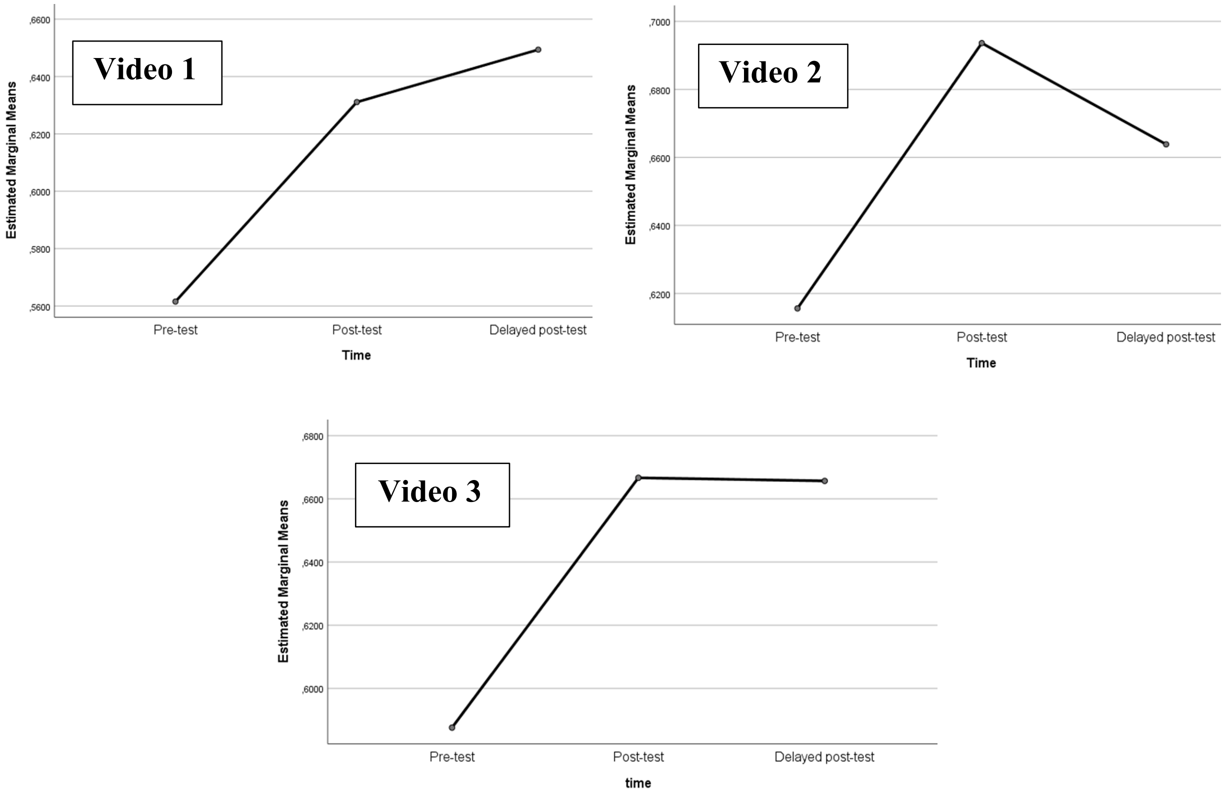


Figure 9 Estimated marginal means of Video 1, 2, and 3 (videos in Cycle 1)

Results Pertaining to Research Question 3

The previous two research questions were concerned with the "old" items, items that the participants were exposed to during the experimental sessions. The analysis of the next two research questions are based on items tested in the grand pre-post-tests participants took in Week 2 and Week 9; items participants were not exposed to in the videos. Grand pre-post-tests tested the "new" items and the "different-speaker" items. In order to answer Research Question 3, "Does their learning get generalized to utterances produced by the same speaker in the video they have not watched?" the results from the "new" items are analyzed. The new items consisted of 15 excerpts taken out from the same show as the old items ("I can see you! Japan"), but from an episode they did not watch. There were 122 morphemes in total with the average being 8.13 morphemes per item (sentence). A paired sample *t*-Test was run using *SPSS version 25* to compare the 12 participants' EIT mean scores from the grand pre-test they took in Week 2 to the grand post-test they took in Week 9. Table 22 shows each participant's pre-post-test scores. Nine out of 12 participants made more than 10% gain, and the average gain score was 12.78%. None of their scores decreased from pre-test to post-test.

Table 22
New Item Pre-Post-Test Scores for All Participants

Participant	Pre-test	Post-test	Gain
1	43.44%	61.89%	18.44%
2	45.08%	45.49%	0.41%
3	76.23%	95.90%	19.67%
4	75.41%	89.34%	13.93%
5	72.95%	83.20%	10.25%
6	30.74%	47.54%	16.80%
7	69.67%	89.34%	19.67%
8	54.51%	61.48%	6.97%
9	56.97%	74.59%	17.62%
10	69.67%	72.13%	2.46%
11	70.49%	80.74%	10.25%
12	74.18%	90.98%	16.80%
Average	61.61%	74.39%	12.78%

Table 23
Descriptive Statistics of the New Items

	Pre-test		Post-test	
	Statistic	Std. Error	Statistic	Std. Error
Mean	.6161	.04376	.7439	.04914
95% CI Mean	.5198		.6357	
	.7124		.8520	
5% Trimmed Mean	.6252		.7480	
Median	.6967		.7766	
Variance	.023		.029	
Std. Deviation	.15160		.17023	
Minimum	.31		.45	
Maximum	.76		.96	
Range	.45		.50	
Interquartile Range	.26		.28	
Skewness	-.923	.637	-.562	.637
Kurtosis	-.354	1.232	-.906	1.232

Before running the paired sample *t*-Test, it was made sure that the same assumptions as in Research Question 2 (except for Assumption 5 regarding sphericity) were met. Assumption 1 and 2 were met because 1) the dependent variable (EIT scores) are measured on a ratio level and 2) the same 12 participants took both the pre-test and post-test. Assumption 3 was also met because there were no data that lied three standard deviation from the mean. A test of normality was conducted to check whether Assumption 4 was met, and we found the scores for the post-test to be normally distributed with *p* = .286 (Table 24). For the pre-test score, however, the *p*-value was *p* = .045, slightly below the set significance level (*p* = .05). Nevertheless, we decided that Assumption 4 was met because there were no outliers in the data, and both the Skewness and Kurtosis were within the range of -1 and 1 in the descriptive statistics (see

Table *23*). This is why it was concluded that the data met all four assumptions to run a *t*-Test.

Table 24
Tests of Normality (Shapiro-Wilk Test) for New Items Pre-Post-Tests

	Statistic	df	Sig.
Pre-test	.857	12	.045
Post-test	.920	12	.286

Table 25 shows the result of the paired sample *t*-Test for new items, *t*(11)=-6.627, *p* = .000. This indicates that there is a significant difference between the mean of the pre-test and post-test for the new items at a 1% level. This result was unexpected, given the fact that the participants were not exposed to the new items at any point between the grand pre-test and post-test. Since they did not watch the corresponding video, participants did not have any context information for the excerpts either. It is also highly unlikely that they had remembered the sentences they heard in the pre-test in Week 2 until they took the post-test in Week 9. In fact, when the researcher told the participants after the post-test that they have actually taken the exact same test in Week 2, all of them said that they did not notice that.

Table 25
Paired Samples t-Test Results for New Items

				95% CI Difference				
	M	SD	Std. Error M	Lower	Upper	t	df	Sig. (2-tailed)
Pre-Post-Test	-.127732	.066770	.019275	-.170156	-.085309	-6.627	11	.000

The result suggests that participants' learning got generalized to utterances produced by the same speaker even without watching the video, resulting in the significant increase in their scores from pre-test to post-test. Participants were exposed to some of the same morphemes in the new items during the 15 video viewing sessions, such as same particles, verbs, verb stems, pronouns, and so on. In fact, 77.05% of the morphemes from the new items appeared at least once in the old items (95 out of 122 morphemes in the new items). It is likely that participants became familiar with the narrator's voice over the course of the experiment as well as how certain morphemes sound when spoken by the narrator, which helped them improve their speech segmentation and word recognition skills. It also seemed that participants themselves became aware of their improvement on their listening skills. When they first took the grand pre-test in Week 2, most participants said they found the task very difficult, especially since that was their first time taking the EIT. At the end of each week's experimental session, the researcher asked the participants how difficult they found that week's EIT. They found it more difficult in some weeks than in some others (for example, they all said the EIT from cycle 5 was difficult and they did indeed score the lowest average in that cycle), but most participants said they did not find the task difficult for the new items post-test in Week 9. When they were given feedback on all their test scores at the end of the experimental sessions, most participants were surprised to see how much their scores increased from the pre-test to post-test on the new items. They were happy to see the results and to be able to feel an improvement themselves.

Results Pertaining to Research Question 4

The items tested in the grand pre-post-tests were used again, this time, to investigate Research Question 4. Research Question 4 asks "Does participants' learning get generalized to utterances produced by a different speaker of a broadly similar accent (i.e. standard Japanese/Tokyo dialect) in the video they have not watched?" The same analysis as in Research Question 3 was carried out. "Different-speaker" items consisted of 15 excerpts taken out from a show participants did not watch ("Good to know the maps"), but the narrator was also a female who spoke standard Japanese, as did the other narrator from the old and new items. A total number of 106 morphemes were in the different-speaker items with an average of 7.08 morphemes per item (sentence). Table 26 shows the pre-post-test scores for all participants. We can see a wide variety of scores, as some participants made over 10% gain while some of their scores slightly decreased from pre-test to post-test. The average gain score was 5.07%. Table 27 shows the descriptive statistics of the different-speaker items.

Table 26
Different-Speaker Items Pre-Post-Test Scores

Participants	Pre-test	Post-test	Gain
1	56.60%	56.13%	-0.47%
2	37.26%	47.64%	10.38%
3	77.36%	82.55%	5.19%
4	78.77%	77.83%	-0.94%
5	72.17%	73.11%	0.94%
6	27.83%	42.92%	15.09%
7	54.25%	70.28%	16.04%
8	42.92%	44.81%	1.89%
9	67.92%	65.57%	-2.36%
10	61.32%	67.45%	6.13%
11	68.87%	74.53%	5.66%
12	77.83%	81.13%	3.30%
Average	60.26%	65.33%	5.07%

Table 27
Descriptive Statistics of the Different-Speaker Items

	Pre-test		Post-test	
	Statistic	Std. Error	Statistic	Std. Error
Mean	.6026	.04885	.6533	.04080
95% CI Mean	.4951		.5635	
	.7101		.7431	
5% Trimmed Mean	.6103		.6562	
Median	.6462		.6887	
Variance	.029		.020	
Std. Deviation	.16923		.14133	
Minimum	.28		.43	
Maximum	.79		.83	
Range	.51		.40	
Interquartile Range	.30		.27	
Skewness	-.725	.637	-.534	.637
Kurtosis	-.566	1.232	-1.188	1.232

The same four assumptions as in Research Question 3 were checked before running the paired sample *t*-Test. Assumption 1 and 2 were met due to the same reason as in Research Question 2 and 3 (the dependent variable (EIT scores)) are measured on a ratio level and the same 12 participants took both the pre-test and post-test). No outliers were identified (Assumption 3) and the Shapiro-Wilk test was used as in Research Question 3 to test the normality (Assumption 4). Table 28 shows the result of the test. The *p*-value for the pre-test score was $p = .235$, and $p = .176$ for the post-test, concluding that both scores were normally distributed. Since all four assumptions were met, we further analyzed the data using paired samples *t*-Test.

Table 28
Tests of Normality (Shapiro-Wilk Test) for Different-Speaker Items Pre-Post-Tests

	Statistic	df	Sig.
Pre-test	.913	12	.235
Post-test	.904	12	.176

Table 29 shows the result from the paired samples *t*-Test for different-speaker items, *t*(11)= -2.903, *p* = .014. This result shows that there is a significant difference in the EIT pre-post-test scores for the different-speaker items at a 5% level. Similarly to the case of the new items, 66.04% of the morphemes from the different-speaker items also appeared at least once in the videos they watched (70 out of 106 morphemes in the different-speaker items). The result here shows that after being exposed to captioned-videos for five weeks, participants made a significant improvement even on items spoken by a different-speaker.

Table 29
Paired Samples t-Test Results for Different-Speaker items

				95% CI Difference				
	M	SD	Std. Error M	Lower	Upper	t	df	Sig. (2-tailed)
Pre-Post-Test	-.050708	. 060511	. 017468	-.089155	-.012260	-2.903	11	. 014

Let us compare the new and different-speaker scores to observe the learners' skill generalization trend. In Table 30 we can see that the group average pre-test scores for these two kinds of items were relatively similar (62.70% for the new items and 58.87% for the different-speaker items). However, the average gain scores differed by 5.88%; participants made much greater gains on the new items compared to the different-speaker items (12.77% vs. 5.80% gain). We believe this difference is due to natural progression occurring in participants' speech segmentation skills. We assume that participants naturally progress from first being able to better comprehend the speech of the speaker they were exposed to, and then to being able to comprehend other people's speech better. If this is the case, we should eventually observe greater gains on the different-speaker items as well.

Table 30
Score Comparison for New and Different-speaker Items for Each Participant

Participants	Pre-test		Post-test		Gain	
	New	DF	New	DF	New	DF
1	43.44%	56.60%	61.89%	56.13%	18.44%	-0.47%
2	45.08%	37.26%	45.49%	47.64%	0.41%	10.38%
3	76.23%	77.36%	95.90%	82.55%	19.67%	5.19%
4	75.41%	78.77%	89.34%	77.83%	13.93%	-0.94%
5	72.95%	72.17%	83.20%	73.11%	10.25%	0.94%
6	30.74%	27.83%	47.54%	42.92%	16.80%	15.09%
7	69.67%	54.25%	89.34%	70.28%	19.67%	16.04%
8	54.51%	42.92%	61.48%	44.81%	6.97%	1.89%
9	56.97%	67.92%	74.59%	65.57%	17.62%	-2.36%
10	69.67%	61.32%	72.13%	67.45%	2.46%	6.13%
11	70.49%	68.87%	80.74%	74.53%	10.25%	5.66%
12	74.18%	77.83%	90.98%	81.13%	16.80%	3.30%
Average	62.70%	58.87%	74.39%	64.67%	12.77%	5.07%

Note. DF = Different-speaker.

Research Question 3 and 4 Summary

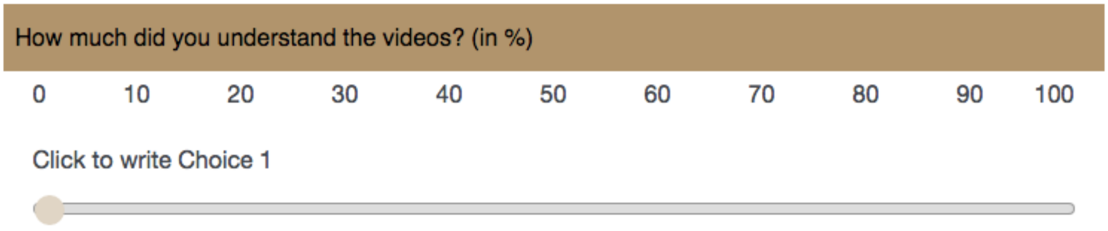
The findings from Research Question 3 and 4 (results from grand pre-post-tests) are quite exciting, suggesting that watching programs with captions may be helpful not only for segmenting the spoken input actually accompanied by captions (as seen from the results in Research Question 1 and 2), but may have a more far-reaching effect on the development of segmentation abilities in a second language. The findings in this study are in line with the results found in Charles and Trenkic (2015), where they found that L2 English learners were better able to segment speech from new non-captioned programs and new speakers after receiving bi-modal input. One of the goals of the present study was to explore the role of bi-modal input in developing lexical segmentation skills amongst learners of languages other than English. The results from this experiment are indeed suggestive that it can help L2 Japanese learners in developing this essential listening skill as well.

While participants made significant gains for both kinds of items, the greater gains on the new items attest the effect of the bi-modal input for the following two reasons; first, if it had been a testing effect (i.e. participants' EIT scores got better simply due to taking the EIT multiple times), we would have expected similar gains between the new and different-speaker items. Second, if the gain in their EIT scores were due to maturation (i.e. naturally occurring changes over time, as in taking the JPNS 302 course during the time of the experiment), we would have expected similar gains on both kinds of items as well. Neither was the case, as we can see from Table 30; 10 out of 12 participants made greater gains on the new items than the different-speaker items.

Results Pertaining to Research Question 5

Next, we will analyze the results obtained from the post-experiment survey to answer Research Question 5, "What are the learners' reactions towards the bi-modal input intervention?" Participants took the post-experiment survey at the end of the experimental sessions in Week 11 using *Qualtrics*, and all 12 participants completed the survey. Participants answered a total of nine items in the survey (including Item 1 which asked for the participant's name), which included number choosing items, yes-no items, five-point Likert scale items, and free response items. A complete list of the items can be found in Appendix B. Responses to the items will be presented below in order.

Items 2, 3, and 4 asked about participants' general reaction regarding the videos they watched during the invention phases. Item 2 asked how much they understood the videos (in percentage). Participants were provided with a scale of 0 to 100 as in Figure 10. The minimum percentage was 60% while the maximum was 100% (two participants said they understood everything), and the average was 80.5%.



How much did you understand the videos? (in %)

0 10 20 30 40 50 60 70 80 90 100

Click to write Choice 1

Figure 10 Post-experiment survey Item 2

In Item 3, participants were asked whether they enjoyed watching the videos or not during the experimental sessions. Eleven participants answered "yes," they enjoyed watching the videos, while one answered "maybe." None of the participants chose the answer "no." The next item asked participants to explain the reason why they chose that response in Item 3, and also to write anything that they did not like about the videos. Table 31 shows all responses from those who chose "yes."

Table 31
Participants' Responses to Item 3.1

- I'm a fan of Kigumiya Rie (cheese's CV).
- The videos are very interesting and informative.
- I liked learning about the different prefectures in Japan, seeing the videos and things of local Japan, and liked the listening practice. The only thing that I didn't like so much was that the videos were for kids, so the style of the video was more directed towards children and was hard to watch sometimes. I would have liked the same kind of videos but maybe with a less animated style.
- The video introduced a lot about Japan's details culture. One thing might can improve is that I prefer the resolution (both video and audio) to be higher.
- I learned new information and it helped with listening comprehension.
- I can learn about famous attractions across different cities in Japan
- Although we have learned about Japan's prefectures in class, there are still lots of places that I didn't know or wanted to know more about. So the videos were pretty fun to watch. They also let me know what to do when travelling in Japan(:
- I can learn a lot of new things about the cities in Japan and I enjoyed doing that.
- The videos were interesting and it covered content that I actually wanted to learn. The videos were only usually 15 minutes long, so it also wasn't as long as I had expected. The only thing I did not like about the videos is that the music sometimes overlapped with what the narrator was saying, so it was difficult to understand at times.
- It is interesting to watch documentary videos that introduce different cultures, food, customs and etc.. I have learned lots of things about Japan that I've never heard before.
- Those videos are very interesting.

Many of those who answered “yes” said although they learned about Japan’s prefectures in their Japanese classes, there were still many places that they did not know about so the content was interesting to them. One participant, however, wrote that she did not like the style of the video so much, as the target audience for the series were Japanese third-graders. She said she would have liked the same kind of videos but with a less animated style. Another participant wrote that the background music made it difficult for him to understand the narrators at times. It is true that there were constantly background music in all videos, and this could have been another way for the video producers to keep the young Japanese audience entertained. One participant wrote that it would be better if the resolution (both video and audio) could be improved. As explained earlier in Chapter 3, the present researcher captured each video with captions and made it into a YouTube video in order to make sure that the participants were watching the correct videos with captions, instead of simply sending them a link of the website with the list of videos. This is why the resolution became lower, and the researcher was aware of this limitation. There was only one participant who answered “maybe” in Item 3, and wrote “I didn't really like them, but they weren't bad either.” All in all, none of the participants disliked the videos.

Item 4 asked if they would like to continue watching the videos on their free time. Participants were instructed to choose one response on a “strongly agree” to “strongly disagree” five-point Likert scale. Five participants chose “strongly agree” and seven chose “somewhat agree.” All participants answered they would like to continue watching the videos to a certain degree, and they received the link to the *NHK for School* website at the end of the experimental session so that they could continue watching the videos on their own.

Item 5 and 6 asked specifically about their reaction to the captions used in the videos. Item 5 asked “How much did you pay attention to the captions in the videos?” and they had to choose one response from 1) Always, 2) Most of the time, 3) About half the times, 4) Sometimes, and 5) Never. Nine out of 12 participants said they paid attention to the captions most of the time, and three responded always. These three participants were in the top five of those who made the most gains in the new items. Item 6 asked whether they found the captions helpful in understanding the videos. Eleven participants answered “yes”, and one answered “maybe.” The next item asked them to explain the reason why they chose that response for Item 6. Table 32 shows the responses from the participants who chose “yes” in Item 6.

Table 32

Participants' Responses to Item 6.1

- It only take me half second to read and I can understand what the narrator gonna say in next, which is helpful when there is a lot information in the video.
- Reading and searching about unknown phrases in captions could help me understand the videos.
- It helped me understand the words that the characters were saying (which were hard to understand sometimes because of their different speech styles). Also, some of the kanji would give me the idea what the word meant which I could not have gotten out of just listening to the word and not reading it at the same time.
- I could see the kanji and sometimes when they are speaking quickly I couldn't understand what they said, so I could look at the captions and see what sounds were made.
- Japanese usually omit some of the work or connecting them making it hard to listen, so reading the caption somewhat showing how the word is written.
- They are very useful for me to understand what the video's talking about. Many times I heard things familiar but couldn't remember its meaning, and reading the caption helped me recall.
- I can verify what I've heard by looking at the captions and also learn about places' names, grammar, etc.
- If I had not heard something the narrator said correctly, I could catch it before it went away since the captions did not appear word for word from what was said, but instead they appeared all at the same time.
- If I didn't understand what they said, I could read what it was instead.
- Because sometimes I am not able to understand what the video is talking about, but the captions have Kanji or Katakana so I can understand the content of the video.
- Because I understand some Kanji, those captions help me understand some sentences.

Most participants said the captions helped them understand when they could not catch the spoken phrases. They said that there were times when a word sounded familiar to them but could not remember its meaning. Times like this, caption helped them recall the meaning from being able to see exactly how it is written. Some participants also wrote that being able to see the kanji gave them an idea of what the word meant, which they could not have gotten out from just listening to the word. This seemed to especially help the L1 Chinese learners, because many of the kanji used in Japanese has the same or similar meaning in Chinese but often sounds completely different. Thus, for many of the L1 Chinese learners, they were able to match the sound of the kanji to its meaning by seeing the captions. Some also said they were able to verify what they heard by looking at the captions, and captions also helped them learn words they did not know before. The one student who answered “maybe” in Item 6 said that he was able to understand most of the content of the video from just listening, so he did not find the captions helpful as the other learners did. He said he preferred watching the images in the video rather than looking at the captions.

In Item 7, we asked the participants to self-assess their EIT performance. We asked them how accurately on average they thought they were able to repeat the sentences *before* (pre-test) and *after* (post-test) viewing the video each time. We asked them to use the scale again as in Figure 10 to assess their own performance in percentage. Participants were quite critical on their pre-test scores, in which the minimum self-assessment score was 16% and the maximum was 65%. The average was 40.83%; much lower than the actual group average of 60.42% for the old items. Compared to the pre-test, all participants believed that they scored much better in the post-test, where the minimum was 45% and the maximum was 87%. The average score was 65.83%, which is relatively close to the actual group average of 69.05%.

From these results, we could see that participants generally had a positive reaction towards the bi-modal input intervention, and seemed to also be aware of their improvement. It could be the case that their positive attitude also helped them improve their listening skills. We will come back to these results again later when we discuss the individual differences in response to the intervention.

Results Pertaining to Research Question 6

In this section, we will discuss the results regarding Research Question 6, "Does the participants'1 L1 affect the effectiveness of the bi-modal input?" There were four L1 English speakers (including one participant who had three L1s: English, Urdu, and Pashto), seven L1 Chinese (Mandarin) speakers (including one participant who spoke both Mandarin and Cantonese as her L1), and one L1 Vietnamese speaker in this study. Table 33 shows the pre-test scores of all three kinds of items (old, new, and different-speaker items) from all participants grouped by their L1 and

Table *34* the post-test scores of all items. Table 35 shows participants' total J-CAT scores, which included scores from listening, vocabulary, grammar, and reading sections. The analysis to answer Research Question 6 was carried out only at the level of descriptive statistics because 1) the N in each group was too small to run a statistical analysis, and 2) the mean of pre-test scores for all three kinds of items were significantly different among groups (i.e. gain scores cannot be compared statistically if the starting points among groups are different). We can see from Table 33 that while the average scores for the L1 English and L1 Vietnamize participants were relatively similar, the L1 Chinese participants scored on average over 20% higher than the L1 English participants on all three kinds of items, and over 25% higher on the old and different-speaker items and over 13% higher on the new items compared to the L1 Vietnamese participant.

Table 33
All Pre-test Scores from All Participants

Participant	Old Items (Average)	New Items	Different-Speaker Items	Average
L1 English				
1	43.05%	43.44%	56.60%	47.70%
2	36.09%	45.08%	37.26%	39.48%
6	34.68%	30.74%	27.83%	31.08%
7	75.90%	69.67%	54.25%	66.61%
Average	47.43%	47.23%	43.99%	46.22%
L1 Chinese				
3	85.87%	76.23%	77.36%	79.82%
4	73.76%	75.41%	78.77%	75.98%
5	67.66%	72.95%	72.17%	70.93%
9	57.58%	56.97%	67.92%	60.82%
10	58.42%	69.67%	61.32%	63.14%
11	69.48%	70.49%	68.87%	69.61%
12	78.09%	74.18%	77.83%	76.70%
Average	70.12%	67.83%	68.99%	68.98%
L1 Vietnamese				
8	44.42%	54.51%	42.92%	47.28%

Table 34
All Post-test Scores from All Participants

Participant	Old Items (Average)	New Items	Different-Speaker Items	Average
L1 English				
1	50.14%	61.89%	56.13%	56.05%
2	43.52%	45.49%	47.64%	45.55%
6	44.41%	47.54%	42.92%	44.96%
7	84.39%	89.34%	70.28%	81.34%
Average	55.62%	61.07%	54.25%	56.98%
L1 Chinese				
3	92.91%	95.90%	82.55%	90.45%
4	81.89%	89.34%	77.83%	83.02%
5	75.83%	83.20%	73.11%	77.38%
9	64.41%	74.59%	65.57%	68.19%
10	70.44%	72.13%	67.45%	70.01%
11	80.29%	80.74%	74.53%	78.52%
12	84.81%	90.98%	81.13%	85.64%
Average	78.65%	79.61%	72.17%	76.81%
L1 Vietnamese				
8	55.57%	61.48%	44.81%	53.95%

Table 35
J-CAT Scores from All Participants

Participant	Listening	Vocabulary	Grammar	Reading	Total
L1 English					
1	43	26	22	39	130
2	50	34	41	33	158
6	48	45	24	46	163
7	51	34	34	45	164
L1 Chinese					
3	66	45	49	47	207
4	72	49	57	60	238
5	67	68	49	73	257
9	54	50	51	82	237
10	60	41	46	56	203
11	67	61	41	53	222
12	71	50	47	53	221
L1 Vietnamese					
8	48	34	50	31	163

Note. Each section had maximum of 100 points so the perfect total score is 400 points.

At a glance, differences in the pre-test scores among groups seemed to be correlated with the total J-CAT scores which participants took at the beginning of the experimental session in Week 1. The Pearson product-moment correlation coefficient (Pearson's correlation, for short) was run to see whether there is a correlation between the J-CAT total score and participants' pre-test scores for all three kinds of items. Before running Pearson's correlation, we made sure for all items that 1) the two variables (J-CAT and pre-test scores) were measured at the interval or ratio level, 2) there was a linear relationship between the two variables, 3) there were no significant outliers, and 4) the variables were approximately normally distributed. We found a strong, positive correlation between the J-CAT scores and the old items pre-test scores ($r = .623$. $n = 12$. $p = .030$), the new items pre-test scores ($r = .699$. $n = 12$. $p = .011$), and the different-speaker items pre-test scores ($r = .755$. $n = 12$. $p = .005$), which were all statistically significant at the .05 level. This means that those who scored higher on the J-CAT (which measured participants' general Japanese language proficiency) generally scored better in the pre-test for all items. In fact, all seven L1 Chinese participants scored higher than the L1 English and Vietnamese participants in the J-CAT, in which they also had the highest average pre-test scores for all three items among the three L1 groups. With these pre-test scores in mind, we will discuss the gain differences participants made from pre-test to post-test for all three kinds of items at the level of descriptive statistics.

Old Items

First we will compare the gain scores from the old items. Table 36 shows the average gain score for the 15 videos for each participant and Table 37 shows the descriptive statistics for the three L1 groups. While the average pre-test scores among the three groups differed significantly, their average gain scores were relatively similar, especially between the L1 English and Chinese participants (8.18% vs. 8.53%). This shows that regardless of their starting point, participants from all three L1 groups made similar gains for the items they were exposed to during the intervention phases.

Table 36

Average Gain Score for Each Participant for Old Items

	Old Items Average Gain Scores					
	L1 English		L1 Chinese		L1 Vietnamese	
	Participant	Gain %	Participant	Gain %	Participant	Gain %
	1	7.08%	3	7.04%	8	11.14%
	2	7.43%	4	8.13%		
	6	9.73%	5	8.18%		
	7	8.49%	9	6.83%		
			10	12.02%		
			11	10.81%		
			12	6.72%		
Average	8.18%		8.53%		11.14%	

Table 37
Descriptive Statistics for Old Items Average Gain Scores

L1	N	Min	Max	M	SD
English	4	.070820	.097300	.081840	.011914
Chinese	7	.067219	.120193	.085319	.020836
Vietnamese	1	.111446	.111446	.111446	.

New Items

Next we will take a look at the gain scores from the new items. Table 38 shows the gain score for each participant and Table 39 shows the descriptive statistics for all three L1 groups. Again, L1 English and Chinese participants made on average similar gains on the new items (13.83% vs. 13.00%), which was much higher compared to the gain the L1 Vietnamese participant made (6.97%). The L1 Chinese group scored on average 20% higher in the pre-test than did the L1 English group, but they made a similar gain despite their high starting point in the pre-test.

Table 38
Average Gain Score for Each Participant for New Items

	New Items Average Gain Scores					
	L1 English		L1 Chinese		L1 Vietnamese	
	Participant	Gain %	Participant	Gain %	Participant	Gain %
	1	18.44%	3	19.67%	8	6.97%
	2	0.41%	4	13.93%		
	6	16.80%	5	10.25%		
	7	19.67%	9	17.62%		
			10	2.46%		
			11	10.25%		
			12	16.80%		
Average	13.83%		13.00%		6.97%	

Table 39
Descriptive Statistics for New Items Average Gain Scores

L1	N	Min	Max	M	SD
English	4	.004098	.196721	.138319	.090249
Chinese	7	.024590	.196721	.129976	.058808
Vietnamese	1	.069672	.069672	.069672	.

Different-Speaker Items

The last item we analyze in this section are the different-speaker items. Table 40 shows the gain score for each participant and Table 41 shows the descriptive statistics for the different-speaker items. Unlike the previous two items types, we can see greater variations in the gain scores among the groups. While the L1 Vietnamese participant made only a gain of 1.89% and the average gain for the L1 Chinese group was only 2.5%, the L1 English group made an average gain of 10.26%. At the level of descriptive statistics, we were able to observe a difference among the groups' gain scores on items produced by a difference speaker in the video they had not watched. It somehow seems as the group who scored the lowest in the proficiency test made the greatest gain for these specific items.

Table 40
Average Gain Scores for Each Participant for Different-speaker Items

	Different-Speaker Items Average Gain Scores					
	L1 English		L1 Chinese		L1 Vietnamese	
	Participant	Gain %	Participant	Gain %	Participant	Gain %
	1	-0.47%	3	5.19%	8	1.89%
	2	10.38%	4	-0.94%		
	6	15.09%	5	0.94%		
	7	16.04%	9	-2.36%		
			10	6.13%		
			11	5.66%		
			12	3.30%		
Average	10.26%		2.5%		1.89%	

Table 41
Descriptive Statistics for Different-Speaker Items Average Gain Scores

L1	N	Min	Max	M	SD
English	4	-.004716	.160377	.102594	.075704
Chinese	7	-.023584	.061320	.025606	.033889
Vietnamese	1	.018867	.018867	.018867	.

Research Question 6 Summary

The discussion focus here will be the comparison between the L1 English and Chinese groups, because there was only one participant in the L1 Vietnamese group. Despite the fact that the starting points (pre-test scores) were different among groups (Table 33), they made very similar gains on average for the old and new items, particularly the L1 English and Chinese groups. This shows that bi-modal input has a positive effect on learners' speech segmentation skills on items spoken by the same speaker which they were both exposed and not exposed to, regardless of their L1. One might have supposed that the L1 English participants would make greater gains than the L1 Chinese participants because they had more room for growth (i.e. lower pre-test scores), but that was not the case. The current researcher has heard from JFL teachers teaching in L1 Chinese environment that they discourage L1 Chinese learners to put captions while watching Japanese programs (in Japanese audio), because they are afraid that Japanese captions (particularly kanji) will take away learners' attention from listening to the Japanese words. However, Charles and Trenkic's (2015) study showed that the 'no sound' group, which watched English programs with captions but without sound, performed consistently more poorly on segmenting speeches than the bi-modal group. Based on this result, they suggest that "the superior performance of the bi-modal group does indeed stem from the simultaneous presentation of sound and [captions], not from the [captions] alone" (p. 195). We saw earlier in Research Question 5 that some of the L1 Chinese participants relied on the kanji for its meaning from time to time when they could not understand what the word was from only listening. However, as seen from the increased post-test scores, watching videos with captions did not seem to take attention away from listening to the words. Although a larger comparison study should be conducted in the future dividing L1 Chinese learners into ‘bi-modal group’, ‘no sound group’, and ‘no caption group’ as in Charles and Trenkic’s study (2015), this study provided positive evidence for the effect of bi-modal input on L1 Chinese learners speech segmentation skills. This result suggests that the practice of disallowing captions may be unwarranted. Both L1 English and Vietnamese learners also made six to 13% gain on the old and new items. Just as the L1 Chinese learners, some learners from these two groups also commented that they made use of the captions including kanji; since the target audience for the videos were Japanese third-graders, most kanji used in the video should have been familiar not only to the L1 Chinese learners, but to all of the participants.

We found a difference among groups in the gain scores for the different-speaker items. Participants in the L1 English group (except for Participant 1) made much greater gains compared to participants from the L1 Chinese and Vietnamese groups. Since we did not find a difference among participants’ L1 in the old and new items, it is hard to believe that this difference in the different-speaker items has to do with their L1. We conjecture that it may be due to their proficiency level. It could be the case that the those who had lower proficiency to begin with (i.e. those who had more room for growth) became able to hear individual morphemes more accurately after receiving bi-modal input, regardless of heard or unheard speech, and speech spoken by same or different speaker they were exposed to. This is, however, only one speculation; we need to test this again with a larger sample, with a control group, and with a longer experimental period. We also need to investigate the case of “outlier” participants (i.e. Participant 1 (L1 English) and 8 (L1 Vietnamese)) who did not score so high in the pre-test for the different-speaker items but still did not make much growth in the post-test. All in all, though, we did not observe a particular difference in the effectiveness of the bi-modal input based on participants' L1. The result suggests that regardless of the learners' L1, bi-modal input is effective in helping L2 Japanese learners improve their speech segmentation skills, and the Orthographic Depth Hypothesis seems to hold true in the case of this particular study.

Results Pertaining to Research Question 7

In this final section of the results, we will analyze the results regarding Research Question 7, "What are some possible reasons why some learners did not respond to the intervention as well as others?" The data from two participants who *did* respond particularly well to the intervention with the two who *did not* will be compared. We chose two participants each who made the most and least gain from the new items, and they happen to be each from L1 English and Chinese groups (see Table 42). We decided to use the scores from the new items in choosing the four participants because 1) the results from the new items were more interesting than the results from the old items since it showed that some listeners became better able in understanding the speaker whose voice they were exposed to even in the absence of captions, and 2) the gain difference was much more clear for the case of new items compared to the different-speaker items.

Table 42
New Items Scores for the Four Participants

	L1 English			L1 Chinese		
	Pre-test	Post-test	Gain	Pre-test	Post-test	Gain
Least Gain		Participant 2			Participant 10	
	45.08%	45.49%	0.41%	69.67%	72.13%	2.46%
Most Gain		Participant 7			Participant 3	
	69.67%	89.34%	19.7%	76.23%	95.90%	19.7%

Participant 2 (least gain) and 7 (most gain) from L1 English Group

Participant 2 made overall the least gain out of the four participants from the L1 English group, while Participant 7 (who had three L1s: English, Urdu, and Pashto) the most. The backgrounds of the two were quite similar; both participants were minoring in Japanese, and they were taking the JPNS 302 course during the time of the experiment. They both started learning Japanese at the university where this study took place, and they both said they regularly watch Japanese shows in Japanese with subtitles in English. Participant 2 said he watches Japanese shows about ten hours per week, while Participant 7 said she watches about three to five hours per week. Only Participant 2 said he regularly reads Japanese other than his course assignments, and that he reads on average about three hours per week. Participant 2's total J-CAT score was slightly lower (158 points) than Participant 7's score (164 points) (see Table 35 for the detailed scores). Participant 2 said he understood the videos (old items videos) about 60% of the time, while Participant 7 answered 65%. One important thing to note is that Participant 2 told the researcher before the start of the experimental session that he has a tinnitus (a ringing sound in the ear that is extremely loud) in his left ear. However, he was highly motivated to improve his Japanese listening skills, and since he was able to hear perfectly fine with his right ear, we had him participate in the research.

Table 43 shows Participant 2's and 7's average pre-post test scores (in proportion scores) for all three kinds of items (old, new, and different-speaker items) and the gain scores. Although Participant 2 made substantial gains in the old and different-speaker items, he made only 0.41% gain for the new items, which was the least gain out of all 12 participants. Participant 7 made substantial gains for all three items, making the most gain in the new items.

Table 43
L1 English Participants (2&7) Average EIT Scores for All Items

	Participant 2			Participant 7		
	Pre-test	Post-test	Gain	Pre-test	Post-test	Gain
Old items	36.09%	43.52%	7.43%	75.90%	84.39%	8.49%
New items	45.08%	45.49%	0.41%	69.67%	89.34%	19.7%
DS items	37.26%	47.64%	10.38%	54.25%	70.28%	16.03%

While these two participants made comparable gains for the old and different-speaker items, their gain score for the new items differed by over 19%. We believe Participant 2's listening condition could have influenced this result. Participant 2 was the only one who had a hearing problem, and he was also the only one who made a comment about the background noise in the videos in the post-experimental survey, saying that the background noise made it difficult for him to hear the narrator at times. It could be the case that Participant 2 could not receive the whole benefit of the bi-modal input, because he was not able to hear the audio as well as Participant 7, who had no hearing problems. Perhaps he was able to make gains for the old items because he could rely on the context information (looking at the images from the videos) even when he could not hear the narrator's voice so well, but it seems like Participant 2's learning did not get generalized to utterance produced by the same speaker he had not watched. It is interesting to note, that Participant 2 was the only one who made the greatest gain for the different-speaker items out of all three kinds of items. We speculate that this could have been due to Participant 2's preference of the speaker. The speaker in the different-speaker item had an animated voice, where she had more distinct intonation when she talked. When an informal interview was conducted after the grand post-test, Participant 2 was the only one who said he thought the speaker from the different-speaker items was easier to understand. All other participants said they found the speaker from the new items easier to understand, naturally because they have been listening to this same speaker's voice for the past five weeks in the old items.

Participant 10 (least gain) and 3 (most gain) from L1 Chinese Group

Next, we will compare the results from the two L1 Chinese participants. Participant 10 made the least gain in the new items among the L1 Chinese group while Participant 3 the most. Here is some background information about the two participants; Participant 3 was minoring in Japanese but Participant 10 was neither majoring nor minoring in Japanese. However, neither of them were taking the JPNS 302 course during the time of the experiment. Participant 3 started learning Japanese by taking an online course from another university, and she tested into JPNS 201 (3rd semester course). She said she has been learning Japanese for a total of two years. Participant 10 took JPNS 101 at this university, but had self-studied beforehand; so in total, he answered he had been learning Japanese for four years (2.5 years of formal education at the university where this study took place). They both answered that they regularly watch Japanese shows. Participant 3 said she watches about 2.3 hours of Japanese shows per week in Japanese with subtitles in Chinese, while Participant 10 said he watches about 11 hours per week with the audio dubbed in Chinese. Only Participant 10 answered that he regularly listens to Japanese music (about 10 hours per week), and also reads Japanese materials (Japanese light novels and through Japanese computer games) about three hours per week. The two participants' total J-CAT scores were also relatively similar, with Participant 3 scoring only four points more (207 points) than Participant 10 (203 points) (see Table 35). Participant 3 said she understood the videos (old items videos) about 77% of the time, while Participant 10 answered 91%.

Table 44 shows Participant 10 and 3's average pre-post test scores for all items and their gain scores. Similarly to Participant 2 in the L1 English group, Participant 10 also made greater gains in the old and different-speaker items than in the new items, while Participant 3 made the most gain for the new items. Participant 10 and 3's gain score for the new items differed more than 17%.

Table 44
L1 Chinese Participants (3&10) Average EIT Scores for All Items

	Participant 10			Participant 3		
	Pre-test	Post-test	Gain	Pre-test	Post-test	Gain
Old items	58.42%	70.44%	12.02%	85.87%	92.91%	7.04%
New items	69.67%	72.13%	2.46%	76.23%	95.90%	19.7%
DS items	61.32%	67.45%	6.13%	77.36%	82.55%	5.19%

Similarly to L1 English Participant 2, Participant 10 may have not been taking the whole advantage of the bi-modal input during the intervention phases. Participant 10 was the only participant who chose the answer "maybe" to the post-experiment survey item that asked whether they found the captions helpful in understanding the videos. He said he preferred watching the images in the video to looking at the captions, because he understood what the narrator was saying only from listening. He was also the only participant who said he watches Japanese shows dubbed in his L1, in comparison to the rest of the participants who said they watch it in Japanese with subtitles in their L1. Although Participant 10 said subtitles do not bother him and that he is extremely comfortable in watching shows with subtitles in the pre-experiment survey, it could be that his attention was more focused on the image itself than the captions because he does not normally watch shows with subtitles. Again, this is one speculation we have; this should be tested using eye-tracking devise for further evidence.

Research Question 7 Summary

In this section, we focused on the analysis of the two participants (Participant 2 (L1 English) and 10 (L1 Chinese)) who made the least gain on the new items, because these two participants made considerably less gain compared to the rest of the participants (the only two participants who made less than 7% gain; see Table 38 for the gain scores for all participants). The two participants with the greatest improvement (Participant 7 (L1 English) and 3 (L1 Chinese)) made a gain of 19.7%, but half the participants made 16.8% or more gain. The background information of the two participants (from the same L1) were quite similar to one another, but they made substantial difference in their gained scores. After looking at their data both quantitatively (EIT scores) and qualitatively (pre-post experiment survey results), we found a possible reason that could explain why Participant 2 and 10 did not make as much gain as the rest of the participants; they may have not been able to take the full advantage of the bi-modal input. Participant 2 may have not been able to hear the audio in the videos as well as the other participants, and Participant 10 may have not been focusing as much on the captions as the others. As mentioned earlier, learners' speech segmentation skills seem to improve the most when they receive bi-modal input (both audio and visual aid), not the visual (caption) nor audio input alone (Charles & Trenkic, 2015). We believe this is one of the reasons that hindered these two participants from making greater improvement for the new items when compared to the rest of the participants, because they were the only ones who seem to have interacted with the videos differently. While participants may be able to make immediate gains on items taken out from the videos they watched, it is suggestive that the improvement in their speech segmentation skills beyond the videos they watched stems from the simultaneous presentation of sound and captions.

Further information is of course needed for a more definitive answer, and it would be worthwhile to investigate what other individual differences might have influenced this result.

CHAPTER 5. CONCLUSION

Overview

This study investigated the effectiveness of bi-modal input in fostering L2 Japanese learners' speech segmentation skills. This final chapter will summarize the research findings and will discuss the implications of this research, as well as acknowledge limitations of the study and offer suggestions for research in the future.

Summary of Findings

Research Question 1 investigated whether repeated exposure to bi-modal input helps L2 Japanese learners improve their ability to segment utterances into words in the video that they *watched*. In other words, we investigated whether a causal relation exists between the introduction of the independent variable (bi-modal input/captioned videos) and change in the dependent variable (EIT scores). The result demonstrated that the post-test scores were constantly higher than the corresponding pre-test scores. While the average group pre-test scores for all 15 videos was only 60.42%, participants were able to score on average 69.05% in the post-test, showing that their ability to recognize words spoken in the video they watched increased by almost 10%. These consistent gains across five weeks are strong evidence for a causal relation between the bi-modal input and the participants' EIT scores, and gave an affirmative answer to Research Questions 1. Perhaps simultaneously listening and reading the speech while viewing the visual images on the screen resulted in participants' improved information processing and boosted their ability to recall words and phrases when they took the post-test.

This finding, however, was anticipated because 1) the post-tests were clearly easier than the pre-tests since participants were able to rely on the contextual information from the video they viewed, and 2) since all post-tests were administered directly after watching the video, what they heard/viewed in the video should have been quite fresh in their memory when they took the post-test. This is why in Research Question 2, we investigated whether learners can segment speech in the videos they watched even if they are tested eight weeks after watching the video just as well as when they were tested immediately after watching the video. Participants took three delayed post-tests in Week 11 on the videos (Video 1, 2, 3) viewed in Week 3, and their immediate and delayed post-test scores from the same videos were compared. The results showed no significant differences between the two test scores for all three videos, meaning that participants were able to retain their improvement even after eight weeks. As Danan (2004) proposed, captions may have helped learners accurately form memory traces of the words, which in turn helped them to later more easily identify identical sounds without textual support.

Research Question 1 and 2 were concerned with the items that the participants were exposed to during the experimental sessions. Research Question 3 and 4, on the other hand, were concerned with items participants were not exposed to. Research Question 3 investigated whether participants' learning get generalized to utterances produced by the *same* speaker in the video they have *not* watched ("new" items), and Research Question 4 examined the case of utterances produced by a *different* speaker in the video they have *not* watched ("different-speaker" items). The results showed that participants' EIT scores increased significantly for both kinds of items. This finding is perhaps even more exciting than the findings from Research Question 1 and 2, because this suggests that watching programs with captions may be helpful not only for recognizing words in continuous speech actually accompanied by captions, but may have a more far-reaching effect on the development of segmentation and word recognition abilities in a second language. The differences between captions-assisted listening and authentic listening may have raised questions about the transferability of skills from a learning context to a real-life context, but since we were able to see that learners' word recognition ability improved even on items they were not exposed to, this shows potential that their speech segmentation skills is transferable to real-life context as well. There was also constantly background noise in all the items tested in the EIT, which is close to a real-life situation.

A greater gain was found from the new items than from the different-speaker items. This outcome was expected, because it is natural to assume that participants will first become better able to segment speech and recognize words spoken by the same speaker they were exposed to during the treatment. Since participants were able to improve their speech segmentation skills even on sentences they were not exposed to, this study showed that captions may have the potential to help build or qualitatively change the phonological representations of the words in the learners' mental lexicon, which in turn helps them in making long-term changes in their speech perception. Perhaps participants adjusted their phonological representation of a word spoken by the speaker in the old items, and was able to apply that knowledge in recognizing the same word spoken by the same speaker in the new items, because over 77% of the morphemes from the new items appeared at least once in the old items. It could be the case that participants may first adjust their phonological representation of a word spoken by a specific speaker, but could later apply that knowledge in recognizing the same word even when it is spoken by a different speaker. At least 66% of the words in the different-speaker items were also used in the old items, and participants also did make some gain for these items as well. In order to examine whether participants can make as much gain as the new items for the different-speaker items, a longitudinal study (longer than our 11-week study) will be needed.

Research Question 5 asked about participants' general reaction towards the bi-modal input, and the results showed that overall, they had a positive reaction to its usage and they seemed to be aware of their improvement. Many participants responded that captions helped them understand the context of the video in general, and found the kanji useful to recall the meaning of the word. This shows that Japanese captions not only help match the orthographical and phonological information of the word, but can also help recall the meaning of the word. Since captions can visually support learners to comprehend the speech they hear, it could lower learners' affective filter by relieving some of the anxiety experienced by them, and a lower affective filter is better for second language acquisition in general. Most participants said they enjoyed watching the videos, and this could also be a result of the low affective filter.

In Research Question 6, we investigated whether participants' L1 affects the effectiveness of the bi-modal input. The scores from all old, new, and different-speaker items between the L1 English and Chinese participants were compared for analysis. Our result showed comparable gains for the old and new items for these two groups, showing that bi-modal input helped learners recognize words in connected speech regardless of their L1. There was, however, quite a difference in the gained score between the groups for the different-speaker items (L1 English group made on group average 10.26% gain and the L2 Chinese group only made 2.5% gain). Since we did not find a difference among participants' L1 in the new and the old item, it is hard to believe that their L1 affected their speech segmentation skills only for the different-speaker items. This puzzling result from the different-speaker items was also found in the next part of the analysis, the results pertaining to Research Question 7.

Since SCD typically focuses on a small group of participants, it allowed the researcher to look more carefully into the individual differences in responses to the bi-modal intervention. Research Question 7 explored the reason why some learners did not respond to the captioned video viewing intervention as well as others. We found that two participants who made the least gain in the new items were the only ones who seem to have interacted with the bi-modal input differently from the rest of the group; one participant had a hearing problem on one of his ear, and the other participant was focusing more on the visual image than on the captions. If these participants were not able to take the full advantage of the bi-modal input, a similar result should have been obtained for the case of the different-speaker items, but the result actually showed otherwise; they both made greater gains in the different-speaker items. One possible reason for this result could be due to the participants' preference of the speaker, but this is not definitive. Further investigation is needed in order to find what other individual difference might have influenced this result.

Limitations and Suggestions for Further Research

Number of Participants

This study was intended to be narrow-focused, but this served both as this study's strength and weakness. By focusing on a very small group of participants, findings of this study are in no way representative of a larger population. However, since we were able to observe a causal relation between the bi-modal input and the participants' EIT scores, the next step of this research would be to replicate the basic research questions in a group-comparison design with a larger number of participants to provide a broader base of data. Future comparison study should divide the participants into three groups as in Charles and Trenkic's (2015) study: 1) bi-modal group (sound and captions), 2) no caption group (sound but not caption), and 3) no sound (with captions but without sound). By dividing the participants this way, we will be able to clearly see whether the advantage in the bi-modal condition stems from the combination of captions *and* sound, and not from the captions nor sound alone. In our study, we were only able to presume that that is the case from looking at the results from the individual differences (see Research Question 7). Also with a larger sample size, we can run a more reliable statistical analysis. For example, our result showed similarities between the L1 Chinese group and the L1 English group at the level of descriptive statistics (in Research Question 6). This result should be verified with a larger sample size for both language groups.

Length of Study

Although we increased the length of the study period compared to Charles and Trenkic's (2015) study, a more prolonged experimental period is desirable since we saw that exposure to captions may lead to long-term changes in listener's speech perception. We saw in Research Question 4 that although participants made significant gains in their EIT scores for the different-speaker items, the gain was not as large compared to the new items. It is worthwhile to investigate whether an even longer exposure to bi-modal input could lead to improved lexical segmentation when listening to speakers that the learner has never heard before (i.e. making as much gain for the different-speaker items as the new items).

Test Items

Although it was made sure that all grammar used in the excerpts were previously learned and that the vocabulary level and the morpheme length were controlled, items tested in some weeks appeared to be more difficult than in others. This is the difficultly of using authentic text for the EIT compared to most research where they carefully construct sentences for the task. For future studies, the test materials need to be better controlled, such as their grammatical difficulty and the amount of noise in the background.

Additionally, the voice of the narrator seems to matter as well. The reasons why we chose the social studies program *Shittoku Chizucho* (知っトク地図帳) “Good to know the maps” for the different-speaker item were because 1) it was also a show provided by *NHK for School,* 2) according to the standard of NHK’s program, this program was also for Japanese 3rd and 4th graders, and 3) the narrator was also a female who spoke standard Japanese. However, although the narrator for the different-speaker items spoke standard Japanese, her voice was more anime-like and was quite different from the narrator of the old and new items. If the narrator’s voice for the different-speaker items was less anime-like and less distinctive, we may have seen a greater improvement for the different-speaker items as well. In any case, better control over the test materials are needed for future studies.

Individual Differences

While we investigated the potential individual differences in Research Question 7 based on participants’ responses from the pre and post experiment survey, we need to further verify these findings with empirical evidence. The next phase in research on this topic would be to use a more direct measure to investigate learners’ use of and attention paid to captions, and one way of doing this is through eye-tracking methodology. Eye-tracking has been used to provide quantitative evidence of the user’s visual attentional processes, as it is hypothesized that eye movements and cognition are linked (Winke, Gass, & Sydorenko, 2013). By implementing eye-tracking, we can investigate learners’ use of captions and see what they attend to when watching foreign language videos with captions.

Another individual difference that should be measured empirically for this type of research is participants' working memory. Working memory provides temporary storage and manipulation of information that is necessary to process complex cognitive activities such as language processing (Baddeley, 1992). Working memory is carried out by two subsystems, which are the phonological loop and a visuo-spacial sketchpad. The phonological loop plays an important role in retention and manipulation of speech as it retains spectral information about the sounds currently being processed, while the visuo-spacial sketchpad holds nonverbal information (Baddeley, 2003). Although the underlining theoretical framework that supports caption research is the benefit of the bi-modal input that helps learners to recall information, the degree to which learners recall information can differ individually based on their working memory capacity. It could be the case that learners with limited working memory capacity are also limited in their ability to use and benefit from captions (Winke et al., 2013). For future studies, participants' working memory capacity should be measured using tests such as listening span test, developed by Daneman and Carpenter (1980).

Implications and Final Conclusion

This study offers both practical and theoretical implications. On the practical side, it provides evidence that captioned videos could be used to train learners' speech segmentation and word recognition skills, and that teachers of Japanese could consider introducing captioned video materials into their class curriculum. Learners can also regularly watch captioned videos outside of the classroom to improve their L2 listening ability. There has been a large increase in captioned videos available in the past couple of years, and thanks to the growing video streaming services, L2 learners have easier access to foreign language videos from anywhere in the world. It is also important for learners to develop phonological knowledge of a word when they are building their word knowledge, and captioning can aid this form-meaning mapping. Most importantly, if participants choose to watch an appropriate captioned video (i.e. i+1: level that is slightly beyond their linguistic ability), this task could be quite enjoyable because they can comprehend the input while they implicitly modify their phonological representation of a word to the correct form if necessary. The design of this study also allowed participants to become aware of their progress by tracking their improvement over one semester. Evidence of progress is inherently motivating for most learners, so the findings from this study were not only valuable for researchers and teachers, but also for the participants themselves.

On the theoretical side, this study provides additional positive evidence for the use of bi-modal input in enhancing L2 learners' speech segmentation skills, and to the researcher's knowledge, this is the first study that investigated the case of L2 Japanese learners. The participants' L1 in this study were typologically and prosodically very different from Japanese, and they failed to recognize around 40% of words in connected speech before the bi-modal input intervention. However, after watching 15 videos with captions for a course of five weeks, they made significant improvement in their speech segmentation and word recognition skills, and the most exciting finding here was that their learning even got generalized beyond the videos watched.

While the findings from this study should be replicated in a larger-scale study, and that future research should be expanded to include the items listed in the limitation section, the results from this study are indeed suggestive that bi-modal presentation of input could benefit L2 Japanese learners to improve speech segmentation skills. Since speech segmentation is a real problem for L2 learners, it is hoped that the results obtained from this study will serve as a springboard to further research projects that can help learners develop this essential skill.

APPENDIX A. BACKGROUND INFORMATION SURVEY ITEMS

Q1 What is your name?

Q2 How old are you?

Q3 What is your sex?

○ Male

○ Female

○ Prefer not to answer

Q4 What is your major?

Q5 What is your minor? (If any)

Q6 What is/are your first language(s)? If you have more than one, write them in the order of what you're most proficient in.

Q7 Which course are you currently enrolled in?

○ JPNS 302 sec.001 (8:30)

○ JPNS 302 sec.002 (4:30)

○ Neither

Q8 If you are currently enrolled in JPNS 302, why are you taking the course?

○ Japanese major requirement

○ Japanese minor requirement

○ Requirement for a degree program (that is not Japanese)

○ For my own interest

○ Other (please specify) ____

Q9 Which Japanese language courses have you previously taken at this University? Please select all the courses you have taken.

□ JPNS 101

□ JPNS 102

□ JPNS 201

□ JPNS 301

□ JPNS 302

Q10 How long have you been studying Japanese?

	0 1 2 3 4 5 6 7 8 9 10
In years	

Q11 Have you lived in Japan before? If yes, for how long?

○ Yes ____

○ No

Q12 Do you speak Japanese regularly at home? If yes, with whom?

○ Yes ____

○ No

Q13 Do you regularly watch Japanese TV shows/dramas/movies/anime/youtube videos at home?

○ Yes

○ No

Q14 How often do you watch Japanese TV shows/dramas/movies/anime/youtube videos every week?

	0	5	10	15	20	25	30	35	40	45	50
In hours											

Q15 How do you watch Japanese TV shows/dramas/movies/anime/youtube videos?

○ In Japanese without subtitles

○ In Japanese with subtitles in your first language

○ In Japanese with Japanese subtitles

○ Dubbed in your first language

○ Other (please specify) ____

Q16 Do you regularly listen to Japanese music?

○ Yes

○ No

Q17 How often do you listen to Japanese music every week?

	0	5	10	15	20	25	30	35	40	45	50
In hours											

Q18 Do you regularly read Japanese *other* than your course assignments?

○ Yes

○ No

Q19 How often do you read Japanese *other* than your course assignments every week?

	0	3	6	9	12	15	18	21	24	27	30
In hours											

Q20 What kind of material(s) do you read in Japanese? (e.g. Japanese novels, newspaper articles)

Q21 How often do you watch TV shows/movies IN ANY LANGUAGE with subtitles?

○ Always

○ Most of the time

○ About half the time

○ Sometimes

○ Never

Q22 How comfortable are you with (or how used are you in) watching TV shows/movies with subtitles?

○ Extremely comfortable (subtitles don't bother me at all)

○ Somewhat comfortable (subtitles don't bother me too much)

○ Somewhat uncomfortable (subtitles bother me a little bit)

○ Extremely uncomfortable (subtitles bother me a lot)

Q23 Of the four skills, please rank which skill you think you are strongest to weakest in Japanese.

_______ Reading
_______ Writing
_______ Speaking
_______ Listening

Q24 You will perform a listen-and-repeat exercise for this experiment. Please estimate on a scale of 0% to 100% how accurately you think you can repeat the sentences you hear.

	0 10 20 30 40 50 60 70 80 90 100
% accuracy	

APPENDIX B. POST-EXPERIMENT SURVEY ITEMS

Q1 Name

Q2 How much did you understand the videos? (in %)

0 10 20 30 40 50 60 70 80 90 100

Q3 I enjoyed watching the Prefecture videos.

○ Yes

○ Maybe

○ No

Q3.1 Please write the reason why you chose the answer 'yes.' If you had anything that you didn't like about, please write that as well.

Q3.2 Please write the reason why you chose the answer 'maybe.'

Q3.3 Please write the reason why you chose 'no.' If you had anything that you did like about, please write that as well.

Q4 I would like to continue watching the videos on my free time. (If you choose neither agree nor disagree, please write your reason why you chose this answer.)

○ Strongly agree

○ Somewhat agree

○ Neither agree nor disagree ____

○ Somewhat disagree

○ Strongly disagree

Q5 How much did you pay attention to the captions in the video?

○ Always

○ Most of the time

○ About half the time

○ Sometimes

○ Never

Q6 I found the captions helpful in understanding the videos.

○ Yes

○ Maybe

○ No

Q6.1 Please explain how the captions were helpful in understanding the videos.

Q6.2 Please explain why you chose the answer 'maybe.'

Q6.3 Please explain why you didn't find the captions helpful in understanding the videos.

Q7 You preformed a listen-and-repeat test before and after you watched each video. On average, how accurately do you think you repeated the sentences both before and after viewing each video?

	0 10 20 30 40 50 60 70 80 90 100
Before video (pretest) in %	
After video (posttest) in %	

APPENDIX C. OLD ITEMS INDIVIDUAL SCORES

	CYCLE 1					
Participant	Video 1 (Yamagata) PRE-TEST	Video 1 (Yamagata) POST-TEST	Video 2 (Aichi) PRE-TEST	Video 2 (Aichi) POST-TEST	Video 3 (Kagawa) PRE-TEST	Video 3 (Kagawa) POST-TEST
1	40.77%	45.77%	40.23%	50.75%	38.58%	47.64%
2	29.23%	38.08%	32.71%	40.98%	40.94%	41.73%
3	84.62%	90.00%	85.71%	95.49%	83.07%	89.37%
4	80.38%	83.46%	78.95%	82.33%	76.38%	82.28%
5	54.23%	57.31%	75.19%	75.19%	64.17%	76.38%
6	34.23%	42.69%	38.35%	45.11%	30.31%	34.65%
7	69.23%	84.23%	73.31%	75.19%	65.75%	81.10%
8	38.08%	43.85%	45.11%	59.40%	43.31%	47.24%
9	53.85%	51.15%	57.14%	66.17%	61.42%	69.29%
10	57.69%	65.77%	72.56%	77.44%	59.45%	72.83%
11	60.00%	75.00%	61.65%	81.20%	63.39%	75.20%
12	71.54%	80.00%	77.82%	83.08%	78.35%	82.28%
AVERAGE	**56.15%**	**63.11%**	**61.56%**	**69.36%**	**58.76%**	**66.67%**
	CYCLE 2					
Participant	**Video 4 (Wakayama) PRE-TEST**	**Video 4 (Wakayama) POST-TEST**	**Video 5 (Tottori) PRE-TEST**	**Video 5 (Tottori) POST-TEST**	**Video 6 (Niigata) PRE-TEST**	**Video 6 (Niigata) POST-TEST**
1	45.24%	59.13%	41.11%	46.30%	46.95%	59.54%
2	38.10%	50.00%	31.85%	38.52%	42.75%	62.98%
3	88.89%	96.03%	87.04%	94.81%	90.84%	99.24%
4	74.60%	82.94%	67.41%	83.70%	90.84%	90.08%
5	70.24%	79.76%	67.41%	75.19%	66.79%	84.35%
6	37.70%	45.63%	24.44%	36.30%	36.64%	60.31%
7	82.94%	92.46%	72.22%	85.19%	82.06%	90.84%
8	41.27%	53.97%	37.78%	61.11%	51.91%	74.43%
9	56.35%	68.65%	50.74%	64.81%	65.27%	75.19%
10	61.11%	73.02%	68.15%	77.41%	66.03%	82.82%
11	69.44%	83.73%	72.22%	84.07%	76.72%	90.46%
12	82.54%	93.25%	75.56%	90.00%	82.44%	94.27%
AVERAGE	**62.37%**	**73.21%**	**57.99%**	**69.78%**	**66.60%**	**80.38%**

	CYCLE 3					
Participant	Video 7 (Aomori) PRE-TEST	Video 7 (Aomori) POST-TEST	Video 8 (Shiga) PRE-TEST	Video 8 (Shiga) POST-TEST	Video 9 (Ooita) PRE-TEST	Video 9 (Ooita) POST-TEST
1	66.27%	69.44%	51.48%	44.44%	44.27%	53.05%
2	55.56%	48.81%	30.74%	44.07%	44.27%	45.80%
3	98.02%	98.41%	88.52%	91.11%	89.31%	95.42%
4	88.10%	90.87%	75.56%	78.52%	79.39%	88.55%
5	77.78%	88.10%	70.74%	77.04%	66.79%	77.10%
6	38.10%	63.10%	35.19%	38.15%	36.64%	45.42%
7	81.75%	94.44%	84.81%	84.44%	84.35%	88.55%
8	55.56%	77.78%	47.04%	57.04%	45.04%	53.44%
9	70.63%	70.63%	63.33%	71.48%	58.40%	66.79%
10	65.87%	78.57%	62.96%	72.96%	54.96%	71.37%
11	86.51%	94.84%	69.63%	75.93%	83.59%	87.02%
12	85.32%	95.24%	72.22%	80.00%	79.77%	84.35%
AVERAGE	**72.45%**	**80.85%**	**62.69%**	**67.93%**	**63.90%**	**71.41%**

	CYCLE 4					
Participant	Video 10 (Ehime) PRE-TEST	Video 10 (Ehime) POST-TEST	Video 11 (Nagasaki) PRE-TEST	Video 11 (Nagasaki) POST-TEST	Video 12 (Nara) PRE-TEST	Video 12 (Nara) POST-TEST
1	37.01%	41.34%	40.80%	45.20%	37.80%	47.97%
2	28.74%	42.91%	36.00%	37.60%	38.62%	47.97%
3	90.94%	95.67%	80.00%	92.40%	92.28%	97.15%
4	70.08%	83.46%	65.20%	79.20%	66.26%	74.39%
5	73.23%	85.43%	68.00%	74.80%	74.39%	74.80%
6	39.76%	50.79%	31.60%	42.80%	34.15%	43.09%
7	81.89%	91.73%	72.80%	85.60%	84.55%	87.40%
8	58.27%	62.20%	35.60%	53.20%	43.50%	53.66%
9	59.45%	68.50%	54.80%	60.80%	55.69%	70.73%
10	54.33%	70.87%	54.40%	72.80%	56.50%	71.95%
11	63.78%	81.10%	71.20%	77.60%	71.54%	86.59%
12	85.04%	88.19%	74.00%	86.40%	82.93%	83.74%
AVERAGE	**61.88%**	**71.85%**	**57.03%**	**67.37%**	**61.52%**	**69.95%**

	CYCLE 5					
Participant	**Video 13 (Kanagawa) PRE-TEST**	**Video 13 (Kanagawa) POST-TEST**	**Video 14 (Tokushima) PRE-TEST**	**Video 14 (Tokushima) POST-TEST**	**Video 15 (Yamaguchi) PRE-TEST**	**Video 15 (Yamaguchi) POST-TEST**
1	44.05%	48.02%	41.48%	51.85%	29.77%	41.60%
2	32.94%	36.11%	34.07%	43.70%	24.81%	33.59%
3	84.13%	90.08%	73.70%	83.70%	70.99%	84.73%
4	69.84%	79.37%	68.89%	77.04%	54.58%	72.14%
5	70.63%	74.21%	62.96%	72.22%	52.29%	65.65%
6	31.35%	33.73%	37.41%	48.15%	34.35%	36.26%
7	78.97%	77.78%	67.78%	76.67%	56.11%	70.23%
8	39.68%	44.84%	54.07%	57.41%	30.15%	33.97%
9	52.38%	57.94%	60.37%	55.56%	43.89%	48.47%
10	55.56%	56.75%	47.78%	69.26%	38.93%	42.75%
11	73.81%	72.22%	60.37%	70.37%	58.40%	69.08%
12	84.92%	82.94%	77.78%	78.52%	61.07%	69.85%
AVERAGE	**59.85%**	**62.83%**	**57.22%**	**65.37%**	**46.28%**	**55.69%**

REFERENCES

Akamatsu, N. (1999). The effects of first language orthographic features on word recognition processing in English as a second language. *Reading and Writing: An Interdisciplinary Journal*, *11*, 381–403.

Anderson, J. R. (1983). *The architecture of cognition*. Cambridge, MA: Harvard University Press.

Anderson, J. R. (2015). *Cognitive psychology and its implications* (8th ed.). New York, NY: Worth Publishers.

Baddeley, A. (1992). Working Memory. *Science, 255*(5044), 556–559.

Baddeley, A. (2003). Working memory and language: an overview. *Journal of Communication Disorders*, *36*(3), 189–208. doi:10.1016/S0021-9924(3)00019-4

Bean, R. M., & Wilson, R. M. (1989). Using closed captioned television to teach reading to adults. *Reading Research and Instruction*, *28*(4), 27–37.

Bird, S. A., & Williams, J. N. (2002). The effect of bimodal input on implicit and explicit memory: An investigation into the benefits of within-language subtitling. *Applied Psycholinguistics*, *23*(4), 509–533. doi:10.1017/S0142716402004022

Bonk, W. J. (2000). Second language lexical knowledge and listening comprehension. *International Journal of Listening*, *14*(1), 14–31.

Borrás, I., & Lafayette, R. C. (1994). Effects of multimedia courseware subtitling on the speaking performance of college students of French. *The Modern Language Journal*, *78*(1), 61–75. doi:10.2307/329253

Brown, G. (1977). *Listening to spoken English*. London, England: Longman.

Buck, G. (2001). *Assessing listening*. Cambridge, England: Cambridge University Press.

Chambers, G. N. (1996). Listening. Why? How? *The Language Learning Journal*, *14*(1), 23–27. doi:10.1080/09571739685200341

Charles, T., & Trenkic, D. (2015). Speech segmentation in a second language: The role of Bi-modal input. In Y. Gambier, A. Caimi & C Mariotti (Eds.), *Subtitles and language learning* (pp. 173–197). Bern, Switzerland: Peter Lang.

Chikamatsu, N. (1996). The effects of L1 orthography on L2 word recognition: A study of American and Chinese learners of Japanese. *Studies in Second Language Acquisition*, *18*(4), 403–422. doi:10.1017/S0272263100015369

Chikamatsu, N. (2006). Developmental word recognition: A study of L1 English readers of L2 Japanese. *The Modern Language Journal*, *90*(1), 67–85. doi:10.1111/j.1540-4781.2006.00385.x

Ching-Shyang Chang, A. (2007). The impact of vocabulary preparation on L2 listening comprehension, confidence and strategy use. *System*, *35*(4), 534–550. doi:10.1016/j.system.2007.06.003

Chung, J. (1999). The effects of using video texts supported with advance organizers and captions on Chinese college students' listening comprehension: An empirical study. *Foreign Language Annals*, *32*(3), 295–308. doi:10.1111/j.1944-9720.1999.tb01342.x

Cowan, N. (1993). Activation, attention, and short-term memory. *Memory & Cognition, 21*, 162-167.

Cutler, A. (2000). Listening to a second language through the ears of a first. *Interpreting: International Journal of Research and Practice in Interpreting*, *5*(1), 1–23. doi:10.1075/intp.5.1.02cut

Cutler, A., & Butterfield, S. (1992). Rhythmic cues to speech segmentation: Evidence from juncture misperception. *Journal of Memory and Language*, *31*(2), 218–236.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1989). Limits on bilingualism. *Nature*, *340*, 229–230.

Cutler, A., Mehler, J., Norris, D., & Segui, J. (1992). The monolingual nature of speech segmentation by bilinguals. *Cognitive Psychology*, *24*(3), 381–410.

Cutler, A., & Norris, D. (1988). The role of strong syllables in segmentation for lexical access. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(1), 113–121.

Danan, M. (2004). Captioning and subtitling: Undervalued language learning strategies. *Meta: Journal Des Traducteurs*, *49*(1), 67–77. doi:10.7202/009021ar

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *Journal of Verbal Learning and Verbal Behavior*, *19*(4), 450–466. doi:10.1016/S0022-5371(80)90312-6

Ellcessor, E. (2012). Captions on, off, on TV, online: Accessibility and search engine optimization in online closed captioning. *Television & New Media*, *13*(4), 329–352. doi:10.1177/1527476411425251

Ellis, K. (2015). Netflix closed captions offer an accessible model for the streaming video industry, but what about audio description? *Communication, Politics & Culture*, *47*(3), 3–20.

Ellis, R. (2005). Principles of instructed language learning. *System*, *33*(2), 209–224. doi:10.1016/j.system.2004.12.006

Field, J. (2003). Promoting perception: Lexical segmentation in L2 listening. *ELT Journal*, *57*(4), 325–334.

Field, J. (2008). *Listening in the language classroom*. Cambridge, England: Cambridge University Press.

Fukada, A. (2013). An online oral practice/assessment platform: Speak Everywhere. *The IALLT Journal*, *43*(1), 64–77.

Fukada, A. (n.d.). Chakoshi: A Japanese text search and collocation extraction application. Retrieved from http://telldev.cla.purdue.edu/chakoshi/public.html

Goh, C. C. M. (2000). A cognitive perspective on language learners' listening comprehension problems. *System*, *28*(1), 55–75.

Graham, S., & Santos, D. (2015). *Strategies for second language listening: Current scenarios and improved pedagogy*. London, England: Palgrave Macmillan Limited.

Greaves, A. E., Camic, P. M., Maltby, M., Richardson, K., & Mylläri, L. (2012). A multiple single case design study of group therapeutic puppetry with people with severe mental illness. *The Arts in Psychotherapy*, *39*(4), 251–261. doi:10.1016/j.aip.2012.03.002

Guichon, N., & McLornan, S. (2008). The effects of multimodality on L2 learners: Implications for CALL resource design. *System*, *36*(1), 85–93. doi:10.1016/j.system.2007.11.005

Hayati, A. M., & Mohmedi, F. (2010). The effect of films with and without subtitles on listening comprehension of EFL intermediate students. *International Journal of Instructional Media*, *37*(3), 301–313.

Huang, H.-C., & Eskey, D. E. (1999). The effects of closed-captioned television on the listening comprehension of intermediate English as a Second Language (ESL) students. *Journal of Educational Technology Systems*, *28*(1), 75–96. doi:10.2190/RG06-LYWB-216Y-R27G

Jarvey, N. (2019, January). Netflix grows subscriber base to 139 Million worldwide. Retrieved from https://www.hollywoodreporter.com/news/netflix-grows-subscriber-base-139-million-worldwide-1176934

J-CAT Japanese Computerized Adaptive Test. (n.d.) Retrieved from http://www.j-cat.org/

Kawamura, Y., Kitamura, T., & Hobara, R. (n.d.). Reading Tutor. Retrieved from http://language.tiu.ac.jp/index.html

Kennedy, C. H. (2005). *Single-case designs for educational research*. London, England: Pearson.

Koda, K. (1987). Cognitive strategy transfer in second language reading. In J. Devine, P. L. Carrell, & D. E. Eskey (Eds.), *Research in reading in English as a second language* (pp. 127–144).

Koda, K. (1989). Cognitive process in second language reading: Transfer of L1 reading skills and strategies. *Second Language Research*, *4*(2), 133–56. doi:10.1177/026765838800400203

Koda, K. (1990). The use of L1 reading strategies in L2 reading: Effects of L1 orthographic structures on L2 phonological recoding strategies. *Studies in Second Language Acquisition*, *12*(4), 393–410. doi:10.1017/S0272263100009499

Krashen, S. D. (1985). *The input hypothesis: issues and implications*. New York, NY: Longman.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskoph, D. M., & Shadish, W. R. (2010). Single-case designs technical documentation. Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf.

Lee, L. (1970). A screening test for syntax development. *Journal of Speech and Hearing Disorders*, *35*(2), 102–112.

Levin, J. R., & Kratochwill, T. R. (2003). Educational/psychological intervention research circa 2012. In I. B. Weiner, W. M. Reynolds, & G. E. Miller (Eds.), *Handbook of Psychology, volume 7, Educational Psychology, 2nd Edition* (pp. 465–492). New York, NY: Wiley.

Lust, B., Chien, Y., & Flynn, S. (1987). What children know: methods for the study of first language acquisition. In B. Lust (Ed.), *Studies in the acquisition of anaphora* (pp. 271–356). Dordrecht, Netherlands: D. Reidel Publishing Company.

Magnuson, J. S. (2016). Mapping spoken words to meaning. In G. Gaskell & J. Mirkovic (Eds.), *Speech perception and spoken word recognition* (pp. 76–96). New York, NY: Routledge.

Markham, P. (1989). The effects of captioned television videotapes on the listening comprehension of beginning, intermediate, and advanced ESL students. *Educational Technology*, *29*(10), 38–41.

Markham, P. (1999). Captioned videotapes and second-language listening word recognition. *Foreign Language Annals*, *32*(3), 321–328.

Mattys, S. L., & Bortfeld, H. (2016). Speech Segmentation. In G. Gaskell & J. Mirkovic (Eds.), *Speech perception and spoken word recognition* (pp. 55–75). New York, NY: Routledge.

Mayer, R. E. (2014). *Introduction to multimedia learning*. Cambridge, England: Cambridge University Press.

McQueen, J. M. (1998). Segmentation of continuous speech using phonotactics. *Journal of Memory and Language*, *39*(1), 21–46. doi:10.1006/jmla.1998.2568

Miller, J. F., & Chapman, R. S. (1975). Length variables in sentence imitation. *Language and Speech*, *18*(1), 35–41. doi:10.1177/002383097501800104

Milton, J., & Hopkins, N. (2006). Comparing phonological and orthographic vocabulary size: Do vocabulary tests underestimate the knowledge of some learners. *The Canadian Modern Language Review / La Revue Canadienne Des Langues Vivantes*, *63*(1), 127–147. doi:10.1353/cml.2006.0048

Ministry of Internal Affairs and Communications. (2018, September). 総務省｜平成 29 年度の字幕放送等の実績. [Ministry of Internal Affairs and Communications｜The report of captioned broadcasting in 2017]. Retrieved from http://www.soumu.go.jp/menu_news/s-news/01ryutsu09_02000217.html

Montero Perez, M., Peters, E., & Desmet, P. (2015). Enhancing vocabulary learning through captioned video: An eye-tracking study. *The Modern Language Journal*, *99*(2), 308–328. doi:10.1111/modl.12215

Montero Perez, M., Van Den Noortgate, W., & Desmet, P. (2013). Captioned video for L2 listening and vocabulary learning: A meta-analysis. *System*, *41*(3), 720–739. doi:10.1016/j.system.2013.07.013

Mori, Y. (1998). Effects of first language and phonological accessibility on kanji recognition. *The Modern Language Journal*, *82*(1), 69–82. doi:10.1111/j.1540-4781.1998.tb02595.x

Munnich, E., Flynn, S., & Martohardjono, G. (1994). Elicited imitation and grammaticality judgment tasks; what they measure and how they relate to each other. In E. E. Tarone, S. M. Gass, & A. D. Cohen (Eds.), *Research methodology in second-language acquisition* (pp. 227–243). New Jersey Hove, England: Lawrence Erlbaum.

Nespor, M., Shukla, M., & Mehler, J. (2011). Stress-timed vs. Syllable-timed Languages. In M. van Oostendorp, C. J. Ewen, E. V. Hume, & K. Rice (Eds.), *The Blackwell Companion to Phonology* (pp. 1147–1159). Oxford, England: Wiley-Blackwell.

Netflix and the National Association of the Deaf reach historic agreement to provide 100% closed captions in on-demand streaming content within two years. (2012, October). *States News Service*. Retrieved from http://link.galegroup.com/apps/doc/A304948701/BIC?u=purdue_main&sid=BIC&xid=9f3530e9

Netflix eyes Europe and Asia for local content. (2019, January). *Broadband TV News Correspondent.* Retrieved from https://www.broadbandtvnews.com/2019/01/08/netflix-eyes-europe-and-asia-for-local-content/

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. doi:10.2307/747835

NHK for School. (n.d.) Retrieved from http://www.nhk.or.jp/school/

Otake, T., Hatano, G., Cutler, A., & Mehler, J. (1993). Mora or Syllable? Speech segmentation in Japanese. *Journal of Memory and Language*, *32*(2), 258–278.

Paivio, A. (1986). *Mental representations: a dual coding approach*. New York, NY: Oxford University Press.

Port, R. F., Dalby, J., & O'Dell, M. (1987). Evidence for mora timing in Japanese. *The Journal of the Acoustical Society of America*, *81*(5), 1574–1585.

Price, K. (1983). Closed-captioned TV: An untapped resource. *Matsol Newsletter*, *12*(2), 1–8.

Robinson, P. (1995). Attention, Memory, and the "Noticing" Hypothesis. *Language Learning, 45*(2), 283–331.

Robinson, P. (2003) Attention and memory during SLA. In C.J. Doughty and M. Long (Eds.) *The Handbook of Second Language Acquisition* (pp. 631-678). Malden, MA: Blackwell Publishing.

Rost, M. (2011). *Teaching and researching listening (2nd ed.).* Harlow, England: Pearson Education.

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

Schmidt, R. (1995). Consciousness and foreign language learning: A tutorial on the role of attention and awareness in learning. In R. Schmidt (Ed.), *Attention and Awareness in Foreign Language Learning* (pp. 1-63). Honolulu: University of Hawai'i Press.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and second language instruction* (pp.3-32). Cambridge, England: Cambridge University Press.

Schmidt, R., & Frota, S. N. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. R. Day (Ed.), *Talking to learn: Conversation in second language acquisition* (pp. 237-326). Rowley, MA: Newbury House.

Segui, J. (1984). The syllable: A basic perceptual unit in speech processing?. In H. Bouma & D. G. Bouwhuis (Eds), *Attention and Performance X: Control of language processes* (pp. 165-181). Hillsdale, NJ: Erlbaum.

Seidenberg, M. S., & Tanenhaus, M. K. (1979). Orthographic effects on rhyme monitoring. *Journal of Experimental Psychology: Human Learning and Memory*, *5*(6), 546–554. doi:10.1037/0278-7393.5.6.546

Suomi, K., Mcqueen, J. M., & Cutler, A. (1997). Vowel harmony and speech segmentation in Finnish. *Journal of Memory and Language*, *36*(3), 422–444. doi:10.1006/jmla.1996.2495

Van Zeeland, H. (2014). Lexical inferencing in first and second language listening. *The Modern Language Journal*, *98*(4), 1006–1021. doi:10.1111/modl.12152

Van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*(4), 457–479. doi:10.1093/applin/ams074

Vandergrift, L. (2004). Listening to learn or learning to listen? *Annual Review of Applied Linguistics, 24*, 3–25. doi:10.1017/S0267190504000017

Vanderplank, R. (1988). The value of teletext sub-titles in language learning. *ELT Journal*, *42*(4), 272–281.

Vanderplank, R. (1990). Paying attention to the words: Practical and theoretical problems in watching television programmes with uni-lingual (CEEFAX) subtitles. *System*, *18*(2), 221–234. doi:10.1016/0346-251X(90)90056-B

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching*, *43*(1), 1–37.

Vanderplank, R. (2014). Thirty years of research into captions/same language subtitles and second/foreign language learning: Distinguishing between 'effects of ' subtitles and 'effects with' subtitles for future research. In Y. Gambier, A. Caimi, & C. Mariotti (Eds.), *Subtitles and language learning* (pp. 19–40). Bern, Switzerland: Peter Lang.

Vinther, T. (2002). Elicited imitation:a brief overview. *International Journal of Applied Linguistics*, *12*(1), 54–73. doi:10.1111/1473-4192.00024

Wang, M., & Geva, E. (2003). Spelling performance of Chinese children using English as a second language: Lexical and visual–orthographic processes. *Applied Psycholinguistics*, *24*(1), 1–25. doi:10.1017/S0142716403000018

Warner, N. (2005). Examples of reduced speech. Retrieved from http://www.u.arizona.edu/~nwarner/reduction_examples.html

Weber, A. (2000). The role of phonotactics in the segmentation of native and non-native continuous speech. In A. Cutler, J. M. McQueen, & R. Zondervan (Eds.), *Proceedings of SWAP, Workshop on spoken word access processes*. Nijmegen, Netherlands: MPI for Psycholinguistics.

Westcott, K., Downs, K., Loucks, J., & Watson, J. (2018, March). A survey of digital media trends. Retrieved from https://www2.deloitte.com/insights/us/en/industry/technology/digital-media-trends-consumption-habits-survey.html?id=us:2el:3pr:4di4479:5awa:6di:032018:&pkid=1005131

Winke, P., Gass, S., & Sydorenko, T. (2010). The effects of captioning videos used for foreign language listening activities. *Language Learning & Technology*, *14*(1), 65–86.

Winke, P., Gass, S., & Sydorenko, T. (2013). Factors influencing the use of captions by foreign language learners: An eye-tracking study. *The Modern Language Journal*, *97*(1), 254–275. doi:10.1111/j.1540-4781.2013.01432.x

Wisniewska, N., & Mora, J. C. (2018, September). *Audio-text synchronization in L2 captioned video*. Paper presented at the meeting of EuroSLA 28, Münster, Germany.