UNIVERSITAT DE BARCELONA

Audio-synchronized textual enhancement in L2 pronunciation teaching and learning with TV series

Valeria Galimberti

UNIVERSITAT DE BARCELONA

Audio-synchronized textual enhancement in L2 pronunciation teaching and learning with TV series

TESI DOCTORAL

Valeria Galimberti

Programa de doctorat en ciència cognitiva i llenguatge

Línia de recerca: Lingüística teòrica i aplicada

Directors: Joan Carles Mora Bonilla

Roger Gilabert Guerrero

Tutor: Joan Carles Mora Bonilla

2023

Valeria Galimberti

ABSTRACT

In the schools of many countries, English is taught as a foreign language with the goal of preparing students for the demands of higher education and for a globalized workforce where English is the lingua franca. However, oral communication skills and pronunciation have long been neglected in many teaching contexts (Derwing et al., 2012; Isaacs, 2018), including the Spanish school system. As a result of an emphasis on vocabulary and grammar in foreign language teaching and assessment, Spanish students' oral proficiency remains very limited even after years of formal instruction (Aliaga-Garcia, 2017). Due to the limited exposure to native or near-native speech in the classroom, language learners at relatively advanced proficiency levels may still mispronounce words that are relatively common in everyday speech (Cook et al., 2016; Van Zeeland, 2017). This can be problematic, as storing accurate phonological representations of words and accessing them in a timely manner is key to producing comprehensible speech (Kormos, 2006). Although sounding like a native speaker is an elusive goal, pronunciation instruction can prevent the fossilization of inaccurate phonological representations, increasing the learner's comprehensibility or ease of understanding (Levis, 2018; Thomson & Derwing, 2014).

Watching L2 captioned videos can provide additional exposure to targetlike L2 speech, with positive effects on vocabulary acquisition, speech segmentation and adaptation to diverse L2 accents (Charles and Trenkic, 2015; Montero-Perez et al., 2014; Mitterer & McQueen, 2009). However, the potential of exposure to captioned video for pronunciation learning is still largely unexplored (Montero Perez, 2022). The two main research gaps regard the effects of audiovisual synchrony (the time lag between learners' processing of captions and spoken dialogue) and of an intervention

focused on pronunciation form. The existing research has shown that proficient learners read captions fast, fixating on target words in captions earlier than their auditory onset (Wisniewska & Mora, 2018). In addition, the pioneering study by Wisniewska and Mora (2020), which specifically targeted pronunciation learning through extended exposure to TV series, yielded mixed findings regarding the effectiveness of a focus on form and the availability of captions.

This dissertation fills these gaps by proposing an intervention aimed at increasing audiovisual synchrony during exposure to captioned video, facilitating the mapping of phonological forms onto orthographic forms and the updating of non-targetlike phonological representations. In study 1, we investigated whether highlighting yellow words in captions right before their auditory onset (audio-synchronized textual enhancement) promoted closer audiovisual synchrony during exposure to multimodal input from TV series, resulting in phonological updatings. In study 2, we interviewed a small group of learners to assess their linguistic focus and depth of processing while viewing unenhanced video and when captions contained audio-synchronized textual enhancement. Study 3 consisted in a classroom intervention that combined exposure to videos containing audio-synchronized textual enhancement with opportunities for pronunciation practice and feedback. The results of study 1 showed that audio-synchronized textual enhancement promoted closer audiovisual synchrony than unsynchronized enhancement, and all enhancement conditions including unsynchronized enhancement promoted the updating of lexical phonological representations. In addition, the enhancement mitigated the effects of proficiency and reading speed by slowing down faster and more proficient readers and speeding up slower and less proficient ones. However, study 2 showed that learners' internal focus

may not necessarily be directed to pronunciation if the target words or features are not perceived as problematic for listening comprehension. Against our predictions, study 3 found that audio-synchronized textual enhancement may not offer substantial advantages within a pronunciation-focused classroom intervention. However, the group of learners who carried out the video-based activities without textual enhancement obtained significant pronunciation gains, regardless of their level of L2 proficiency and learning aptitude.

Taken together, these results partially support the use of audio-synchronized enhancement in pronunciation teaching and learning but also highlight the challenges of determining the effects of this semi-incidental learning condition (Leow & Martin, 2017). While the more sensitive measures obtained from eye-tracking and from an auditory word recognition test revealed the impact of exposure to audio-synchronized enhancement, this advantage was not reflected in the results of the classroom intervention, where the enhancement did not lead to significant accuracy gains in the production of the target feature. This finding is in line with the hypothesis that audio-synchronized enhancement may facilitate noticing of the enhanced target features, but other factors such as depth of linguistic processing, treatment length, and learners' previous experience with the target features have a crucial impact on interlanguage restructuring (Han et al., 2008; Leow, 2015). Although semi-incidental pronunciation learning was difficult to quantify due to learners' individualized response to input enhancement, the classroom intervention may nevertheless have contributed to gradual improvements in pronunciation accuracy during subsequent exposure and practice. The methodological proposals in this dissertation, from the use of eye-tracking and verbal recall to the inclusion of different pronunciation tests, provide a

valuable foundation for further research on this topic. Future studies on pronunciation learning from captioned video may delve further into learners' processing of auditory input and explore how different audiovisual manipulations can enhance the audiovisual integration of spoken dialogues and captions.

RESUM

L'expressió oral i la pronunciació s'han deixat de banda en molts contextos d'ensenyament, inclòs el sistema educatiu espanyol. L'èmfasi en l'ensenyament de vocabulari i gramàtica, així com l'exposició limitada a la llengua estrangera dins de l'aula, fa que hi hagi estudiants que encara pronunciïn malament paraules relativament comunes en el llenguatge quotidià després de molts anys d'ensenyament reglat. Això és problemàtic, perquè emmagatzemar representacions fonològiques correctes i accedir-hi ràpidament és vital per a produir un discurs comprensible. L'aprenentatge de la pronunciació i l'exposició a un llenguatge autèntic pot prevenir la fossilització de representacions fonètiques incorrectes. Les sèries de televisió, amb diàlegs que reflecteixen converses espontànies i presenten una varietat d'accents i parlants, tenen un gran potencial en aquest àmbit. Tanmateix, l'ús de vídeos subtitulats per a l'aprenentatge de la pronunciació ha estat poc investigat.

Aquesta tesi proposa una intervenció per incrementar la sincronia audiovisual durant l'exposició a vídeos subtitulats, per a facilitar el mapatge de formes fonològiques en formes ortogràfiques i actualitzar les representacions fonològiques incorrectes. La intervenció implicà exposició a input multimodal (sèries de televisió), on una selecció de paraules es van ressaltar visualment just abans que fossin pronunciades (realçament textual amb sincronització d'àudio). Els resultats mostraren que aquesta tècnica promogué una sincronia audiovisual més propera, tot alentint els lectors més competents i agilitzant els més lents. També promogué l'actualització de representacions lèxiques fonològiques, si bé en un grau similar al realçament textual sense sincronització. Tanmateix, el realçament textual amb sincronització d'àudio no

va oferir avantatges substancials dins d'una intervenció a l'aula centrada en la pronunciació que també va implicar activitats de producció oral.

En conjunt, els resultats donen suport parcial a l'ús de realçament textual amb sincronització en l'ensenyament i aprenentatge de la pronunciació, però alhora posen en relleu el repte de determinar els efectes d'aquesta situació d'aprenentatge semi-incidental. El realçament textual amb sincronització pot ajudar a dirigir l'atenció als aspectes lingüístics realçats, però altres factors com ara el nivell de processament lingüístic, la durada del tractament i les experiències prèvies amb aquests aspectes lingüístics en particular poden ser crucials per activar la reestructuració de la interllengua. Les propostes metodològiques d'aquesta tesi, des de l'ús de seguiment ocular (*eye-tracking*) i recol·lecció verbal fins a la inclusió de diversos tests de pronunciació, proporciona fonaments valuosos per a futures investigacions.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors, Dr. Joan Carles Mora and Dr. Roger Gilabert, for their guidance, support and encouragement in the completion of this thesis. Thank you for always showing interest in my ideas and offering continuous support. I really appreciated how you took the time to walk me through the various stages of my research, especially the most difficult aspects of study design and data collection. Thank you for proposing crucial revisions that significantly improved the quality of this dissertation and of the manuscripts we published. Outside of this project, I am grateful for the opportunity to work side by side in the organization of international events such as the EuroSLA and New Sounds conferences. Thank you for giving us students responsibilities but never leaving us alone to deal with difficult situations. You have shown me what being a fair and supportive supervisor and project leader really means, and I am now going to continue my journey with a very high standard in mind.

I am extremely thankful to Joan Carles for inviting me to join the Barcelona L2 Speech Group and for involving me in the ministerial projects he directed (PhonLex-iSLA, grant PID2019-107814GB-I00) and currently directs (L2E-PPT, grant PID2022-138129NB-I00). Thank you for giving me the opportunity to conduct research on various aspects of L2 speech acquisition with excellent researchers.

I owe special gratitude to Dr. Andrea Révész, for hosting me during my study abroad semester at University College London and providing crucial insights regarding research methodology and analysis. Your extraordinary generosity in sharing your expertise was greatly appreciated.

A big thank you to Dr. Carme Muñoz and to all the members of the GRAL group at the University of Barcelona for their support. I am grateful for the opportunity to have met such a dedicated and talented group of researchers. Your enthusiasm for research is contagious, and your feedback really helped me improve the quality of my work.

Thank you to the professors and graduate students in the L-Slarf research group, for giving me constructive feedback and engaging me in very interesting discussions. Heartfelt thanks to Dr. Andrea Révész, Dr. Kazuya Saito, Dr. Ana Pellicer-Sánchez, and Dr. Jean-Marc Dewaele, for the extraordinary passion with which they inspire junior researchers.

I would like to express my deepest gratitude to the teachers (Aaron and Mariarosa), the students, and the staff at Santa Teresa Gandutxer for generously allowing me to collect data amid the challenges of two demanding school years (2020-2021 and 2021-2022). It would not have been possible to conduct research at school without the help of each and every one of you.

Thank you to the university students who willingly participated in the laboratory study. A special thank you to those who later asked me questions and showed as much enthusiasm about research as I do.

Thank you from the bottom of my heart to my fellow students in the applied linguistics MA and PhD programs at the UB, for the excellent company throughout this long, difficult journey. I won't forget our time together and the reason we made it through so many Mondays: Tuesday board game nights. Thank you Georgia and Ferran for looking out for us like a big sister and brother, and for setting a good example of how to be a successful early career researcher.

Special thanks to Anastasia, Gisela, Ingrid, Matthew, Miren, Natasha, and Radha for being so helpful and patient throughout my seemingly endless data collection. Anastasia, thanks a million for always having simple solutions to my most daunting problems, I admire you for being very professional but at the same time nice and generous. Gisela, we both know that collecting data at school would have been impossible without you, thank you for your incredible kindness and helpfulness. Ingrid, I have learned so much from seeing you deal with massive projects with extreme patience and precision, thank you for helping me figure out many things about research. Natasha, thank you for being there for me through the good and bad times. You showed me that life is about finding your next adventure and jumping on the boat. All of you guys… you are amazing and have a special place in my heart.

I extend my gratitude to the institutions that provided funding for this thesis: The Spanish Ministry of Science, Innovation and Universities [grant PID2019-107814GB-I00] and the Secretary of Universities and Research of the Government of Catalonia, along with the European Social Fund [grant FI_2019].

Although words cannot really express how much I owe my family, I would like to thank them. Thank you mum and dad, for being my rock day after day. To my brother, thank you for the endless banter and for the serious advice. I wish our grandmas and grandpas were here too.

Thank you Giacomo, for being the smartest and most patient teacher of statistics in the world. I cannot thank you enough for your company and all the special memories we have together. In light of my very narrow confidence intervals, I have to reject the hypothesis that you do not make my life significantly better.

It is thanks to Monique, Morgana and Sabrina that I felt at home in the suburbs of London. I was looking for a room and I found a life-long group of fun and kind hearted friends.

To Chiara, Laura, Federica, Carlos, Alice, Alice, Rossana, Giulia, Ste, Uri, Hamish, Yashar, all the friends I made in Barcelona and those who have been a part of my life since I was a school girl… I am a better person for having known you, and I will keep all of you with me always, wherever I go.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AOI: Area of interest

DSR: Delayed sentence repetition task

EFL: English as a foreign language

ESO: Educación secundaria obligatoria (secondary school, 12-16 years)

FFI: Form-focused instruction

FL: Foreign language

FoF: Focus on form

FoFs: Focus on forms

FoM: Focus on meaning

IQR: Interquartile range

L1: First language

L2: Second language

LDT: Lexical decision task

OPT: Oxford Placement Test

RIDT: Reading index of dynamic text

R2c: Conditional R-squared value

R2m: Marginal R-squared value

RT: Reaction time

SLA: Second language acquisition

T1: Time 1 (Pre-test); T2: Time 2 (Post-test)

TW: Target word

CHAPTER 1. INTRODUCTION

Speech is an essential means of communication, and learning how to effectively convey meaning through spoken language is arguably one of the most important goals of language learning (Kormos, 2006). Nevertheless, oral communication skills and pronunciation have long been neglected in second and foreign[1] language teaching and testing (Isaacs, 2018). As a result, it is not unusual for language learners to retain doubts regarding the pronunciation of words that are relatively common in everyday speech, even after years of formal instruction (Cook et al., 2016; Van Zeeland, 2017). Although listeners tend to adapt to minor deviations, frequent mispronunciations can lead to misunderstandings and communication breakdowns, potentially affecting the interlocutor's view of the speaker's linguistic and professional abilities (Marslen-Wilson & Welsh, 1978; Sheppard et al., 2017). In Catalunya, where this dissertation is set, the students' English oral proficiency often remains very limited even after years of formal instruction (Aliaga-Garcia, 2017). The struggles of high school students transitioning into post-secondary education have been linked to the limited importance of oral production and pronunciation in the FL curriculum across the country (Aliaga-Garcia, 2017).

To further aggravate the imbalance between general proficiency and phonological competence, foreign language learners are typically exposed to limited authentic input and plenty of non-targetlike speech inside the classroom (Darcy et al., 2021; Muñoz, 2008). In the schools of many countries including Spain, foreign language instruction

[1] This dissertation centers on the foreign language (FL) context - where a language not commonly spoken in a country is taught in a classroom. However, in the following sections the term *second language (L2) learning* will be used as an umbrella term to refer to the processes involved in learning any language other than one's native language.

amounts to no more than 2-4 hours a week, with exposure to the target language often coming from textbook exercises and non-native teacher talk (Henderson et al., 2012; Muñoz, 2008). Consequently, young learners come in contact with authentic second language (L2) input mainly outside the classroom, carrying out leisure activities that involve the use of technology such as watching TV and playing videogames. In a survey among 865 children living in seven European countries, Lindgren and Muñoz (2013) found that television and movies are the preferred source of L2 input for young learners, and that out-of-school exposure is one of the strongest predictors of FL performance. Investigating the effects of exposure to video in language learning has become increasingly important since the rise of streaming platforms such as Netflix and YouTube has made available an incredible variety of video clips, TV series and movies in many different languages. Nowadays, this content can be personalized to match one's interests and accessed with incredible ease virtually everywhere and at all times on smaller mobile devices such as phones and tablets. The large output of anglophone countries, such as the US and Britain, and the fact that new episodes are often available in the original language days or weeks before being dubbed (and sometimes are never dubbed), provide English learners with significant incentives to watch videos in the foreign language.

A genre like TV series has a high potential for language learning and pronunciation learning, because comprehension is supported by visual elements and contextual knowledge, and the dialogues closely mirror spontaneous conversation and generally feature a variety of speakers and accents (Ghia, 2021; Rodgers, 2011). Research on listening comprehension, vocabulary, and grammar has provided abundant evidence

of the benefits of exposure to captioned video[2], including video from TV series, thanks to the contribution of different methodologies including eye-tracking, classroom research, corpus research and surveys (among others, Lee & Révész, 2020; Pattemore & Muñoz, 2022; Pujadas & Muñoz, 2019; Rodgers & Webb, 2011; Webb & Rodgers, 2009; Winke, 2013). These studies have shown that the availability of verbatim L2 captions supports L2 comprehension and facilitates vocabulary and grammar learning through exposure to video, although several factors like corpus frequency and the learners' vocabulary size can influence the size of the learning gains (for a comprehensive review, see Montero Perez, 2022). In addition, since vocabulary learning involves accumulating knowledge about the different aspects involved in knowing a word (Nation, 2001), the congruency between input and test modality may greatly influence the learning outcomes (Montero Perez, 2022). A handful of studies have attempted to direct learners' attention to novel vocabulary and grammar features by visually highlighting these features in the captions, with mixed results regarding the long-term gains from single exposure to enhanced video (Cintrón Valentín & García-Amaya, 2021; Lee & Révész, 2020; Pattemore & Muñoz, 2022). The use of online data collection methods such as eye-tracking has proven crucial in the establishment of a link between the allocation of attentional resources and learners' gains (Lee & Révész, 2020). The insights provided by research on vocabulary and grammar learning have significant methodological implications for this dissertation and will inform the discussion of the learning potential of captioned video. However,

[2] A *caption* is the verbatim transcription of the dialogue of a video, generally displayed at the bottom of the screen in short lines of text. While the term *caption* is conventionally used to refer to text in the same language as the spoken dialogue (the learner's L2), *subtitle* refers to its translation into the learner's L1.

the focus of this dissertation is limited to the investigation of captioned video and pronunciation learning, and the literature review will closely reflect this scope. In particular, a more detailed review will be provided of studies on the effects of captions on the updating of learners' phonological representations of *known* words, rather than the acquisition of new words. Ultimately, we aim at providing an overview of the processes involved in pronunciation learning through the independent viewing of TV series as well as through a teaching intervention with materials based on TV series.

Factors like the learners' perception of the viewing activity (exclusively recreative vs primarily educational), the amount of attention paid to the language during the viewing, and the specific focus of this attention all concur to moderate the possible learning gains from watching TV. While extensive viewing of TV programs maximizes out-of-class exposure to L2 input and increases the opportunities for incidental learning, "gist watching" may not be sufficient to learn specific properties of a language (Vanderplank, 2019). Although watching videos at home gives learners more freedom over the type and amount of content, it is also easier to be distracted than in the classroom, where the focus is more openly placed on learning. Even when the learner is very interested in understanding the dialogues, they may focus on meaning without paying close attention to linguistic form. In particular, pronunciation accuracy tends to be difficult to learn without a focus on form, due to the unfamiliarity of FL learners with authentic aural input mentioned above and to the intrinsic difficulty of processing fast speech (Vanderplank, 2016). Learners' ability to integrate information coming from auditory and visual sources also plays a crucial role in the processing of captioned video. Learners may focus exclusively on reading captions or switch back and forth between listening and reading, unable to map the auditory form

of words onto their written form. As a consequence, they may not take full advantage of the potential of captions to support auditory processing. Even when they detect words of which they know the meaning but not the pronunciation, the attention they pay to the auditory form of these words may be insufficient to trigger deeper processing. As a result, watching large amounts of TV series with a primary focus on meaning may not improve learners' ability to recognize or pronounce L2 words.

This dissertation introduces a technique that we have called *audio-synchronized textual enhancement*, which consists in highlighting selected target words in the captions of a video just before their auditory onset to direct learners' attention to these words right before they are spoken in the dialogue. This technique, used here to enhance the pronunciation learning potential of exposure to captioned video, was inspired by similar techniques successfully implemented in reading-while-listening studies (Gerbier et al., 2018; Stenton, 2013). Ideally, due the synchronization of the word's enhancement with its auditory onset, the learners should hear the word pronounced by the speaker in the video immediately after mentally pronouncing it in their own mind. The resulting audiovisual synchrony is expected to encourage learners to compare the phonological representation of the word that they have stored in their mental lexicon with its targetlike auditory form, as produced by a native speaker. While noticing of specific auditory forms can occur thanks to this input-based intervention, providing opportunities for production practice and feedback is likely to trigger deeper processing, interlanguage restructuring and, ultimately, more accurate use of these features in production (DeKeyser, 2007; Leow, 2015).

Concurrently with the increased popularity of undubbed TV series among language learners, teachers have also started to implement activities involving the viewing and

elaboration of video content (Alonso-Perez & Sánchez-Requena, 2018; Kaderoğlu & Romeu Esquerré, 2021). To facilitate the investigation of activities involving learners' manipulation of the captions and spoken dialogues in a video, Zabalbeascoa et al. (2012) provided a framework in which these activities are categorized based on their design features and expected learning outcomes. Although research based on Zabalbeascoa et al.'s (2012) audiovisual framework has mostly targeted overall L2 proficiency and measured the effectiveness of these activities in terms of *reported* learning using surveys and interviews, some studies have specifically focused on L2 pronunciation adopting a pre- post-test design (e.g., Lima, 2015a; Sánchez-Requena, 2017; Zhang, 2016; Zhang & Yuan, 2020). In this dissertation, the input enhancement technique discussed above is implemented in combination with two audiovisual activities that are most likely to facilitate bottom-up processing of speech: Captioning, or the insertion of text matching the spoken dialogue in a video, and dubbing, or the revoicing of a muted clip while imitating as close as possible the original dialogue. The aims and contributions of this dissertation are presented in detail in the following sections.

1.1. Aims of this dissertation

This dissertation has three main aims, which guided the design of all the studies involved and the discussion of the results obtained:

1) The overarching aim of this doctoral dissertation is the investigation of a novel technique, audio-synchronized textual enhancement, in pronunciation teaching and learning. This technique, which will be extensively described and contextualized in the next chapters, consists in highlighting target words in the captions of a video in synchrony with their auditory onset, e.g., 500 ms before auditory onset.

2) In addition, we will assess the effectiveness of audio-synchronized textual enhancement within a pronunciation teaching intervention containing pronunciation-focused activities that also involve a focus on speech processing and on the integration of auditory and orthographic input.

3) Finally, we will examine the contextual and individual factors that may moderate the effectiveness of audio-synchronized textual enhancement as a standalone teaching technique or in combination with other video-based activities.

1.2. Contributions of this dissertation

The first main contribution of this dissertation lies in its investigation of the acquisition of **pronunciation through captioned video exposure**, an area that has remained underdeveloped compared with the wealth of research on listening comprehension and vocabulary learning (Montero Perez, 2022). To this end, learners' attention allocation and noticing of pronunciation during video viewing will be investigated through several online and offline methods, thereby expanding on existing research on the effects of caption availability on speech processing (Charles & Trenkic, 2015; Mitterer & McQueen, 2009; Wisniewska & Mora, 2020).

While previous research has found evidence that the availability of word-by-word L2 captions may help learners process L2 spoken dialogue in a video, the mixed results obtained with a form-focused intervention (Wisniewska & Mora, 2020) warrant further research into the effects of a specific focus on pronunciation. To address this issue, this dissertation investigates how **textual enhancement**, a technique commonly used by teachers to direct learners' attention to specific language features, can increase the pronunciation learning potential of captioned video. The second main contribution

of this dissertation is that through the investigation of pronunciation acquisition with and without target word enhancement and the analysis of moderating factors, a complex and nuanced picture is provided regarding the effectiveness of audio-synchronized textual enhancement to direct learners' attention to pronunciation.

The third main contribution lies in the **creation of meaningful activities** aimed at fostering L2 comprehension and pronunciation learning, which responds to the need to promote L2 pronunciation proficiency in a communicative context. Due to the relative novelty of the research topic, a significant effort was required to create learning materials through the selection, linguistic analysis and manipulation of video clips. In addition, to gauge the impact of the interventions on initial and long-term pronunciation learning, as well as its transferability to real life speech (Saito & Plonsky, 2019), a number of different tests had to be created and carefully piloted.

Finally, one further contribution regards collecting data in the **secondary school classroom**, with a population that has historically received scant attention, especially in pronunciation research, compared to more accessible participants such as undergraduate students (Kruk & Pawlak, 2021). In addition, most studies specifically using video-based activities to teach pronunciation have collected questionnaire and interview data on learners' perceptions of learning from these activities, instead of assessing objectively any possible learning gains. In this dissertation, a considerable effort was made to safeguard ecological validity by using learning materials largely familiar to the learners and implementing the learning activities in the classroom, except when eye-tracking data had to be collected individually. At the same time, internal validity was increased by offering a quantitative analysis of learning outcomes both in laboratory and classroom settings, and by taking into consideration

a number of individual factors that may have influenced the results. Overall, this dissertation aims to contribute to the SLA discussion by offering generalizable insights and replicable results on the use of video with enhanced captions and video-based activities to teach L2 pronunciation.

1.3. Organization of the dissertation

This doctoral dissertation is organized into five chapters. Chapter one has introduced the background, motivations, and overarching aims of this dissertation, highlighting its contributions to the field of second language acquisition. Chapter two reviews the relevant theoretical frameworks and research studies, with each section dedicated to a different topic. Section 2.1 begins by discussing the role of phonology in speech processing and its importance for language learning. It then highlights some of the challenges learners typically encounter when acquiring L2 pronunciation, and proceeds to define viable pronunciation teaching goals, describing the approach adopted in this dissertation. Section 2.2 focuses on multimodal input, which is input that consists of a combination of auditory, textual, and pictorial stimuli. After describing the path of multimodal input from the moment it enters memory to the various stages involving deeper processing, this section reviews two influential theories of multimodal learning and the related cognitive load theory, with an emphasis on their application to the processing of captioned video. Tightly linked to the concept of cognitive load is the role of attention in input processing, which is the subject of Section 2.3. This section outlines the development of the concept of *noticing* as a facilitative or even necessary condition for second language acquisition. Then, it reviews a number of theoretical models of attention and their application to learning through exposure to video. Section 2.4 introduces the concept of input

enhancement as a specific type of form-focused instruction, which refers to a variety of pedagogical approaches in which learners' attention is deliberately directed towards specific linguistic features. This section provides an overview of empirical studies on textual enhancement and pronunciation learning, highlighting the main methodological issues and implications. Turning to pronunciation learning through exposure to L2 captioned video, section 2.5 describes the characteristics of TV series and the type of visual and auditory processes involved in viewing video. This section also concludes with an overview of studies that have investigated different aspects of auditory processing, such as speech segmentation, speed of lexical access, and phonetic discrimination. Section 2.6 examines how audiovisual activities, a type of video-based activities, can enhance the learning potential of L2 captioned video. After introducing the audiovisual framework and narrowing down the discussion to specific activities fostering accurate perception and production of the target pronunciation feature, a number of studies that have used these activities to teach pronunciation are reviewed.

Chapter three is organized into three sections, each reporting the specific aims and research questions of a study, followed by its pronunciation target, methodology, results and discussion. Section 3.1 corresponds to an eye-tracking study that featured 58 first-year university students who are native speakers of Spanish and Catalan, and learners of English as a foreign language. Their eye movements were recorded as they watched two L2 video clips under one of five conditions: with words in captions enhanced in synchrony with the words' auditory onset (300 or 500 ms before onset), with words enhanced from caption appearance, with unenhanced captions or without captions. Potential changes in the phonological representation of target words in

learners' interlanguage, i.e., each learner's unique L2 language system (Selinker, 1972), were assessed in terms of the speed and accuracy of rejection of mispronunciations in a lexical decision task. The results of this study suggested that all enhancement conditions directed learners' attention to the target words in captions, with positive effects on speed of lexical access, but the synchronized conditions led to higher synchrony in audiovisual processing. The study described in section 3.2 featured eleven 15 year old high school students who are also L1 Spanish/Catalan learners of English. This study also involved recording eye movements while exposing learners to audio-synchronized textual enhancement in video clips but combined it with verbal recall interviews aimed at assessing learners' level of processing of the enhanced words, detection of their auditory form, and elaboration of the relevant pronunciation rule. The results confirmed that textual enhancement directs learners' attention to the target words, but also showed that learners tend to notice the words' lexical or grammatical properties rather than their pronunciation. Therefore, we designed an intervention that combined the viewing of video clips containing audio-synchronized enhancement with video-based pronunciation-focused activities. Section 3.3 describes the design and implementation of this intervention in three classes of a secondary school and provides an overview of the learning outcomes and learners' perceptions of the intervention.

After discussing the results of each study separately in Chapter three, in Chapter four we will provide a comprehensive discussion of the effects of audio-synchronized textual enhancement in pronunciation teaching and learning. Finally, Chapter five will conclude this dissertation by outlining a summary of its key findings, highlighting its pedagogical implications, and formulating recommendations for future research.

CHAPTER 2. LITERATURE REVIEW

2.1. Pronunciation in language teaching and learning

2.1.1. Introduction

The goals of pronunciation teaching have changed radically over time, from early instructional approaches aimed at the development of nativelike pronunciation instead of meaningful communication practice, to communicative methods centered on the development of fluent, rather than accurate, L2 speech (DeKeyser, 2010; Thomson & Derwing, 2015). Nowadays, a native accent is considered unattainable, except for very few talented learners, and also unnecessary to be functional in most professional and personal settings (Isaacs, 2018; Levis, 2018). However, striving for fluency at the expenses of accuracy has also been found to have detrimental effects on intelligibility. As a consequence, the goal of pronunciation teaching has gradually shifted towards developing learners' ability to be understood clearly and with little effort on the part of the listener (Derwing & Munro, 2005; Murphy, 2014). Possibly due to concerns about increased noise levels and the perceived loss of class control associated with speaking activities, teachers often gravitate toward exercises centered on exposure to written input and on the production of written output (Ellis, 2020). As a result of these pedagogical challenges and of the historical emphasis on written testing, it is not unusual for pronunciation to be sidelined in textbooks and language programs, in favor of other linguistic components (Darcy, 2018; Darcy et al., 2021). In this section, we will begin by describing how phonology is involved in speech perception and production, drawing from the psycholinguistic model proposed by Ramus et al. (2010). Then, we will discuss the broader impact of L2 pronunciation on the development of different components of L2 proficiency, such as listening, speaking,

and reading. Finally, we will discuss the challenges generally encountered in the acquisition of L2 pronunciation, as well as the benefits and objectives of pronunciation instruction.

2.1.2. Ramus et al.'s model of speech perception and production

The mechanisms involved in L2 phonological acquisition are described in Ramus et al.'s (2010) information-processing model of speech perception and production. In this model, a basic distinction is operated between an underlying level of mental representations operating with abstract units, such as a series of phonemes and features, and a surface level in which words are represented as more detailed sequences of sounds made up of phones, the physical realizations of phonemes. The mechanism that mediates between these two levels is the phonological grammar, a set of rules based on the phonetic properties and phonotactics constraints of the language which maps underlying representations onto surface representations. The focus of Ramus et al.'s (2010) model is on the *word*, from the smaller units that compose words to the long-term inventory of words stored in a speaker's memory, known as *mental lexicon*. The mental lexicon contains phonological representations (for example, the auditory form of the word /maʊs/) linked to orthographic representations (the written form *mouse*), and semantic representations (the small rodent, but also the input device for computers).

Speech perception and production require the synergy of these different levels and types of representations (Figure 2.1). The processes involved in the interpretation of spoken language are shown in the left side of the diagram, with the first step at the bottom and the last step at the top. The acoustic properties of speech sounds are first

encoded into abstract, language-specific sub-lexical phonological representations. These representations then provide access to the phonological representations of word forms (hence phono-lexical representations) previously stored in the listener's mental lexicon. The links between phonological representation and semantic and orthographic representations may now be activated, leading to full or partial comprehension of the meaning conveyed by the speaker. Speech production follows a symmetrical but distinct output pathway from the selection and retrieval of words at the semantic and phonological level to their articulation. The processes involved in speech production are shown in the right side of Figure 2.1, starting from the top.

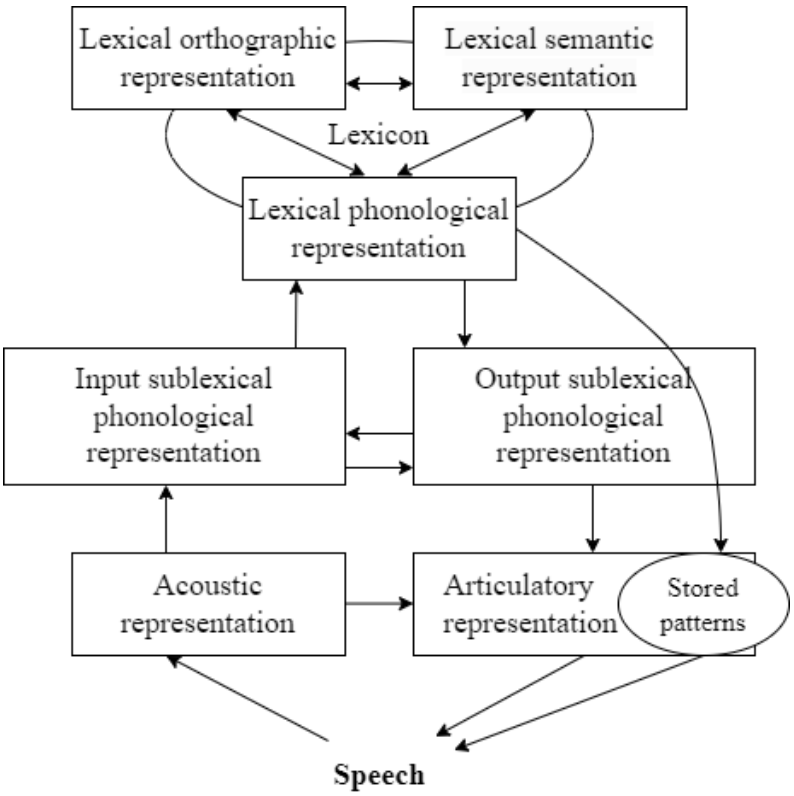


Figure 2.1. Ramus et al.'s (2010) information-processing model.

Production involves the efficient selection of semantic representations, the retrieval of the corresponding phonological representations, their combination into an utterance

consisting of sub-lexical phonological representation, and the conversion of this utterance into an articulatory representation. However, an exception to the regular encoding route from lexical phonological representation to articulatory representation is represented by frequent words, which may follow a "direct articulation route" (Ramus et al., 2010, p. 326). According to this hypothesis, frequent words are initially stored in a simplified form that does not require complex articulatory gestures such as the pronunciation of consonant clusters. This is consistent with data showing that, when learning a native language, children retain mispronunciations in frequent words for longer, compared to less frequent words (e.g., they keep saying *poon* for *spoon*). Just like these representations are not automatically updated once the child's articulatory skills develop, L2 speakers may take longer to adjust the pronunciation of known words once they learn how to pronounce specific phonemes, clusters and stress patterns (Darcy & Holliday, 2019). By the time native speakers of a language reach the adult age, the decoding and encoding of spoken language has become fully automatized, making speech perception (including word recognition and segmentation) and speech production seem effortless (Kormos, 2006). However, for L2 learners, L2 speech perception and L2 word recognition may be inefficient due to the tendency to use L1 patterns in the decoding and encoding of L2 sounds. The role of phonology in language learning and the challenging aspects of L2 pronunciation acquisitions are discussed in more detail in the following subsections.

2.1.3. The importance of L2 pronunciation learning

The acquisition of L2 pronunciation plays a crucial role in the development of other skills such as listening, speaking and reading (Darcy, 2018). To begin with, accurately perceiving the segmental aspects of a language (such as specific phonemes) and its

suprasegmental aspects (such as sentence stress) enhances the listener's ability to decode auditory input. In particular, learning about the speech sound contrasts, sequence of phonemes and rhythm patterns in a foreign language helps the listener segment connected speech into words and access the corresponding lexical representations (Cutler, 2000). Pronunciation development also quite obviously supports the development of speaking skills, as target-like pronunciation of words and phrases tends to result in clearer and more understandable speech (Darcy, 2018; Levis, 2018). In particular, oral communication is more effective when L2 speakers are fluent, i.e., able to retrieve and encode grammatical, lexical and phonological information with a certain degree of automaticity, and accurate, i.e., their speech is closely aligned with the production of native speakers in terms of the articulation of phonemes and segmental sequences, as well as intonation and prosody (Mora & Levkina, 2017).

The importance of accurate pronunciation extends beyond spoken communication to the process of decoding and understanding written text. The interaction between pronunciation and reading involves the interplay between orthographic representations and phonological representations, which are stored in separate but interrelated subsystems of the mental lexicon (Charoy & Samuel, 2019; Ramus et al., 2010). Research shows that orthographic information may be activated during auditory word recognition, and phonological information tends to be automatically activated in written word recognition, compatibly with the frequently reported phenomenon of hearing a voice in one's head during silent reading (Burton et al., 1996; Charoy & Samuel, 2019; Eiter & Inhoff; 2010; Erdener & Burnham, 2005). It has been hypothesized that reading involves the activation of abstract, not fully

specified representations of words based on their phonological structure (Frost, 1998). As a result, beginner readers are slower because they need to convert each grapheme into a phoneme in order to generate a detailed phonological representation, whereas as the reader's phonological decoding skills evolve, undetailed phonological structures can be efficiently linked to their meaning, making lexical access faster and less taxing (Frost, 1998). Given the established importance of pronunciation in language processing, we will now focus on the challenges encountered by L2 learners, discussing how their language learning background may have contributed to the development of non-targetlike pronunciation.

2.1.4. Challenges in the acquisition of L2 pronunciation

Several models have been proposed to explain the influence of the L1 phoneme inventory and phonotactics on L2 perception and production. For example, the perceptual assimilation model of L2 speech learning (PAM-L2 model - Best & Tyler, 2007) attributes learners' difficulty in perceiving L2 contrasts to the assimilation of two L2 phonemes to the same L1 category (e.g., assimilating the English vowels in *ship* and *sheep* to Spanish /i/). On the contrary, according to PAM-L2 it is easier to learn contrastive L2 sounds assimilated to distinct L1 categories (e.g., pronouncing *cap* and *cup* like /kap/ and /kup/, respectively). According to Flege's (1995) speech learning model (SLM), the pronunciation errors observed in L2 learners' production derive from their inability to form a new phonetic category, leading to an excessive reliance on pre-existing L1 categories (L1-based processing). Although the SLM model originally assumed that accuracy in perception precedes accuracy in production, in the revised SLM-r model (Flege & Bohn, 2021) these two aspects develop simultaneously and are mutually beneficial. Overall, according to these

models, L2 pronunciation errors arise from the inability to perceive and produce sounds and patterns that are absent from their L1 phonemic inventory, or to distinguish them from other L1 or L2 sounds.

However, issues in L2 perception and production can also be caused by uncertainties regarding the sequence of phonemes constituting the phonological form of a lexical item. These uncertainties may in turn lead to unstable form-to-meaning mapping and imprecise or "fuzzy" phono-lexical representations, which undermine the speed and accuracy of lexical access (Kormos, 2006; Cook et al., 2016; Darcy et al., 2013). The imprecision of phono-lexical representations may be independent from difficulties in perceiving phonemic contrasts, but it can similarly increase the number of activated lexical candidates, delaying word recognition due to the activation of "phantom words" which were not present in the speech signal (Broersma & Cutler, 2011; Cook et al., 2016; Llompart & Reinisch, 2018). Therefore, to achieve effective communication in the target language, learners need to focus not only on accurately distinguishing and producing single sounds and features, but also on developing target-like phonological representations of words and retrieving them efficiently during listening and speaking.

On the one hand, learning about the phonemes and stress patterns of a foreign language supports the segmentation of speech into smaller units, facilitating the recognition of words and phrases. On the other hand, the acquisition of a substantial amount of spoken vocabulary enhances bottom-up processing of aural input and frees up working memory for top-down processing, with positive effects on L2 listening comprehension (Vafaee & Suzuki, 2020; Zhang & Zhang, 2020). However, in education systems that put considerable emphasis on reading and writing, imbalances

typically arise between L2 learners' written and spoken vocabulary (Van Zeeland, 2017). While their written vocabulary is broad enough to allow them to efficiently recognize words while reading, their limited spoken vocabulary leads to difficulties in word recognition while listening (Van Zeeland, 2017). The next subsection focuses on how pronunciation instruction can help learners face these challenges by defining its goals and present relevant pedagogical strategies for achieving these objectives.

2.1.5. Pronunciation teaching goals: Comprehensibility, automaticity and generalizability

Focused instruction regarding segmental and suprasegmental features of speech has been repeatedly found to have positive effects on learners' overall linguistic abilities (Saito & Hanzawa, 2016; Saito & Lyster, 2012; Thomson & Derwing, 2015). Nevertheless, pronunciation is often overlooked in textbooks, teacher training, and school curricula in favor of other linguistic aspects like vocabulary and grammar (Celce-Murcia et al., 2010; Henderson et al., 2012; Isaacs, 2018). Several factors may have contributed to the neglect of pronunciation in language teaching, including the traditional association of pronunciation teaching with rote exercises such as decontextualized drills and repetitions (Darcy et al., 2021; Isaacs, 2018). In addition, questions regarding the attainability of a native accent have discouraged many learners from attempting to improve their oral communication skills (Murphy, 2014). Nowadays, there is general consensus that the goal of pronunciation instruction should be to develop the learners' *intelligibility* (the ability to be understood) and *comprehensibility* (the listener's ease of understanding), rather than to eradicate L1 accents (Darcy, 2018; Levis, 2018; Mora & Levkina, 2017; Murphy, 2014; Thomson & Derwing, 2015). In fact, it has been noted that heavily accented L2 speech produced

by both L2 speakers and L1 speakers with strong regional accents may nevertheless be highly comprehensible and intelligible (Murphy, 2014).

Setting goals and priorities for pronunciation instruction involves recognizing that not all phonological deviations are equally problematic or, indeed, suitable for teaching at specific proficiency levels (Darcy et al., 2012). For example, at lower levels PI should focus on word-based features such as word stress and congruent sound-spelling correspondences, at least until learners can comprehend and formulate sentences (Celce-Murcia et al., 2010; Darcy et al., 2012). As learners advance and develop their speaking skills, it is important to review regularly basic features that may affect intelligibility like stress and intonation patterns, but also segmental aspects like final consonant and clusters, and especially those with a grammatical function (Celce-Murcia et al., 2010; Darcy et al., 2012; Jenkins, 2002). It is also important to consider that the impact of non-targetlike phonological realizations on comprehensibility and intelligibility depends on several aspects including the listener's familiarity with accented speech, the salience of the mispronounced features, and the potential for misinterpretation (Levis, 2018, Murphy, 2014). While it is common to adjust to small deviations in L2 speech production, a large number of mispronunciations may affect speech intelligibility, especially if the listener cannot easily interpret the message with the help of contextual knowledge (Marslen-Wilson & Welsh, 1978). Mispronunciations that lead to semantic or grammatical ambiguity may be particularly disruptive and lead to a loss of intelligibility (Levis, 2018). An example of problematic morphophonological feature is the past <-ed> ending of a regular past verb form, which will be the target of two of the studies in this dissertation. When pronouncing a verb in the regular past, learners may incorrectly simplify the final

consonant cluster (*jumped* pronounced */ʤʌmp/ instead of /ʤʌmpt/) or change the syllable structure by pronouncing a non-syllabic ending as a separate syllable (*/ˈʤʌmpɪd/). Frequently mispronouncing relatively common words may not only result in communication breakdowns, but also affect the listener's perception of the speaker's language proficiency and of their professional skills (Sheppard et al., 2017).

While accurate production of the phonological form of words is crucial for learners to produce comprehensible speech, the ability to retrieve lexical representations with a certain *automaticity* is also key to producing fluent speech that is easy to understand (Darcy, 2018; Jarosz, 2019; Pellicer-Sánchez, 2015; Tavakoli, 2019). As in spontaneous speech the focus is typically on lexis and on the conceptualization of the intended message, learners who fail to proceduralize declarative knowledge of L2 phonology need to direct conscious attention to pronunciation to achieve pronunciation accuracy, with negative effects on speech rate and increased pause frequency and duration (Kormos, 2006; Levis, 2018; Skehan, 2009). If the focus on accuracy takes place after an L2 utterance is spoken, the speaker may not pause as frequently, but rather self-correct and repair, nevertheless resulting in increased effort on the part of the listener and reduced comprehensibility (Kormos, 2006; Levis, 2018). On the other hand, automatic recognition and retrieval of auditory word forms frees up attentional resources, allowing the speaker to invest additional effort and attention into different aspects of language processing, such as macro-planning aspects of message generation and monitoring (Leow, 2015; Levelt, 1989). Focused practice can strengthen the connection between concepts and words, leading to faster and more efficient retrieval of target-like phonological representations, as shown in the faster and more accurate performance of L2 learners in auditory lexical decision tasks at

advanced proficiency levels (Darcy & Holliday, 2019). Acquiring the targetlike phonological realization of difficult words and features from the initial stages of acquisition and regularly revise them may prevent fossilization, i.e., the process by which learners habitually use an inaccurate linguistic item or rule (Selinker, 1972). On the contrary, the fossilization of non-targetlike phono-lexical representations may be reinforced by regular access over time, impairing their updating (Darcy & Holliday, 2019).

To foster the development of automatic phonological processing and, consequently, enhance fluency, it is crucial to incorporate repetition and controlled practice components into pronunciation instruction (Celce-Murcia et al., 2010; Darcy, 2018; DeKeyser, 2010). While traditional decontextualized techniques such as drills, minimal pair repetition, and discrimination tasks may provide opportunities for repetition, they do not promote the transfer of classroom practice to other contexts of language use (Darcy et al., 2012; Darcy, 2018). To enhance both automaticity and generalizability, effective pronunciation instruction should combine exposure to diverse sources of input in the target language, explicit feedback on challenging phonological forms, and contextualized perception and production practice (Darcy, 2018).

2.1.6. Pedagogical approaches

The importance of pronunciation in language pedagogy has historically been a controversial topic, with some approaches glorifying it and others purposefully disregarding it (Isaacs, 2018). Up until the second half of the 19th century, pronunciation has served a marginal role in favor of classical methods concerned with

the study of grammar and rhetoric (Murphy & Baker, 2015). These methods, which have been grouped under the label of *grammar translation method*, identified language proficiency as the ability to read, understand and translate original versions of written texts. Although language teaching approaches based on similar principles are still practiced today in some contexts, there is no evidence supporting the effectiveness of the grammar translation method (Murphy & Baker, 2015). A breakthrough was represented by the creation circa 1887 of the International Phonetic Alphabet, an alphabetic system devised to represent graphically the speech sounds of any language and still universally adopted today in phonetic transcription. The publication of the IPA by the International Phonetic Association, paired with the declared primacy of the spoken form of a language, led to the application of an *analytic-linguistic approach* to pronunciation instruction (Celce-Murcia et al., 2010). This approach involved explicit instruction of the target language phonology with the help of physical tools, such as the IPA and diagrams of the vocal tract, as well as articulatory descriptions and contrastive information. On the contrary, the *intuitive-imitative approach* consisted in carefully listening to target speech models and imitating them without being given any explicit information. With the rise of the analytic approach, the intuitive method was not rejected, but rather incorporated into lessons as a separate phase of pronunciation practice.

The distinction between analytic and intuitive approaches has been crucial in the development of pronunciation pedagogy and can be used to categorize most of the language teaching approaches recorded over the centuries (Celce-Murcia et al., 2010). The *audiolingual method*, dominant in the 50s and early 60s, is considered an analytic approach which placed a substantial emphasis on the development of accurate

pronunciation. This method focused on developing learners' speaking and listening competence by exposing them to large amounts of dialogue in the target language, having them memorize and repeat sentences and minimal pair drills, and correcting errors to prevent the formation of bad habits (Celce-Murcia et al., 2010). In reaction to the audiolingual method and in stark contrast with the grammar translation method, the *natural approach* (Krashen & Terrell, 1983) rejected the deliberate study of grammar and the explicit correction of learner errors, including their pronunciation. This approach, which stemmed from the observation of child L1 acquisition, exposed learners to plenty of L2 input from a model (the teacher or recordings) but allowed language output to emerge naturally and gradually, rather than forcing it. Similarly, the *communicative language teaching* approach emphasized the development of fluent communication skills rather than the mastery of language forms, leaving little room for decontextualized pronunciation exercises (Murphy & Baker, 2015).

Over the past 30 years, with the rise of form-focused approaches aimed at drawing learners' attention to linguistic form within communicative activities, the importance of pronunciation has been gradually reevaluated in applied linguistics research (Isaacs, 2018). Some of the most common form-focused techniques in pronunciation instruction involve exposure to structured input requiring processing of form for meaning comprehension, exposure to textually enhanced input in which target structures have been highlighted, and/or production activities requiring accurate pronunciation for successful completion[3] (Saito & Lyster, 2012). To direct learners' attention to commonly mispronounced words and foster the updating of their

[3] The theoretical framework of form-focused instruction will be dealt with in detail in section 2.4 when illustrating the enhancement technique used in this study.

phonological representations, it is important to include plenty of listening practice with materials that are meaningful and interesting for learners while also featuring a variety of voices, contexts, and speech rates (Darcy, 2018). Integrating multimodal input from sources like captioned video in language teaching and promoting extracurricular exposure to L2 video can theoretically help learners associate the auditory form of words with their meaning, although research in this area is limited (Darcy, 2018).

2.1.7. Summary

Section 2.1 has highlighted the importance of pronunciation in language processing and acquisition, drawing from Ramus et al.'s (2010) lexicon based L2 speech model. While storing accurate lexical phonological representations (the mental representations of spoken words) and accessing them in a timely manner is essential for perception and production, learners often face challenges in acquiring targetlike L2 lexical phonological representations. The limited importance traditionally assigned to pronunciation in the foreign language curriculum and the pursuit of unattainable goals, such as eradicating non-native accents, have relegated pronunciation to a minor role in language learning. After proposing more practical goals for pronunciation teaching, such as helping learners being understood with ease, the section has traced the history of pronunciation instruction and discussed how the various approaches have influenced the form-focused instruction approach that is the focus of this dissertation. The next sections formulate a proposal of pronunciation-focused teaching approach that incorporates the above mentioned recommendations by integrating meaningful comprehension and production activities based on captioned video into the foreign language classroom.

2.2. Multimodal input in language learning

2.2.1. Introduction

Most of the input we receive every day is multimodal, in that it combines information in different modalities (Montero Perez, 2020). For example, watching a video involves exposure to visual stimuli, such as moving images and captions, and auditory stimuli, such as the dialogue and sound effects. In a similar way, understanding our interlocutor's words, gestures and facial expressions is key to achieving effective and meaningful communication. The pervasive nature of multimodal stimuli in human perception of reality has prompted extensive research on the impact of exposure to such input. In particular, exposure to materials that combine multiple modalities, such as videos or PowerPoint presentations, has been found to lead to deeper processing and more durable learning of subject content compared to exposure to single-modality input, such as written text without any accompanying visuals (Mayer, 2005). This section begins with a brief introduction of the structure and functions of human memory, as defining the relevant concepts and terminology is essential for understanding in what ways memory is implicated in the processing of captioned video. Then, it reviews Paivio's (1986) and Mayer's (2005, 2009) cognitive theories of multimodal learning, with a focus on the simultaneous processing of information in the auditory and visual modalities. By describing how information can be concurrently processed through multiple sensory channels, these theories provide the conceptual underpinning for the research conducted in this dissertation. Finally, the importance of cognitive load in the design of instructional materials will be discussed, highlighting how exposure to multimodal input can be optimized to enhance foreign language learning outcomes. Given the focus of this dissertation on exploring

techniques that can enhance pronunciation learning through exposure to multimodal input, it is crucial to discuss possible methods to reduce unnecessary cognitive load and promote deep processing of the input.

2.2.2. An overview of memory

During exposure to a source of multimodal input such as captioned video, a wealth of auditory and visual stimuli reach our senses and begin their trajectory within our memory, subject either to successful retention through rehearsal and elaboration, or to decay and extinction. These stimuli are processed through three main types of memory: Sensory memory, short-term memory or working memory, and long-term memory, which is further categorized into explicit and implicit memory (Atkinson & Shiffrin, 1968). *Sensory memory* is a large-capacity storage of sensory information (e.g., visual, auditory, tactile) which is only retained for a few seconds. Short term memory is often referred to as working memory to avoid the confusion arising from the different uses of the term "short term memory" in psychology (Hall & Stewart, 2010). However, the two terms have slightly different meanings, as *short-term memory* is the memory component concerned with the temporary *storage* of a limited amount of information (seven items +/- 2) for a maximum time of approximately 30 seconds. Working memory, on the other hand, is the multimodal, limited capacity system responsible for the *processing* of this information during cognitively complex tasks such as speech comprehension (Baddeley & Hitch, 1974; Robinson, 2003). The duration of short-term memory can be extended by the process of rehearsal through which auditory items are mentally repeated, usually through subvocalization (Hall & Stewart, 2010). In Baddeley & Hitch's (1974) model, working memory consisted of three elements: the phonological loop, which allows the brain to hold verbal

information for 1-2 seconds unless it is refreshed by rehearsal, a visuospatial sketchpad that stores visual information, and a central executive that controls attentional processes. Of interest to the explanation of multimodal input processing is the fact that Baddeley's (2000) updated working memory model also included an episodic buffer, which integrates information coming from both the phonological loop and visuospatial sketchpad, providing a temporary interface with long-term memory (Figure 2.2).

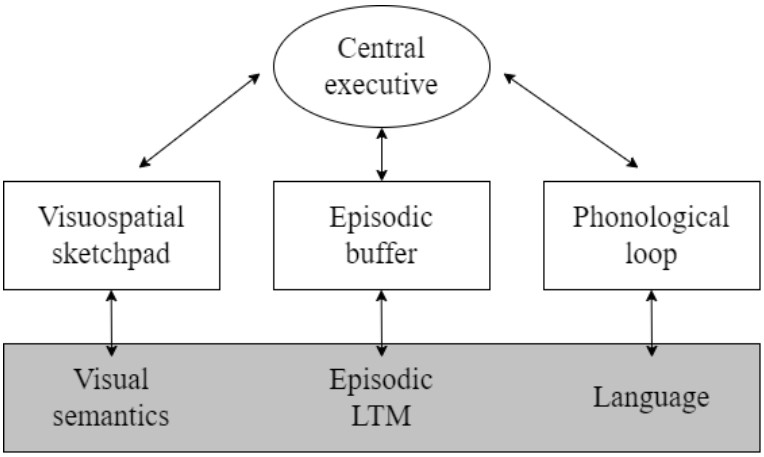


Figure 2.2. Working Memory Model (Baddeley, 2000, p. 421)

Finally, *long-term memory* has a potentially unlimited capacity and is further divided into two categories: explicit (or declarative) memory, which involves conscious recollection of facts and events, and implicit (or non-declarative) memory, which is revealed through changes in behavior or performance without conscious awareness of the underlying information (Hall & Stewart, 2010). Explicit memory includes *episodic memory*, the ability to recall personal experiences with contextual details such as time and place, and *semantic memory*, which stores general knowledge and facts about the world, such as the meaning of words.

To illustrate how the memory system is involved in the processing of captioned video, let us consider a concrete example of an individual watching an episode of a TV series. As the viewer follows the unfolding narrative, segments of spoken dialogues and visual sequences are encoded by their sensory memory and enter short-term memory. The caption lines, usually placed at the bottom of the screen, may also be processed as part of the visual input. The episodic buffer, a component of working memory, facilitates the coherent integration of the verbal and non-verbal information processed via the phonological loop and visuospatial sketchpad. Later, the individual may discuss the series with friends, recalling various aspects of the show including character interactions, plot twists, and even specific lines from the dialogues. While the storyline and the language used in the TV series are initially stored in episodic memory, this information can gradually stabilize into long-term memory. Just like previous knowledge of the characters can aid comprehension of subsequent episodes, it is possible that repeated encounters with words and expressions in the TV series may facilitate the processing of related input in novel contexts.

2.2.3. Paivio's dual coding theory

Having described how information from multiple auditory and visual sources can be managed and stored by the human mind, we will now review the cognitive theories that explain how language processing and retention can be facilitated specifically by exposure to multimodal input. Paivio's (1986) dual coding theory postulated the existence of two separate but interconnected coding systems for words and images in the human mind, where verbal and non-verbal stimuli can be processed simultaneously and independently. The verbal representational units, called *logogens* in the dual coding theory, roughly correspond to spoken language, whereas non-verbal

representations or *imagens* correspond to natural objects. Figure 2.3 shows how these units can be activated by external elements (representational connections), by another unit in the same modality (associative connections) or in a different modality (referential connections). While the spoken form of a word activates its written form through associative connections, referential processing allows a spoken word to activate not only the representation of that word, but also of an image, and vice versa. The positive impact of referential processing on recall has been established in a number of studies where the association of L2 vocabulary to non-verbal stimuli (pictures and objects) was achieved in fewer trials and with fewer errors than the association of L1 and L2 vocabulary (Paivio, 1986). From this perspective, exposure to multimodal input such as captioned video may support language learning as it activates both the verbal and non-verbal processing channels and affords opportunities for mental connections within and between systems (Montero Perez et al., 2013). Recent evidence supports this hypothesis, as the co-occurrence of a word in the soundtrack of a video together with its relative image has a facilitative effect on L2 vocabulary learning (Pujadas & Muñoz, 2023). Elements of Paivio's dual coding theory have been incorporated in Mayer's cognitive theory of multimedia learning (reviewed in subsection 2.2.4), which similarly emphasizes the value of presenting information through multiple modalities to enhance learning.

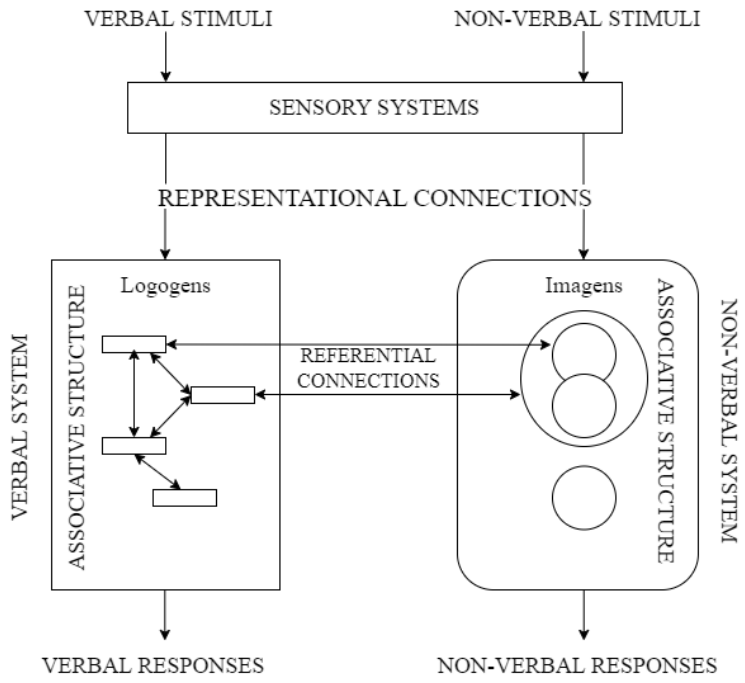


Figure 2.3. Dual coding theory (Paivio, 1986, p. 67)

2.2.4. Mayer's cognitive theory of multimedia learning

Mayer's (2005, 2009, 2014) cognitive theory of multimedia learning is based on three assumptions: The dual-channel assumption, the limited capacity assumption and the active processing assumption. According to the *dual-channel* assumption, people have separate channels for processing visual and auditory information, and learning is enhanced when these channels are simultaneously engaged (as in Paivio, 1986). The *limited capacity* assumption refers to the limited nature of working memory, which can only process a limited amount of information at once. The *active processing* assumption challenges traditional views of memory as a passive storage in which information can be arbitrarily accumulated, by stating that learning involves actively constructing and connecting mental representations (Mayer, 2014).

Mayer's (2005, 2014) model explains how the cognitive processes of selecting, organizing, and integrating information occur between and within the three memory components (sensory memory, working memory and long-term memory) to achieve long-term retention of knowledge (Figure 2.4). Learning requires three fundamental steps: the selection of words and images that enter working memory through the senses, the organization of selected material into a meaningful model, and the integration of novel information with relevant prior knowledge in long-term memory. Interestingly, while spoken sounds only need to be selected and organized into verbal units in order to activate mental representations, written words take a more complex route through the system because they enter the brain as images and are converted into sounds to be processed through the phonological system. Therefore, in line with Baddeley's view of working memory (2007), the multimodal learning theory assumes that a single structure is responsible for processing verbal information (phonological loop and auditory channel, respectively), whether from direct auditory input or from the subvocal articulation of a visually presented word. The assumption that subvocal articulation may be necessary, at least for less skilled readers, is consistent with the frequent finding that L1-biased phonological representations reflecting the grapheme-phoneme correspondences and phonological patterns of the learner's L1 interfere with L2 phonological development (Bassetti et al., 2015; Bassetti & Atkinson, 2015).

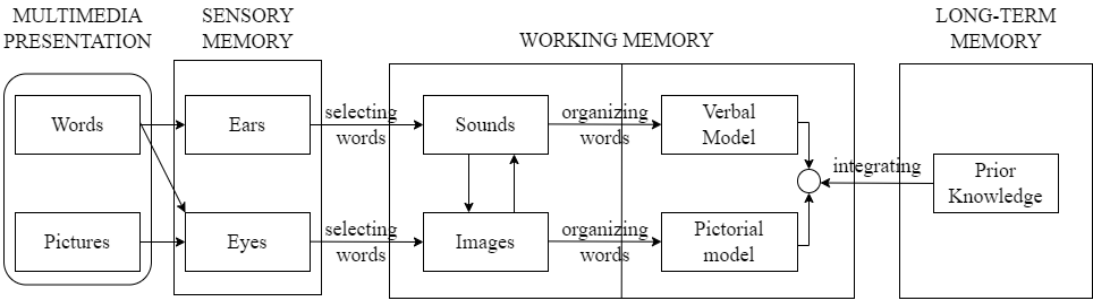


Figure 2.4. Cognitive theory of multimodal learning (Mayer, 2005, p. 52)

Mayer's theory hinges on a number of principles which clarify under which circumstances multimodal presentation can be more advantageous than single modality presentation. However, the multimodal learning theory was initially devised to assist educators in designing more effective instructional materials such as PowerPoint presentations composed of diagrams, images, and narration. While some principles are closely related to this objective and might have limited relevance outside the domain of presentation design, others may be critically applied across various educational scenarios to optimize learners' comprehension and retention of information through exposure to multimodal input.

On the one hand, the *redundancy principle* states that the concurrent presentation of images and narration is more effective than the combination of images, narration *and* printed text, which can overload learners and potentially impair learning (Mayer, 2009). The *modality principle* also emphasizes the importance of using information processed by different channels, such as spoken narration and images (processed by the auditory and visual channel, respectively) instead of multiple sources of information in the same modality, like text and images (both visual). On the other hand, Mayer concedes that not all forms of redundancy are detrimental to learning and some are actually beneficial. For example, the inclusion of a very short text (rather

than a long text) next to an image is less likely to split learners' attention between these two sources of visual information, especially if the text is in close proximity of the corresponding image. The *spatial principle* and *temporal contiguity principles* also highlight the importance of presenting related words and pictures close together and simultaneously rather than successively. Finally, the *personalization principle* relates to another potential benefit of L2 captioned video, i.e., that learning materials are more effective when a sense of personal connection between the learner and the input material is fostered through, for example, the use of conversational language and realistic scenarios. Overall, these principles align with the pedagogic use of L2 captioned video, as the word-by-word transcription of the soundtrack provided by captions potentially supports learners' comprehension without being too taxing for working memory.

2.2.5. Cognitive load in multimedia learning

Mayer's emphasis on the idea of limited capacity in learning can be traced back to Sweller's (1988) cognitive load theory, which posits that learners have a limited capacity to process new information and that cognitive overload can occur when that capacity is exceeded. Applying his cognitive load theory to multimedia learning, Sweller (2005) pointed out that the design of instructional materials must take into account the limited capacity of working memory. Information processing is seen by Sweller as largely dependent on pre-existent *schemas*, i.e., patterns of thoughts that organize pieces of information. For example, reading requires first the construction of schemas for individual letters, then schemas that combine letter strings into words, then words into sentences. After storing countless schemas into long-term memory, we are able to understand and process language fluently. With practice, reading and

learning in general become more *automatic* and require less conscious effort, since most information falls within pre-existent schemas. On the contrary, processing a large amount of novel information taxes working memory, which can only contain about two to four elements at a time. The role of instruction is to provide learners with missing schemas, so that they become more fluent at processing and do not need to engage constantly in the time-consuming process of random hypotheses generation and effectiveness testing (Sweller, 1988).

A key concept of cognitive load theory is the distinction between extraneous, intrinsic, and germane cognitive load (Sweller, 2005). *Extraneous* cognitive load is detrimental to learning, because it diverts working memory resources from schema construction and automation, the main processes that guide learning. Extraneous cognitive load is caused by elements external to the learner and related to the design of learning materials and can be manipulated by the material designer. *Intrinsic* cognitive load is essential to learning and depends on the natural complexity of the learning materials, namely the number of elements contained and their level of interaction. For example, learning the translation of a single word has a lower intrinsic cognitive load than learning word order, as attending to the translation of each word is not sufficient to understand the interactions between them. Lastly, *germane* cognitive load is induced by learners' efforts to process the material and is effective because it results in schema construction and automation. An effective way to increase germane cognitive load is to provide several examples from different contexts to illustrate a rule, as this would support the process of schema construction and therefore learning. According to Sweller (2005), instruction should reduce extraneous cognitive load to allow for an

increase in germane cognitive load, unless intrinsic cognitive load is very low and the total cognitive load remains low, as is the case with very simple materials.

2.2.6. Summary

This section has provided an overview of human memory and of the processing and retention of information obtained from visual and auditory sources of input, drawing upon the foundational theories of multimodal learning proposed by Paivio (1986) and Mayer (2005). The concepts of multimodality and cognitive load have been discussed in the context of designing effective instructional materials, highlighting the need to balance the amount and complexity of information presented to learners to avoid overwhelming their working memory (Sweller, 2005). Despite the relative complexity of cognitive theories of multimodal learning, theoretically sound learning materials involving multimodal input may be designed with relative ease following the relevant principles. For example, designing completion activities featuring a partial solution that helps learner solve the problem is an effective way to reduce extraneous cognitive load and promote germane cognitive load (Gesa, 2019). In the case of video-based activities, this may involve asking learners to fill in the missing words in captions rather than writing the entire caption line or having them repeat shorter utterances that would not require processing of an overwhelming amount of information. As we move forward, the next section will delve into a more in-depth exploration of attention as a limited capacity resource and of its importance in the context of language learning.

2.3. Attention, noticing and L2 input processing

2.3.1. Introduction

Research on multimodal learning aligns with the assumption that the input available to be perceived by the human senses is the fundamental source of data for language learning, and that input processing represents the first step towards L2 acquisition (Leow, 2015; VanPatten, 2004). However, learners with limited knowledge of the foreign language may encounter challenges in processing the large amount of information available in the input. This is because at any moment in time, listeners need to take, in a relatively automatized manner, many quick decisions in order to concentrate on the chunks of speech or text that will help them extract meaning from the input. The cognitive control mechanism that allows the human mind and its limited processing capacity to focus on specific information, ignoring countless competing stimuli, is *attention* (Chun et al., 2011). This section describes the mechanism of input processing from an attentional perspective, focusing on the role of attention in filtering and selecting relevant input, and on its implications for language learning.

Questions regarding the nature and allocation of attentional resources have stimulated a wealth of research in the fields of instructed and naturalistic language acquisition (Robinson et al., 2012). Increasingly more detailed accounts of how attention modulates SLA processing have been developed over time, inspired by cognitive psychology views of attention as a limited-capacity information-processing mechanism (Robinson et al., 2012). Starting from the debate on the role of consciousness and awareness in input processing (Schmidt, 1990, 1995, 2001), this section will review relevant models of attention (Leow, 2015; Robinson, 1995, 2003;

Tomlin & Villa, 1994; VanPatten, 2004) and their implications for language learning and teaching.

2.3.2. The noticing hypothesis

Theories highlighting the importance of attention in SLA (Schmidt, 1990; Sharwood Smith, 1991; Tomlin & Villa, 1994) emerged in response to Krashen's (1982) monitor model, specifically to Krashen's radical rejection of the role of conscious awareness in language acquisition. The monitor model postulated that the unconscious language learning system (acquisition) was superior to the conscious one (learning), and that learners simply pick up a language if they are exposed to a sufficient amount of comprehensible input. Schmidt (1990), however, argued that while language comprehension and production may revolve around unconscious processes, language *acquisition* can happen unconsciously in the sense of unintentionally or without metalinguistic understanding, but not unless some awareness has been gained of a specific aspect of the target language. Schmidt's noticing hypothesis (1990, 1995, 2001), building on the distinction between input (what is perceived at a sensory level) and intake (what is cognitively registered), claimed that noticing input, that is, becoming aware of an aspect of the target language, is necessary to convert it into intake. *Noticing* was originally defined as availability for verbal report, for example in written diaries (Schmidt & Frota, 1986), but recent research has relied on more sophisticated and sensitive methods such as monitoring eye movements through eye-tracking and stimulated recall immediately following an experimental session (e.g., Godfroid et al., 2010).

2.3.3. A functional analysis of attention: alertness, orientation and detection

While acknowledging the role of Schmidt's noticing hypothesis in establishing the importance of attention in SLA, Tomlin and Villa (1994) argued against the necessity of conscious awareness for learning. To explain the rationale behind their model of attention, they reviewed the psychological theories that have influenced SLA research, identifying four main characteristics of attention and related implications:

1) Attention is a limited capacity system which is not able to handle all the stimuli perceived by the senses at any given time.

2) Relatedly, attention is the mechanism responsible of selecting a subset of stimuli in the input for further processing.

3) Attentional mechanisms require effort and are therefore limited, in contrast with more automatic processes that can be carried out in parallel with little interference.

4) Unlike automatic processes, attention can be at least partially controlled and directed to specific aspects of the presented stimuli.

In light of the limitations of this "coarse-grained" analysis of attentional mechanisms for the investigation of SLA processes, Tomlin and Villa (1994) proposed a "finer-grained" analysis involving three distinct but interrelated functions: Alertness, orientation and detection. While *alertness* is a general readiness to register and respond to incoming stimuli in a timely manner, the *orientation* towards a specific type of sensory information facilitates further processing of selected stimuli at the expense of others. Lastly, *detection* is the most effective and resource-consuming mechanism of selection of sensory information for further processing at the expense of other undetected information. Detected linguistic items are maintained in short term

memory and become available for further processes leading to language learning, such as hypothesis formation and hypothesis testing. Tomlin and Villa (1994) argue that awareness of a linguistic phenomenon is not necessary for its detection, but it can be facilitative by increasing the learner's alertness and directing their orientation to specific linguistic aspects.

To illustrate how Tomlin and Villa's (1994) model of attention can shed light on the attention allocation patterns observed in certain learning and testing contexts, let us consider the common scenario of a language learner watching a video clip in English, the foreign language they study at school. If the viewing takes place in a language laboratory as part of a research study, the learner is likely to exhibit greater *alertness* than when watching the same video in the classroom where, in turn, they will be more alert than when casually watching a video at home. A learner who finds reading easier than listening in the foreign language may, counterintuitively, be more *oriented* towards the spoken dialogues when watching videos in a classroom setting, because they believe that reading captions can impair the development of listening skills (Kruger et al., 2015; Vanderplank, 2019). On the contrary, when watching L2 video clips at home, the same learner might rely heavily on captions to follow the dialogues. Due to the informal viewing context and the primary focus on meaning, the detection of L2 phonological patterns would largely happen without *awareness*. Suppose that, before the viewing, the learner is asked to record themselves saying a list of words containing a target phonemic contrast (e.g., /iː/ - /ɪ/). As a result of the increased orientation towards this specific feature, the learner may detect (cognitively register) several occurrences of the target vowel sounds during the viewing. In this case, the learner's focus on the pronunciation of these sounds in the dialogues may be more

deliberate than when they are paying attention to the auditory input without a specific focus, and this may immediately translate into higher accuracy in subsequent testing. However, due to the exclusive focus generated by the attentional mechanism of detection, other aspects concerning the content and language used in the video may be overlooked.

2.3.4. Robinson's model of attention

Building on Schmidt's (1990) hypothesis that learning requires noticing, Robinson (1995, 2003) put forward a model to explain the nature of the attentional mechanisms involved in noticing and their relationship to memory. Robinson (1995) pointed out that Tomlin and Villa's (1994) detection function is comparable to Schmidt's (1990) concept of noticing in that they both indicate the attentional level required for learning, although noticing, unlike detection, requires awareness. In Robinson's model of attention, awareness is achieved when the information detected and stored in short-term memory is activated beyond a certain threshold. This activation is then triggered by an increase of the attentional resources allocated voluntarily by a central executive to perform a specific task.

In an attempt to reconcile Schmidt's and Tomlin and Villa's positions on the need for conscious awareness in learning, Robinson (1995) proposed that noticing happens when detection, i.e., attentional allocation due to task demands, is accompanied by rehearsal in short-term memory. Robinson (2003) further elaborates on the distinction between the two processes by describing noticing as "selective focal attention and rehearsal in working memory" and detection as "recognition outside of awareness in passive short-term memory" (p. 655). The nature of the rehearsal and elaboration of

the detected information depends on whether the task stimulates data-driven processing or conceptually-driven processing. Data-driven processing, intended as the accumulation of many instances of small portions of the input, leads to *maintenance rehearsal*, whereas conceptually-driven processing or the interpretation of encoded stimuli according to mental rules or "schemata" triggers *elaborative rehearsal* (Robinson, 2003). Robinson (2003) points out that the general agreement is that noticing is facilitative, if not necessary, for learning, which calls for further research into pedagogical approaches aimed at promoting noticing, such as input enhancement and focus on form.

Applying Robinson's model of attention to the scenario outlined in subsection 2.3.3, his definition of *detection* would imply that the learner attends to the spoken dialogues in the video without focusing on linguistic form and, by successfully processing auditory word forms for meaning comprehension, unconsciously assimilates knowledge regarding a diversity of novel and previously encountered phonological aspects. If, however, they have been recently introduced to a phonemic contrast in class (e.g., /iː/ - /ɪ/), or they simply start observing a specific pattern regarding these vowel sounds, the rehearsal of the corresponding words in working memory would result in *noticing*. When a linguistic feature is noticed following instruction of the relative rule, it is likely to trigger *elaborative rehearsal*. Conversely, the spontaneous accumulation of word stimuli containing the /iː/ - /ɪ/ contrast in short-term memory would lead to *maintenance rehearsal*. Notably, further processing of the auditory input beyond the detection stage may also occur without awareness, when watching video clips for recreational purposes. However, the incidental development of pronunciation skills through exclusively meaning-focused input processing may

require regular exposure to substantial amounts of video content over extended periods of time.

2.3.5. VanPatten's model of input processing

VanPatten's (1996, 2004) model of input processing also examines the effects of attention allocation in SLA, focusing on the early stages of the conversion of input into intake (Leow, 2015). In line with the models described above, the learner's mind is seen as a limited capacity processor that, when exposed to L2 input, can only process a subset of this input. Differently from Robinson (1995), however, VanPatten (2004) considers both noticing (conscious) and perception (unconscious) as the registration of a form which does not necessarily imply the mapping of this form to its meaning, and postulates that, in the absence of a form-meaning connection, noticed linguistic data does not undergo further processing. Input processing is regulated by a series of principles, all derived from the Primacy of Meaning Principle, which states that learners first engage with the input to decode its meaning and only then process linguistic form, and the First Noun Principle, according to which learners assign the role of subject/agent to the first noun or pronoun in a sentence (VanPatten, 2004). Here we will focus on the subprinciples derived from the Primacy of Meaning Principle, in light of its implications for the creation and implementation of input-based learning activities.

The Primacy of Content Words Principle describes learners' tendency to process content words before function words and other elements of the input, and the Lexical Preference Principle points out learners' preference for lexical items rather than grammatical form when both encode the same meaning. In other words, when exposed

to the sentence "Yesterday, we walked to the beach", learners are expected to process the semantic notion of "past" by paying attention to the adverb, rather than the grammatical past tense marker *-ed*, regardless of whether they voluntarily pay attention to language in order to extract meaning from the input. The Preference for Nonredundancy Principle and the Meaning-Before-Nonmeaning Principle describe learners' preference for nonredundant grammatical forms carrying semantic information that cannot be expressed through lexical forms, and meaningful grammatical forms, such as English progressive marker *-ing* to express the on-going nature of an actions, as opposed to, for example, noun-adjective agreement in Spanish when the corresponding noun is an inanimate object (e.g., "El libr*o* antiguo"). To explain how learners still process some of these "unnecessary" forms, the Availability of Resources Principle postulates that redundant and nonmeaningful forms can be processed if meaning comprehension does not drain processing resources. This principle is in line with Krashen's (1982) notion that increased comprehensibility, associated with slow L2 speech made up of short sentences containing high frequency vocabulary, enhances acquisition, although the end product of VanPatten's (1996, 2004) input processing model is not acquisition but, as specified above, the initial stage of input processing that involves the connection of a grammatical form with its meaning.

The implications of VanPatten's (1996, 2004) model for this dissertation primarily relate to the selection of learning materials and enhanced target features for the research studies. The Primacy of Meaning Principle implies that unless learners can comfortably understand the language used in the video clips, they are unlikely to process linguistic form. In particular, as the initial stage of language learning involves

connecting form and meaning, it is crucial to ensure that learners are familiar with the meaning of the target words, before encouraging the processing of their auditory form. Most of VanPatten's (2004) input processing principles seem to indicate that learners might not spontaneously focus on the pronunciation of English regular past <-ed>, which is the target of two studies in this dissertation, due to its low salience compared to time adverbials. However, in line with the Availability of Resources Principle, redundant forms can be successfully processed if meaning comprehension is not overly taxing on the learner's attentional resources.

2.3.6. Leow's model of the L2 learning process in instructed SLA

Leow's (2015) model draws on VanPatten's (1996, 2004) model and other models of attention described above (Robinson, 1995, 2003; Schmidt, 1990, 1995; Tomlin & Villa, 1994) to offer an account of the cognitive processes specific to the instructed L2/FL learning context. In Leow's (2015) model of L2 learning in instructed SLA, the learning process is broken down into three stages, during which linguistic information is processed and produced until it becomes new L2 knowledge: The input processing stage, the intake processing stage, and the knowledge processing stage (Figure 2.5). In each of these stages, attention allocation is determined by depth of processing, cognitive registration and/or awareness of the new information in the input.

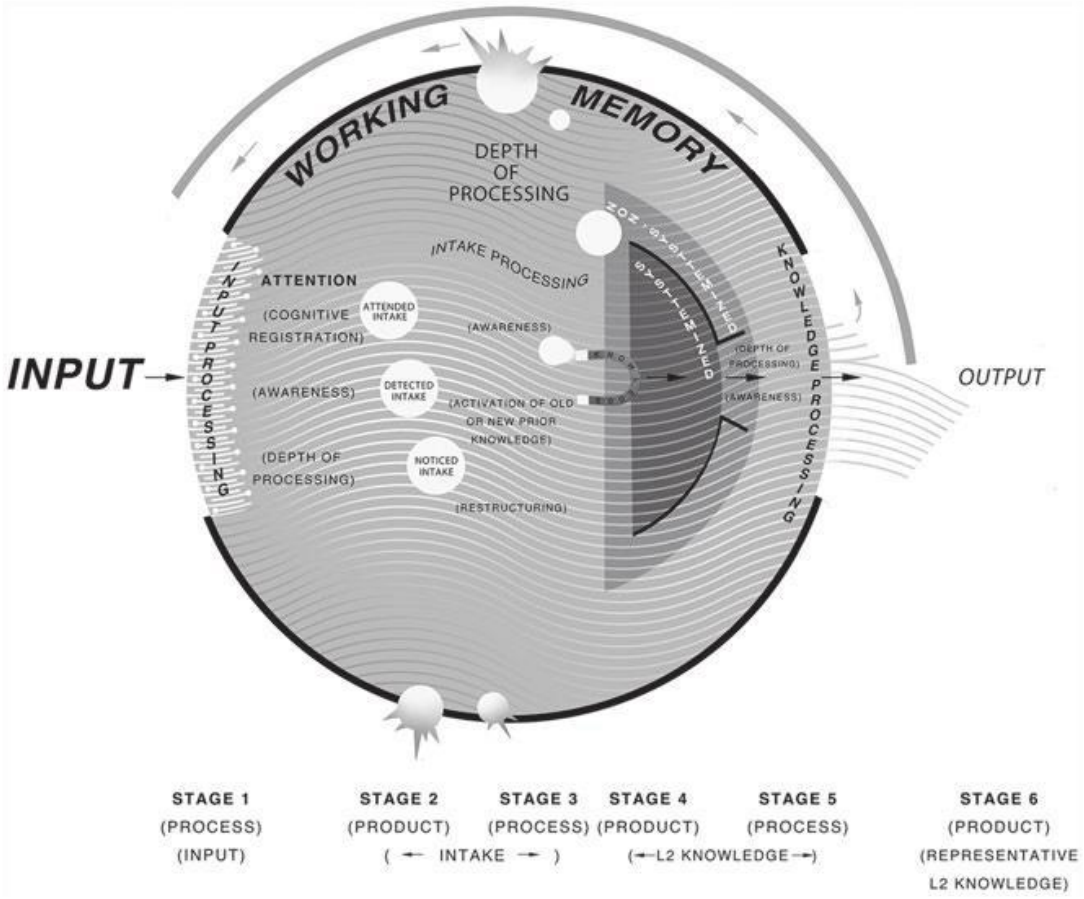


Figure 2.5. Leow's (2015) model of the L2 learning process in instructed SLA

The level of attention paid to new information plays a crucial role during *input processing*, when perceived information in the input is converted into intake and stored in working memory. Peripheral attention unaccompanied by higher levels of processing, cognitive registration, or awareness of the linguistic data generates *attended intake* that is usually discarded from working memory (Leow, 2015). *Detected intake* is also characterized by a very low level processing, but it is associated with some amount of selective attention, resulting in cognitive registration in the absence of linguistic awareness. Storage in working memory is more likely for detected than attended intake, but may depend on working memory capacity and on the level of processing allocated to detected intake. The input processing phase most

likely to generate intake that remains available for further processing is *noticing*, i.e., the selective attention allocation to linguistic data which leads to cognitive registration in the presence of some awareness. Detected and/or noticed intake may remain stored in working memory long enough to be tested receptively (recognition tests), but, in the absence of further processing, it is destined to fade away.

Depending on depth of processing and cognitive effort allocation, *intake processing* can take the path of data-driven processing or conceptually-driven processing. *Data-driven processing* initially involves the non-systemized accumulation of instances of a linguistic feature, in line with the notion of maintenance rehearsal proposed by Robinson (2003). Repeated reactivation of this un-systemized data through further exposure to related exemplars may be accompanied by lower or higher levels of processing and awareness, resulting in implicit or explicit restructuring and systemized learning, respectively (Leow, 2015). Implicit learning requires a longer period of time and a larger number of exposures compared to explicit learning, which, in turn, is driven by mechanisms that involve greater cognitive effort, such as hypothesis testing and rule formation. The higher depth of processing associated with explicit learning, however, is only effective when "awareness at the level of understanding" is achieved, i.e., when the learner is able to identify the correct underlying rule and master it. Both implicit and explicit processing are facilitated by multiple exposures to the target linguistic elements in meaningful contexts and by the provision of opportunities for meaningful practice. Data-driven learning, also known as item learning, can be tested through recognition or "simple controlled production assessment", such as fill-in-the-blanks for target grammar features (Leow, 2015, p. 243). *Conceptually-driven processing* involves a higher level of cognitive effort, just

as Robinson's (2003) elaborative rehearsal mechanism, and is facilitated by a high level of awareness that maintains the new information in working memory. The key element of this type of processing is the connection of novel linguistic data with prior knowledge, which assists the encoding and decoding of preliminary intake and ultimately its incorporation into the learner's systemized grammatical system. For example, L2 Spanish learners encountering the verb *morir* (to die) in its third person singular form *muere* may notice a vocalic stem change similar to the one in the irregular preterit form *murió*. Over time, any type of data-driven and conceptually-driven intake processing is expected to require lower levels of awareness and depth of processing, thanks to the automatization of prior knowledge activation for the same linguistic elements.

As a result of intake processing, then, two types of product are stored in the *L2 developing system*: Un-systemized chunks of language deriving from minimal data-driven processing, and systematized data that has undergone internalization and restructuring. The following stage, *knowledge processing*, involves the elaboration of knowledge accumulated in the L2 developing system with the aim to produce auditory and written output. Deeper processing, higher levels of awareness and the ability to activate the relevant linguistic data may all contribute towards the achievement of accurate and fluent *L2 production*, which implies the timely and target-like activation of L2 representations. Leow summarizes the fundamental characteristics of his model in a 13-point list:

1. The postulation that it is not the limited attentional capacity that is responsible for any potential breakdown in processing the L2 (at the input, intake and knowledge

processing stages) but learners' limited processing capacity, hence the potential roles of depth of processing and awareness at all three processing stages.

2. The postulation of three phases of intake that may be taken into learners' working memory.

3. Awareness does not play an important role at the input-to-intake stage.

4. All phases of intake may disappear from working memory unless further processed.

5. The shift in the centrality of both attention and awareness to the role of depth of processing taking place in the intake processing stage.

6. Higher depth of processing may lead to higher levels of awareness.

7. Activation of two types of prior knowledge (old and new).

8. High depth of processing does not necessarily lead to awareness at the level of understanding.

9. Learning occurs in the internal system.

10. Both implicit and explicit learning are possible, even during the same exposure, with the former dependent upon specific conditions.

11. There are two types of learning: item learning and system learning.

12. The view of the L2 learning process as both processes and products, and

13. This representation of the learning process is not viewed as linear given that learners' output may also serve as additional input. (Leow, 2015, p. 246)

Leow's (2015) model, which holds particular significance for this dissertation due to its focus on the instructed language learning context, will set the grounds for the formulation of the research questions and for the discussion of the results obtained. To illustrate how the attention mechanisms described above might contribute to pronunciation learning through exposure to multimodal input, let us revisit one last time the scenario of a learner watching a TV series episode. Although plenty of auditory and visual stimuli reach the learner's senses, many details regarding background elements (such as the pictures hanging behind the characters in their house as they speak in the foreground) tend to be peripherally attended to and readily discarded from working memory (*attended intake*). If the learner only scans captions quickly as an additional support to understand the dialogue, they may not notice the presence of filler words such as "you know" and "well" (*detected intake*). However, when the learner's attention is drawn to specific words in captions, for example due to listening comprehension issues, the *noticed intake* is likely to remain available for further processing. For example, upon hearing the word "pitch" multiple times in subsequent episodes of a TV series, the learner may successfully differentiate it from the word "peach" and map its form onto its meaning thanks to the un-systemized accumulation of knowledge regarding its pronunciation (*data-driven processing*). If, on the other hand, the learner focuses on words like "pitch" and "peach" due to previous instruction on the /iː/ - /ɪ/ vowel contrast, they are likely to apply their knowledge of the underlying pronunciation rule to decode preliminary intake, resulting in *conceptually-driven processing*. Over time, extended and repeated exposure to this feature may lead to the restructuring of the learner's L2 developing system, with positive effects on the perception and production of the target feature.

2.3.7. Summary

To sum up, this section has reviewed a number of SLA models of attention, highlighting that paying attention to input is essential for learning and that noticing, intended as a mechanism of selective attention leading to cognitive registration, is necessary for the restructuring and systemization of L2 knowledge. Due to the challenges associated with tapping into learners' subjective experience of linguistic data, and despite recent methodological advances that offer increasingly sensitive measures of awareness (e.g., stimulated recall, eye-tracking, EEG), to date the role of awareness in SLA remains controversial (Robinson, 2012). Some of the pedagogical and methodological implications derived from the models of attention and learning described in this section are that (a) noticing a target feature in the input supports learning, whereas attended or detected intake generated through lower level processing is less likely to remain available for further processing; (b) a conscious focus on linguistic form guided by prior knowledge may be more effective for learning, but it requires greater cognitive effort than unaware data-driven processing, as it goes against learners' tendency to process meaning first; (c) repeated exposures to exemplars of the target feature and sustained practice embedded in meaningful activities lead to more automatic processing; (d) since more fluent and accurate L2 production and perception require timely and appropriate activation of L2 mental representations, learners' progress in L2 processing skills can be estimated with tests tapping into the speed and accuracy of access to linguistic representations.

2.4. Input enhancement

2.4.1. Introduction

The previous section has illustrated the attentional mechanisms involved in the selection of input, the maintenance of selected intake in working memory and its connection with previous knowledge stored in long-term memory. In this section, we will address the question of why certain aspects of the input are more likely to be perceived than others, depending on various factors such as the physical properties of the input, its potential to convey meaning, and the perceiver's intentional allocation of attention. We will discuss the concept of *salience*, the property by which some elements stand out from the context and are more likely to be attended to (Ellis, 2017), and explore how salience can be manipulated through interventions such as textual enhancement to improve language learning outcomes. The review will first describe the concept of form-focused instruction proposed by Ellis (2012, 2016), tracing its development from the concept of focus on form formulated by Long (1991). Then, it will deal with input enhancement, a specific type of form-focused instruction grounded in the framework developed by Sharwood Smith (1991, 1993). To conclude, it will review a selection of SLA studies, narrowing down the focus on *textual* enhancement, and on the linguistic target relevant to this dissertation, L2 pronunciation/auditory word form.

2.4.2. Form-focused instruction

The term *focus on form* (FoF) was coined by Long (1991) to describe an approach in which the learners' attention is drawn to formal linguistic aspects while they use the target language communicatively. Focus on form involves briefly directing learners'

attention to language elements in context as they arise during meaning-focused lessons, aiming to promote *noticing* of forms (Schmidt, 1990), even without immediate comprehension of their meaning or function (Long, 1997). Therefore, the language focus of each lesson is not selected a priori, but rather emerges through the provision of reactive feedback, i.e., in response to linguistic issues affecting communication during primarily meaning-based tasks. Long (1991) argued in favor of the adoption of FoF and against: a) a focus on forms (FoFs) approach involving rote learning of grammar rules sequenced according to their linguistic complexity, and b) a focus on meaning (FoM) approach in which L2 acquisition is expected to occur spontaneously as a result of meaning-focused communicative practice.

Over time, a separate research strand emerged in which it was accepted that a focus on form can also be achieved pre-emptively (as opposed to reactively) and non-interactively within a communicative lesson, by directing learners' attention to specific linguistic elements through pre-planned instruction or awareness-raising activities (Spada, 1997). The concept of form-focused instruction (FFI) was used by several authors, notably Spada (1997) and Ellis (2001), to refer to any intervention aimed at drawing learners' attention to form, regardless of whether the language is the object of *intentional* learning activities, such as rule-based controlled practice activities, or, following Long's (1991) original notion of FoF, a tool used to carry out communicative tasks that may result in *incidental* learning. As such, FFI includes both FoFs and FoF, bringing together traditional approaches where each lesson in the syllabus has a pre-planned linguistic target with communicative approaches where the linguistic focus is a consequence of meaningful interaction. As Ellis (2012, 2016) points out, it is difficult to define FFI classroom interventions purely as FoF or FoFs,

since actual learner performance of an activity may not even match teachers' intentions. It is not unusual that, within the same lesson, learners may initially interpret a FoF activity as FoM, have their focus redirected to forms by the teacher (FoFs), and end up carrying out controlled practice activities to ensure that some attention has been paid to linguistic form.

Ellis (2012) proposes classifying classroom activities in terms of the methodological procedures used in their design and implementation. Figure 2.6 shows a visual representation of FFI options, with the types of activities that were used in one or more studies in this dissertation highlighted in bold. Among the *proactive* techniques, that is, those involving the performance of linguistic tasks on the part of the learner, the author distinguishes between consciousness-raising options, corresponding with direct (deductive) or indirect (inductive) explicit instruction, and language-processing options aimed at inducing processing of target L2 features through exposure to input or by prompting the production of output containing these features. In particular, input-based instruction can be exposure-based, if learners simply read or listen to L2 passages in interventions involving input flooding or input enhancement, or response-based if learners are required to show that they have processed the input, as in structured-input activities requiring a non-verbal or minimally verbal response. Production activities can be placed on a continuum in the distinction between text-manipulation and text-creation, according to whether they involve minimal changes to a text, such as fill-in-the-blank, or the creation of original content, such as writing sentences in response to a prompt asking for the learner's opinion. The studies conducted within this dissertation will either exclusively feature input-based activities involving input enhancement or combine input-based activities with production-based

activities involving the manipulation of auditory input and written text. The rest of this section will focus on textual input enhancement, a semi-incidental FFI technique that can direct learners' attention to linguistic form during primarily meaning-focused activities such as reading a book (Pellicer-Sánchez & Boers, 2019).

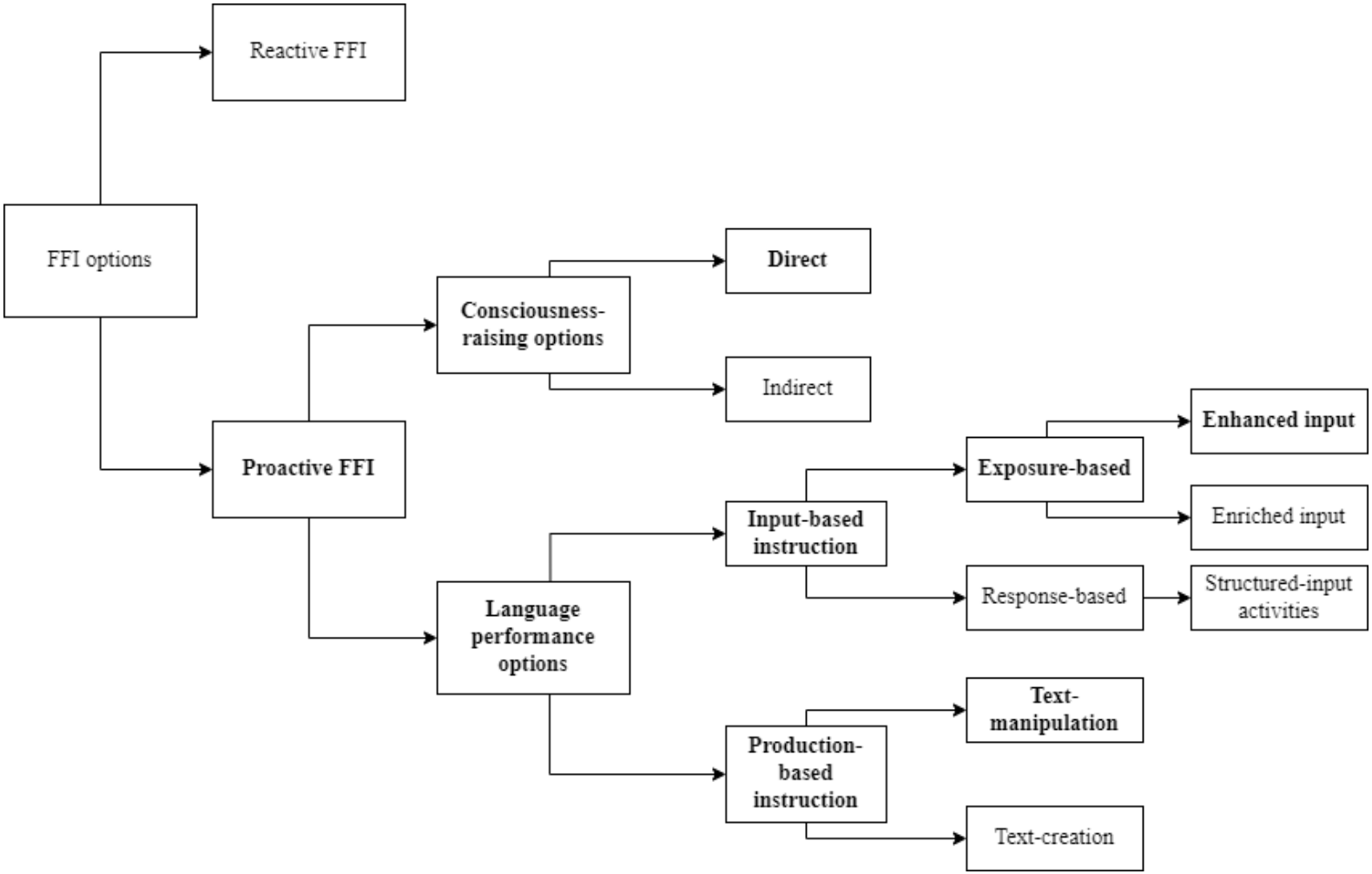


Figure 2.6. Ellis's (2012) methodological classification of FFI.

2.4.3. The input enhancement hypothesis

Sharwood Smith's (1981, 1991, 1993) input enhancement hypothesis aimed at explaining why learners pay attention to specific elements in the input, and what types of input manipulation can successfully direct learners' attention to linguistic form. Sharwood Smith (1981) originally used the term *language consciousness-raising* to refer to teacher- or learner-initiated instances of focus on linguistic form with varying degrees of explicitness and elaboration. According to Ellis (2012), consciousness-raising as initially described by Sharwood Smith was a form of explicit instruction aimed at facilitating the rule-formation process resulting in explicit knowledge and was therefore compatible with a focus on form*s* approach. However, Sharwood Smith (1991) abandoned the notion of consciousness-raising due the controversy surrounding the definition and measurement of consciousness, which is an ambiguous term that can refer to awareness, intention or knowledge (Schmidt, 1990). In addition, he pointed out that even when teachers try to direct learners' attention to some target features, these may not be perceived or processed any further than at a perceptual level, in line with the hypothesis that not all input becomes intake (Corder, 1967). As a consequence, Sharwood Smith (1991, 1993) adopted the term *input enhancement*, shifting the focus from the learners' internal processes to the manipulations the teacher can apply to the input, without assuming that these manipulations will necessarily assist acquisition. In fact, Sharwood Smith (1991) acknowledges that the perceptual salience created *externally* through visual or aural enhancement of target features can sometimes clash with the *internal* mechanism by which learners naturally pay increased attention to a specific linguistic feature once they have become ready to learn it.

Sharwood Smith's dichotomy between externally and internally created salience presents analogies with Chun et al.'s (2011) categorization of attentional mechanisms as belonging to an external or internal system, depending on the source and nature of the information that needs to be processed. In the context of reading a text where single words are highlighted, external attention would be responsible for the selection and further processing of enhanced written words, provided the reader perceives them as more visually salient than the adjacent unenhanced words. In the same context, internal attention refers to the encoding and rehearsal in working memory of phonological, orthographic, and/or semantic information that is guided by previously stored information, such as the learners' previous knowledge of the words in the text and the instructions received before starting the task (Chun et al., 2011). While there is evidence that external manipulations such as textual enhancement can affect learners' eye movements, a reflection of overt attentional allocation (Issa & Morgan-Short, 2019), language development can only take place if internal attention is also directed to the target features (Chung & Révész, 2021). In other words, paying attention to a word or phrase, i.e., its *selection* among competing stimuli, does not guarantee *modulation*, defined as successful processing resulting in improved accuracy on behavioral tests (Chun et al., 2011). The main implication for this dissertation is that relying on a single data collection method, even one as advanced as eye-tracking, provides limited insights into the learner's depth of processing of enhanced target words. The analysis of eye-tracking data may serve to gauge whether visual attention was directed towards enhanced words (attended intake or *selection*), and whether the enhanced words were fixated for longer compared to the typical fixation duration on unenhanced words, which could reveal detection or noticing

(Godfroid, 2019; Kruger, 2013). However, triangulating data from language tests and verbal reports can reveal whether increased attention leads to *modulation*, facilitating further processing of the enhanced target features.

2.4.4. Key aspects of research on input enhancement

The complexity of operationalizing learners' level of processing of externally manipulated features is reflected in the variety of research designs adopted in studies on input enhancement. A first important distinction was made by Leow (2009), who categorized input enhancement studies as *non-conflated* if learners are only exposed to enhanced input (mainly through the enhancement of words in reading passages), or *conflated* if other learning opportunities are also provided through explicit instruction or teacher feedback. Although conflated input enhancement research presents the disadvantage that the effects of enhancement cannot be teased out independently from those of feedback or instruction, this type of enhancement has been consistently found to promote superior cognitive processing and language development compared to non-conflated input enhancement (Han et al., 2008; Leow, 2009). Non-conflated input enhancement studies have consistently reported mixed results or non-significant language gains (Leow & Martin, 2017), possibly due to factors such as the target feature's salience, the learner's previous knowledge of the target feature and their overall L2 proficiency (Ellis, 2016). For example, learners are more likely to notice the regularities emerging from the enhanced exemplars of a feature they have at least partially acquired, even if the relevant knowledge has fossilized and they do not notice errors in their own production (Han et al., 2008; Sharwood Smith, 1993). Lower proficiency learners, however, tend to devote most of their cognitive resources to meaning comprehension while reading or listening and may therefore struggle to

simultaneously process the formal aspects highlighted through input enhancement (Ellis, 2016).

Besides the issues regarding the nature of the enhanced linguistic feature and the learners' L2 proficiency mentioned above, other factors may determine whether input enhancement leads to language development (Leow & Martin, 2017). In particular, input enhancement studies traditionally measured learners' accuracy gains after a reading task, typically through grammaticality judgement tasks, rather than how they respond to the enhancement during the task (Leow, 2009). Recently, however, there has been an increase in studies measuring not only offline performance but also online processing through eye-tracking (e.g., Lee & Révész, 2020; Winke, 2013) or verbal reports, in which learners are asked to verbalize their thoughts during or after exposure to input enhancement (e.g., Leow, 2001; Leow et al., 2019). While concurrent think-aloud protocols have offered valuable insights into the learners' cognitive processes, this methodology was criticized due to potential reactivity, i.e., the risk that the verbal recall interfered with the primary cognitive processes required for the main task (Leow, 2015). Retrospective think aloud protocols, on the other hand, require learners to verbalize their thoughts after the main task has been completed. In a specific type of retrospective protocol called *stimulated recall*, a videotape or voice recording of the learners' own performance is shown to them immediately after the task to support the recall of their thoughts during the task, with the aim of increasing the validity and reliability of the information elicited (Gass & Mackey, 2000). Although only few textual enhancement studies have combined offline language tests with eye-tracking and stimulated recall, the results show that enhancement influences learners'

processing of target forms, although its effects on language development may not be captured by immediate post-tests (Lee & Jung, 2021; Jung et al., 2022).

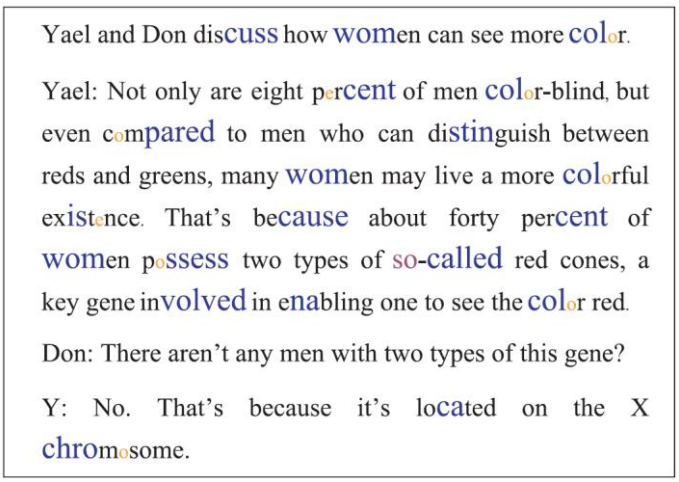
The analysis of eye tracking data is largely based on the eye-mind assumption, which interprets fixation location as the focus of cognitive processing and longer fixation duration as deeper processing and higher cognitive load (Godfroid, 2019; Kruger, 2013). However, the discrepancy between gaze data and test outcomes is in line with Leow and Martin's (2017) observation that even when eye-tracking data shows substantial attention directed to the enhanced forms, language development from simple exposure to input enhancement is rarely registered. Leow and Martin (2017) explained this finding by pointing out that attention (*what* is noticed) is less important than awareness (*how* noticed input is processed) and recommended that learners should be encouraged to engage in higher level processing (such as formulating rules based on the enhanced exemplars) rather than shallower, perceptual processing. However, it must be noticed that the grammaticality judgement tasks used in many input enhancement studies may fail to register the data-driven processing of a linguistic feature, i.e., the un-systemized accumulation of occurrences of this feature without a deliberate attempt to deduce the underlying rule. Subsection 2.4.5 presents an overview of studies on textual input enhancement focusing on L2 pronunciation learning and the mapping of auditory word forms onto written forms. This review will also delve into research design and other methodological issues in light of their impact on assessing the effectiveness of input enhancement.

2.4.5. Studies on textual enhancement and pronunciation learning

In early studies on textual enhancement, the most common modality of exposure was through printed text (e.g., Leow, 1997; Leow, 2001; Shook, 1994; Simard, 2009) and the L2 target feature was usually grammatical or, less frequently, lexical. A handful of recent textual enhancement studies have featured multimodal input from videos, but the focus has remained on grammar and vocabulary learning (e.g., Cintrón Valentín et al., 2019; Cintrón Valentín & García-Amaya, 2021; Lee & Révész, 2020; Winke, 2013). To the best of our knowledge, only a limited number of textual enhancement studies have focused on L2 pronunciation acquisition or on the development of phonological awareness. Stenton (2012, 2013) proposed visually annotating lexical stress in reading-while-listening to promote the acquisition of spoken intelligibility, whereas Showalter (2019) and Alsadoon and Heift (2015) have used textual enhancement to increase learners' awareness of L2 sound-to-symbol mappings during exposure to word or sentence stimuli.

Stenton (2012, 2013) reports on the results of the project SWANS (Synchronised Web Authoring Notation System), aimed at improving the intelligibility of L1 French learners of English through reading-while-listening. Figure 2.7 provides an example of text annotated for lexical stress, where primary stress is indicated in blue, secondary stress in purple and reduced vowels in orange. This type of audio-synchronized textual enhancement was designed to help learners identify errors in the phonological representation of L2 words as they subvocalize the text, by disrupting automatic reading patterns based on L1 sound-symbol mappings (Stenton, 2012; cf. Woore, 2018). In addition, synchronizing the enhancement line-by-line as each line was spoken in the soundtrack was assumed to reduce cognitive load and increase ocular

comfort by directing learners' attention to the relevant linguistic information at each point in time. The decision to visually represent sound through the enhancement of orthographic information and not, for example, through graphic models of sound waves obtained with Praat, was based on learners' familiarity with written word representations as well as the comparative ease with which enhanced texts could be generated by teachers (Stenton, 2012, 2013).


Yael and Don discuss how women can see more color.

Yael: Not only are eight percent of men color-blind, but even compared to men who can distinguish between reds and greens, many women may live a more colorful existence. That's because about forty percent of women possess two types of so-called red cones, a key gene involved in enabling one to see the color red.

Don: There aren't any men with two types of this gene?

Y: No. That's because it's located on the X chromosome.

Figure 2.7. Example of text manually annotated through SWANS (Stenton, 2013, p. 153).

To test the SWANS methodology, a number of interventions spanning 10 weeks were carried out with several hundreds of students in French high schools and universities. The training sessions consisted in viewing a video of a fluent French learner of English speaking English with inconsistent realization of lexical stress, and carrying out form and meaning focused activities with various levels of explicitness, thus described by Stenton:

1. Reading out loud from the unannotated script in pairs, with mutual correction.

2. Annotating scripts on paper after listening to the sound, with self-correction.

3. Carrousel activities: repeated 3-minute oral summaries on research topics in constantly changing pairs. The same presentation is made 4 times to different partners. After the first presentation the student stops worrying about content which frees brain resources for concentrating on spoken form.

4. Distance teacher correction of student annotated keywords, followed up by "PowerPoint" oral presentations in class.

5. EXPLICS Internet case studies where oral performance was continually monitored (Website managed by CercleS in Goettingen).

(2012, p. 222)

Learning gains were measured perceptually through multiple-choice listening tests on isolated words and words in context, as well as in production through recording individual words and phrases. The learners' production was also assessed by analyzing a 3-minute classroom presentation based on a script in which the lexical stress of selected keywords had been annotated by the student and corrected by the teacher. Although English lexical stress was confirmed to be a problematic feature for learners at different proficiency levels, students' perception and controlled production improved after the intervention. However, the gains did not transfer to spontaneous speech, suggesting that longer exposure may be needed in order to restructure fossilized phonological representations. Finally, despite initial concerns about the cognitive demands of multimodal, synchronized textual enhancement, students reported a high level of satisfaction with the SWANS multimodal learning

environment and preferred it to unimodal techniques involving the use of audio and text separately (Stenton, 2012).

Reading-while-listening research has also investigated the use of textual enhancement synchronized with auditory input to support children's development of L1 reading skills (e.g., Gerbier et al., 2018). Based on extensive piloting, Gerbier et al. (2018) hypothesized that highlighting each word in a static text (i.e., presented on the screen all at once) 300 ms ahead of a voice reading the text would support L1 children's reading comprehension and incidental learning of new words. Therefore, they visually highlighted each word 300 ms before its auditory onset, providing a word-by-word, stricter audiovisual alignment compared to Stenton (2012), who synchronized the enhancement line-by-line. The results showed that the synchronized word enhancement had no effect on incidental orthographic learning of pseudowords embedded in the texts, and a detrimental effect on the memorization of the pseudowords' semantic category. However, poorer readers reported a preference for the synchronized modality over the unsynchronized modality, possibly because, as revealed by eye-tracking data, the audiovisual synchronization helped them keep up with the voice reading the text. On the other hand, children with higher reading proficiency were more likely to prefer the unsynchronized modality, as the absence of enhancement allowed them to process the text at their naturally faster reading pace, irrespective of the synchronization with the auditory input. Despite having a learning target unrelated to L2 pronunciation, Gerbier et al.'s (2018) study offers initial support to the use of textual enhancement synchronized at the word level and offers insights into the variables that may affect learners' preference for this exposure modality.

L2 pronunciation was the focus of Showalter (2018, 2019), who included a textual enhancement component into her studies on the effects of orthographic input, specifically grapheme-phoneme correspondence congruence and grapheme familiarity, on the processing of Russian L2 phonological word forms by English L1 speakers. In Showalter (2018), naïve participants were trained through exposure to the auditory form of target nonwords, together with a picture and either a meaningless sequence of graphemes (no orthography condition) or a Cyrillic word form (orthography condition). In the testing phase, they were asked to match the auditory form of each nonword to the corresponding image. Incongruent grapheme-phoneme correspondences (e.g., <PAT> pronounced [rɑt]) were found to interfere with the development of phono-lexical representations in naïve learners, with no effect of grapheme familiarity. To investigate whether focusing their attention to grapheme-phoneme correspondences would benefit the L2 phono-lexical development of naïve participants, Showalter (2019) included two training interventions (textual enhancement with or without explicit instruction). As a control, different naïve participants as well as beginner and experienced learners were exposed to the same nonword stimuli under an orthography or no orthography condition with no enhancement. The results obtained by Showalter (2018) with naïve participants were replicated, and experienced learners were equally accurate on congruent and incongruent grapheme-phoneme correspondences, indicating that a longer experience with the language helped learners overcome the effects of orthographic input. No significant effects were registered in consequence of the brief interventions featuring textual enhancement, but there was a nearly significant difference between naïve participants who were explained grapheme-phoneme correspondences rules in

advance and experienced learners. The (descriptively) negative effects of instruction are tentatively explained in terms of cognitive overload, as having to remember phonological rules during meaning-focused exposure to the input may have overwhelmed learners' working memory. On the other hand, simple enhancement of target graphemes may have encouraged learners to attend more carefully to auditory input and select important information in the attempt to figure out phonological rules on their own.

Alsadoon and Heift (2015) compared the effects of textually enhanced and unenhanced unimodal input on the noticing and decoding of English vowels by native speakers of Arabic, a language in which vowel sounds convey redundant information and vowel graphemes are represented by non-salient diacritics. In the learning phase, the participants' eye movements were recorded as they read short sentences in which one target word was underlined and its vowel graphemes were bolded and colored. The control group read the same sentences without enhancement. In this phase, participants only selected each word's meaning from three word choices and received feedback, whereas in the pre-, post- and delayed post-tests they also had to recognize accurate orthographic word forms, choosing among three competing options with the same consonantal structure (e.g., "wanter, winter, wentir" for *winter*). Despite achieving ceiling performance on word meaning recognition, all participants performed poorly on written word form recognition at pre-test, confirming that English vowel decoding was affected by the interference of L1 Arabic processing strategies based on consonantal cues (Alsadoon & Heift, 2015). The significantly higher word form recognition scores of the experimental group at immediate and delayed post-test, and their strong positive correlation with refixation duration and

total fixation duration pointed at a beneficial effect of textual enhancement. By using high frequency words that were familiar to the participants, this study promoted a sequential processing of form after meaning, which is expected to be more beneficial than simultaneous form and meaning processing for L2 grammar acquisition (Han et al., 2008). However, Alsadoon and Heift (2015) acknowledged that the strong effects of textual enhancement observed after exposure to short simple sentences may not be replicated with longer and more complex texts, which tend to generate a heavier cognitive load and are therefore naturally associated to less explicit processing, due to the increased efforts devoted to meaning comprehension. In addition, due to the use of the same sentences for both the learning and testing phases, it is unclear whether the gains should be attributed to episodic memory or to restructuring of the phonological system. It remains an empirical question whether the learning gains observed in the recognition of familiar written word forms after exposure to short sentences would have generalized to non-familiar words, especially if these were enhanced in longer texts and tested in the auditory modality.

2.4.6. Summary

Overall, input enhancement has been shown to be an effective way to direct learners' attention to form during meaning-focused activities, but its effectiveness may depend on factors such as the inclusion of instruction and feedback in the input enhancement intervention, the learners' L2 proficiency, and their previous knowledge of the enhanced target feature. Research on textual enhancement and L2 pronunciation learning is scarce and has produced mixed or inconclusive results, possibly due to the challenges that learners face when integrating visual and auditory modalities, as well as methodological difficulties associated with operationalizing learning outcomes. A

range of different measures have been used to examine the effects of textual enhancement, including analyzing readers' eye gaze behavior, assessing awareness of target feature pronunciation rules, and evaluating pronunciation gains achieved in offline tests. To the best of our knowledge, no study has used textual enhancement to direct the viewers' attention to L2 pronunciation in the context of authentic multimodal input (such as TV programs), which is the focus of the research conducted in this dissertation. However, encouraging results have been obtained for other aspects of language learning through video exposure, such as grammar and vocabulary (e.g., Lee & Révész, 2020; Montero Perez et al., 2014; Pattemore & Muñoz, 2022). When exposed to captioned video with no instructions or manipulations, learners are naturally inclined to focus on meaning comprehension rather than linguistic form (Vanderplank, 2015). This dissertation will explore the effects of enhancing target words in the captions of video clips in synchrony with the words' auditory onset as a means to direct learners' attention to the auditory form of those words. The visual enhancement of words in captions is expected to interfere with the automatic generation of L1-biased phonological representations, encouraging the comparison of pre-existent representations and upcoming auditory information (Stenton, 2013). The next section will focus on the specific type of multimodal input used in this dissertation (L2 captioned video from TV series) and on its potential for the acquisition of various aspects of L2 speech.

2.5. Pronunciation learning through L2 captioned video

2.5.1. Introduction

The advent of streaming services like YouTube and Netflix has provided users with virtually unlimited content of their choice, in a format that allows them to rewind and replay videos, as well as add captions in several languages. A growing body of research has focused on L2 learning through exposure to captioned videos, finding evidence that watching movies and TV series is not only a popular extracurricular activity but that it also has a very positive impact on the development of learners' listening and reading skills (Lindgren & Muñoz, 2013). While much of the research on second language (L2) learning through exposure to captioned videos has focused on vocabulary and grammar acquisition, recent studies have found evidence of a positive impact on L2 pronunciation development. This section explores the potential of TV series as a source of comprehensible input and reviews relevant studies on audiovisual processing and pronunciation development through exposure to L2 captioned video.

2.5.2. TV series as a source of L2 input

TV series represent a rich source of input which offers opportunities for language learning inside and outside the classroom (Pujadas & Muñoz, 2019). First, the dialogues (scripts) of TV series intentionally simulate natural speech and may contain a fair amount of inter-speaker variability in terms of speech rate, accent and sociolinguistic variation (Ghia, 2012). In addition, speech comprehension can be facilitated by the availability of visual cues, the recurrence in subsequent episodes of themes and characters, and the repetition of L2 words and phrases (Rodgers & Webb,

2011). Finally, although this review will primarily focus on the linguistic factors that make TV series a good source of L2 input, it is likely that learners are naturally drawn to this type of input due to a range of psychological factors. For example, watching TV in the foreign language has been found to be more enjoyable and motivating than other classroom activities, with positive effects on the amount of attention paid by students in class and on their feeling of learning (Dizon, 2018; Pujadas & Muñoz, 2017).

The accessibility of TV programs for beginner and intermediate learners is related to their low vocabulary demands in terms of coverage (Webb & Rodgers, 2009). The coverage of a text refers to the percentage of words that are familiar to a learner and is directly related to the vocabulary size necessary to understand the text. Knowing the 3,000 most frequent word families is sufficient to achieve listening and reading comprehension of most TV programs and movies, because it provides 95% coverage of their script (Rodgers & Webb, 2011; Webb & Rodgers, 2009). To watch TV for pleasure, it is generally necessary to reach 98% coverage, which requires a larger vocabulary size of 5,000 to 10,000 word families plus proper nouns and marginal words (Webb & Rodgers, 2009). When selecting a TV series for a teaching intervention, matching the vocabulary demands of the script with the learners' vocabulary size ensures comprehension of the video and facilitates the incidental acquisition of words in the text. To further test the learning potential of authentic multimodal input from TV series, Rodgers (2013) conducted a series of studies with ten episodes of a TV series and found evidence that this type of input satisfies Nation's (2007) conditions for input to be suitable for a *meaning-focused* intervention. Specifically, the input provided should be abundant, interesting and familiar to the

learners both in terms of topics and language used, and it should contain context clues that, together with background knowledge, help learners guess the meaning of new words and expressions. Therefore, watching TV series potentially exposes language learners to large amounts of comprehensible input, intended as input slightly more complex than what they can comfortably understand at the current proficiency level (Krashen, 1982).

While integrating TV series into the language classroom and encouraging learners to watch undubbed TV out of school may boost language learning (Pujadas & Muñoz, 2019), viewing whole episodes during class time may be impractical in settings where the classroom contact is reduced to 2-4 hours a week, as in most European countries (Muñoz, 2008). There is evidence that language teachers have positive perceptions of L2 videos and use video clips from TV series, YouTube videos and TED talks quite regularly to teach specific linguistic aspects or increase the learners' overall listening comprehension (Alonso-Perez & Sánchez-Requena, 2018; Kaderoğlu & Romeu Esquerré, 2021). Therefore, it is not surprising that an increasing number of SLA studies have used shorter clips from TV shows to direct learners' attention to linguistics aspect in the input, e.g., by exposing learners to captioned video with enhanced target words, asking them to fill in missing words to complete the captions or dub short video excerpts (Lee & Revesz, 2020; Lima, 2015; Sánchez-Requena, 2018). In this type of *form-focused* interventions, the input needs to engage learners' deliberate attention to language features, enable deep processing of language features, allow for spaced and repeated processing of the same features, and focus on simple features that do not require advanced developmental knowledge (Nation, 2007). After determining that TV series can provide an engaging and effective source of L2 input

for language learners, it is now possible to delve into specific aspects of multimodal input processing, i.e., the effects of the availability of written information on learners' processing of auditory information.

2.5.3. Audiovisual processing of L2 captioned video

While the amount of attention paid to auditory and visual input when viewing L2 captioned video can vary greatly for different people under different conditions, research has revealed certain patterns that may apply to all learners or to specific groups. To begin with, reading captions is a largely incidental and automatic activity, as evidenced by the fact that most viewers tend to read captions regardless of the availability of the soundtrack, the viewer's proficiency and whether they usually watch subtitled videos or dubbed videos without captions (D'Ydewalle, 2002; D'Ydewaelle & De Bruycker, 2007; D'Ydewaelle & Van de Poel, 1999). The viewing experience can be affected by factors related to the material, such as the familiarity and complexity of the content of the video, and factors related to the learner, notably, their L2 proficiency and/or reading skills, working memory, language learning aptitude and viewing preferences (Montero Perez, 2022). In general, learners tend to pay substantial attention to the caption area when the content is unfamiliar and reliance on top-down processing is impaired, i.e., learners cannot resort to previous knowledge of the topic to achieve comprehension (Bisson et al., 2014; Winke, 2013). Learners may also more readily resort to captions when the video contains faster dialogues, as higher speech rate increases the complexity of the listening task and impairs bottom-up processing, i.e., the process of decoding and processing individual words and phrases in the captions (Ghia, 2012). In this case, the presence of captions may support the learners' listening process, preventing the cognitive overload

typically associated with the limited short-term memory capacity and dominant use of bottom-up auditory processing by low proficiency learners (Kruger & Steyn, 2014; Yang & Chang, 2014). However, reading captions may not aid comprehension, for example when the video contains a fast-paced narration on an unfamiliar topic and learners struggle to read captions fast enough to support listening comprehension (Mayer et al., 2014).

The analysis of individual differences can contribute to explaining the heterogeneity of audiovisual behaviors L2 learners exhibit when watching captioned video, and the amount of language learning gains they may achieve. While Montero Perez et al.'s (2013) meta-analysis found that L2 captions supports listening comprehension and vocabulary acquisition regardless of the viewer's L2 proficiency, learners with poor L2 reading skills (due to young age or limited proficiency) may not benefit from the presence of captions due to the speed of presentation of the text, which requires fairly automatic processing of written input (Muñoz, 2017; Winke et al., 2013). On the contrary, learners at an intermediate proficiency level have been consistently found to pick up new vocabulary and formal linguistic aspects from exposure to L2 captioned video with relative ease compared to learners with limited proficiency (Gesa, 2019; Pattemore & Muñoz, 2020; Pujadas & Muñoz, 2019; Rodgers, 2013). Therefore, watching L2 movies with L1 subtitles may be more effective to help learners move from the lowest proficiency levels to intermediate proficiency, when they will benefit the most from using L2 captions, finally switching to no captions as they achieve the highest proficiency levels (Araujo & Costa, 2013; Markham et al., 2001; Vanderplank, 2010).

In line with the hypothesis that L1 or L2 text may be more beneficial depending on the learners' proficiency level, Hutchinson and Dmitrieva (2022) found that L1 subtitles may support L2 pronunciation learning in monolingual speakers with no previous experience of the target language. In Hutchinson and Dmitrieva (2022), exposure to a 45-minute episode of a documentary with L2 French audio had a small but significant effect on L1 English speakers' pronunciation of French vowel /y/. The authors attribute the gains to the high visual saliency of the French high rounded vowel /y/ and to the considerable perceptual distance with its English counterpart /i/. On the contrary, the lack of improvement in the pronunciation of the vowel /u/ is associated with the confusion arising from perceptual assimilation to its L1 counterpart. Although the study by Hutchinson and Dmitrieva (2022) on pronunciation learning through L1 subtitling has yielded promising results, the small effect sizes hint at the difficulties encountered by beginner learners during exposure to authentic L2 video, as well as at the detrimental effects on L2 speech processing of constantly reading L1 subtitles to achieve a minimal level of comprehension. As this dissertation focuses on intermediate learners who have received formal L2 instruction for many years, L2 captions were used in the attempt to simultaneously facilitate speech comprehension and pronunciation learning.

While the role of proficiency in learners' audiovisual behavior has received some attention, the role of working memory and aptitude is still under-researched (Pattemore & Muñoz, 2020). The availability of captions has the potential to maximize working memory capacity thanks to the presentation of redundant on-screen text in short segments of one or two lines and the synchronization with auditory input, which facilitates sequential allocation of attention and the efficient integration

of information in different modalities (Kruger et al., 2013; Kruger & Doherty, 2016). In line with this hypothesis, there is evidence that, although a higher working memory may be crucial for learning L2 grammar from *uncaptioned* video, the presence of captions can reduce cognitive load and equalize learning gains among learners with varying working memory abilities (Pattemore & Muñoz, 2020). Interestingly, the effects of working memory may intertwine with those of viewing preference, as revealed by Kam et al.'s (2020) study of the effects of modality preference and working memory capacity on listening comprehension during exposure to captioned video. Among learners with high working memory but, notably, not among those with low working memory, the learners who preferred the visual modality had higher comprehension scores when they watched a video with captions, whereas auditory learners performed best without captions. Finally, the impact of aptitude in vocabulary and grammar learning through captioned video may be observed when the tests involve greater cognitive involvement and target more explicit aspects, but more research is needed to draw solid conclusions about the role of aptitude in language learning through exposure to captioned video (Pattemore & Muñoz, 2020; Suarez & Gesa, 2019).

Most of the findings on language learning through exposure to audiovisual input discussed so far come from research on vocabulary and grammar learning. In fact, research on language learning through captioned video has only recently focused on pronunciation learning, intended as the process of acquiring and improving the ability to produce and perceive the sounds and intonation patterns of a language (Montero Perez, 2022). The rest of this section will narrow down the focus to the specific target of this dissertation, i.e., the update of nontarget-like phonological representations of

known L2 words of which the learner has already developed a semantic representation. In this sense, although knowledge of the auditory form of a word is one of the different aspects involved in knowing a word (Nation, 2001), studies on vocabulary learning through exposure to L2 captioned video rarely provide relevant insights regarding pronunciation learning due to a number of factors. To begin with, most researchers have used written tests tapping into learners' recognition or recall of written word forms (Jelani & Boers, 2018), which gives very little information on the learners' knowledge of auditory word forms, given the discrepancies between L2 written and auditory vocabulary knowledge (Van Zeeland, 2017). In the studies that included testing of L2 auditory forms (e.g., Galimberti & Miralpeix, 2018; Pujadas & Muñoz, 2023; Suarez & Gesa, 2019), the aim was vocabulary teaching, so the target words were purposedly low frequency words that were expected to be unknown at pre-test, rather than words that present pronunciation issues for L2 learners. In addition, the vocabulary tests generally required learners to write down on a sheet of paper the orthographic form of a word after listening to it, which provided no information on whether they would be able to pronounce the word and/or on the relative speed and automaticity of lexical access. Therefore, our review will now focus on the small number of studies that to date have tested the development of pronunciation from exposure to L2 captioned video.

2.5.4. Studies on pronunciation learning through L2 captioned video

The methodology and results of the studies investigating pronunciation learning through L2 captioned video is summarized in Table 2.1. The samples are quite homogeneous, as most studies were conducted in European countries with local university students learning English as a foreign language (with the exception of the

L1 Chinese learners in Charles, 2017). While some studies have adopted a pre- and post-test design (Birulés-Muntané & Soto-Faraco, 2016; Charles, 2017; Mitterer & McQueen, 2009; Wisniewska & Mora, 2020), others consisted in a single-session assessment of L2 speech processing abilities and their correlation with learners' exposure to TV programs (Kusyk & Sockett, 2012; Wisniewska & Mora, 2018; Yibokou, 2023). This subsection provides a description of each study, focusing on its methodological approach and on the implications for the investigation of pronunciation learning through TV exposure.

Bird and Williams (2002) was the first study to the best of our knowledge to investigate the effects of written text availability on the processing and acquisition of auditory word forms. A total of 56 native and advanced L2 learners were exposed to familiar and unfamiliar word stimuli presented in one modality (sound or text) or bimodally (sound and text), then tested on implicit word recognition with a series of lexical decision tasks (LDT) or rhyme monitoring tasks, and, lastly, on explicit recognition with a memory decision test. The results showed that training in the bimodal presentation condition promoted both implicit and explicit auditory form recognition more than single modality presentation training, confirming the potential of captions to support bottom-up processing of the soundtrack. In contrast with the hypothesis that learners perform better when captions are available because they only read the captions and ignore the soundtrack, Bird & Williams (2002) found that the availability of text does not impair processing of auditory information and the use of captions can support implicit and explicit learning of L2 spoken word forms. Although this study has often been cited as evidence of the effectiveness of captions to support auditory processing and the acquisition of spoken word forms (e.g., Bisson et al.,

2014; Lee & Révész, 2020; Montero Perez et al., 2015; Wisniewska & Mora, 2020), it must be pointed out that Bird and Williams (2002) did not expose their participants to captioned video. In fact, the tests only included single-word items in all training and testing phases, and deliberately removed the semantic context to show the effect of modality at a pure phonological level.

Building on the influential work of Bird & Williams (2002), the investigation of L2 speech perception during exposure to captioned video moved from the recognition of single spoken words to the recognition of words within utterances, and to speech segmentation, or the ability to identify the boundaries between words when listening to a continuous stream of speech (Charles, 2017; Charles & Trenkic, 2015; Mitterer & McQueen, 2009). In Mitterer and McQueen (2009), 121 L1 Dutch advanced learners of English unfamiliar with Scottish and Australian English watched one of two 25-minute videos with Scottish or Australian English soundtrack and with English, Dutch or no subtitles. After the viewing, participants did an auditory-only sentence repetition task with 40 excerpts from each video and 80 excerpts from similar but novel materials, i.e., they heard 160 sentences in total, some in the accent they had been exposed to and some in a different accent, and repeated them back. Whereas English captions enhanced perceptual adaptation and the recognition of foreign-accented speech for both familiar and unfamiliar items, as determined by the percentage of words accurately repeated in each excerpt, Dutch subtitles only improved recognition of old items and actually led to worse performance on new materials. While both captions and subtitles supported word recognition, only L2 captions promoted lexically-guided perceptual learning, intended as learning to "interpret ambiguous phonemes on the basis of disambiguating lexical contexts",

resulting in a beneficial transfer of this learning to the comprehension of new utterances from the same speaker (Mitterer & McQueen, 2009, p. 1). The authors conclude that the orthographic information in subtitles can facilitate or inhibit learning in speech perception, depending on how consistent it is with respect to the spoken language.

Adaptations of the sentence repetition task created by Mitterer and McQueen (2009) were used by Charles (2017) and Charles and Trenkic (2015) to investigate the general L2 speech segmentation ability of non-native speakers of English and the potential benefits of watching L2 videos with captions, which provide visually segmented speech units and may facilitate word recognition in a continuous stream of sounds. Although Charles and Trenkic (2015) is often cited as the study that established the superiority of captioned video over uncaptioned video for the development of L2 learners' speech segmentation skills (Gesa, 2019; Muñoz, 2017; Pujadas & Muñoz, 2023; Vanderplank, 2019; Wisniewska & Mora, 2020), the study was small-scale, as only 10 participants were tested in the first experiment on general ability and 12 in the second experiment (divided into 3 conditions) on the effects of video exposure on speech segmentation. More solid evidence in support of the beneficial role of captions on the development of speech segmentation skills is offered by one of the studies in Charles's (2017) dissertation, in which 48 L1 Chinese and nine L1 English speakers repeated the same treatment and testing procedure as in Charles and Trenkic (2015). In the first week, the L2 English learners were pre-tested with a sentence repetition task containing short excerpts from English documentaries. In the two following weeks they watched two 30-minute documentaries and did two immediate post-tests on old and new utterances (present or not present in the documentaries to which they

had been exposed, respectively), and in the fourth week they did a sentence repetition post-test. The findings confirmed that the simultaneous presentation of oral and written sentences helps learners not only segment speech they have already heard before, but also generalize this learning to novel utterances. This finding is crucial because, as Charles (2017) observed in a follow-up experiment, improvements in speech segmentation skills can be associated with overall more efficient processing and comprehension of spoken dialogue.

The effects of captioning on perceptual learning was also the target of Birulés-Muntané and Soto-Faraco (2016), who exposed 60 intermediate learners of English to a 1-hour-long episode of a TV series with either L2 English, L1 Spanish or no subtitles. The pre- and post- test included a meaning recognition vocabulary test, a plot comprehension test and a listening cloze test featuring a 1-minute-long excerpt of a conversation from the same TV series. The gains in listening skills of the English captions group were significantly larger than those of the Spanish and no subtitles groups, whereas the vocabulary test showed no differences and plot comprehension scores were highest under native, Spanish subtitles. As the listening test was a non-timed cloze test and the word clues or other contextual knowledge may have helped participants complete the sentences, it cannot be excluded that better performance reflected more efficient top-down processing rather than the development of perceptual skills. However, the findings of Birulés-Muntané and Soto-Faraco (2016) are consistent with Mitterer and McQueen's (2009) account of the superiority of L2 captions to support the mapping of auditory word forms onto written forms and, ultimately, the "retuning" of L2 learners' phonetic categories.

Among research studies on the potential of captioned videos for pronunciation learning and teaching, Wisniewska and Mora (2018, 2020) and Wisniewska (2021) stand out for having conducted a thorough analysis of L2 learners' processing of auditory and visual input, and of the effects of extensive viewing on pronunciation development. Wisniewska and Mora (2018) investigated whether the ability to match auditory and orthographic representations during exposure to captioned video affects L2 speech processing, potentially leading to the acquisition of more target-like phonological representations. They recorded the eye movements of 38 L1 Spanish/Catalan learners of English to find the extent to which they processed the captions using the Reading Index of Dynamic Text (Kruger & Stein, 2014) and, for ten selected target words, the degree of synchronization between the fixation on a word and its auditory presentation in the soundtrack. To assess individual differences in speech processing, participants were also administered an elicited imitation proficiency task, a speech segmentation (word spotting) task, an auditory statistical learning task and two audio-text integration tasks in English (the participants' L2) and Basque (an unknown language). Results showed that the viewers processed from as little as 16% to 78% of the words in captions, and mostly fixated on the selected target words before (m = 200 ms) their auditory onset, but post-fixations were recorded on 30% of the target words (on average 500 ms after auditory onset). As expected, English proficiency was positively correlated with the ability to segment L2 speech and detect text-sound mismatches in English, but this advantage did not hold in Basque, a language of which participants had no previous knowledge. Neither proficiency nor any of the English speech processing measures correlated with the eye-tracking measures, after controlling for the effect of L2 proficiency. However,

better performance in the Basque modality integration task correlated with a shorter fixation distance, suggesting that efficient text-sound integration may promote closer synchronization of audio and dynamic text processing. The authors conclude that the ability to map auditory forms on written forms, and identify mismatches between the two, may play an important role in the processing of audiovisual input, and point out that further research on the role of individual factors such as reading speed is needed.

Wisniewska's doctoral dissertation (2021), which had been partly published as Wisniewska and Mora (2020), investigated whether pronunciation learning through TV series could be enhanced by the availability of captions in combination with a focus on phonetic form or meaning comprehension. Ninety university students participated divided into five groups, one control group and four experimental groups who watched a total of 5 hours of TV series over 8 weeks and answered questions after each viewing, in a 2 x 2 experimental design (captions/no caption x focus on form/meaning). First, the effects of the treatment on L2 speech processing were tested with a sentence repetition task (Mitterer & McQueen, 2009; Charles & Trenkic, 2015), an animacy judgement task assessing participants' speed of lexical access, and a sentence verification task assessing speed of processing at the sentence level. All groups started from comparable speech segmentation scores, but only the experimental groups achieved significant gains in speech segmentation (regardless of the presence of captions or of a focus on phonetic form) and these gains generalized to untrained materials from a different TV series. The participants watching TV series without captions achieve significant gains in L2 speech processing at the sentence level, but no between-group differences were found in processing gains at the word level. Participants' performance in an ABX phonetic discrimination task and in a

sentence repetition task (foreign accent ratings) revealed no clear treatment effects on phonological accuracy.

The analysis of eye-tracking data (Reading Index of Dynamic Text or RIDT) showed a high degree of individual variability in the amount of caption reading, in line with previous eye-tracking studies on learners' use of captions (Muñoz, 2017; Winke et al., 2013; Wisniewska & Mora, 2018). Despite an overall reduction of RIDT scores after the intervention, no correlation was found between participants' changes in RIDT and their gains in speech processing or phonological accuracy. However, when considering only the RIDT scores at T1 (before the start of the intervention) viewers in the meaning-focused uncaptioned condition who usually relied on captions had higher gains in speed of lexical access and lower gains in phonological accuracy. In addition, accent reduction was negatively related to the viewers' reliance on captions in the pronunciation-focused uncaptioned condition and positively related in the meaning-focused captioned condition, suggesting that, in the meaning-focused condition, the availability of captions supported phonological learning for learners who usually read captions, whereas in the focus on phonetic form condition, the absence of captions impaired learning for learners who usually read more.

Overall, Wisniewska (2021) found evidence that extensive exposure to TV series with and without captions can lead to pronunciation development, and that the availability of captions may enhance the incidental acquisition of L2 pronunciation in the absence of a focus on phonetic form. The learning gains were more marked when considering general aspects of L2 speech processing, such as the ability to segment and process speech in a fast and efficient manner, rather than the updating of specific phonetic categories. Finally, Wisniewska (2021) investigated the effects of attention,

phonological short-term memory and proficiency on the observed pronunciation gains, but only found weak correlations between speech segmentation gains and the ability to select auditory cues in the input, and between phonetic discrimination gains and the ability to switch attentional focus between modalities. As Wisniewska and Mora (2020) and Wisniewska (2021) pointed out, the absence of an effect of proficiency may have been due to the very narrow range of participants' proficiencies, and the overall advanced level of proficiency at the start of the intervention limited the extent of the gains and between-group differences that could be observed.

Research conducted within a distinct strand of investigation focused on informal English language learning highlights a connection between patterns of extracurricular engagement with L2 input (such as TV series) and accent acquisition. This line of inquiry builds upon existing evidence that L1 speakers of different languages unintentionally imitate specific L1 accents featured in popular TV series (Ota & Takano, 2014; Stuart-Smith et al., 2013). For English L2 learners, it is not unusual to mix Received Pronunciation and General American accents in everyday speech (Maeda, 2009; Navrátilová, 2013; Rindal and Piercy, 2013). The imitation of a mix of accents may be intentional or unintentional and driven by factors such as the learners' pursuit of intelligibility, their sense of identity and their geographical origin (Markham, 1997). However, when learners have only received formal instruction from British teachers or teachers with near-British accents, their close imitation of a GA accent provides evidence of the impact of informal learning on the acquisition of L2 pronunciation (Yibokou, 2023). Among the extracurricular sources of L2 input, TV series are considered to influence more strongly learners' accents due to the substantial viewing time reported, which frequently exceeds the duration of their

language classes (Navrátilová, 2013; Rindal and Piercy, 2013). In support of the pivotal role played by TV series, regular viewers exhibit higher accuracy in the recognition and translation of auditorily presented phrases that frequently occur in TV programs, compared to casual viewers (Kusyk & Sockett, 2012). The investigation of L2 accent acquisition from informal learning, which has explored the correlation between speech proficiency and reported TV series exposure without following the trajectory of pronunciation development over time, offers a perspective that can be considered complementary to the studies presented above, which have adopted a pre- and post-test design.

2.5.5. Summary

To sum up, when learners' L2 proficiency is sufficiently developed to allow them to read fast and integrate different sources of information efficiently, reading L2 captions does not seem to impair the processing of the L2 soundtrack in videos. On the contrary, the availability of redundant text improves speech segmentation, supporting the identification of single words in the stream of speech and facilitating the acquisition of novel spoken words. While gains in overall L2 speech processing from TV series exposure have been found to transfer to novel contexts, offering tangible benefits for real-life language use, there is to date scarce evidence that incidental exposure to TV series leads to the updating of specific phonetic categories. The generalizability of findings regarding the effects of captioned video on pronunciation development is limited by the extreme individual variability in caption reading behavior and by the possible influence of other individual differences in auditory and textual input processing. As the analysis of online data on learners' use of captions during the viewing may provide useful insights into their performance in offline

pronunciation tests, online data collection techniques such as eye-tracking will be employed in this dissertation.

Among the various factors influencing multimodal input processing, the degree of audiovisual synchrony may be a crucial factor, as the availability of text identical to the soundtrack provides visual cues about upcoming auditory information, but the misalignment of visual and auditory information may impair processing (Conklin et al., 2021; Wisniewska & Mora, 2018). Moreover, since the text in captions competes for attention with the moving image in the background, the reading may be interrupted at any point to pay attention to the image or the auditory input, and there is little time to read and reread single words (Kruger et al., 2015). L2 viewers who consistently read captions tend to rest their gaze briefly on the image area, especially on the speaker's face, then move linearly along the text lines and finally return to the image area (Bisson et al., 2014; D'Ydewaelle and Van de Poel, 1999). As a result, for some language learners watching L2 videos can involve a great deal of automatic, meaning-focused caption reading, and the corresponding generation of phonological forms based on the learner's interlanguage. Unless the focus is on the comparison of generated phonology and auditory input, inaccurate phonological representations may automatically enter the phonological loop for processing, "silencing" the soundtrack (Mitterer & McQueen, 2009). This dissertation investigates whether manipulating captions to provide a focus on specific target words may direct learners' attention to the corresponding auditory forms, enhancing the comparison with pre-existent phonological representations.

Table 2.1. Studies on pronunciation learning through L2 captioned video.

Study (by publication date)	Participants	Learning target	Tasks	Number and length of sessions	Measurement of learning	Results
Bird and Williams (2002)	A total of 16 native speakers and 40 advanced L2 learners of English (2 studies)	Auditory word forms	Lexical decision tasks (LDT), rhyme monitoring, familiarity judgement	One session of unspecified length	Reaction times, hit rate and false alarm rate	Bimodal presentation promoted both implicit and explicit auditory form recognition. Therefore, text availability did not impair but aided auditory information processing.
Mitterer and McQueen (2009)	121 L1 Dutch advanced learners of English	Foreign accent adaptation	Auditory-only sentence repetition task	One 25-minute video	Percentage of words accurately repeated	The speech learning benefits of exposure to video transferred to novel sentences only with English captions.
Kusyk & Sockett (2012)	45 L1 French learners of English	Incidental learning of common L2 phrases in TV shows	Vocabulary knowledge scale (auditory input), TV viewing habit survey	One testing session (no training)	Errors in L1 translation of L2 phrase, analysis of questionnaire responses	Regular viewers were more accurate in translating the L2 phrases and self-evaluated their oral comprehension skills to be higher than casual viewers.
Charles (2017)	A total of 112 (mostly L1 Chinese) L2 learners of English across 5 studies	Speech segmentation	Auditory-only sentence repetition task, phonological working memory test	In each study, 0 to 4 sessions with 30-minute documentaries	Percentage of words accurately repeated	The studies found evidence that L2 learners' speech segmentation skills can be limited and highlighted the benefits of captioned video.

Birulés-Muntané and Soto-Faraco (2016)	60 L1 Spanish/Catalan learners of English	Listening skills	Meaning recognition vocabulary test, listening cloze test	A 1-hour episode of a TV series	Percentage of accurate responses	Listening scores improved with English captions and transferred to novel items. The vocabulary test showed no differences between conditions.
Wisniewska and Mora (2018)	38 L1 Spanish/Catalan learners of English	Audiovisual processing during exposure to L2 captioned video	Eye gaze analysis, tests of individual differences in speech processing	Two 90-second clips	Eye-tracking measures of visual processing and audiovisual synchrony	Viewers read captions 200 ms before auditory onset on average. No significant correlations were found between eye gaze and speech processing measures.
Wisniewska and Mora (2020)	90 L1 Spanish/Catalan learners of English	Speech segmentation, speed of lexical access, sentence processing, and phonological accuracy	Sentence repetition task, animacy judgment task, sentence verification task, ABX discrimination, accentedness ratings	5 hours of TV series over 8 weeks	1) Percentage of words accurately repeated; 2-3-4) reaction times; 4) percentage of accurate responses; 5) native speaker ratings	Both captioned and uncaptioned viewing led to gains in L2 speech processing. Mixed effects were found for phonological accuracy, as only viewing with captions but without a focus on form led to gains in production, and no gains in perception were found.
Yibokou (2023)	20 L1 French learners of English	Foreign accent adaptation	Word read-aloud task, questionnaire about English learning and accent imitation	One testing session (no training)	Acoustic analysis to identify accent (RP or GA), quantitative analysis of questionnaire responses	Despite having teachers with British or near-British accents, the learners pronounced English words with a mixture of RP and GA accents, suggesting that they unconsciously picked up GA accent from media such as films.

2.6. Enhancing pronunciation learning with video-based activities

2.6.1. Introduction

The previous sections have reviewed the language learning potential of multimodal input from video, with a focus on input in which target features have been textually enhanced. The main challenge in promoting language development through exposure to enhanced input is that new knowledge accumulates gradually over time and is susceptible to factors unrelated to the enhancement technique, such as the difficulty of the input, the learners' internal focus, and their cognitive abilities (Han et al., 2008). The aim of this section is to discuss how combining exposure to enhanced video with post-viewing activities can enhance the learning experience by providing a targeted focus on pronunciation and promoting a more purposeful engagement with materials that may otherwise be seen as exclusively recreational (Sánchez-Requena, 2017; Vanderplank, 2015, 2019). While certain instructional activities, such as introducing the context and characters before the viewing and pre-teaching target vocabulary, have been proven to be effective (Rodgers & Webb, 2011; Vanderplank, 2015), there is an increasing interest in investigating the language learning benefits of individual and collaborative video-based activities, among which activities involving repeated rehearsal and practice under teacher monitoring (Vanderplank, 2010). In this chapter, we will review a variety of studies that have used video-based activities to teach pronunciation or foster the development of skills closely connected to pronunciation learning. The discussion will be centered on the framework developed by Zabalbeascoa et al. (2012) to assist teachers in the creation and implementation of audiovisual activities, such as dubbing and captioning, that involve the manipulation

of the soundtrack and captions in L2 videos. These activities were deemed particularly effective because they require learners to pay close attention to L2 speech while carrying out a meaning-focused task.

2.6.2. The audiovisual framework

The audiovisual framework was devised within the *Clipflair* project, which aimed to develop and disseminate among language teachers a range of activities based on watching, manipulating and discussing videos in the foreign language (Zabalbeascoa et al., 2012). These activities, described and categorized in the *Clipflair* conceptual framework and collected in the project's platform, enable learners to not only practice the four traditional language skills but also develop their "audiovisual skills", i.e., watching, captioning and revoicing (Zabalbeascoa et al., 2012). In the *Clipflair* project, these terms represent umbrella concepts encompassing various types of activities. While *watching* simply means interpreting a video using a combination of verbal and non-verbal clues, *captioning* refers to various activities that involve producing a written version of the spoken dialogues, as well as annotations and free commentaries, and *revoicing* refers to any form of voice recording, including audio description of the scene, free commentary and dubbing. Adding another layer of complexity to the categorization of audiovisual activities, these can be interlingual, such as when an L1 video clip is translated and dubbed into the learner's L2, or intralingual, i.e., involving exclusively the foreign language. While interlingual activities have been successfully used to train translators and teach L2 vocabulary (Alonso-Perez & Sánchez-Requena, 2018; Danan, 2010), this dissertation will only feature intralingual activities that build on the dual coding potential of L2 captioned

video for pronunciation learning. Finally, activities within the audiovisual framework are also categorized based on the possible actions that learner perform in response to the video, which can lead to more or less spontaneous oral and written production. The learner could be *repeating* the soundtrack as literally as possible, *rephrasing* it to write a summary, or *reacting*, if they record their comments regarding the plot, setting and actors. According to this classification, producing same language captions for an L2 video is an *intralingual-repeating-captioning* activity aimed at practicing both listening and writing. Silencing the turns of a character in a clip leads to an *intralingual-reacting-revoicing* activity, which could turn into *rephrasing* or *repeating* if the learner is allowed to listen to the character's turn just before filling in the pause.

As a comprehensive framework intended to encompass a wide range of activities, the proposal of the *Clipflair* project is inherently broad. However, for the purpose of this dissertation, it was deemed necessary to limit the focus to specific types of activities, given our constraints of time and scope. In particular, we will focus on captioning, in the sense of a faithful transcription of words in dialogues, and dubbing, in the sense of silencing the actor's voice and repeating the exact same words imitating the actor's speed of delivery and intonation, along with a limited number of secondary activities instrumental to learners' successful completion of dubbing activities. From a theoretical point of view, these activities potentially represent valuable tools in pronunciation teaching, due to the use of authentic materials in a meaningful context, their collaborative nature, and the focus on verbal production. Moreover, audiovisual activities provide an inherent focus on form by requiring learners to listen closely to authentic L2 speech and practice imitating target models, an essential ability for L2

speech acquisition (Ragni, 2018; Vaquero et al., 2017). The possibility to listen to and repeat the same utterances as many times as needed, even if this aspect of the task may not be very communicative in nature, assists the development of automaticity of phonological processing (Darcy, 2018; DeKeyser, 2010; Tavakoli, 2019). Further, when various actors are featured in the selected dialogues, learners are exposed to a variety of voices and accents, in line with recommendations for listening practice to enhance pronunciation learning (Darcy, 2018). Finally, the collaborative nature of these activities involves both giving and receiving feedback from peers during their preparation, as well as receiving feedback from the teacher during the presentation of the final product (the complete captions or dubbed clip) to the class. In this review, we will focus on studies that have used *intralingual captioning* and *dubbing* to teach various aspects of L2 pronunciation, listening comprehension and oral communication. Although these activities may be designated with different names in the original papers, once we have clarified what the activities entailed, we will use the terms captioning and dubbing to refer to them.

2.6.3. Audiovisual activities in pronunciation teaching

In recent years, intralingual audiovisual activities have been increasingly incorporated into the language classroom, with positive effects on the development of L2 listening, speaking and pronunciation skills (Campbell, 2016; Lima, 2015a; Sánchez-Requena, 2017; Zhang, 2016; Zhang & Yuan, 2020). Table 2.2 presents an overview of studies employing captioning and dubbing activities to teach aspects of a language related to L2 pronunciation, intended as the ability to produce and perceive the sounds and intonation patterns of a language (Montero Perez, 2022). It can be observed that most

studies in this area have employed an ecological approach to both teaching and assessment, integrating the audiovisual activities into the curriculum and analyzing their effects qualitatively through data from interviews and questionnaires. Accordingly, these interventions were usually implemented with intact classes and in the absence of a control group, although some included a comparison group that received some form of instruction but did not carry out audiovisual activities (Campbell, 2016; Chiu, 2012; Zhang & Yuan, 2020). Learning outcomes were primarily evaluated in terms of learners' perceived gains and attitudes towards the activities, as well as teachers' observations of their linguistic development. Only five studies included an objective measure of learners' linguistic development such as ratings of their speech before and after the intervention, and only one assessed delayed effects (Zhang & Yuan, 2020). In four studies, the activities were combined with general pronunciation instruction or, alternatively, instruction focused on a specific target, such as segmental features or prosody (Chiu, 2012; Lima, 2015a; Martinsen et al., 2017; Zhang & Yuan, 2020). The rest of this section presents an overview of the findings of research on intralingual captioning and dubbing, categorized according to the language skills targeted by the intervention.

2.6.4. Studies on audiovisual activities and L2 listening development

In relation to the *listening* potential of audiovisual activities, several studies have documented perceived gains in the ability to perceive and understand auditory information after the intervention, based on data from learner and teacher questionnaires (Martinsen et al., 2017; Sánchez-Requena, 2017; Sokoli, 2018; Zhang, 2016). However, only Campbell (2016) quantitively assessed the effects of

intralingual captioning on the development of listening skills adopting a pre- and post-test design. The study took place in an English for a Specific Purpose context, with members of the military taking an intermediate level of English, and the pre- and post-test was a standardized test that specifically targeted their need to understand radio-based communication. During the test, participants were presented with military video footage and simultaneously listened to authentic radio communication. Their task was to complete a written transcript in which some words were missing that could not be inferred from context alone (Campbell, 2016). After the pre-test, the experimental group carried out individually a subtitling activity, whereas the control group engaged in other online activities using the same audiovisual materials. Participants in the experimental group were first taught how to use the subtitling platform, then did pre-viewing activities designed to activate their knowledge regarding the content of the video, and finally wrote the captions for a 3-minute video clip. The results revealed that engaging in the intralingual captioning activity had a significant and positive effect on the experimental group's listening comprehension performance, but watching the same video without performing the captioning activity did not lead to improvement in listening skills.

2.6.5. Studies on audiovisual activities and L2 pronunciation learning

Several studies featuring dubbing activities have documented improvements in overall speaking skills, including comprehensibility, fluency and speech rate. As previously mentioned, these conclusions were often derived from data obtained through questionnaires and interviews, which provide insights into learners' or teachers' *perceptions* of improvement in these areas (Chiu, 2012; Navarrete, 2013;

Sokoli, 2018). In some cases, however, objective measures of comprehensibility, intelligibility and fluency were also obtained through quantitative speech analysis and raters' assessment of the learners' performance in oral tasks (Lima, 2015a; Martinsen et al., 2017; Sánchez-Requena, 2017; Zhang & Yuan, 2020). As all of these studies had the primary objective of teaching the pronunciation of specific segmental and suprasegmental aspects, they are of particular relevance to this dissertation. Therefore, the remaining of this section will be devoted to describing more in detail the approaches adopted in these studies and their outcomes.

Lima's (2015a) doctoral dissertation, partly published as Lima (2015b) and Lima and Zawadzki (2019), centered on the development and piloting of a four-week online pronunciation tutor targeting suprasegmental aspects of L2 speech. The Supra Tutor included diagnostic quizzes, a variety of training activities involving exposure to authentic multimodal input, such as identifying the stress in words in the dialogue of a sitcom scene and imitating the actors by dubbing muted excerpts, eventually receiving perception and production feedback. Twelve international teaching assistants, coming from different L1 backgrounds and working at a US university, used the tutor over four weeks and their oral production was rated for comprehensibility by a total of 178 naïve and six expert raters before and after the intervention. While less than half the participants showed significant improvement in comprehensibility, low gains seemed to be related to irregular and suboptimal use of the learning materials. Despite enjoying the variety and appeal of materials and the flexibility of scheduling with an online tutor, in the follow-up questionnaire learners expressed a desire for human interaction and feedback, which could be fulfilled in a classroom setting through pair work activities and learner-teacher interaction.

Martinsen et al. (2017) examined the effects of shadowing and tracking exercises on L2 French pronunciation development. Although tracking will not be considered, because it refers to repeating what a speaker says simultaneously or with as little delay as possible, shadowing allows for a longer pause between listening and repeating and may therefore be considered a synonym of dubbing[4]. The participants were 19 L1 English high school students with varying levels of French proficiency. The intervention lasted ten weeks and included whole-group sessions three times per week and individual weekly sessions. In class, students watched a 2- to 3-minute video clip and imitated the speaker with the help of the captions, leaving as little time as possible between hearing and speaking. The teacher stopped the video whenever the class made evident mistakes, corrected their pronunciation, and asked them to repeat the corrections before resuming the tracking exercise. In the individual sessions, however, the students were allowed to stop the video, listen again and repeat as many times as they wanted. A read-aloud pre- and post-test showed significant improvements in learners' pronunciation in controlled production, whereas data from a picture description task showed no evidence of a transfer to a more spontaneous context. Although students' beliefs on the usefulness of tracking and shadowing were not very positive before the intervention, their perceptions improved steadily over time and were mostly positive by the end of the intervention.

[4] Other authors use *shadowing* to refer to immediately vocalizing an auditory text upon hearing it, leaving no delay between hearing a syllable and repeating it aloud (Foote & McDonough, 2017; Hamada, 2016). By allowing no time to access meaning, this different technique encourages a decontextualized focus on pronunciation form. Since this dissertation focuses on dubbing as the temporary storage in working memory and subsequent repetition of meaningful information, this alternative, simultaneous form of shadowing will not be discussed.

Sánchez-Requena's (2017) doctoral dissertation consisted of three studies, two of which were also published in academic journals (2016, 2018), on the use of dubbing to improve L1 English L2 Spanish learners' speech rate, intonation and pronunciation. The treatment involved practicing the spoken dialogue of muted video clips using the transcript and on-screen captions for support and receiving feedback on pronunciation by the teacher. Additionally, if time permitted at the end of the class, the learners had the opportunity to record their performance on the muted clip using Movie Maker. Although in the pilot study the pre- and post-test consisted of interviews, in the two latter studies a total of six 3-minutes podcast on a topic of choice were recorded. The podcasts were analyzed quantitatively by measuring speech rate in words per minute, as well as qualitatively through expert raters' assessment of learners' performance. The qualitative analysis did not involve anonymization and randomization of the samples. The results of the learner and teacher surveys indicated that the learners experienced improvement in various areas. These improvements included the ability to self-correct, increased comprehensibility (referred to as "ease of understanding" by the author), and enhanced motivation. Furthermore, the benefits extended to gains in speech rate (both quantitatively and qualitatively assessed), pronunciation, and intonation, in that order.

The most recent study, Zhang and Yuan (2020), did not focus exclusively on audiovisual activities but rather used them as a supplementary tool to investigate the impact of explicit pronunciation instruction on L1 Chinese L2 English pronunciation development. The study compared three experimental conditions: Explicit focus on segmentals, explicit focus on suprasegmentals, and no specific pronunciation focus. Over a period of 18 weeks, both instruction groups received 35 minutes of explicit

instruction per session, twice per week, through teacher explanations and textbook exercises. Then, the segmental group continued practicing the target sounds by studying word lists and doing additional exercises such as sound discrimination, listening and repeating, and reading aloud texts. The suprasegmental group initially studied the target features through listening and imitation exercises from the textbook, but their classes concluded with a movie-dubbing activity for oral practice. The pre-post- and delayed post-test consisted of a sentence reading task and a narrative task with picture prompts, rated for comprehensibility by six trained raters. The results indicated that both instruction groups improved their comprehensibility in the controlled production task, but these gains transferred to the spontaneous narrative task and were retained at delayed post-test for the suprasegmental group only. The authors attributed this finding to the fact that successful completion of the prompted narrative task required the proceduralization of knowledge due to the heavier cognitive load associated with spontaneous speech production compared to controlled production. Although the dubbing activities were only one of the two components of the intervention and explicit instruction likely played a significant role, Zhang and Yuan (2020) provided some evidence of the benefits of practicing target pronunciation features by imitating models in L2 captioned video.

2.6.6. Summary

To summarize the research findings on intralingual captioning and dubbing, interview data shows that both learners and teachers believe that the learners' increased motivation and engagement with audiovisual activities enhance language learning (Alonso-Perez & Sánchez-Requena, 2018). However, when it comes to assessing the

learners' speaking and pronunciation gains objectively, the results paint a more complex picture. Regarding captioning, while only one quantitative study has focused specifically on its use, the results have been positive and seem to support the perceptions gathered in other studies that did not use a pre- post-test design. Captioning offers the practical advantage of being more easily implemented than dubbing in larger classes, since it does not involve oral production and feedback can be as simple as showing learners the original script. Captioning also lends itself to being used as diagnostic self-test for specific target features or general auditory perception, and diagnostic analysis is the foundation of any pronunciation teaching intervention (Benitez Correa et al., 2020; Lima, 2015; Zhang & Yuan, 2020).

Regarding dubbing, several elements seem to be associated with its successful implementation. To begin with, independent work on these activities carries the risk of missed assignments, possibly because dubbing may be initially seen by learners as a "fun" but not particularly effective activity (Lima, 2015a; Martinsen, 2017). Therefore, it may be more beneficial to introduce dubbing in the classroom under the supervision of a teacher who explains the rationale behind the activity and provides continuous feedback on learners' performance. Secondly, since dubbing is a meaning-focused activity with incidental learning potential, its effectiveness has usually been enhanced by combining it with explicit instruction, which can greatly accelerate acquisition. Although the inclusion of explicit instruction makes it challenging to tease apart the effects of the "dubbing" variable, it is reasonable to assume that teachers would use dubbing as supplementary rather than main activity in the classroom, so this combined approach may be considered more ecologically valid. Explicit instruction, which should be thoroughly described in research papers to allow

for result replication, need not be anything more burdensome than directing learners' attention to the difficulty of a pronunciation feature and to the underlying rule determining its phonological realization. Finally, dubbing whole clips, even as short as 3-4 minutes, can become a daunting task for learners who are not familiar with the task. Dubbing is a specialized skill performed by professionals for a reason - it is a challenging job. Therefore, especially with beginning and intermediate learners, it is recommended to work with shorter excerpts from simple video clips containing highly intelligible dialogue spoken at an appropriate pace (Danan, 2010).

To address the research gaps identified in our review of research on pronunciation teaching with audiovisual activities, in this dissertation the effectiveness of audiovisual activities will be pre- and post-tested using speaking tasks involving both controlled and spontaneous L2 production. By including a delayed post-test, we expect to be able to draw more robust conclusions regarding the durability of any documented pronunciation gains. To further improve the internal validity and generalizability of the study, a specific feature (the pronunciation of the past tense <-ed> ending) will be targeted. Finally, the inclusion of a control group will allow for an objective evaluation of the effectiveness of intralingual audiovisual activities compared to conventional English instruction in the absence of such activities. While research on captioning and dubbing is currently limited, investigating the use of audiovisual input in task design is crucial to make the most of the availability of new technologies in the EFL classroom (Carless, 2012).

Table 2.2. Studies on L2 speech acquisition with intralingual audiovisual activities.

Study (by publication date)	Participants	Target feature	Learning task	Number and length of sessions	Measurement of learning	Results
Chiu (2012)	83 university students	Pronunciation and intonation	Dubbing one muted clip of 10’	Self-paced preparation at home with the script	Questionnaire; Semi structured interviews	Participants in the dubbing group reported higher satisfaction with L2 use and awareness than the control group.
Navarrete (2013)	20 students in 9th grade	Listening and speaking skills	Dubbing a 1’30” clip into L2, after a pre-task (comprehension questions and ordering sentences from the L2 script)	105’ workshop	Teachers’ observations	This meaning-focused activity developed students’ language and technology skills. Dubbing took longer than expected.
Lima (2015a)	12 International Teaching Assistants	Word Stress, Rhythm, Intonation	A variety of tasks, including imitating a model and recording one’s voice, role-play, sing along	Self-paced for four weeks	Rating of a 7’ lecture; Questionnaire	Students who completed the online program improved their comprehensibility.
Campbell (2016)	46 students in the military	Listening comprehension	Captioning L2 clips	Independent work on one video clip (3’)	Bimodal (auditory and written) cloze test; questionnaire	The captioning group had larger gains than control group. Students enjoyed the task and reported language gains.
Zhang (2016)	120 university students	Listening and speaking skills	Dubbing L2 clips, songs and other contents into L2	No less than 8 dubbings over one semester	Questionnaire	1/3 of the students did more than 30 dubbings.

						Most students reported L2 improvement.
Martinsen et al. (2017)	19 students in 10th grade	Pronunciation	Whole-group and individual video-based tracking and shadowing tasks	5-10’ classroom practice, 3 times x 10 weeks; 30’ individual practice x week	Ratings of picture read-aloud task and description task; weekly questionnaires	Read-aloud performance improved at post-test and there was no difference in picture description performance.
Sánchez-Requena (2017)	17 + 47 students in 11th grade; 30 university students	Fluency and pronunciation	Dubbing clips into L2 with L2 script and captions	80’ each week for six to twelve weeks	Ratings of pre-post-test recordings (podcast or interviews); student questionnaires; teacher’s notes	The ratings showed improvements in speed, intonation and pronunciation. Students reported gains in oral skills, vocabulary and increased motivation.
Sokoli (2018)	1.250 university students	Audiovisual skills (writing, speaking, reading, listening)	91% of participants carried out one audiovisual activity, e.g., captioning, dubbing, audio description	At least one audiovisual activity	Questionnaire	Most participants found audiovisual activities engaging and useful for learning. There were some technical issues with the *Clipflair* platform.
Zhang and Yuan (2020)	90 university students	Pronunciation of segmental vs suprasegmental features	Suprasegmental group used movie-dubbing activities	35’ twice a week x 18 weeks	Ratings of comprehensibility in sentence reading task and prompted narrative task	Both the segmental and suprasegmental groups improved.

CHAPTER 3. RESEARCH STUDIES

Our overarching aim was to examine the effectiveness of an innovative input enhancement technique, audio-synchronized textual enhancement, in second language pronunciation teaching and learning. Building upon previous research on pronunciation acquisition with similar synchronized enhancement techniques in reading-while-listening (Gerbier et al., 2018; Stenton, 2013), we expanded the investigation to L2 captioned video from TV series. Study 1 and 2 (reported in sections 3.1 and 3.2) examined the impact of this technique by exposing learners to videos with enhanced and unenhanced captions, without conflating the viewing with explicit instruction. By analyzing data collected through pronunciation tests, eye-tracking, and (in study 2) verbal recall, we aimed at establishing a causal relationship between input enhancement and the processing of phonological form, while controlling for possible confounding variables. However, while comparing two variables in isolation improved internal validity, the findings obtained were not directly generalizable to the classroom context, where the viewing of video clips is usually combined with complementary learning activities. Therefore, study 3 (section 3.3) was conducted within secondary school classrooms and involved the implementation of audio-synchronized textual enhancement alongside other activities aimed at enhancing audiovisual processing, such as captioning and dubbing. This chapter describes the aim of each study, the research questions and methodology employed, and the results obtained.

3.1 Audio-synchronized textual enhancement: Which time-lag(s) promote audiovisual synchrony and pronunciation learning?[5]

3.1.1. Study aims

Before implementing audio-synchronized textual enhancement in a classroom intervention (section 3.3), we recognized the need to conduct a controlled laboratory experiment to test different synchronization time-lags and to explore L2 learners' visual processing of captioned video with and without enhancement. Specifically, we aimed to determine the potential of audio-synchronized enhancement to promote synchrony between auditory and written input processing, i.e., a reduced time lag between the moment a learner reads a target word in the captions and the moment that word is spoken in the soundtrack. In addition, we investigated whether audio-synchronized textual enhancement would facilitate learners' processing of the phonological form of the target words, leading to more automatic and accurate retrieval of the corresponding phonolexical representations. To control for possible confounds, we included in the analysis factors related to the presentation of the target words, such as their presentation duration and frequency of occurrence in the clips, and learner factors such as proficiency and reading speed.

[5] An article based on this work has been previously published as: Galimberti, V., Mora, J. C., & Gilabert, R. (2023). Audio-synchronized textual enhancement in foreign language pronunciation learning from videos. *System, 116,* 103078. https://doi.org/10.1016/j.system.2023.103078

3.1.2. Research questions

The study addresses the following research questions:

RQ1: What are the effects of audio-synchronized and unsynchronized textual enhancement on the synchrony of visual and auditory word processing in L2 captioned videos?

RQ2: What are the effects of audio-synchronized and unsynchronized textual enhancement in L2 captioned videos on the updating of target phonolexical representations?

RQ3: Is learners' processing of L2 captioned videos with and without textual enhancement moderated by learner proficiency and reading speed?

Our hypotheses are that:

HP1: Textual enhancement right before auditory onset, but not unsynchronized enhancement, will enhance the simultaneous processing of visual and auditory word forms. In particular, the 500 ms time-lag interval will be the most effective in promoting closer audiovisual synchrony compared to the unenhanced condition.

HP2: The audio-synchronized textual enhancement of words in L2 captions will promote phonolexical update to a larger extent than unsynchronized enhancement and no enhancement. In particular, the 500 ms synchronized condition will promote larger gains in lexical decision accuracy and response times to mispronounced target words than the other conditions.

HP3: The higher the L2 proficiency level and the faster the learners' general reading speed, the larger the time-lag will be between the learner's pre-fixation on a word and

the word's auditory onset. However, we expect textual enhancement to level out the effect of proficiency and reading speed, promoting more homogeneous audiovisual processing of target words compared to the unenhanced condition. As a result, we do not anticipate a significant effect of proficiency on phonolexical update.

3.1.3. Pronunciation target

This study targeted a selection of English words that were anticipated to be difficult for L1 Spanish/Catalan learners of English and that are typically mispronounced. The difficulty of these words was not necessarily related to the presence of phonemes that are not present in the L1 inventory (such as specific English vowels), but rather to the tendency of Spanish speakers to store imprecise phonolexical representations of these words, mainly due to L1-biased decoding of their written form. As a result, these imprecise representations may hinder automaticity in L2 speech perception and production due to the low speed and accuracy of lexical access and to unstable form-to-meaning mappings in the mental lexicon (Cook et al., 2016). After describing the participants, section 3.1.4 outlines the target word selection process and presents the list of words and error categories targeted in this study.

3.1.4. Methodology

In this study, 58 L1-Spanish/Catalan learners of English watched two videos with target words (TWs) highlighted 500 ms or 300 ms before auditory onset, highlighted from caption onset, or under a control condition (either unenhanced or uncaptioned). While a 300 ms time-lag interval had been previously used in reading-while-listening (Gerbier et al., 2018), we anticipated a longer time lag of 500 ms to be beneficial in the context of video. This is because, if learners mostly pay attention to the moving

image at the center of the screen, noticing the enhanced word in the peripheral visual field, planning and executing the saccade towards the word would take an additional 200 ms on average (Godfroid, 2019). Therefore, a time lag of 500 ms would allow them to activate the stored phonological representation before hearing the word's auditory onset, promoting greater synchrony between auditory and visual processing. Figure 3.1 presents an overview of the study design, and Figure 3.2, included below in the *Procedure* subsection, illustrates the viewing conditions. The learners' eye movements were recorded to gauge whether the synchronized conditions had a positive effect on the level of attention directed to the target words and on the synchrony of audiovisual processing. Instances of phonolexical form update were identified through a lexical decision task measuring the accuracy and speed of rejecting mispronunciations of the target words at pre- and post-test. The learners' reading speed and L2 proficiency were also tested to explore the relationship between these variables and the learners' eye gaze patterns and learning outcomes.

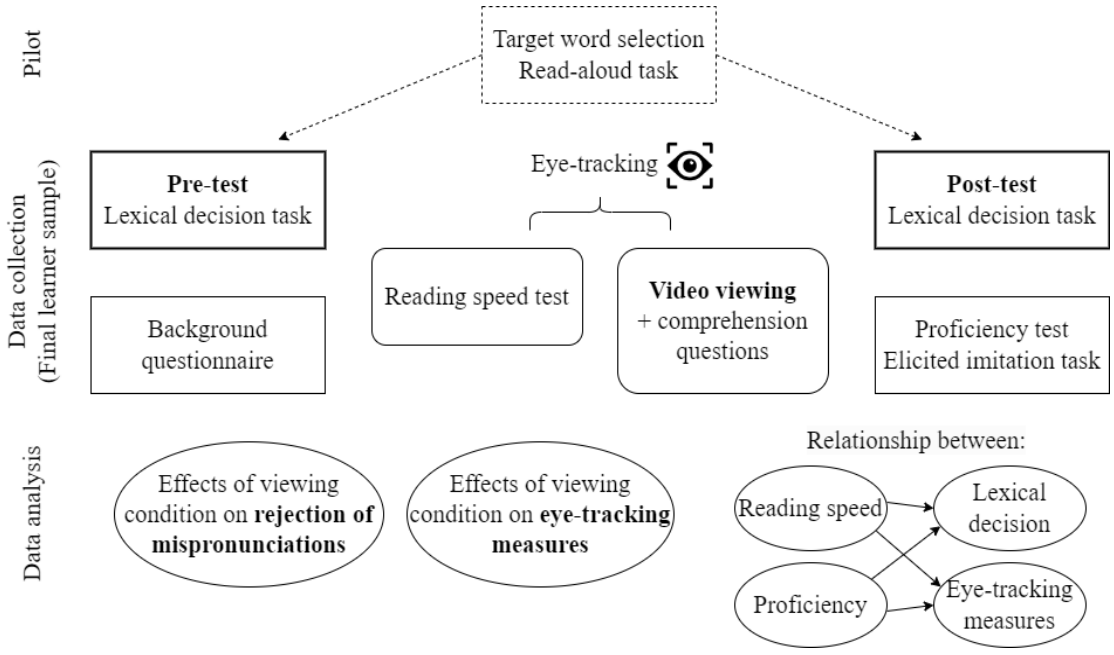


Figure 3.1. Overview of the methodology employed in study 1.

Participants

A total of fifty-eight first-year university students, 51 of which female, enrolled in an English degree program at a public university in Spain were enlisted for this study. All students reported that their first languages were Spanish and Catalan, and their English proficiency was estimated to be lower- to upper-intermediate, based on language certificates and self-assessment. In addition, the elicited imitation task in Ortega et al. (2002) was used as a validated instrument to obtain an L2 proficiency measure (Kostromitina & Plonsky, 2021). The task involved listening to and repeating 30 sentences increasing in word length and structural complexity and was assessed following the rubric provided by Ortega et al. (2002), available in the IRIS digital repository (Marsden et al., 2016). The average score obtained was 97 out of 120 (range 64-118, *SD* = 13.20), indicating an intermediate to upper-intermediate level of proficiency. The participants were randomly divided into four groups with comparable proficiency ($F(3, 54) = .981$, $p = .41$). Most participants had never spent more than a month in an English-speaking country, with the exception of a small number of participants who had lived abroad for up to 15 months, combining all their visits to English-speaking countries. Most reported regularly watching English-language TV shows and videos, averaging five hours per week, both with and without captions. For further details regarding the demographic characteristics of the participants, please refer to Table 3.1.

Table 3.1. Participants' demographics by group.

	500 ms synchronized (*n* = 21)			300 ms synchronized (*n* = 21)			Unenhanced captions (*n* = 8)			Uncaptioned (*n* = 8)		
	M	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Age at testing	20.65	3.03	[19.23, 22.07]	20.26	2.50	[19.09, 21.43	19.56	0.85	[18.85, 20.27]	23.70	10.25	[15.13, 32.27]
L2 proficiency (0-120 points)	97.80	13.47	[91.50, 104.10]	98.50	13.78	[92.05, 104.95]	89.50	9.55	[81.52, 97.48]	97.38	15.85	[84.13, 110.62]
Extracurricular classes (years)	2.95	3.91	[1.12, 4.78]	2.28	3.70	[0.54, 4.01]	5.63	4.14	[2.17, 9.08]	4.13	4.12	[0.68, 7.57]
Time spent in an English-speaking country (months)	2.33	4.03	[.50, 4.17]	.93	2.96	[-.42, 2.28]	.19	.35	[-.10, .48]	.75	1.39	[-.41, 1.91]
Estimated spoken L2 input[a]	28.55	16.75	[20.71, 36.39]	22.35	11.23	[17.09, 27.61]	18.50	7.95	[11.86, 25.14]	25.00	10.00	[16.64, 33.36]
Estimated L2 output[b]	10.90	12.48	[5.06, 16.74]	10.48	11.57	[5.06, 15.89]	5.75	1.98	[4.09, 7.41]	12.63	12.76	[1.96, 23.29]
Exposure to L2 videos and TV (hours per week)	8.27	5.99	[5.47, 11.08]	5.26	3.19	[3.76, 6.75]	4.21	1.56	[2.90, 5.51]	5.71	3.81	[2.52, 8.89]
Self-estimated L2 proficiency (1 = very poor – 9 = proficient)[c]	6.41	1.66	[5.63, 7.19]	6.49	1.42	[5.82, 7.16]	6.47	0.81	[5.80, 7.14]	5.40	2.54	[3.27, 7.53]

[a] English input from L1 and L2 speakers in hours per week. [b] Oral L2 use with L1 and L2 speakers in hours per week. [c] Averaged self-estimated reading, writing, listening, speaking and pronunciation proficiency.

Materials

Clips

Four video clips were selected from the initial episode of the television series *The Good Place*, in which the main character named Eleanor wakes up in the afterlife and discovers that she has been sent to the good place instead of the bad place by mistake. This TV series had previously been used in a language acquisition study involving a similar group of learners (Pattemore & Muñoz, 2020). For clip 1, which had a duration of 1 minute and 40 seconds, the captions were presented without any enhancements and contained a total of 19 target words (see Table 3.2). Following that, a sample clip lasting 35 seconds was used to familiarize participants in the experimental groups with audio-synchronized textual enhancement. In the last two clips, each featuring 9 target words (see Table 3.3) and lasting 1 minute and 50 seconds, the target words were highlighted in yellow at different time-lags with auditory onset, depending on the specific experimental condition (see Files 1-3 in the Supplementary Materials for an example of the enhancement conditions). The captions were manually synchronized and enhanced using Aegisub, a software that incorporates a spectrum analyzer, and were hardcoded as one- or two-line captions in Arial font size 20. The Vocabprofile Compleat program (Cobb, 2015) was used to analyze the script of the clips. The most frequently occurring word families, specifically the 1,000-, 2,000-, and 3,000-word families, accounted for 90%, 95%, and 96% coverage of the script, respectively. According to previous studies by Van Zeeland and Schmitt (2013) and Rodgers (2013), a coverage level of 95% from the 2,000-word families was expected to provide intermediate L2 learners with sufficient comprehension.

Target words

Forty target words were selected based on a pilot survey with six L1 Spanish/Catalan speakers taking an intermediate English course at a language academy. After the exclusion of three words, due to their mispronounced versions being too similar to real words in English (e.g., /ˈpɛrsən/ for *person*), the final set included 37 target words (TWs). A subset of 19 words was presented unenhanced in the clips (Table 3.2), whereas 18 words (enhanced subset) were enhanced (Table 3.3). The final list of 37 words contained five error categories: Word stress (*n* = 8), vowel (*n* = 9), diphthong (*n* = 11), or consonant sounds (*n* = 4), and the insertion of an extra vowel in the regular past tense <-ed> form (*n* = 5). These features are problematic for L2 learners of English and can potentially impact intelligibility (Jenkins, 2002; Levis, 2018). Despite the variability naturally occurring in videos not specifically designed for learning, efforts were made to ensure comparable datasets for both the enhanced and unenhanced target words. The majority of the words in both subsets were nouns (9 in the enhanced subset and 7 in the unenhanced subset, respectively), followed by verbs (4 and 8, respectively), adjectives (2 and 4), and finally, 2 adverbs and a conjunction.

Table 3.2. Linguistic properties of the target words (unenhanced subset).

	Word class	Orthographic length	Phon length[a]	Occurrences in clips	Lexical frequency[b]	Error category
Allow	verb	5	3	1	44.37	diphthong
Area	noun	4	3	1	74.92	diphthong
Australia	noun	9	7	1	8.43	diphthong
Betray	verb	6	5	1	9.14	vowel
Come	verb	4	3	2	3140.98	vowel
Control	noun	7	7	1	130.63	stress
Cottage	noun	7	5	1	5.29	diphthong
Earth	noun	5	2	1	99.49	vowel
Embarrassing	adjective	12	9	1	22.84	stress
Ended	verb	5	5	1	29.63	past <-ed>
Existence	noun	9	9	1	11.69	stress
Fundamental	adjective	11	9	1	3.27	vowel
Plowed	verb	6	4	1	0.65	diphthong
Question	noun	8	7	2	198.35	consonant
Raised	verb	6	4	1	25.73	past <-ed>
Returned	verb	8	6	1	24.76	past <-ed>
Rolled	verb	6	4	1	8.47	past <-ed>
Special	adjective	7	5	1	148.57	consonant
Traumatic	adjective	9	8	1	2.71	diphthong

[a] Phonological length, i.e., number of phonemes forming the word, transcribed using the IPA notation system for American English. [b] Frequency per million words in the $SUBTLEX_{US}$ database (Brysbaert & New, 2009).

Table 3.3. Linguistic properties of the target words (enhanced subset).

	Word class	Orthographic length	Phon length	Occurrences in clips	Lexical frequency	Error category
Actually	adverb	8	6	2	322.33	consonant
Adorable	adjective	8	8	1	10.53	stress
Arizona	noun	7	7	3	11.06	vowel
Basically	adverb	9	7	1	26.02	diphthong
Clown	noun	5	4	4	15.82	diphthong
Happened	verb	8	6	1	490.08	past <-ed>
Interior	noun	8	8	1	5.24	vowel
Language	noun	8	7	1	35.1	vowel
Lawyer	noun	6	3	3	79.51	diphthong
Mission	noun	7	5	1	47.06	consonant
Nigeria	noun	7	6	1	0.71	diphthong
Overwhelming	adjective	12	9	1	4.92	vowel
Phoenix	noun	7	6	2	10.88	vowel
Promise	verb	7	6	2	153.12	diphthong
Pursuit	noun	7	6	1	7.04	stress
Rescued	verb	7	7	1	5.41	stress
Review	verb	6	5	1	14.8	stress
Whereas	conjunction	7	5	1	3.55	stress

Regarding the word presentation properties, Fisher's exact tests with Monte Carlo estimations of the *p* values (two-tailed) showed that there was no significant difference between the two subsets in terms of number of caption lines (1 or 2) displayed on-screen ($p = .41$), the occurrence of the TW in caption line 1 or 2 ($p = .52$), and the position of the TW (initial, medial or final) ($p = .16$). Similarly, there were no notable variations between the subsets concerning the number of target words appearing in each line ($p = .18$). In the enhanced subset, there was typically one target word per line, with only one instance where it appeared alongside a word from the unenhanced subset (albeit at opposite ends). A T-test revealed no differences in

presentation time, i.e., the time-lag between caption appearance and offset ($t(35) = -1.67$, $p = .11$). For further details regarding the auditory duration of the words and the size and position of the rectangular areas of interest (AOIs) surrounding each TW, please refer to Table 3.4.

Table 3.4. Presentation properties of the target words in the enhanced subset.

	Caption lines	Line with TW	TW position	Presentation time (ms)	Auditory duration (ms)	AOI position[a] (px)	AOI size (px)
Adorable	2	1	Medial	3220	580	481, 593	184 x 52
Basically	2	2	Initial	3820	560	375, 644	184 x 55
Clown	2	2	Medial	4700	500	602, 647	145 x 50
Happened	2	2	Medial	2650	350	470, 648	207 x 52
Interior	2	2	Medial	3100	620	401, 646	148 x 51
Lawyer	2	1	Medial	4500	470	593, 597	141 x 53
Mission	2	2	Initial	3000	410	409, 648	162 x 50
Review	2	1	Medial	2650	380	549, 594	141 x 52
Whereas	2	2	Initial	3220	390	415, 645	186 x 53
Actually	1	1	Medial	2500	460	415, 643	164 x 55
Arizona	2	2	Initial	1600	640	369, 647	167 x 50
Language	2	1	Medial	2250	350	435, 594	196 x 54
Nigeria	2	1	Final	3420	670	746, 592	151 x 55
Overwhelming	2	2	Final	4960	680	524, 646	296 x 55
Phoenix	1	1	Medial	3650	590	684, 642	174 x 55
Promise	2	1	Medial	4950	420	512, 596	171 x 55
Pursuit	2	1	Final	3400	520	728, 595	142 x 51
Rescued	2	2	Medial	3260	530	505, 647	171 x 50

[a] Horizontal and vertical coordinates (respectively) of the area of interest containing the target word, with reference to the upper-left corner of the screen.

Lexical decision task

L2 learners' performance in lexical decision tasks, which test learners' speed and accuracy in recognizing L2 auditory word forms, reflects the degree of automaticity in L2 lexical access resulting from lexical acquisition (Darcy et al., 2013; Darcy & Holliday, 2019; Harrington, 2006; Llompart & Reinisch, 201; Williams & Paciorek, 2016). The rationale behind using a perceptual task to test pronunciation development

is that the performance on such tasks reflects the degree of development of the learners' L2 phonological system, which is involved in both language perception and production (Mora, 2007; Ramus et al., 2010). In a lexical decision task (LDT) where nonwords reflect L1-based mispronunciations of real L2 words, lower speed and accuracy in rejecting mispronounced forms reflects the instability of imprecise L2 phonological representations, whereas higher speed and accuracy are a sign of stable lexical representations and automaticity of lexical access (Cook et al., 2016; Pellicer-Sánchez, 2015). In the current study, improved accuracy and faster response times in rejecting mispronounced forms indicated that the corresponding representations had been updated in the mental lexicon.

The 157 test items included 40 correctly pronounced target words (e.g., /əˈlaʊ/ for *allow*), 37 corresponding nonwords (L1-biased versions, e.g., /əˈloʊ/ for *allow*), 40 unrelated word distractors and 40 nonword distractors. The word distractors, which did not appear in the script of any of the video clips, were chosen to match the same orthographic and phonological length and lexical frequency as TWs. Nonword distractors were selected from the CLEARPOND database to match the orthographic length, phonological length, and neighborhood size of each TW (Marian et al., 2012). An ANOVA revealed no significant differences in orthographical length and phonological length between real and nonword distractors and TWs ($F(3, 116) = .385$, $p = .76$) and $F(3, 116) = .524$, $p = .66$). All the stimuli were recorded twice by an English L1 speaker of the same variety of American English spoken by the characters in the clips. One recording of each item was saved as a separate sound file in Praat (version 6.0.13), then the audio files were low-pass filtered (60 Hz) and normalized for mean amplitude using the *filter* function in GSU Tools 1.9 (Owren, 2008). The

stimuli were presented randomly and once only, with an inter-stimulus-interval of 2000 ms and a time out of 2500 ms, using the software DMDX 6.0.0.1 (Forster & Forster, 2003). An instruction screen informed participants that they should press a key with their right index finger when they believed the stimulus they heard was an English word. A different key had to be pressed with their left index finger when they believed the stimulus was not an English word. Before the test, they practiced with eight items that did not belong to the set of target words. Two L1 English speakers (different from the speaker who recorded the stimuli) achieved ceiling performance on the task (92% and 94%).

Reading speed test

A short text, followed by a comprehension question, was used to assess the participants' reading speed (Appendix A.1). The paragraph was 93 words long and had been selected from a longer text about the International Space Station found in an EFL textbook. The first 1,000-, 2,000-, and 3,000-word families provided 89%, 93%, and 95% coverage, respectively.

Procedure

Participants were tested individually in the university laboratory, within one session of approximately one hour. After signing a consent form and completing a language background questionnaire, each participant was randomly assigned to one of the viewing conditions. They did the lexical decision task, then read the text and watched the clips as their eye-movements were recorded using a Tobii T120 eye-tracker integrated into a 17" monitor with a resolution of 1024 x 768 pixels. The eye-tracker has a sampling rate of 120 Hz, an accuracy of .5°, and a resolution of .2°. The

participant were seated at a distance of 60 to 64 cm from the screen, which allowed Tobii T120 to accurately track fixations within the areas of interest, as AOI height subtended a visual angle of ~2°, and their width was larger. A 9-point calibration and validation procedure was performed prior to the reading task and repeated before casting the video clips. After watching the first clip with unenhanced captions and the sample clip with target words highlighted at different time intervals, participants in the experimental groups watched one clip with audio-synchronized enhancement (500 ms or 300 ms before auditory onset), and one clip with TWs highlighting at caption onset (henceforth "unsynchronized enhancement"), which served as a within-subject control condition (Figure 3.2). To ensure that the participants would pay attention to the content of the video clips, a multiple choice comprehension question followed each clip (Appendix A.2), and participants were not informed that there would be a post-test. Finally, they did the LDT post-test and the elicited imitation task. The purpose of the study was not disclosed to the participants until the end of the session.

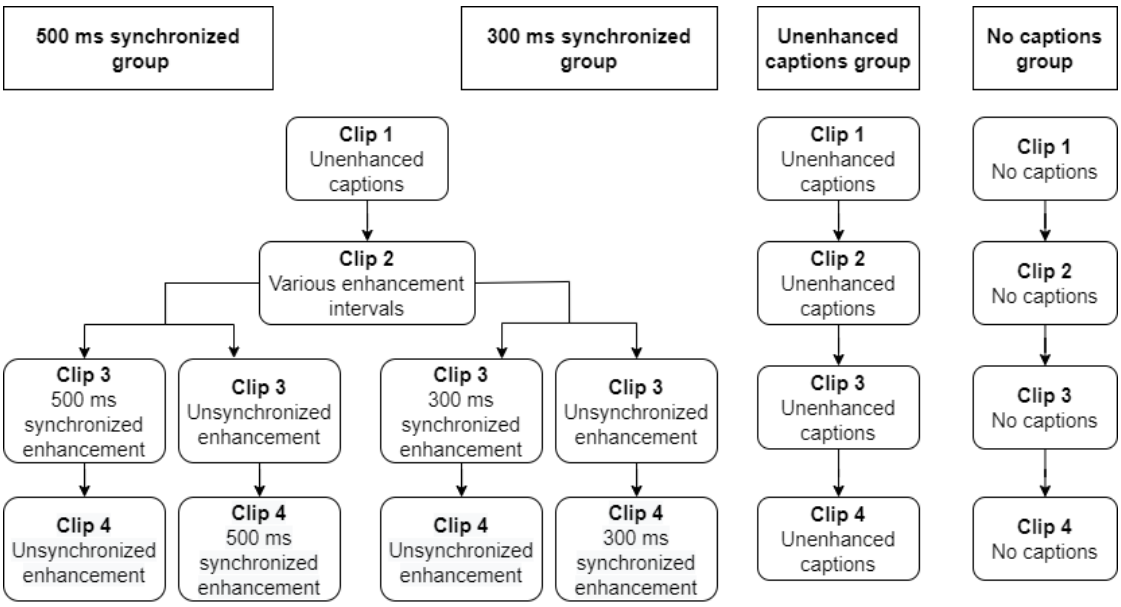


Figure 3.2. Viewing phase flowchart.

Analyses

The eye-tracking data, obtained from Tobii Studio using the I-VT fixation filter, showed that less than 75% of the data was available for three participants, who were excluded from the analyses. Additionally, a fourth participant who did not have fixations in the caption area was also excluded. For the 46 participants included in the analysis, 92.6% of eye-tracking data was available on average. Before analyzing fixation duration and fixation distance, fixations shorter than 50 ms and longer than 800 ms were removed, as 50 ms is the threshold for a fixation to trigger noticing and fixations longer than 800 ms may signal a lapse in attention (Godfroid, 2019). This resulted in the exclusion of 3.5% of the fixations, leaving 1043 fixations out a total of 1702 fixated and skipped items. The LDT data was not screened or transformed, as any response recorded within the available timeframe of 2500 ms was deemed theoretically acceptable, and a-priori data trimming of outliers is unnecessary and potentially detrimental in the analysis of reaction time with mixed models (Baayen & Milin, 2010; Lachaud, & Renaud, 2011). The statistical models were built in RStudio using the *glmer* and *lmer* functions of the *lme4* package, and Bonferroni-adjusted significance tests for pairwise contrasts were obtained using the *lsmeans* function of the *emmeans* package. Model performance and effect sizes (marginal and conditional R-squared values – R2m and R2c) for linear mixed models were assessed using the *performance* package. Pseudo-R-squared values based on the delta method for generalized linear mixed models were obtained using the *r.squaredGLMM* function from the *MuMIn* package. Based on established guidelines for the analysis of human behavior in the social sciences, R-squared values between 0.1 and 0.5 are considered good, provided that one or more variables are statistically significant (Ozili, 2022).

As the assumptions underlying the computation of the asymptotic 95% CIs (and therefore of the *p* values) did not hold for some of the models, basic confidence intervals were calculated from the empirical distribution of the parameter estimate and independently from model assumptions using non-parametric bootstrapping with replacement (*n* = 1e3 simulations). Bootstrapping the regression coefficients of all the models allowed us to provide theoretically valid estimates of the true population parameter even when there was no evidence against the validity of the model assumptions.

We first analyzed the subset of enhanced target words. A mixed effects model based on a gamma distribution and a log-link function was used to compare the effects of viewing condition on the total duration of all fixations within an AOI. To investigate the effects of viewing condition on skipping probability (the proportion of words that were not fixated in relation to the total number of words in the subset), we run a mixed effects logistic regression model based on a binomial distribution and a logit link function. Finally, a linear mixed model was used to assess the effects of viewing condition on fixation distance, or the synchronization between visual and auditory processing. Fixation distance was computed by subtracting the timestamp of the onset of a target auditory form in the soundtrack from the timestamp of the first fixation on the word in the caption (Wisniewska & Mora, 2018). In the analysis of fixation distance, any first fixations occurring more than 3000 ms after the auditory onset of a word (*n* = 12) were replaced with missing values. To control for potential confounds, all eye-tracking models included fixed effects for *Presentation Time* and *Frequency of Occurrence* (across the four clips) alongside *Viewing Condition*, and random intercepts for participants and items. The baseline condition was the unenhanced

condition, unless stated differently in the *Results* section regarding a specific model. If the effects of a covariate did not reach statistical significance, the model was re-run without the covariate.

The responses of all 58 participants were included in the lexical decision task analysis. The dependent measures were accuracy, intended as correctly identifying mispronounced target words (TWs) as nonwords, and response time or the latency between the auditory presentation of a word and participant response. To compare the effects of viewing condition and testing time (T1-T2) on participants' accuracy on the subset of enhanced nonwords, we ran a logistic mixed model based on a binomial distribution and logit link function. The accuracy gains reported in the descriptive tables refer to items that elicited an inaccurate response at T1 and an accurate response at T2. A mixed effects gamma regression was run to measure the effects of viewing condition and testing time on reaction times (RTs). Both models were run including *Viewing Condition*, *Time* and the interaction of *Viewing Condition* * *Time* as fixed effects, and random intercepts for participants and items. RT gains only refer to the items eliciting accurate responses at T2 and were calculated by subtracting each absolute RT at T1 from the corresponding one at T2 and discarding negative gains.

Turning to the control subset of unenhanced words, we re-ran the models replacing the variable *Viewing Condition* (5-level variable) with *Group* (4-level variable), because all participants encountered the words in this subset under the same unenhanced condition. The purpose was to examine whether participants from each group demonstrated different behavior compared to the other groups regardless of the presence of enhancement. For example, we investigated whether participants in certain groups typically skipped more words or fixated on words for longer.

Following Charles (2017), reading speed per minute was calculated by dividing the number of characters in the text (429) by reading time and multiplying the results x 60. Reading time was computed by subtracting the timestamp when the text was displayed from the timestamp of the mouse click that closed the text and opened the comprehension question. Manual inspection of the recordings confirmed that each participant actually read the text. Additionally, Pearson correlations were run by word subset to highlight any relationship between participants' proficiency and reading speed and their eye gaze behavior, and between proficiency and LDT gains. The strength of the correlation was considered weak if the coefficient was between .10 and .30, medium if r was between .30 and .50, and strong if it was between .50 and 1 (Cohen, 1988).

3.1.5. Results

Video comprehension

Participants' responses to the comprehension questions were 87% correct on average (*SD* = 7.56), indicating that overall participants understood the clips and primarily focused on meaning.

Eye gaze results

The absence of fixations on any areas of interest under the uncaptioned condition confirmed that, under the captioned conditions, the participants' attention was not drawn to the areas of interest due to extraneous factors. Figure 3.3 shows the average fixation duration on enhanced and unenhanced target words by viewing condition and group, respectively.

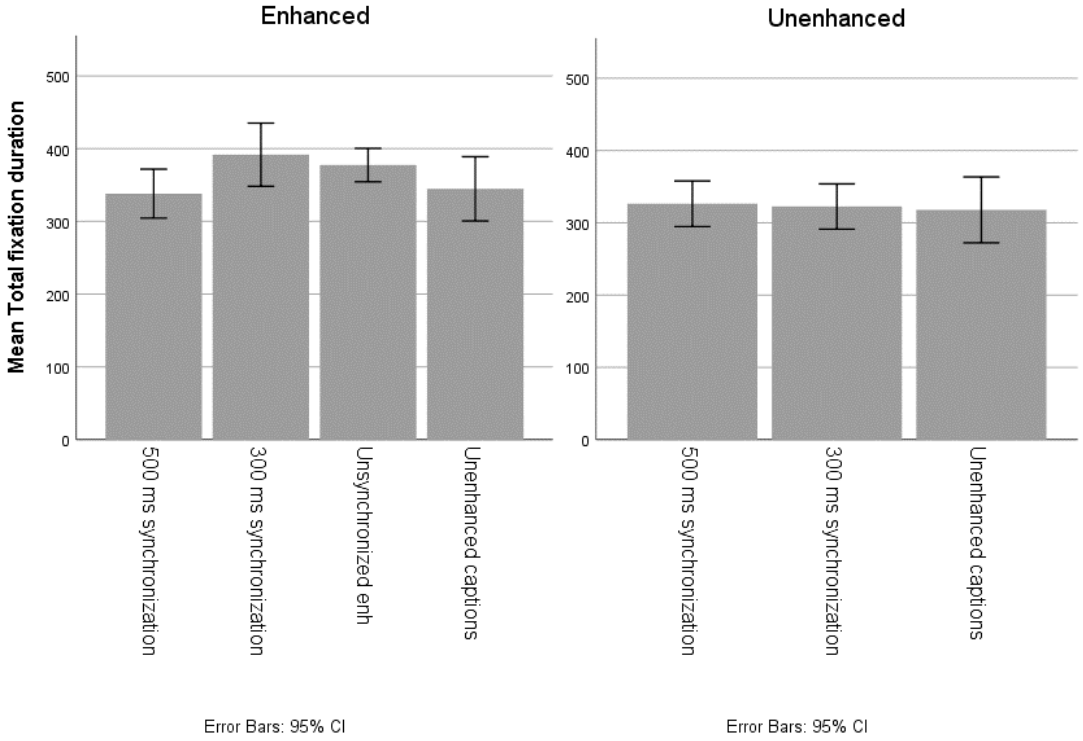


Figure 3.3 Total fixation duration by viewing condition (enhanced subset) and group (unenhanced subset).

The total fixation duration descriptive data for the subset of target words with enhanced captions can be found in Table 3.5. Although there was no statistically significant effect of *Viewing Condition* or *Presentation Time* on total fixation duration based on asymptotic CIs, the bootstrapped confidence intervals showed that *Presentation Time*, as well as the 300 ms synchronized and the unsynchronized enhancement conditions, had a significant and positive effect on the total fixation duration (Table 3.6). However, the discrepancy between asymptotic and bootstrapped intervals, along with the small *R2m* and *R2c* values, suggest that while there are statistically significant effects on total fixation duration for certain variables, the actual size of these effects is relatively small. Pairwise contrasts did not detect any difference between the conditions (Table 3.7).

Table 3.5. Total fixation duration by viewing condition (enhanced subset).

				95% Confidence Intervals	
	N	*M (ms)*	*SD (ms)*	Lower	Upper
500 ms synchronization	129	338.14	193.35	304.46	371.82
300 ms synchronization	128	391.64	248.43	348.19	435.09
Unsynchronized enhancement	281	377.39	195.81	354.40	400.38
Unenhanced captions	94	344.70	215.61	300.54	388.86

Table 3.6. Gamma regression examining total fixation duration on the enhanced subset of TWs.

							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	*SE*	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	5.72	0.12	46.16	< .001***	0.04	0.22	5.47	5.96	5.63	5.84
500 ms synchronization	0.03	0.13	0.26	.79			-0.21	0.28	-0.10	0.16
300 ms synchronization	0.14	0.13	1.07	.29			-0.11	0.38	0.00	0.26
Unsynchronized enhancement	0.13	0.12	1.10	.27			-0.10	0.37	0.01	0.24
Presentation time	0.10	0.06	1.81	.07			-0.01	0.21	0.07	0.14

*** $p < .001$

Table 3.7. Results of pairwise contrasts for total fixation duration (enhanced subset).

					95% Confidence Intervals	
	Contrast estimate	*SE*	*z ratio*	*p*	Lower	Upper
500 ms synchronization - 300 ms synchronization	-0.10	0.08	-1.36	1.00	-0.30	0.10
500 ms synchronization - unsynchronized enhancement	-0.10	0.06	-1.79	.44	-0.24	0.05
300 ms synchronization - unsynchronized enhancement	0.00	0.06	0.06	1.00	-0.14	0.15
Unenhanced captions - 500 ms synchronization	-0.03	0.13	-0.26	1.00	-0.37	0.30
Unenhanced captions - 300 ms synchronization	-0.14	0.13	-1.07	1.00	-0.47	0.20
Unenhanced captions - Unsynchronized enhancement	-0.13	0.12	-1.10	1.00	-0.45	0.19

To analyze the control subset of *unenhanced* words, which participants in all groups watched under unenhanced condition, *Group* was used as a predictor instead of *Viewing Condition*. The results indicated that *Presentation Time* had a significant impact on the total fixation duration, but *Group* and *Frequency of Occurrence* did not show significant effects (Table 3.8). Additionally, pairwise comparisons conducted on the unenhanced subset revealed no differences among any of the groups.

Table 3.8. Fixed coefficients for the gamma regression examining total fixation duration (unenhanced subset).

							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	*SE*	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	5.63	0.12	46.08	< .001***	0.10	0.29	5.39	5.87	5.51	5.79
500 ms synchronization	-0.08	0.11	-0.73	.47			-0.31	0.14	-0.24	0.08
300 ms synchronization	-0.14	0.11	-1.21	.23			-0.36	0.09	-0.28	0.03
Presentation time	0.19	0.08	2.46	.01*			0.04	0.35	0.12	0.27

*** $p < .001$, * $p < .05$

Figure 3.4 shows the average skipping probability for enhanced and unenhanced target words by viewing condition and group, respectively. The descriptive data regarding the probability of skipping words in the enhanced subset of target words is reported by enhancement condition in Table 3.9. The logistic mixed-effects model showed a significant and negative effect of *Viewing Condition* on skipping probability for the unsynchronized enhancement condition ($p = .04$), based on both asymptotic and bootstrapped confidence intervals (Table 3.10). *Presentation Time* and *Frequency of Occurrence* were not found to have an effect on skipping probability. Pairwise contrasts revealed no difference between any of the conditions (Table 3.11).

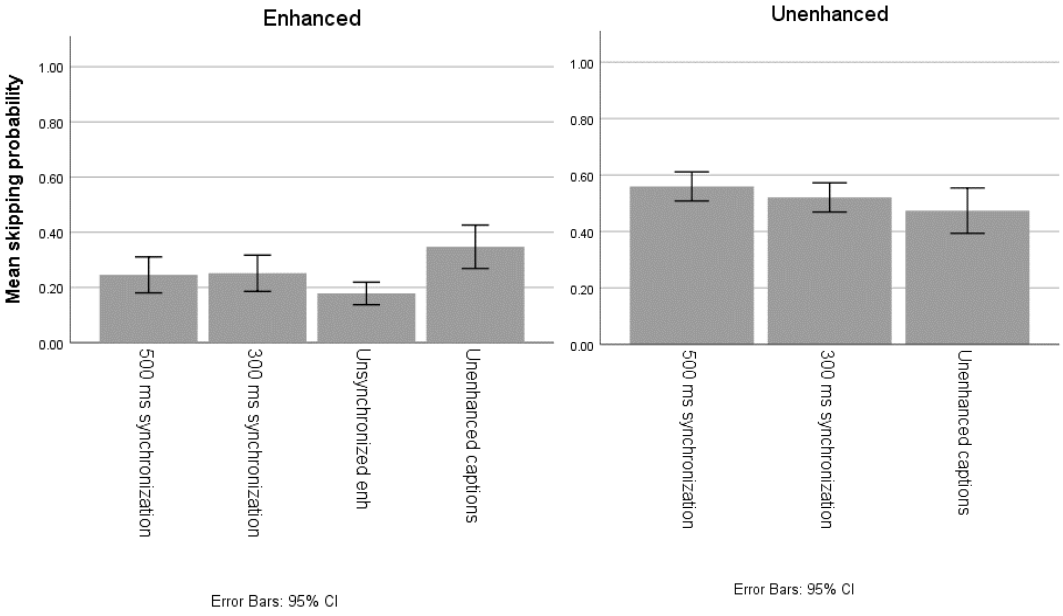


Figure 3.4. Skipping probability by viewing condition (enhanced subset) and group (unenhanced subset).

Table 3.9. Skipping probability by viewing condition (enhanced subset).

				95% Confidence Intervals	
	N	*M*	*SD*	Lower	Upper
500 ms synchronization	171	24.56%	43.17	18.04	31.08
300 ms synchronization	171	25.15%	43.51	18.58	31.71
Unsynchronized enhancement	342	17.84%	38.34	13.76	21.91
Unenhanced captions	144	34.72%	47.77	26.85	42.59

Table 3.10. Fixed coefficients for the logistic regression examining skipping probability (enhanced subset).

							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	*SE*	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	-0.84	0.46	-1.81	.07	0.03	0.32	-1.74	0.07	-1.15	-0.29
500 ms synchronization	-0.58	0.54	-1.07	.29			-1.64	0.48	-1.14	0.16
300 ms synchronization	-0.54	0.54	-0.99	.32			-1.60	0.52	-1.08	0.18
Unsynchronized enhancement	-1.04	0.52	-2.02	.04*			-2.06	-0.03	-1.46	-0.37

* $p < .05$

Table 3.11. Results of pairwise contrasts for skipping probability (enhanced subset).

					95% Confidence Intervals	
	Contrast estimate	*SE*	*z ratio*	*p*	Lower	Upper
500 ms synchronization - 300 ms synchronization	-0.04	0.36	-0.11	1.00	-0.98	0.91
500 ms synchronization - unsynchronized enhancement	0.47	0.27	1.74	.49	-0.24	1.17
300 ms synchronization - unsynchronized enhancement	0.50	0.27	1.84	.39	-0.22	1.23
Unenhanced captions - 500 ms synchronization	0.58	0.54	1.07	1.00	-0.85	2.01
Unenhanced captions - 300 ms synchronization	0.54	0.54	0.99	1.00	-0.89	1.97
Unenhanced captions - Unsynchronized enhancement	1.04	0.52	2.02	.26	-0.32	2.41

The logistic regression analysis conducted on the subset of unenhanced words revealed a significant effect of *Presentation Time* on skipping probability, but not of *Frequency of Occurrence* (Table 3.12). Although, when considering asymptotic confidence intervals, there was no significant effect of *Group*, the bootstrapped confidence intervals for the 500 ms synchronized group did not contain 0, i.e., participants in this group were found to be more likely to skip words. Pairwise contrasts did not show any significant differences between the groups (*p*s = 1.00).

Table 3.12. Fixed coefficients for the logistic regression examining skipping probability (unenhanced subset).

							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	*SE*	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	-0.11	0.54	-0.20	.84	0.06	0.42	-1.17	0.95	-0.52	0.29
500 ms synch	0.53	0.65	0.82	.41			-0.74	1.80	0.03	0.97
300 ms synch	0.31	0.65	0.47	.64			-0.96	1.57	-0.18	0.80
Presentation time	-0.57	0.09	-6.69	< .001***			-0.74	-0.40	-0.71	-0.33

*** $p < .001$

The descriptive data for the last eye-tracking measure, fixation distance, is reported in Table 3.13 and Figure 3.5. As described in the *Methodology* subsection, fixation distance was calculated by subtracting the timestamp of a word auditory onset from the timestamp of the first fixation on the word. Therefore, the prevalence of positive values, i.e., pre-fixations on the target words, indicates that the target words were generally fixated before their auditory onset (AO). Since the standard deviation was larger than the mean values, we report median and interquartile range (IQR) instead (Baayen & Milin, 2010), where interquartile range is the range of values that lies between the upper quartile and lower quartile.

Table 3.13. Fixation distance by viewing condition (enhanced subset).

				95% Confidence Intervals	
	N	*Mdn* (ms)	*IQR* (ms)	Lower	Upper
500 ms synchronization	128	136.90	615.95	4.48	264.11
300 ms synchronization	126	239.70	740.10	118.97	398.94
Unsynchronized enhancement	277	368.80	831.90	419.81	609.33
Unenhanced captions	93	307.60	687.60	187.79	506.14

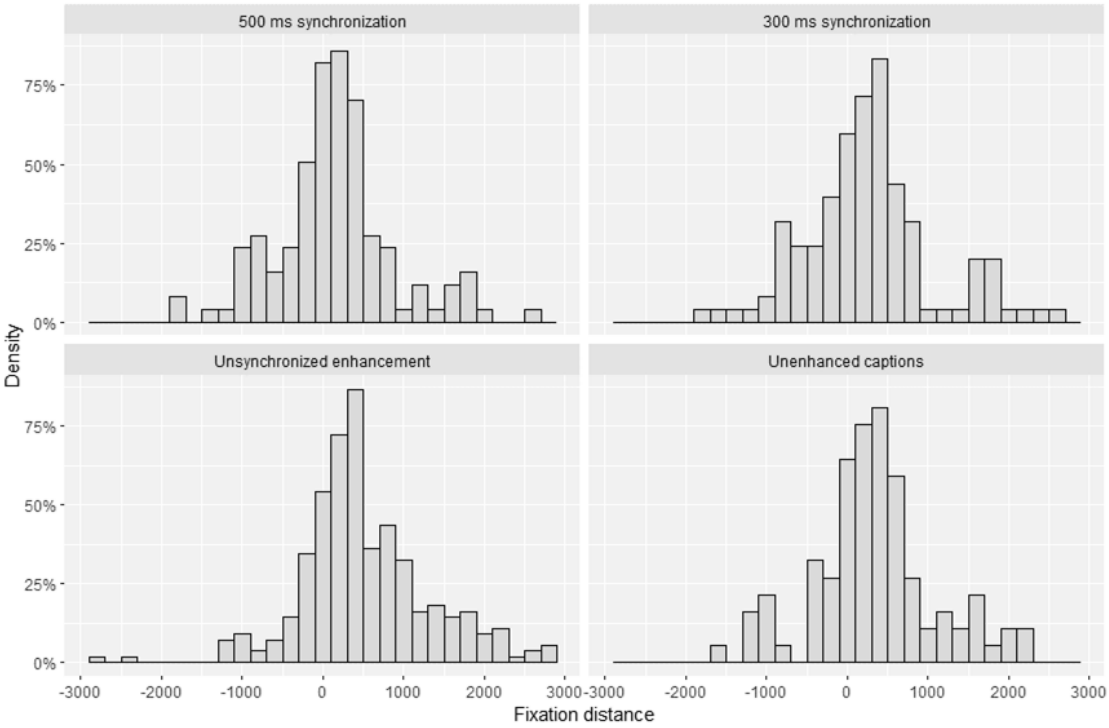


Figure 3.5. Distribution of pre-fixations and post-fixations in relation to word auditory onset.

For the enhanced subset of words, the linear regression analysis did not yield significant effects of *Viewing Condition* on fixation distance when the unenhanced condition was used as baseline. However, when considering the *500 ms synchronized enhancement condition* as the reference level for the predictor (Table 3.14), the model showed that the 300 ms synchronized enhancement condition and the unsynchronized enhancement condition had a significant effect on fixation distance (p = .02 and p <

.001, respectively). While *Presentation Time* had a statistically significant effect (p = .05), *Frequency of Occurrence* did not. According to pairwise contrasts (Table 3.15), both the 500 ms and 300 ms synchronized conditions were associated with smaller fixation distances compared to the unsynchronized condition ($p < .001$ and $p = .03$, respectively). For the unenhanced subset of words, the linear regression model showed a significant effect of *Presentation Time* on fixation distance, but no significant effects of *Frequency of Occurrence* or *Group* (Table 3.16). Pairwise contrasts indicated no significant differences between the groups (ps = 1.00).

Table 3.14. Fixed coefficients for the linear mixed model examining fixation distance.

								95% Confidence Intervals		
								Asymptotic		Bootstrapped	
	B	*SE*	*df*	*t value*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	93.74	131.71	39.69	0.71	.48	0.10	0.55	-164.42	351.90	-33.33	223.94
300 ms synchronization	220.27	93.38	588.20	2.36	.02*			37.25	403.30	31.73	398.69
Unsynchronized enhancement	410.47	67.89	599.81	6.05	< .001***			277.40	543.50	266.09	554.42
Unenhanced captions	159.29	174.16	51.02	0.91	.37			-182.06	500.60	-30.45	332.04
Presentation time	221.74	102.94	15.96	2.15	.05*			19.98	423.50	165.69	277.96

Table 3.15. Results for pairwise contrasts for fixation distance (enhanced subset).

						95% Confidence Intervals	
	Contrast estimate	*SE*	*df*	*t ratio*	*p*	Lower	Upper
500 ms synchronization - 300 ms synchronization	-220.00	94	590	-2.35	.12	-469.00	28.20
500 ms synchronization - unsynchronized enhancement	-410.00	68	600	-6.03	< .001***	-591.00	-230.30
500 ms synchronization - unenhanced captions	-159.00	174	55	-0.91	1.00	-636.00	317.80
300 ms synchronization - unsynchronized enhancement	-190.00	69	600	-2.77	.03*	-372.00	-8.60
300 ms synchronization - unenhanced captions	61.00	175	55	0.35	1.00	-417.00	538.50
unsynchronized enhancement - unenhanced captions	251.00	168	48	1.50	.84	-210.00	712.40

Table 3.16. Fixed coefficients for the linear model examining fixation distance (unenhanced subset).

								95% Confidence Intervals			
								Asymptotic		Bootstrapped	
	B	*SE*	df	*t value*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	17.03	140.83	30.9	0.12	.91	0.11	0.56	-258.98	293.00	-87.32	144.60
500 ms synchronization	9.24	98.01	30.28	0.09	.93			-182.85	201.30	-154.03	131.70
300 ms synchronization	25.81	97.29	29.88	0.27	.79			-164.86	216.50	-120.70	175.50
Presentation time	272.01	119.78	17.73	2.27	.04*			37.25	506.80	217.86	326.60

*** $p < .001$, * $p < .05$

Summary of eye gaze results

The 300 ms and unsynchronized enhancement conditions resulted in longer total fixation duration on the target words compared to the unenhanced condition, and only the unsynchronized enhancement condition was associated with less skipping. However, significance should be interpreted cautiously due to discrepancies between asymptotic and bootstrap confidence intervals and the relatively small proportion of variance explained by the predictor variables (*R2m* and *R2c*). The fixation distance model, on the other hand, had higher R-squared values, indicating that the model accounted for a larger proportion of the variability in the dependent variable. Under the unsynchronized enhancement condition, the time gap between first fixation and auditory onset was larger compared to other enhanced conditions, suggesting that both synchronized enhancement conditions promoted stricter audiovisual synchronization compared to the unsynchronized condition, although not compared to the unenhanced condition. Interestingly, while presentation time affected total fixation duration and fixation distance, frequency of occurrence did not have a significant effect on eye gaze behavior.

To verify that any difference obtained for the subset of enhanced TWs were contingent on the enhancement condition, we repeated the analyses on the subset of words that were watched under unenhanced condition by all participants. No differences were found between the groups for total fixation duration and fixation distance, confirming that the results obtained for the subset of enhanced TWs depended on the enhancement condition. However, the higher skipping probability found for the 500 ms synchronized group on the unenhanced subset of words indicated a natural inclination for participants in this group to skip more target words regardless of enhancement.

Lexical decision task results

As all tests were conducted in a single session, previous knowledge of the meaning of the target words was not tested to avoid interference with the results. However, the participants' ability to accurately recognize the correctly pronounced version of each word indicated that they were already familiar with each item. The data provided in Table 3.17 shows that the accuracy of their responses reached ceiling for both the enhanced subset (*m* = 93%, *SD* = 26) and unenhanced subset (*m* = 90%, *SD* = 30). The high accuracy in identifying correctly pronounced target *words* was expected and did not undermine the improvements observed for target *nonwords*, because L2 speakers who accept mispronounced versions of words are thought to have stored "unstable" representations of those words, regardless of their ability to recognize the correctly pronounced versions (Cook et al., 2016). Consequently, accepting an accurately pronounced word does not guarantee the rejection of incorrect mispronunciations of that word, as supported by the low accuracy rates obtained on the nonword items. The average gain in response time to distractor items, which served as an indicator of test practice effect, was 230.88 ms (*SD* = 209.42, 95% CI [224.58, 237.19].

Table 3.17. Average scores at time 1 for accurately pronounced target words.

Item	Subset	Mean	Std. Deviation
Allow	Unenhanced	.93	.26
Area	Unenhanced	.90	.31
Come	Unenhanced	.91	.28
Control	Unenhanced	.98	.13
Earth	Unenhanced	.91	.28
Embarrassing	Unenhanced	.90	.31
Ended	Unenhanced	.95	.22
Existence	Unenhanced	1.00	.00
Plowed	Unenhanced	.64	.49
Question	Unenhanced	.98	.13
Returned	Unenhanced	1.00	.00
Rolled	Unenhanced	.66	.48
Traumatic	Unenhanced	.95	.22
Cottage	Unenhanced	.64	.49
Special	Unenhanced	.98	.13
Australia	Unenhanced	.97	.18
Betray	Unenhanced	.93	.26
Fundamental	Unenhanced	.95	.22
Raised	Unenhanced	.88	.33
Adorable	Enhanced	.98	.13
Basically	Enhanced	.98	.13
Clown	Enhanced	.93	.26
Happened	Enhanced	1.00	.00
Interior	Enhanced	.91	.28
Lawyer	Enhanced	.97	.18
Mission	Enhanced	1.00	.00
Review	Enhanced	.97	.18
Whereas	Enhanced	.83	.38
Actually	Enhanced	.95	.22
Arizona	Enhanced	.88	.33
Language	Enhanced	.98	.13
Nigeria	Enhanced	.72	.45
Overwhelming	Enhanced	.93	.26
Phoenix	Enhanced	.72	.45
Promise	Enhanced	1.00	.00
Pursuit	Enhanced	.95	.22
Rescued	Enhanced	.98	.13

The descriptive data on participants' accurate recognition of enhanced target words is reported in Table 3.18 by time and word enhancement condition. The logistic regression failed to converge when the baseline was set as the *uncaptioned condition*,

so the baseline was set as the *500 ms synchronized condition* (Table 3.19). The results indicate a significant effect of Time (p = .002), suggesting that there were changes in accuracy over time. However, none of the conditions or *Time * Condition* interactions reached statistical significance, considering either asymptotic or bootstrapped confidence intervals, and based on the pairwise comparisons, only the *unsynchronized condition* showed a significant improvement in accuracy (Table 3.20). For the unenhanced subset of words, there was a significant effect of Time (p < .001) on accuracy, but the model did not yield any significant effects of *Group* or the interaction between *Group* and *Time* (Table 3.21). When examining the pairwise contrasts between pre- and post-test for the unenhanced subset, significant accuracy gains were found only for the *500 ms synchronized group* (Table 3.22).

Table 3.18. Accuracy averaged scores (max 1) and gains for enhanced target nonwords.

					95% Confidence Intervals	
	Time	*N*	*M*	*SD*	Lower	Upper
500 ms synchronization	1	189	0.37	0.48	0.30	0.44
	2	189	0.50	0.50	0.43	0.57
	Gains	378	0.21	0.41	0.17	0.25
300 ms synchronization	1	189	0.38	0.49	0.31	0.45
	2	189	0.48	0.50	0.40	0.55
	Gains	378	0.16	0.37	0.13	0.20
Unsynchronized enhancement	1	378	0.37	0.48	0.32	0.41
	2	378	0.49	0.50	0.44	0.55
	Gains	756	0.21	0.40	0.18	0.24
Unenhanced captions	1	144	0.31	0.46	0.23	0.38
	2	144	0.44	0.50	0.36	0.53
	Gains	288	0.22	0.41	0.17	0.26
Uncaptioned	1	144	0.42	0.49	0.34	0.50
	2	144	0.47	0.50	0.39	0.55
	Gains	288	0.17	0.38	0.13	0.22

Table 3.19. Fixed coefficients for the logistic regression examining accuracy (enhanced subset).

							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	*SE*	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	-1.33	0.52	-2.54	.01*	0.03	0.46	-2.36	-0.30	-2.16	-0.37
300 ms synchronization	-0.18	0.60	-0.31	.76			-1.36	0.99	-1.29	1.05
Unsynchronized enhancement	-0.32	0.50	-0.63	.53			-1.31	0.67	-1.38	0.76
Unenhanced captions	-0.67	0.78	-0.86	.39			-2.20	0.86	-1.92	0.67
Uncaptioned	0.61	0.76	0.80	.42			-0.88	2.09	-0.80	1.87
Time	0.78	0.25	3.05	.002**			0.28	1.28	0.19	1.24
300 ms synch*time	-0.15	0.36	-0.41	.68			-0.86	0.56	-0.86	0.56
Unsynch enhancement*time	0.00	0.31	0.01	.99			-0.61	0.61	-0.62	0.66
Unenhanced captions*time	0.08	0.39	0.20	.84			-0.68	0.84	-0.71	0.86
Uncaptioned*time	-0.47	0.38	-1.25	.21			-1.20	0.27	-1.27	0.39

** $p < .01$, * $p < .05$

Table 3.20. Results of pairwise contrasts for accuracy (enhanced subset).

							95% Confidence Intervals	
	Testing time	Contrast estimate	*SE*	df	*z*	*p*	Lower	Upper
500 ms synchronization	1 - 2	-0.78	0.25	Inf	-3.05	.10	-1.61	0.05
300 ms synchronization	1 - 2	-0.63	0.26	Inf	-2.45	.65	-1.47	0.21
Unsynchronized enhancement	1 - 2	-0.78	0.18	Inf	-4.35	<.001***	-1.36	-0.20
Unenhanced captions	1 - 2	-0.86	0.29	Inf	-2.91	.16	-1.81	0.10
Uncaptioned	1 - 2	-0.31	0.28	Inf	-1.12	1.00	-1.21	0.59

*** $p < .001$

Table 3.21. Fixed coefficients for the logistic regression examining accuracy (unenhanced subset).

							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	*SE*	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	0.91	-0.42	2.15	.03*	0.02	0.41	0.08	1.74	-1.42	-0.25
300 ms synchronization	-0.62	-0.48	-1.30	.19			-1.56	0.31	-0.21	1.39
Unenhanced captions	0.33	-0.65	0.51	.61			-0.94	1.59	-1.41	0.73
Uncaptioned*time	-0.16	-0.63	-0.26	.80			-1.41	1.08	-0.97	1.15
Time	-0.55	-0.17	-3.24	< .001***			-0.88	-0.22	0.15	0.89
300 ms synch*time	0.13	-0.24	0.53	.59			-0.34	0.60	-0.62	0.37
Unenhanced captions*time	-0.09	-0.32	-0.28	.78			-0.72	0.54	-0.57	0.77
Uncaptioned*time	-0.02	-0.31	-0.05	.96			-0.63	0.60	-0.61	0.67

*** $p < .001$, * $p < .05$

Table 3.22. Results of pairwise contrasts for accuracy (unenhanced subset).

							95% Confidence Intervals	
	Testing time	Contrast estimate	*SE*	df	*z*	*p*	Lower	Upper
500 ms synchronization	1 - 2	-0.55	0.17	Inf	-3.24	.03*	-1.08	-0.02
300 ms synchronization	1 - 2	-0.42	0.17	Inf	-2.48	.37	-0.95	0.11
Unenhanced captions	1 - 2	-0.64	0.28	Inf	-2.34	.55	-1.50	0.22
Uncaptioned	1 - 2	-0.57	0.27	Inf	-2.13	.92	-1.40	0.26

* $p < .05$

The descriptive data for reaction time in response to the enhanced subset of TWs can be found in Table 3.23. Figure 3.6 displays the average response times along with 95% confidence intervals by condition and testing time, with the minimum value set at 1000 milliseconds for clarity. The gamma regression analysis revealed a significant effect of Time (p = .04) and of the interaction between the *500 ms condition* and *Time* (p = .003) on response times. These findings were further confirmed through bootstrapping as shown in Table 3.24. Pairwise contrasts showed that only the enhancement conditions led to significant RT gains, with p values < .001 (Table 3.25). In the analysis of RT for the unenhanced word subset (Table 3.26), no significant effects of *Group* or the interaction between *Group* and *Time* were observed, but there was a significant effect of *Time* (p < .001). When examining the pairwise comparisons between pre- and post-test for the unenhanced subset, significant RT gains were found for the *300 ms group* and *unenhanced group* only (Table 3.27).

Table 3.23. Reaction time averages and gains for enhanced target nonwords.

					95% Confidence Intervals	
	Time	*N*	*M*	*SD*	Lower	Upper
500 ms synchronization	1	70	1560.94	397.98	1466.05	1655.84
	2	94	1312.65	235.28	1264.46	1360.84
	Gains	114	416.72	357.86	350.32	483.12
300 ms synchronization	1	71	1588.56	343.46	1507.27	1669.86
	2	90	1417.15	334.08	1347.18	1487.12
	Gains	124	326.05	240.62	283.28	368.82
Unsynchronized enhancement	1	138	1470.75	328.36	1415.48	1526.03
	2	187	1392.63	327.52	1345.38	1439.88
	Gains	236	314.39	278.59	278.67	350.12
Unenhanced captions	1	44	1462.33	251.56	1385.85	1538.82
	2	64	1463.59	340.02	1378.65	1548.52
	Gains	74	243.74	214.83	193.96	293.51
Uncaptioned	1	60	1466.55	356.43	1374.47	1558.62
	2	68	1373.97	290.32	1303.70	1444.24
	Gains	76	278.78	248.12	222.09	335.48

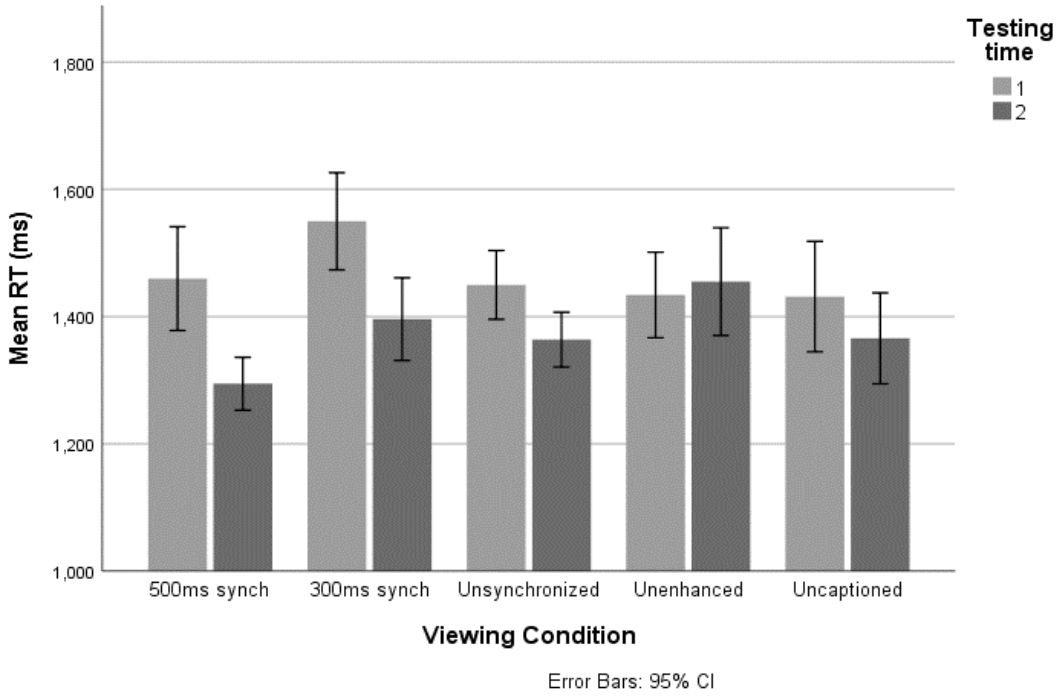


Figure 3.6. Average reaction times by time and condition.

Table 3.24. Fixed coefficients for the fixed effects gamma regression examining reaction time.

							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	*SE*	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	7.45	0.09	85.04	< .001***	0.10	0.32	7.28	7.62	7.35	7.55
500 ms synch	0.12	0.10	1.27	.20			-0.07	0.31	-0.02	0.27
300 ms synch	0.04	0.10	0.40	.69			-0.15	0.23	-0.09	0.17
Unsynch enhancement	-0.01	0.09	-0.10	.92			-0.19	0.17	-0.13	0.12
Unenhanced captions	0.00	0.12	-0.02	.98			-0.23	0.22	-0.15	0.13
Time	-0.06	0.03	-2.08	.04*			-0.12	0.00	-0.12	-0.01
500 ms synch*time	-0.12	0.04	-2.99	.003**			-0.20	-0.04	-0.20	-0.04
300 ms synch*time	-0.05	0.04	-1.27	.20			-0.13	0.03	-0.12	0.03
Unsynch*time	-0.03	0.03	-0.98	.33			-0.10	0.03	-0.10	0.03
Unenhanced captions*time	0.04	0.04	0.87	.39			-0.05	0.12	-0.04	0.13

*** $p < .001$, ** $p < .01$, * $p < .05$

Table 3.25. Results of pairwise contrasts for reaction time (enhanced subset).

							95% Confidence Intervals	
	Testing time	Contrast estimate	*SE*	df	*z*	*P*	Lower	Upper
500 ms synchronization	1 - 2	0.18	0.03	Inf	6.80	< .001***	0.09	0.27
300 ms synchronization	1 - 2	0.11	0.03	Inf	4.23	< .001***	0.03	0.20
Unsynchronized enhancement	1 - 2	0.10	0.02	Inf	5.07	< .001***	0.03	0.16
Unenhanced captions	1 - 2	0.02	0.03	Inf	0.72	1.00	-0.08	0.13
Uncaptioned	1 - 2	0.06	0.03	Inf	2.08	1.00	-0.03	0.16

*** $p < .001$

Table 3.26. Fixed coefficients for the RT gamma regression (unenhanced subset).

							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	*SE*	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	7.49	0.09	86.85	< .001***	0.08	0.28	7.32	7.66	7.40	7.58
500 ms synch	-0.03	0.09	-0.37	.71			-0.21	0.14	-0.14	0.08
300 ms synch	-0.01	0.09	-0.15	.88			-0.19	0.16	-0.11	0.10
Unenhanced captions	0.03	0.11	0.29	.77			-0.18	0.24	-0.10	0.16
Time	-0.08	0.03	-2.90	< .001***			-0.14	-0.03	-0.14	-0.03
500 ms synch*time	-0.03	0.03	-0.93	.35			-0.10	0.03	-0.10	0.03
300 ms synch*time	-0.01	0.03	-0.16	.87			-0.07	0.06	-0.07	0.06
Unenhanced captions*time	0.00	0.04	0.09	.93			-0.08	0.08	-0.07	0.09

*** $p < .001$

Table 3.27. Results of pairwise contrasts for reaction time (unenhanced subset).

	Testing time	Contrast estimate	*SE*	df	*z*	*p*	95% Confidence Intervals	
							Lower	Upper
500 ms synchronization	1 - 2	0.08	0.03	Inf	2.90	.10	-0.01	0.17
300 ms synchronization	1 - 2	0.11	0.02	Inf	6.26	< .001***	0.06	0.17
Unenhanced captions	1 - 2	0.09	0.02	Inf	5.17	< .001***	0.03	0.14
Uncaptioned	1 - 2	0.08	0.03	Inf	2.55	.30	-0.02	0.17

Summary of lexical decision task results

To sum up, the analysis of accuracy scores yielded inconclusive results, as only the unsynchronized condition led to significant gains in participants' rejection of mispronounced target words, but the amount of variance explained by the fixed factors was minimal. Only the viewing conditions involving enhancement were associated with significant gains in reaction time to the mispronounced TWs, suggesting that lexical access was facilitated by the previous exposure to enhanced captioned video. However, the significant effect of time observed for both the enhanced TWs and the control subset of unenhanced words pointed at the possibility of a practice effect, and the significant gains achieved by some of the groups on the unenhanced subset represent a confound.

Effects of proficiency and reading speed

No significant correlation was found between participants' proficiency and total fixation duration in the enhanced subset ($r(46) = -28$, $p = .06$) or unenhanced subset ($r(42) = -.09$, $p = .59$). Similarly, non-significant correlations were found for skipping probability in the enhanced subset ($r(46) = -.28$, $p = .06$) and unenhanced subset ($r(46)$

= .18, p = .23). As the p values were close to significance for words in the enhanced subset, we ran again the correlations excluding participants in the unenhanced group who were exposed to all words without enhancement (including words in the enhanced subset). By limiting the analysis to groups that watched the videos with enhancement, we found that total fixation duration was, in fact, correlated with proficiency (r(38) = -.34, p = .04), while skipping probability was not (p = .44). Based on these further analyses, participants at different proficiency levels were equally as likely to skip (not fixate) enhanced target words, but the lower the proficiency level, the longer a participant fixated on these words (Figure 3.7). Vice versa, the higher the proficiency level, the less attention a participant paid to enhanced words.

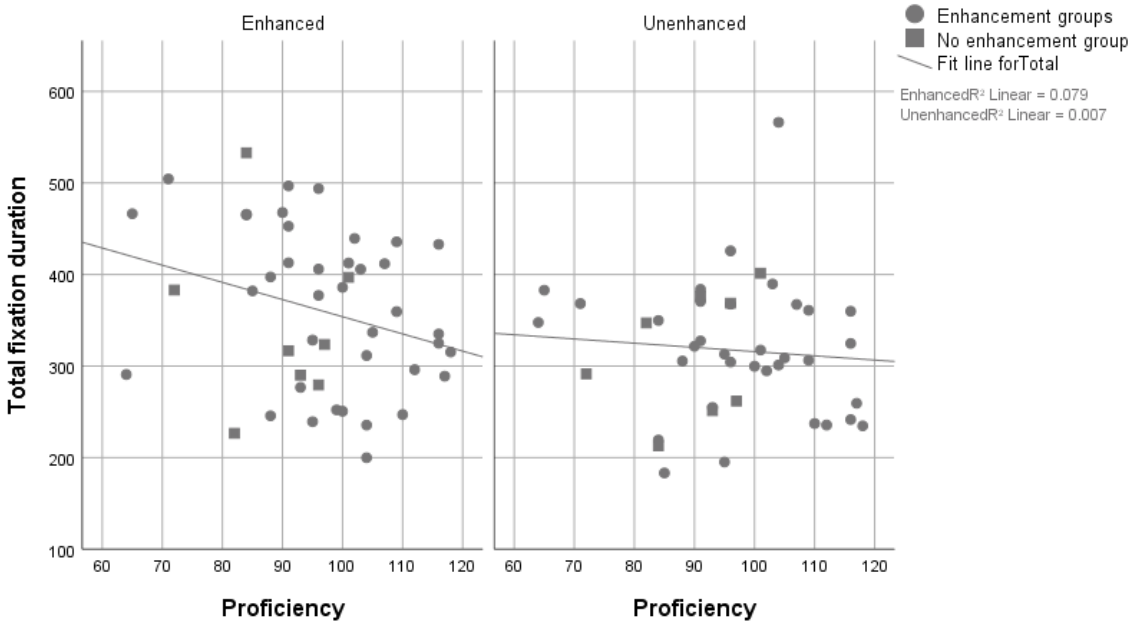


Figure 3.7. Correlation between proficiency and total fixation duration by word subset.

Regarding fixation distance, although the correlation with proficiency was not significant for the enhanced subset (r(46) = .07, p = .64), the correlation was strong and significant for the unenhanced subset (r(42) = .52, p < .001). The visual

representation of fixation distance data for the unenhanced subset, representing participants' baseline behavior, shows that an increase in participants' proficiency corresponded to an increase in pre-fixation distance, and a decrease in proficiency corresponded to an increase in post-fixation distance (Figure 3.8). In other words, the most proficient learners fixated on words in captions way earlier than their auditory onset, whereas the opposite was true for less proficient learners, who fixated on words long after these words were pronounced. However, the enhancement appeared to successfully direct participants' attention to the target words in captions shortly before their auditory onset regardless of proficiency.

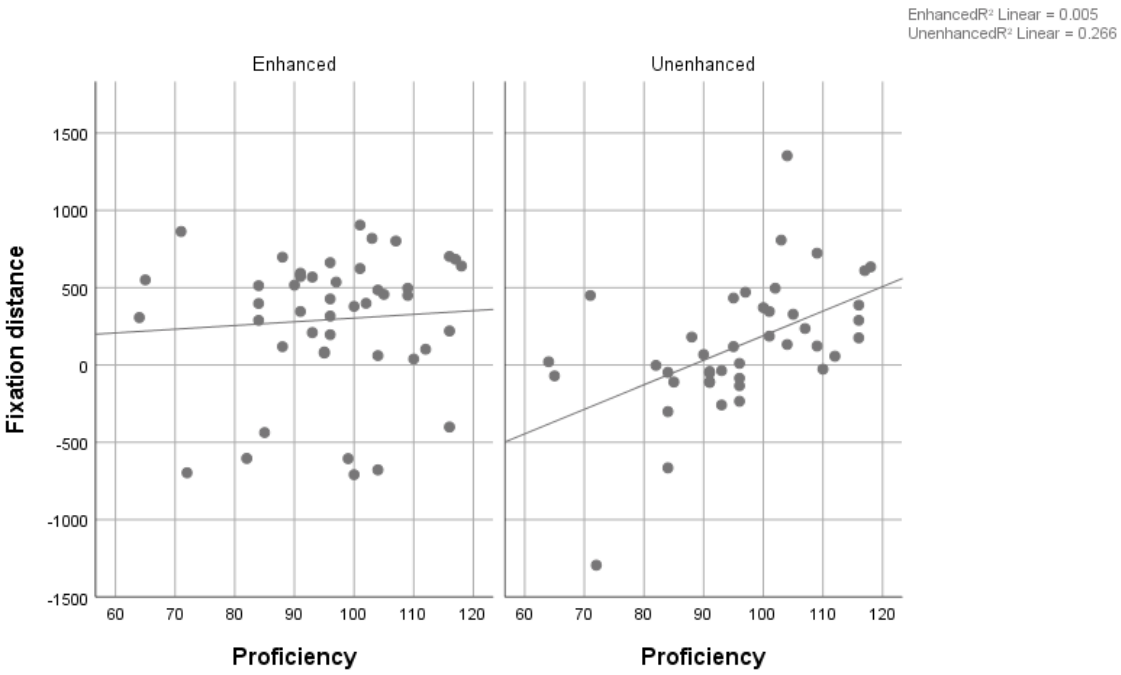


Figure 3.8. Correlation between proficiency and fixation distance by word subset.

There were no significant correlations between participants' proficiency and accuracy gains in the lexical decision task for the enhanced subset ($r(58) = .04$, $p = .79$) or unenhanced subset ($r(58) = -.08$, $p = .54$). Similarly, proficiency was not correlated to reaction time gains for the enhanced subset ($r(53) = -.01$, $p = .93$) or unenhanced

subset ($r(56) = -.04$, $p = .76$). The absence of a correlation is evident from the graphs reported in Figures 3.9 and 3.10.

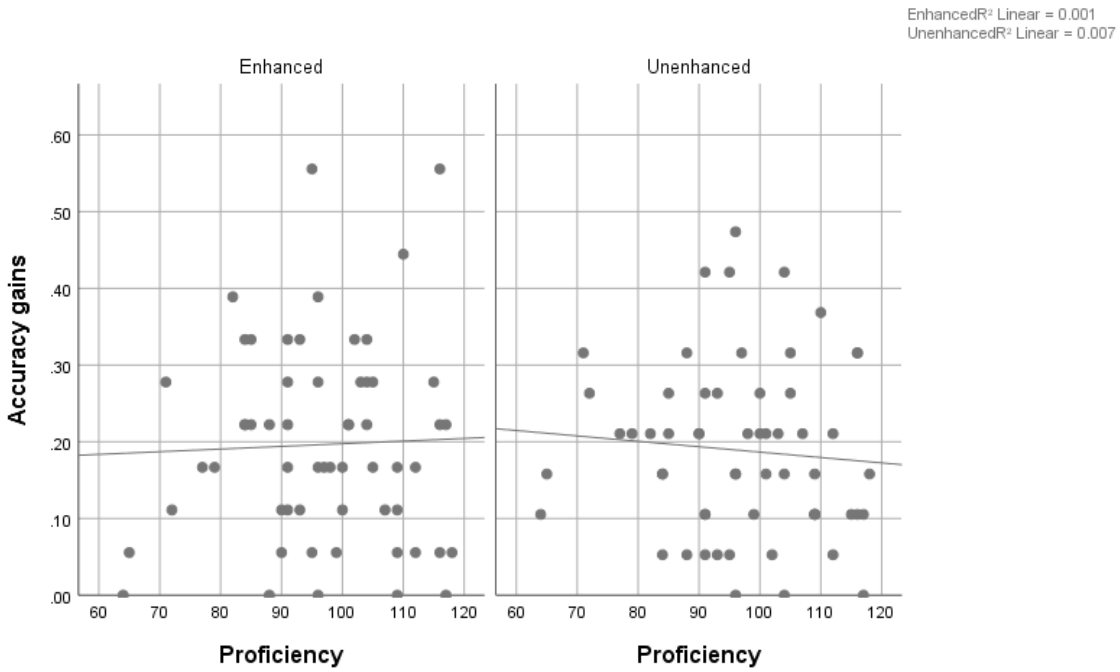


Figure 3.9. Correlation between proficiency and accuracy gains by word subset.

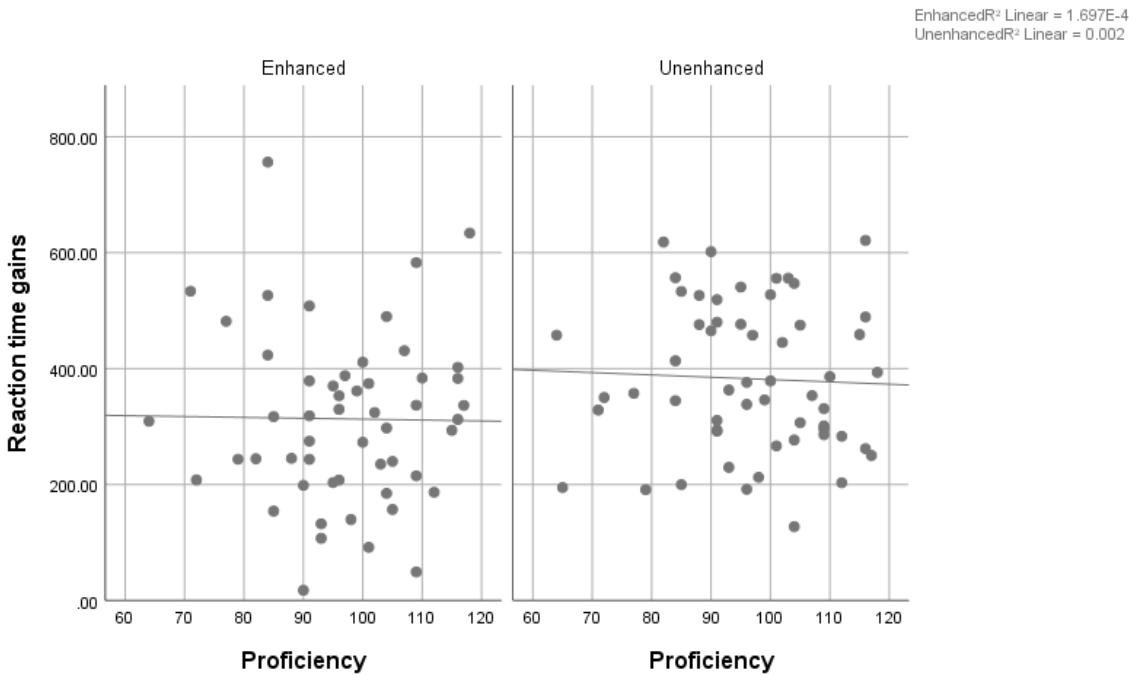


Figure 3.10. Correlation between proficiency and reaction time gains by word subset.

A positive correlation of medium strength was found between reading speed and proficiency ($r(46) = .45$, $p = .002$), leading to a similar pattern of correlations with the eye-tracking and lexical decision measures for the two variables. In particular, a significant correlation was observed between reading speed and fixation distance for the unenhanced subset ($r(42) = .37$, $p = .02$), but not for the enhanced subset ($p = .81$), which highlighted the moderating effect of enhancement in leveling out reading speed differences across different proficiency levels. Despite the significant correlation between reading speed and fixation distance, no significant correlations were found with total fixation duration ($p = .29$) and skipping probability ($p = .57$) in the unenhanced subset, indicating that both fast and slow readers paid similar attention to unenhanced words in captions. Following the pattern found for proficiency, total fixation duration on enhanced words also correlated with reading speed after excluding the unenhanced group ($r(38) = -.42$, $p = .01$), but not with skipping probability ($p = .83$). This finding confirmed that participants who read captions fast tend to spend less time focusing on enhanced words in captions compared to slower readers. Reading speed was not correlated with accuracy and reaction time gains in the enhanced or unenhanced subset ($ps > .47$).

Summary of the effects of proficiency and reading speed

To sum up, the analysis of proficiency and reading speed showed that the higher a participant's proficiency level, the faster they read captions. As a result, higher proficiency participants fixated for the first time on the unenhanced target words way earlier than their auditory onset, and lower proficiency participants fixated on words way after the auditory onset. However, the textual enhancement had an equalizing effect on participants' fixation distance, as most of them first fixated on the target

words right before auditory input, regardless of differences in proficiency and reading speed. Relatedly, the higher the proficiency and reading speed, the shorter time a participant spent fixating target words in the enhanced subset. However, participants at different proficiency levels were equally as likely to fixate or skip target words in the unenhanced subset. Finally, the participants' proficiency and reading speed did not appear to affect their pronunciation learning outcomes.

3.1.6. Discussion

This study explored learners' visual processing of L2 videos with audio-synchronized and unsynchronized textual enhancement of target words in captions, unenhanced captions, or no captions, and the effects of exposure to these viewing conditions on phonolexical update. With our first research question, we aimed to determine whether textual enhancement audio-synchronized at 300 ms, 500 ms or unsynchronized improved the synchrony between visual and auditory word processing in L2 captioned videos. The 300 ms and unsynchronized enhancement conditions were associated with a longer total fixation duration on the target words compared to the unenhanced condition, and learners were more likely to attend to (as opposed to skip) words under the unsynchronized enhancement condition. Increased attention to enhanced forms is consistent with previous research on multimodal input, where input enhancement was used to direct learners' attention to specific L2 words and constructions (Alsadoon & Heift, 2015; Lee & Révész, 2020). However, the robustness of these findings is weakened by some limitations in terms of the consistency of statistical significance and the amount of variance explained by viewing condition in the model. In addition, the 500 ms group had a significantly higher skipping probability than the other groups for the unenhanced subset of words, which could indicate that the viewing behavior

of this group differed from the other groups at baseline. Due to these possible confounds, it is difficult to derive clear implications from the analysis conducted on the total fixation duration and skipping probability data.

In contrast, the analysis of fixation distance demonstrated greater robustness in terms of the proportion of variance explained by the predictors. In addition, the model conducted on the unenhanced subset data found no differences between the groups at baseline, supporting the validity of the findings obtained for the enhanced subset. Participants in all enhanced and unenhanced conditions displayed a tendency to first fixate visually on the target words before the corresponding auditory onset, but, under the unsynchronized enhancement condition, pre-fixations happened with a significantly longer lag than under the synchronized conditions. This finding suggests that, under the unsynchronized enhancement condition, participants felt compelled to visually attend to the enhanced word immediately upon the appearance of the caption lines. Subsequently, they may have either shifted their attention back to viewing the image while listening, resulting in a lack of audiovisual synchrony, or returned to reading the caption after fixating on the target words well in advance of their auditory onset. It is possible that some participants managed to read the remaining words in captions, including the TW, in synchrony with the corresponding auditory onsets. However, given the dynamic nature of the captions, which were displayed onscreen for an average of three seconds, and the relatively short duration of the auditory form of the target words (*m* = 506.66 ms, 95% CIs [445.34, 559.36]), the cognitive effort involved in shifting attention to the enhanced word, processing it, and then returning to reading the caption is unlikely to have facilitated audiovisual synchrony.

In partial confirmation of our first hypothesis, the synchronization of textual enhancement, whether at 300 ms or 500 ms before the word's auditory onset, appeared to be more effective in promoting audiovisually synchronized allocation of attention to the TWs, compared to unsynchronized enhancement. The timely visual processing of the orthographic form of the TWs likely triggered the activation of corresponding mental representations immediately before hearing their spoken form, allowing for a comparison with stored phonolexical representations (Stenton, 2012). As a result, the improved audiovisual synchrony was expected to provide an advantage in the lexical decision task at post-test, as learners transitioned from processing input to processing intake. This stage of learning involves the formulation and testing of hypothesis about language properties and generates an initial product that is held in working memory (Leow, 2015). This preliminary product is mainly accessible via receptive testing and can be later revisited, further processed, and incorporated into the learner's internal system (Leow, 2015).

Our second research question investigated the effects of audio-synchronized and unsynchronized textual enhancement in L2 captioned videos on the updating of target phonolexical representations. Although the analysis of the enhanced subset of target words indicated that only the unsynchronized condition yielded significant improvements in *accuracy*, the validity of this finding is limited by the significant impact of time (possibly reflecting a practice effect), the relatively low amount of variance explained by the fixed factors in the model, and the significant accuracy gains observed in the 500 ms synchronized group when presented with unenhanced words.

The analysis of reaction times indicated that only the viewing conditions involving textual enhancement resulted in faster rejection of mispronunciations from pre- to

post-test. This finding, along with the smaller fixation distance observed in the eye-tracking analysis, provides partial support for our hypothesis that audio-synchronized textual enhancement would facilitate a comparison between the target realization of enhanced L2 words and the learners' stored representations, leading to their update. Contrary to our hypothesis, the unsynchronized enhancement condition also led to significant reaction time gains, in line with the assumption that skipping fewer words and fixating for longer on the orthographic form of words would increase attention to the corresponding auditory forms. While the intended threshold for the proportion of variance explained by the fixed factors was met in the model built for the enhanced subset of words, the significant effect of time observed in both models and the significant decrease in reaction times in the 300 ms group and unenhanced group when presented with unenhanced words constitute limitations in the interpretation of the reaction time data.

In particular, the significant gains in reaction times obtained by the 300 ms group and the unenhanced group for the unenhanced subset of words introduce a confound, as the former group may have employed a different response strategy prioritizing speed over accuracy. In contrast, the learners in the unenhanced group may have naturally allocated more attention to the unenhanced subset of words. Since words in the unenhanced subset were not expected to be less salient than the target words for the group that saw both subsets in the unenhanced condition, this could explain their faster reaction times compared to the other groups. When considering the enhanced subset of words, however, the absence of reaction time gains in response for the unenhanced and uncaptioned conditions is indicative of a "speed-accuracy trade-off" affecting decision-taking under time pressure (Heitz, 2014). In other words, learners who

watched the L2 videos with unenhanced captions and no captions may have relied on more explicit and less readily accessible knowledge, at the cost of automaticity (Williams & Paciorek, 2016).

Our third research question aimed to establish whether proficiency and reading speed affected pronunciation learning from exposure to enhanced and unenhanced L2 captioned video. The results supported our hypothesis that learners with higher proficiency levels would normally pre-fixate on the target words in captions too early to process them in synchrony with their auditory onset. In addition, the strong positive correlation found for reading speed and proficiency indicated that, in the absence of enhancement, the lower the proficiency, the greater the time-lag between a word's auditory onset and the learners' first fixation on that word. These findings support Wisniewska and Mora's (2018) hypothesis that individual factors such as reading speed may affect the processing of auditory and written input in captioned video. However, it must be noticed that in this study, the enhancement of target words successfully mitigated the effects of reading speed and proficiency, greatly improving audiovisual synchrony. Moreover, proficiency and reading speed were not correlated with skipping probability but proficiency was negatively related to total fixation duration in the enhanced subset.

Taken together, these findings suggest that less proficient learners used captions to support the processing of spoken dialogue (Kruger & Steyn, 2014; Yang & Chang, 2014), whereas more proficient learners skimmed through captions quickly, possibly to have more time to process the moving image (D'Ydewaelle & De Bruycker, 2007; D'Ydewaelle & Van de Poel, 1999). However, synchronized textual enhancement attracted the attention of learners at any proficiency level right before word auditory

onset, suggesting that the intervention provided every learner an equal chance to notice the target auditory word forms (Stenton, 2012, 2013). More proficient learners spent a shorter amount of time on the enhanced target words compared to less proficient learners, possibly indicating that their more advanced reading skills allowed them to process bimodal input more efficiently. The finding that proficiency and reading speed were not correlated with learners' accuracy and reaction time gains in the lexical decision task supports our hypothesis that textual enhancement facilitated phonolexical update regardless of learners' proficiency. Therefore, the "rich get richer" effect often observed in studies on the acquisition of new vocabulary through captioned video (Gesa, 2019; Pattemore & Muñoz, 2020; Pujadas & Muñoz, 2019; Rodgers, 2013) did not emerge when assessing phonolexical update, defined as more accurate and faster rejection of mispronunciations of the target words. Finally, as reading speed was highly correlated to proficiency, we conclude that including proficiency alone in subsequent studies should provide sufficient information on factors, such as reading speed, that are generally related to the learners' proficiency level (Muñoz, 2017).

3.1.7. Conclusion and limitations

Overall, the results of this study indicate that exposure to multimodal input impacts pronunciation positively, and that pronunciation learning from L2 video could be further stimulated by a planned form-focused intervention. Our results provide initial support to the hypothesis that audio-synchronized textual enhancement can be used to direct learners' attention to auditory forms and facilitate the updating of L2 phonolexical representations at the intermediate to upper-intermediate proficiency level. Consistent with our predictions, exposure to audio-synchronized textual

enhancement resulted in a greater degree of audiovisual synchrony during the viewing of L2 video than unsynchronized enhancement. In other words, the enhancement of target words from the onset of the caption line appeared to negatively impact audiovisual synchrony during the viewing. However, no significant differences were observed with the degree of synchrony obtained through exposure to any enhancement condition and unenhanced captions. All enhancement conditions promoted significant gains in word recognition response times, indicating that both synchronized and unsynchronized textual enhancement facilitated phonolexical updating compared to the unenhanced and uncaptioned viewing conditions.

This study presents some limitations, starting from the absence of a clear language learning advantage of the audiovisually synchronized conditions compared to unsynchronized enhancement, which contradicts our hypotheses. It is worth noting that while unsynchronized enhancement effectively directed learners' attention to the target words, leading to significant gains in auditory form recognition, the analysis of eye-tracking data revealed that learners immediately fixated on the enhanced target words upon caption appearance and spent a substantial amount of time focusing on them. We interpreted this pattern as potentially disruptive to the reading process and concluded that the focus on form associated with unsynchronized enhancement might have impaired the processing of video content. However, learners' awareness of the enhanced features and their perception of enhancement was not investigated in this study, and the comprehension questionnaire used was relatively short, limiting its ability to provide insights into potential differences in content comprehension between the two groups. Another limitation arises from the significant effect of time in the accuracy and reaction time models for the unenhanced subset of words, which

may be indicative of practice effect. However, conducting the study in one session allowed us to rule out confounding factors such as intentional and incidental exposure to the target words outside the study, increasing internal validity. In addition, the small size of the gains obtained through the intervention are also related to the few occurrences of each target word in the video clips. While the amount of exposure required for phonological representations to update is still under investigation, input frequency has been hypothesized to play a crucial role in lexical encoding (Charoy & Samuel, 2019; Gor et al., 2021). Therefore, it is reasonable to assume that a larger number of exposures to targetlike phonolexical representations would be associated with more precise and automatic encoding. The small number of participants in the unenhanced and uncaptioned groups may also have affected the results, despite the inclusion of random effects in the statistical models and the large number of test items. The relatively short duration of the video clips (less than 2 minutes per clip) represents another limitation, as the eye-tracking data collected within this brief timeframe might not accurately reflect learners' natural viewing behavior. The ecological validity of the study was limited by the necessity to collect the data in a language laboratory instead of a more natural setting to watch TV, such as the participants' own house, to limit the interference of external factors and allow for the use of an eye-tracker. Nevertheless, care was taken to avoid giving away the aim of the study, for example by including a high number of distractors in the lexical decision task. In addition, after the initial calibration, the eye-tracker became relatively invisible to the participants, who were sitting in front of the monitor displaying the clip, away from the screen with the recording software. Although a sampling rate of 120 Hz (refresh rate of ~8.33ms) was considered appropriate for analyzing fixations ranging between 50 and 800 ms, the relatively low sampling rate of the eye-tracker may not have provided sufficient

resolution in the context of dynamic presentation of text (but see Lee & Révész, 2020 for a discussion of this limitation).

To address these limitations, a follow-up study should involve a larger participant sample, distributed across balanced groups of equivalent size, and a longer treatment allowing for longer time spans between pre- and post-test. Collecting verbal recall data may provide valuable information regarding participants' allocation of attention and level of processing of auditory and visual stimuli. The combination of synchronized enhancement with explicit instruction may prove beneficial, as previous research has shown that it results in larger gains compared to incidental exposure to enhanced input (Han et al., 2008). The subsequent studies in this dissertation aimed to address these limitations while investigating the pronunciation learning potential of audio-synchronized enhancement within a classroom context. In study 2 (section 3.2), we conducted a preliminary analysis of how high school students process captioned video with and without enhancement, using both eye-tracking and a stimulated recall protocol. Building upon the insights from Study 1 and 2, Study 3 (section 3.3) integrated audio-synchronized enhancement within a pronunciation teaching intervention. To further boost the learning potential of audio-synchronized enhancement, the classroom intervention combined exposure to enhanced video with other interactive learning activities as well as an explicit teaching component aimed at raising learners' awareness of the enhanced target feature. To address the limitations of study 1 concerning participant sample size, distribution, and treatment duration, study 3 was conducted with three classes of learners of equal size, involved multiple sessions and included both a post-test and a delayed post-test.

3.2. L2 learners' perception and use of audio-synchronized textual enhancement in L2 captioned videos: Insights from eye-tracking and stimulated recall.

3.2.1. Study aims

The results of study 1 supported the hypothesis that synchronizing the enhancement of target words in captions both 300 ms and 500 ms before auditory onset was more likely to promote synchrony in audiovisual processing than enhancement in the absence of synchronization. Although we found initial evidence of a positive effect of audiovisual synchrony on speed and accuracy of lexical access, we did not gather information on learners' depth of processing of the enhanced target words. The aim of study 2 was to analyze whether language learners noticed audio-synchronized textual enhancement and at what level they processed information from different sources in captioned video with and without synchronized enhancement. While study 1 involved first-year university students and was conducted in a laboratory setting, study 2 was carried out in the language classroom with a group of 15-year-old high school students, who constituted the target population for our longitudinal intervention (study 3). As a consequence of its setting, study 2 had a limited number of participants, making it impractical to form experimental groups of sufficient size and leading to the adoption of a within-subject design. However, by triangulating data from eye-tracking, offline pronunciation and comprehension tests, and stimulated recall, we expected to gain valuable insights into the learners' level of processing of information in the captions, moving images and soundtrack.

3.2.2. Research questions

The study addresses the following research questions:

RQ1: Do learners notice audio-synchronized textual enhancement?

RQ2: What aspects of the second language do learners attend to while watching videos:

a) With unenhanced captions?

b) With audio-synchronized textual enhancement?

RQ3: What is learners' perception of audio-synchronized textual enhancement?

Our hypotheses are that:

HP1: Learners will notice audio-synchronized textual enhancement and register enhanced target words with various levels of processing depth.

HP2: In the absence of enhancement, we expect learners to attend to aspects of the language that are essential to comprehension, such as semantics. On appearance of enhanced target words, we expect the learners to focus on the corresponding written and auditory form, and specifically on the pronunciation of the target feature.

HP3a: We expect positive perceptions if, while processing L2 captioned, learners' internally established salience video aligns with the externally generated salience through the enhancement.

HP3b: We expect mixed perceptions regarding the usefulness of enhancement in case of a mismatch between internal and external salience.

3.2.3. Pronunciation target

Although the target words in study 1 contained a variety of phonological features that Spanish learners of English could find challenging, study 2 and 3 focused on a single linguistic aspect. This decision aimed at increasing the generalizability and comparability of findings with previous research and future studies, as well as the internal validity of the study, thanks to a stricter control over potential confounds such as learners' previous knowledge and use of the target feature. The selected feature was the pronunciation of the English regular past tense ending, which can be realized as three different allomorphs depending on the preceding context. The rule for pronouncing the <-ed> ending can be easily explained (Celce-Murcia et al., 2010):

- If the verb ends in a voiced sound other than /d/, such as /b/, /g/, /v/, /z/, /ð/, or /ʒ/, the ending is pronounced as /d/.
- If the verb ends in a voiceless sound other than /t/, such as /p/, /k/, /f/, /s/, /θ/ or /ʃ/, the ending is pronounced as /t/ due to progressive assimilation (devoicing).
- If the base form of the verb ends in /d/ or /t/, an additional vowel is inserted, and the ending is pronounced as /ɪd/ or /əd/.

Despite the relative simplicity of the underlying rule, this morphophonological feature is acquired late by L1 and L2 learners of English and presents difficulties for English learners of all proficiency levels (Bell et al., 2015; Solt et al., 2003; Strachan & Trofimovich, 2019). L2 morphological features are not usually acquired from naturalistic exposure, without explicit instruction, due to their low saliency and redundancy, and due to learners' automated processing of language based on L1-tailored patterns and strategies (Ellis, 2017). These issues concur to impair the

phonological processing of inflectional morphemes, which are pronounced as one or two phonemes in a non-prominent position (word ending), and express the same meaning as other, more salient chunks of language such as time adverbials (Strachan & Trofimovich, 2019; VanPatten, 2004). Language-related factors such as the relatively low frequency of regular verbs in the input, compared to the irregular verbs, also concur to impair the acquisition of targetlike pronunciation of the regular past <-ed> ending (Bell et al., 2015; Strachan & Trofimovich, 2019). Finally, the presence of consonant clusters in English poses difficulties for learners with simpler syllable structures in their native language, which would explain why the <-ed> ending is more likely to be omitted when it is preceded and/or followed by a consonant sound than a vowel sound (Bell et al., 2015; Ernestus et al., 2017). The non-targetlike simplification of final clusters, including the omission of the <-ed> ending and the erroneous addition of an epenthetic vowel, affects comprehensibility and should be addressed early on, as it becomes harder to correct once it has become fossilized (Celce-Murcia et al., 2010; Haslam & Zetterholm, 2019; Kruk & Pawlak, 2021; Levis, 2018).

Teaching the pronunciation rule, or at least raising learners' awareness of the different allomorphs and their use, may have positive effects on the pronunciation of past <-ed> endings, but automatizing this declarative knowledge and applying it in real-time processing requires extensive form-focused practice (Kruk & Pawlak, 2021; Solt et al., 2003). To encourage noticing of past <-ed> pronunciation, learners should be exposed to authentic (rather than artificially simplified) input containing salient (/ɪd/) and less salient (/d/, /t/) allomorphs in various contexts, including contexts that favor reduction or indistinct articulation of <-ed> endings (Bell et al., 2015). Further opportunities for perception and production practice can be offered through

decontextualized read aloud, listen and repeat, and multiple-choice exercises (Benitez-Correa et al., 2020; Kruk & Pawlak, 2021), as well as more communicative activities such as telling stories in the past or predicting the pronunciation of regular past verbs in English songs and videos (Celce-Murcia et al., 2010). Providing repeated exposure to target exemplars of regular past verbs is crucial, as late L2 learners have been found to rely less on morphological structure and more on lexical storage during processing of inflected forms, compared to native speakers who, on the contrary, tend to store regularly inflected verbs as new forms of the same word or stem and do not require new lexical entries (Bassetti & Atkinson, 2015; Clahsen et al., 2010; Ullman, 2005). Embedding the target exemplars in meaning-based tasks requiring comprehension of the context helps learners develop their bottom-up acoustic abilities by using contextual cues and previous knowledge to enhance auditory perception (Strachan & Trofimovich, 2019).

3.2.4. Methodology

Taking a similar approach to study 1, study 2 focused on the initial stages of processing and acquisition of a pronunciation feature by pre- and post-testing learners before and after exposure to L2 captioned video content (Figure 3.11). However, study 2 adopted a within-subject design, comparing the test outcomes and viewing patterns observed when exposing learners to an unenhanced clip and a clip in which the words containing the target feature highlighted in synchrony with their auditory onset. With a smaller sample size compared to study 1, it was possible to conduct a more in-depth analysis of learners' perceptions regarding this enhancement technique, as well as their awareness of the enhanced target feature. To help learners recall whether they noticed the textual enhancement, they were immediately shown their own eye-

tracking recording and interviewed regarding their thoughts at the time of the viewing. The discussion of the pronunciation outcomes, eye-tracking data and stimulated recall data was enriched through the insights provided by learners' descriptions of the target pronunciation rule in the final questionnaire. In addition, a vocabulary size test and a narrative task were administered at pre-test to assess learners' L2 proficiency and their accuracy when using the target pronunciation feature in semi-spontaneous speech.

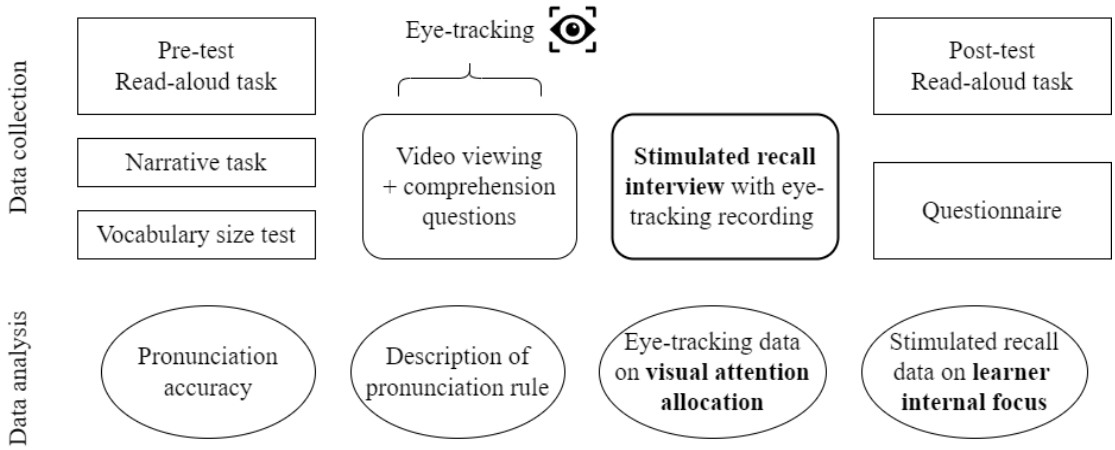


Figure 3.11. Overview of the methodology employed in study 2.

Participants

Eleven L1 Spanish/Catalan participants, 9 of which female, (age *m* = 14.5 years, *SD* = 0.5) were recruited (Table 3.28). They were secondary school students in the fourth year of secondary school (ESO) in a privately owned, government-funded high school in Barcelona (Spain), where they took three hours of English classes per week. Although only two participants reported having received a B2 (Cambridge First) certificate and one participant had a B1 (PET) certificate, based on the textbook used in class and on the participants' vocabulary size (*m* = 2900, *SD* = 803), their proficiency was estimated to range from lower-intermediate to upper intermediate (B1 to B2 in the CEFR) (Milton, 2010). In addition, all participants except two reported

having attended extracurricular English classes for at least 7 months (*m* = 28.73, *SD* = 24.71). Five of the eleven participants had been in an English-speaking country (the UK or USA) for one to two weeks, one for eight weeks, and one for four years. In order to ensure participant anonymity, a unique identifier was generated using a combination of alphanumeric characters, e.g., the second student to be tested was named S_02.

Table 3.28. Participants' demographics.

	M	*SD*	95% CI
Age at testing	14.50	3.03	[14.19, 14.90]
Vocabulary size[a]	2900	803.12	[2360.46, 3439.54]
Extracurricular classes (years)	2.39	2.06	[1.01, 3.78]
Time spent in an English-speaking country (months)	1.43	3.55	[-.96, 3.82]

[a] X-Lex (Meara and Milton, 2003).

Materials

Video clips

The two video clips in this study consisted of various excerpts from the first episode of the TV series *The Good Place* and largely overlapped with the clips used in study 1. Clip 1 (2' 17") was presented with unenhanced captions, whereas clip 2 (6' 05") contained seven target words that highlighted 500 ms before auditory onset. The video clips mainly featured dialogues between two or three characters in quiet settings and were selected to both provide an adequate number of occurrences of the target feature and allow participants to follow the sequence of events with ease despite not watching the whole episode. The vocabulary profile of the script was almost identical to the

script of study 1, and the captions were formatted in the same way, except for the way target words were enhanced. Written forms were enhanced in the captions by highlighting the entire word in yellow and underlining the <-ed> morpheme and the letter representing the vowel or consonant preceding it (Figure 3.12).

Figure 3.12. Screenshot of video with captions containing an enhanced target word.

On the one hand, the past <-ed> pronunciation rule cannot be entirely explained in terms of spelling due to instances when a voiced phoneme and a voiceless phoneme are mapped onto the same grapheme (e.g., /s/ and /z/ in *ba**s**ed* and *clo**s**ed*) and vice versa two or more graphemes are based on the same phoneme (e.g., *pa**c**ed* and *ba**s**ed*). On the other hand, enhancing the preceding letter may at least reduce the most common mispronunciations involving the erroneous addition of epenthetic vowels by directing learners' attention to the fact that verbs ending in <-t> and <-d> in their present form take the /əd/ or /ɪd/ pronunciation, while other spelling endings take either /d/ or /t/ (Brutten et al., 1986). Bolding and enlarging were not used to avoid changes in caption size and position, so that the caption with the enhanced target word smoothly replaced the unenhanced line 500 ms before the target word's auditory onset. The comprehension questions, six related to clip 1 and ten to clip 2, included a mix of multiple choice and true/false items (Appendix B.1).

Target words

A total of 21 target verbs were selected in the clips, 7 verbs in the past tense that were highlighted in captions and a mix of base form and past tense verbs that were presented unenhanced (Table 3.29). Fisher's exact tests with Monte Carlo estimations of the *p* values (two-tailed) determined that there was no significant difference between enhanced and unenhanced words, in terms of number of caption lines (1 or 2) on-screen at the time of TW presentation ($p = .18$), whether the TW occurred in line 1 or 2 ($p = .25$), and the TW position within the caption line ($p = 1.00$). No test was run on the number of TW per line, as no more than one target word was included per line. Lastly, a T-test found no differences between the two subsets in presentation time ($t(19) = 2.03$, $p = .057$).

Table 3.29. Linguistic and presentation properties of the target words.

	Enhanced	Orthographic length	Phonological length[a]	Occurrences in clips	Lexical frequency[b]	Caption lines	Line with TW	TW position	Presentation time (ms)	Auditory duration (ms)	AOI position[c] (px)	AOI width[d] (px)
Collected	Yes	9	8	1	5.75	2	1	Medial	3650	580	650, 850	290
Created	Yes	7	8	1	23.43	2	1	Final	2950	380	1060, 840	250
Decorated	Yes	9	10	1	2.82	1	1	Medial	2520	540	860, 920	320
Happened	Yes	8	6	2	490.08	2	2	Medial	2500	340	630, 940	320
Lived	Yes	5	4	3	66.04	2	1	Final	3320	350	1280, 840	190
Moved	Yes	5	4	1	69.33	1	1	Medial	2990	340	615, 940	220
Raised	Yes	6	5	2	25.73	2	2	Initial	3370	310	660, 940	230
Believe	No	7	5	1	625.14	1	1	Final	1750	610	1195, 920	240
Calculate	No	9	10	1	2.08	1	1	Medial	2660	670	480, 920	290
Died	No	4	4	2	157.22	1	1	Final	1080	360	1070, 920	180
Dropped	No	7	5	1	48.67	2	1	Medial	2030	380	740, 840	270
Ended	No	5	5	2	29.63	1	1	Final	1530	420	1198, 920	212
Hear	No	4	3	1	555.35	1	1	Final	1290	350	1250, 920	200
Know	No	4	3	4	5721.18	2	1	Final	2150	210	1100, 840	200
Love	No	4	3	2	1114.98	1	1	Medial	3980	290	750, 940	150
Pick	No	4	3	1	198.39	2	2	Medial	1540	190	840, 920	150
Promise	No	7	6	2	153.12	2	1	Medial	4570	660	770, 850	260
Rolled	No	6	5	1	8.47	2	1	Initial	3170	360	650, 840	200
Sound	No	5	5	1	143.39	1	1	Medial	820	280	740, 920	240
Take	No	4	4	1	1891.04	1	1	Medial	1950	170	800, 920	150
Want	No	4	4	2	2759.18	1	1	Medial	1910	170	870, 850	160

[a] IPA notation system for American English. [b] Frequency per million words in the $SUBTLEX_{US}$ database (Brysbaert & New, 2009). [c] Horizontal and vertical coordinates (respectively) of the area of interest containing the target word, with reference to the upper-left corner of the screen. [d] AOI height was maintained constant at 110 pixels.

Pronunciation tests

a. Read-aloud task

In the read-aloud task, each target word was presented in standard orthography on the screen once only, for 1.5 seconds. A "speak now" logo then replaced the written form, and a "beep" sound prompted the participant to say the word into the microphone. The whole task lasted about 2.5 minutes and was administered using DMDX 6.0.0.1 to present the stimuli in random order. This test was designed to tap into a type of knowledge that could be considered explicit/declarative (Saito & Plonsky, 2019).

b. Narrative task

A narrative task was used to assess procedural knowledge of past <–ed> forms in a more spontaneous context, as controlled and spontaneous L2 pronunciation performance weighs on distinct types of L2 knowledge that require separate assessment (Saito & Plonsky, 2019). Two wordless comic strips were used as prompts for narration. These sequences of pictures were selected for containing several actions, which was expected to elicit a large number of verbs in participants' production, and for having a "clear climax and resolution" (Gilabert, 2007, p. 223). In an effort to obtain comparable speech samples from different participants, a list of English regular verbs was included with the instructions (see Appendix B.2 for task 1). Participants were given one minute to look at the pictures before telling the story. To encourage the use of the past tense in a context that may have otherwise elicited present tense narration, they were given the first sentence of the story, e.g., "Yesterday Lucy was in the garden, when…" (Kruk & Pawlak, 2021).

Questionnaire

The questionnaire consisted of three sections, targeting the participants' language background, their knowledge of the past <-ed> pronunciation rule, and their perceptions of the video clips (Appendix B.3). After providing information on their L1(s) and extracurricular exposure to English, the participants were asked to describe how the <-ed> ending of regular past verbs is pronounced and what its pronunciation depends on. Then they indicated to what extent they agreed with statements about the videos: 1) I understood the videos; 2) The videos were fun; 3) I read the subtitles; 4) I learned some English pronunciation from the videos; 5) I learned some English grammar or vocabulary from the videos. Finally, they were asked whether any letters were enhanced in the subtitles, what those letters had in common, and whether they believed that the enhancement was useful for learning or distracting. Two versions of the questionnaire were created, one in Spanish and one in Catalan, to enable participants to complete it in the language with which they feel most comfortable.

Stimulated recall protocol

The interviews aimed at helping learners recall which input sources (auditory, written, pictorial) they attended to during the viewing, and what effect enhancement had on their allocation of attentional resources. A stimulated recall protocol was created to elicit information on learners' processing of audiovisual information from enhanced and unenhanced video, based on the protocols in Révész et al. (2019) and Révész and Brunfaut (2013). The protocol was semi-structured, in that it contained prompts regarding specific events in the video recording but also aimed at reducing as much as possible researcher interference by providing precise instructions on what to say and not to say to a participant (Gass & Mackey, 2000). The following excerpt from

the stimulated recall protocol, available in full in Appendix B.4, illustrates these objectives:

(Pause clip when some of this happens)

- Unenhanced target words
- Enhanced target words
- Very long fixations
- Re-reading (regressions)
- (Occasionally, other areas to avoid giving away aim.)

(As appropriate, ask one of the following questions)

- Why were you watching this area?
- What made you watch this area for a long time?
- What made you go back to this area?

The interview was conducted in English to comply with the school's request to provide language learning opportunities during the allotted class time. However, the learners were informed that they could revert back to the L1 to express difficult concepts. In addition, it was ensured that they were familiar with the technical terms *skipping*, a high frequency verb in English, and *fixating*, which is similar to the Spanish "fijar"/"fijarse" (focus/pay attention to).

Procedure

The participants were contacted through their English teacher, and their parents signed a consent form with detailed information on the study. Data collection took around 50 minutes per participant, during which the students were called to a quiet room within the school and did the tasks individually under the supervision of two researchers. The pre-test included the paced reading task, narrative task and X-Lex vocabulary test. Then, each participant watched the clips while their gaze was recorded on a Tobii Pro Spectrum eye-tracker unit (1200 Hz sampling rate, .3° accuracy and .06° RMS resolution at optimal conditions) integrated into a 23.8" monitor (1920 x 1080 pixels resolution). The eye-tracker was set up on one desk and connected to the computer running the Tobii Pro Lab software. The interview was conducted by replaying the eye-tracking recording immediately after the end of the viewing, in the attempt to provide the participants with a strong stimulus that would support the recall of learners' thought processes during the viewing (Gass & Mackey, 2000). To maximize time availability, while the researcher interviewed one participant, another participant responded to the questionnaire and did the pronunciation and proficiency tests on another desk about four meters away.

Analysis

The analysis of the read-aloud task involved listening to each audio recording and assigning an accurate (1) or inaccurate (0) score (Benitez Correa et al., 2020). A score of 0 was awarded when the speaker:

a) Added an epenthetic vowel before a -d or -t ending when not necessary (*roll* pronounced /roled/) or omitted the vowel when necessary (*wanted* pronounced /wɑnt/ with a marked /t/);

b) Said an unintelligible word or remained silent;

c) If the speaker did not pronounce the past <-ed> allomorph when the stimulus in the read-aloud task was in past <-ed> form (*rolled* pronounced /rol/).

On the other hand:

a) inaccurate pronunciation of other segments of the word was not penalized, i.e., *paused* pronounced as /paʊzd/ was considered correct, just like /pɔzd/;

b) pronouncing the /d/ allomorph as /t/ and vice versa was not penalized, i.e., /roʊld/ and /roʊlt/ were both considered correct pronunciations of rolled (Dickerson, 2015). Since L1 English speakers tend to devoice word final obstruents, especially before a pause and before a voiceless consonant (Edge, 1991), the voicing distinction tends to be absent or not particularly marked in the input, and the two phonological variants function more like allophones than phonemes (Levis, 2018).

In the narrative task, regular past verbs in obligatory contexts, defined relative to the participants' production, were analyzed for pronunciation accuracy (Godfroid & Kim, 2021). All verbs with a regular past form that were pronounced as base forms were considered incorrect, because the participants knew that they were supposed to tell the story in the past using the past tense. Each verb was scored only once for each story, even if it was repeated by the speaker within the story (Kim & Tracy-Ventura, 2011). However, when the speaker changed their mind halfway through a sentence (false

start) the verb was not counted, so for example in “then, she *start… she, she *sprint”, only the verb “sprint” was analyzed, not “start”. If a verb was pronounced correctly once in the same story, it was considered correct despite previous or later mispronunciations.

To analyze the eye gaze data, AOIs were drawn manually around each target word, and fixation data was extracted using the I-VT fixation filter in Tobii Pro Lab. On average, 88.68% of the data per participant was available (*SD* = 6.39). Following Godfroid (2019), we eliminated fixations shorter than 50 ms or longer than 800 ms (5%) from the analysis of fixation duration and fixation distance (but not skipping probability), leaving 88 fixations on a total of 226 fixated and skipped items. The analysis of fixation distance was conducted on 64 data points, as 24 data points had to be excluded due to a technical issue. Zero fixations, i.e., skipped items, were excluded from the analysis of fixation duration. To analyze total fixation duration, skipping probability and fixation distance data, three statistical models were built in RStudio using the same packages, underlying distributions and log-link functions as in study 1. However, since this study adopted a within-subject design, we compared the effects of enhancement by running only one mixed model per independent variable, with *Viewing Condition* (enhanced/unenhanced), *Presentation Time* and *Frequency of Occurrence* as fixed effects, and random intercepts for participants and items. When the effects of presentation time and frequency did not reach significance, the model was re-run excluding the non-significant covariate(s). As the fixation duration model failed to converge, a different optimizer was used (Bound Optimization BY Quadratic Approximation). Non-parametric bootstrapping with replacement (*n* = 1e3 simulations) was then used to calculate basic confidence

intervals from the empirical distribution of the parameter estimate and independently from model assumptions.

The author listened to the recordings of the stimulated recall interviews and transcribed the excerpts relevant to the study, e.g., excluding greetings and task instructions. In the transcription, turns were defined as the utterances of one speaker delimited by another speaker's utterances, whereas one or more interviewer-participant turns regarding a specific scene or caption line were considered "events". On a total of 622 interviewer-participant turns embedded in 95 events (Appendix C.1), 97 turns within 76 events were included in the analysis for containing information regarding participants' processing of the video. Upon re-reading the transcription several times, seven categories were identified according to the participants' reported focus of attention: Noticing of word or phrase and its sub-category "language aspect noticed", general comprehension, audiovisual processing, mention of test, captions attract gaze, face/mouth attracts gaze, does not know (Appendix C.2). For example, the turn "That was one word that was actually in the other exercise that we've done before, so I thought "well maybe, that's why the yellow words are in yellow", cause they were like... vocabulary that we had in the other test?" was coded as *noticing + aspect noticed: vocabulary + mentions test*. Turns were coded as "Does not know" when a participant reported not remembering what they were thinking at the time of the viewing. In addition, due to the low proficiency of some of the participants and to the limited time available, the protocol had not been respected at all times and some questions were more explicit than expected. Therefore, one further variable was added to classify utterances either as explicitly prompted by the interviewer or as spontaneously produced by the participant following minimal prompting.

3.2.5. Results

Video comprehension

Participants' responses to the comprehension questions were accurate overall (*m* = 88.64%; *SD* = 17.64) and for each video separately (*m* = 86.36%; *SD* = 16.36 and *m* = 90%; *SD* = 20.98).

Pronunciation accuracy

Participants' scores in the production tasks illustrate their ability to pronounce past <-ed> endings accurately in controlled and spontaneous contexts (Table 3.30). Due to the almost identical performance at pre- and post-test, no statistical model was run on the read-aloud data. In the narrative task, participants' pronunciation of regular past endings was generally accurate in at least 50% of the obligatory contexts relative to their speech, but two participants correctly pronounced only 1 and 0 past verbs, in 6 and 7 obligatory contexts, respectively (Figure 3.13).

Table 3.30. Averaged accuracy scores (max 1) for regular past verbs.

					95% Confidence Intervals	
Verb form	Time	*N*	*M*	*SD*	Lower	Upper
Read-aloud task	1	121	0.79	0.41	0.72	0.87
Read-aloud task	2	121	0.78	0.42	0.70	0.85
Narrative task	1	92	0.64	0.48	0.54	0.74

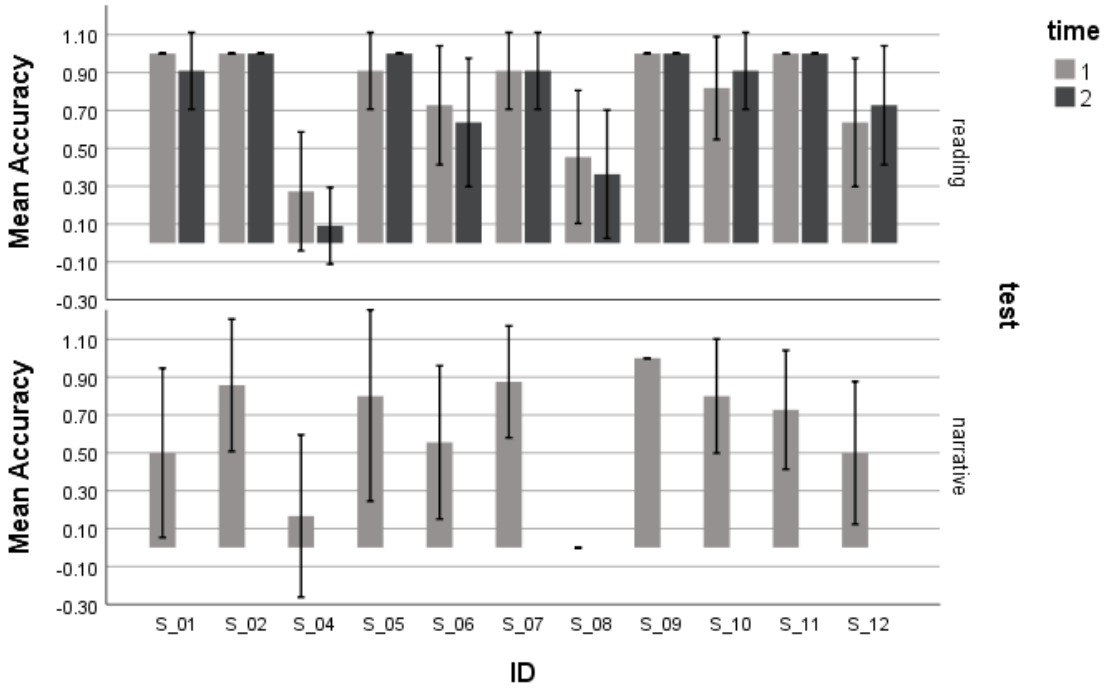


Figure 3.13. Accuracy scores by participant for regular past verbs, with error bars = 95% CI.

Questionnaire

In the questionnaire, five participants indicated that the enhancement was useful for learning, whereas four said it was distracting and two had a neutral opinion. Seven out of 11 participants indicated that the yellow words were highlighted because they ended in <-ed>, and four also or alternatively mentioned that they were in the past tense. While one participant mentioned that the words appeared in the pre-test, the remaining two participants did not provide a response. Almost all participants were able to point out some elements of the rule to pronounce <-ed> endings, in particular that a dental stop is always pronounced (Table 3.31). In some cases, more complex aspects such as the existence of different variants depending on the presence of a vowel and the difference between the /d/ and /t/ allomorphs was mentioned. However, the participants' responses contained several errors and mainly consisted of hints

rather than formal explanations, denoting limited knowledge of the underlying pronunciation rule. The feeling of learning was generally quite low for pronunciation and even lower for grammar and vocabulary, even though all participants reported understanding the videos and liking them (Figure 3.14).

Table 3.31. Participants' description of the past <-ed> pronunciation rule.

Participant	Response (translated from Spanish)
S_01	The "e" is pronounced like "i", there are times when the "e" is not pronounced because it is accompanied by another letter that is pronounced similarly. The "d" is sometimes pronounced and sometimes not.
S_02	It is not pronounced in a very noticeable way, for example, walk in the past is walked, but its pronunciation is more like walkd.
S_04	The final d is pronounced but not the ed.
S_05	It's usually pronounced like a "t", but a "t" similar to a "d", like a mix I would say. For example: "opened" --> "opent" // "closed" --> "clost". I think the pronunciation depends on the word. I am not aware of another more specific explanation.
S_06	It is pronounced as it sounds -ed, you pronounce the whole word.
S_07	I know it but I can't explain it.
S_08	Depends on how you say it and how the sentence is structured.
S_09	It is pronounced like a d or t, it depends on the word and what letters are in front of it.
S_10	It is usually pronounced as a -d at the end of the verb. For example: Talked is pronounced tokd, lived is pronounced livd. I don't think the pronunciation will change, since all the verbs end in the same way and are pronounced the same.
S_11	In some cases, the "e" is not pronounced, only the final "d" is pronounced.
S_12	It depends on the word, example: expected (here it is pronounced more) is not the same as bored.

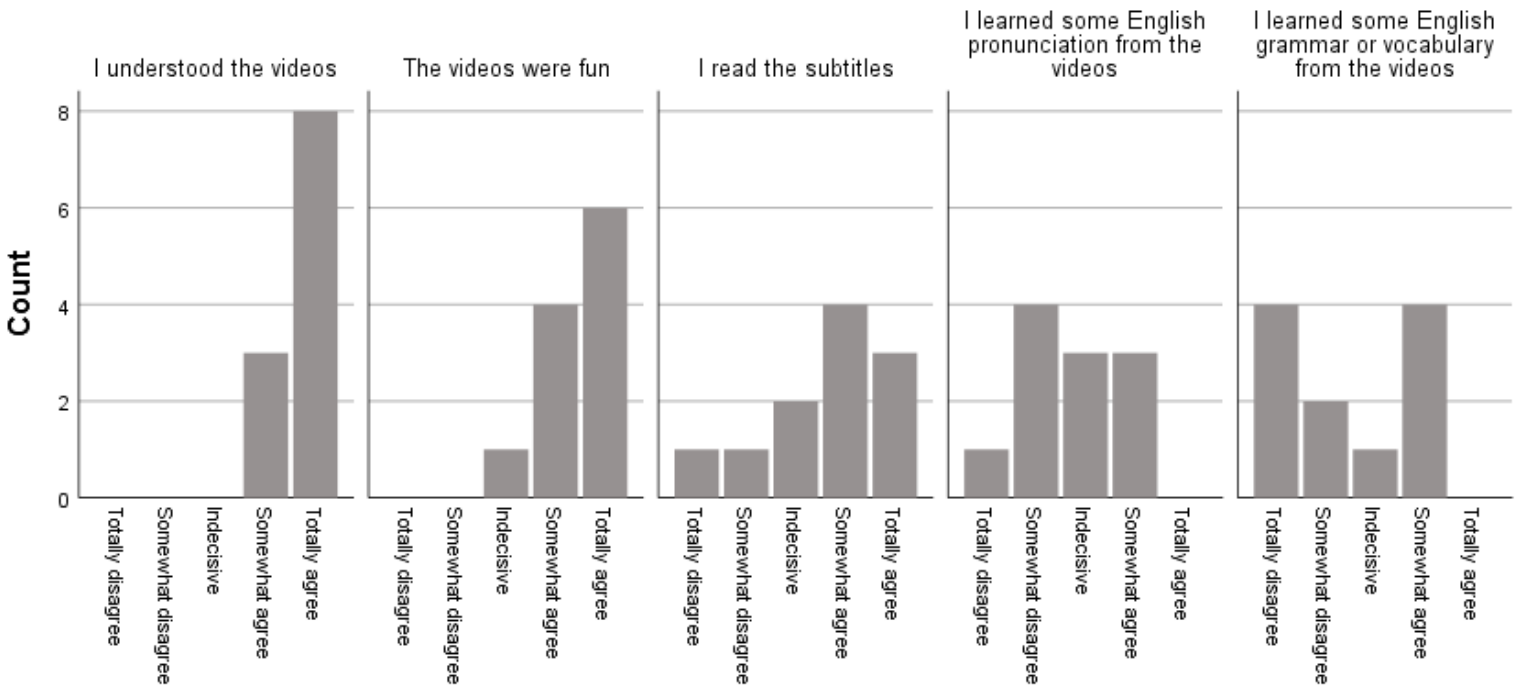


Figure 3.14. Participants' perceptions of the video clips and learning perceptions.

Eye-tracking

The bar graph shows the average total fixation duration time by participant, illustrating at a glance the difference in viewing behavior when reading enhanced and unenhanced words in captions, as well as the high degree of individual variability (Figure 3.15).

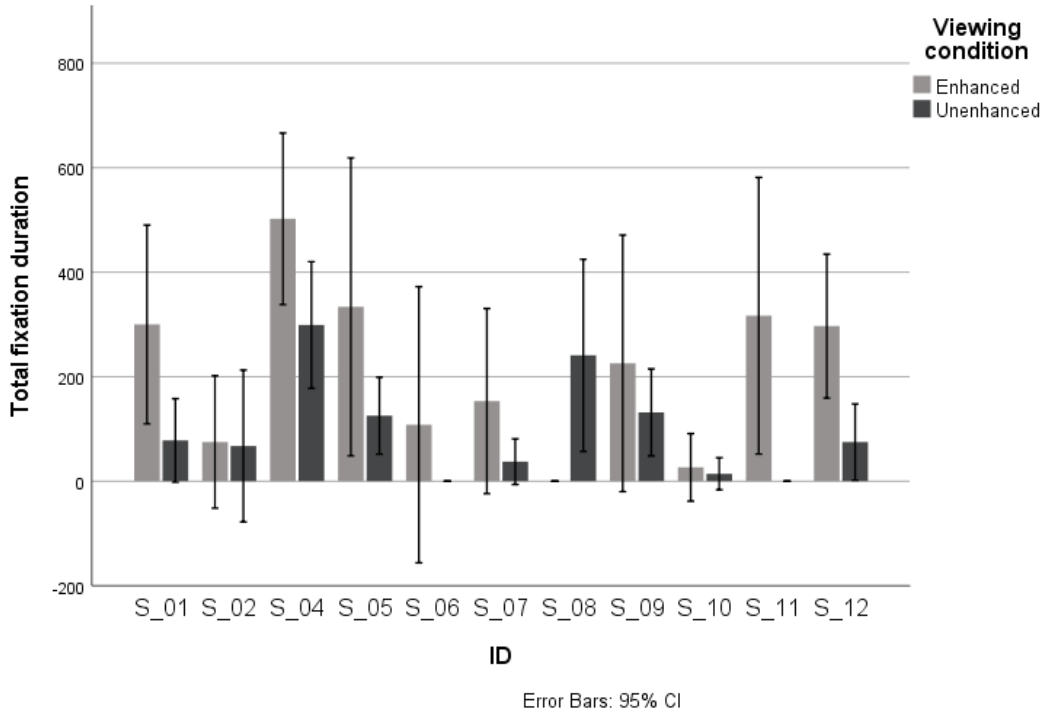


Figure 3.15. Total fixation duration by participant and viewing condition.

The descriptive statistics for the eye-tracking metrics are reported in Table 3.32. Due to the high standard deviation in the fixation distance data, we also obtained the median and interquartile range by viewing condition (*Mdn* = -286; *IRQ* = 604.50 for enhanced TWs and *Mdn* = -85; *IRQ* = 824, for unenhanced TWs). Figure 3.16 reports the fixation distance histograms by condition, with pre-fixation defined as the temporal interval that spans from a participant's first fixation on a specific word in captions to the auditory onset of that word. On the contrary, post-fixations refer to

fixating on a written form only after the auditory onset of the corresponding auditory form.

Table 3.32. Eye-tracking descriptive statistics by target word enhancement condition.

				95% Confidence Intervals	
	N	*M (ms)*	*SD (ms)*	Lower	Upper
Total fixation duration					
Enhanced	42	377.67	206.26	313.39	441.94
Unenhanced	46	314.33	195.61	256.24	372.42
Skipping probability					
Enhanced	76	0.45	0.50	0.33	0.56
Unenhanced	150	0.69	0.46	0.62	0.77
Fixation distance					
Enhanced	35	-392.63	807.79	-670.12	-115.14
Unenhanced	29	72.52	694.62	-191.70	336.74

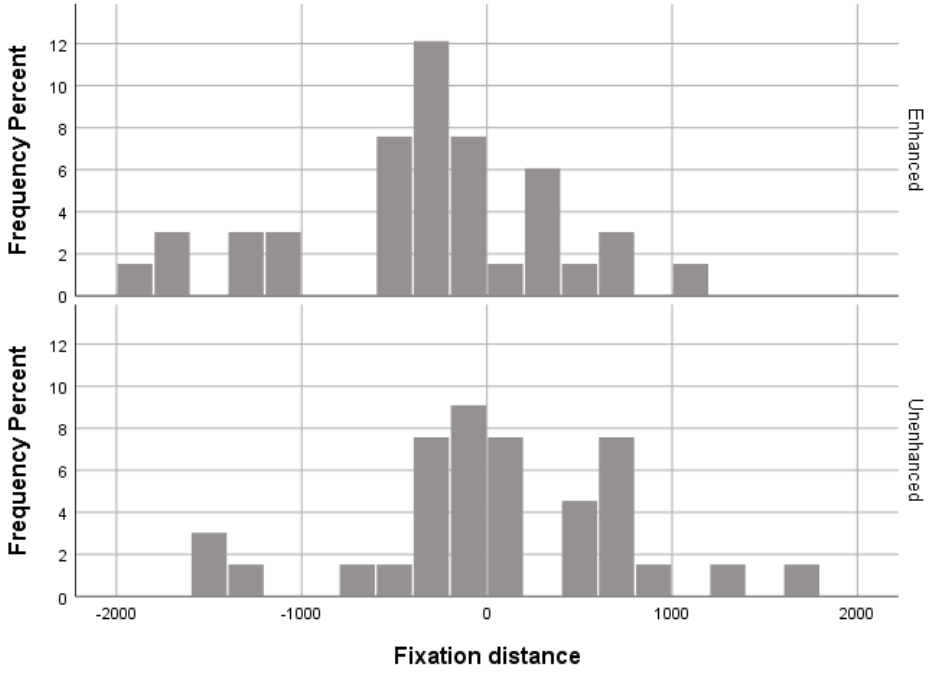


Figure 3.16. Distribution of pre-fixations (positive values) and post-fixations in relation to word auditory onset by enhancement condition.

In a GLMM based on a gamma distribution with a log link, *Viewing Condition* had a significant effect on total fixation duration, pointing at longer fixation intervals on enhanced than unenhanced target words (*B* = 0.29, *SE* = 0.14, 95% CI [0.03, 0.56]). Since in the model that included *Presentation Time* and *Frequency of Occurrence*, these two variables did not reach significance, the model reported did not include covariates (Table 3.33). However, the significant effect of enhancement was not confirmed through bootstrapping, as in this case the confidence intervals crossed the zero, although the lower bound was very close to zero (-0.02). Similarly, a GLMM based on a binomial distribution with a logit link found a significant effect of *Viewing Condition* on skipping probability (*B* = -1.82, *SE* = 0.72, CI [-3.24, -0.40]), but the null hypothesis could not be rejected based on the bootstrapped confidence intervals.

The LMM for fixation distance that included participant and item as random effects found no effect of enhancement. As bootstrapping returned a few singularity errors, indicating that some dimensions of the variance-covariance matrix had been estimated as zero and suggesting that the model had been overfitted, the model was re-run and reported in Table 3.33 after eliminating the random intercept for participant, which explained less variance in the original model (variance = 188808, *SD* = 435 for item vs variance = 17130, *SD* = 131 for participant). The results of bootstrapping did not align with the significance established using asymptotic CIs, as in this case the bootstrapped 95% CIs did not contain 0, indicating a significant effect of viewing condition on fixation distance. The negative value obtained for the beta coefficient pointed at a longer fixation distance when the target words were enhanced, compared to the unenhanced baseline.

To sum up, the analysis of eye-tracking data pointed at longer fixations on the enhanced target words, less skipping, and longer fixation distance in the presence of audio-synchronized enhancement, but it was difficult to establish the significance of the effects of these variables due to the discrepancy between the two methods of estimating the confidence intervals (asymptotic and bootstrapped). Due to the small size of the original sample and the potential violation of assumptions regarding the distribution and variance of the parameter estimates, bootstrapped confidence intervals were considered more reliable. In addition, the high conditional R2 values and low marginal R2 values obtained especially for the skipping probability and fixation distance model showed that a large proportion of the outcome variation could be found at the level of individual variables (participant and item) represented by the random effects. Therefore, in this study, the effects of audio-synchronized enhancement on learner's noticing and processing of the target words could not be established solely based on the analysis of eye-tracking data.

Table 3.33. Fixed coefficients for the eye-tracking models.

							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	*SE*	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Total fixation duration										
Intercept	5.68	0.15	36.94	<.001***	0.06	0.26	5.38	5.98	5.53	6.04
Viewing condition	0.29	0.14	2.15	.03*			0.03	0.56	-0.02	0.56
Skipping probability										
Intercept	1.39	0.69	2.01	.04*	0.08	0.58	0.03	2.75	0.07	1.68
Viewing condition	-1.82	0.72	-2.51	.01**			-3.24	-0.40	-2.33	0.23
Fixation distance										
Intercept	104.68	191.22	0.55	0.59	0.08	0.39	-270.11	479.46	-183.46	278.42
Viewing condition	-460.45	278.84	-1.65	0.12			-1006.97	86.08	-736.35	-96.87

*** $p < .001$, ** $p < .01$, * $p < .05$

Stimulated recall

Crucial information on the participants' level of processing of enhanced and unenhanced captions was obtained through the analysis of stimulated recall data. In about half of the turns[6], participants either reported simply watching for general comprehension or claimed that they did not know what they were focusing on at the time of the viewing. This finding suggested that the answers were truthful and supported the validity of the recall protocol. In those cases when the noticing of specific words or phrases was reported (40% of the turns), the language aspect attended by the participants seemed to depend on the viewing condition (Figure 3.17). While participants recalling the processing of unenhanced words exclusively mentioned attending to lexical or phonological aspects, enhanced words were mainly associated with morphological or grammatical aspects. Essentially, participants recalled automatically fixating unenhanced words in captions when they did not know their meaning or did not recognize the corresponding auditory form in the soundtrack. However, enhanced words were most often associated with their grammatical function (e.g., the regular past form and the recurring <-ed> ending of the enhanced verbs), and only rarely did participants report noticing their auditory form. It must be noticed however that the turns related to the noticing of a linguistic aspect in combination with an *enhanced* word were always explicitly prompted by the interviewer (e.g., "Would you have read [this word] or would you not have read it if it wasn't in yellow?" "No, I wouldn't have read it... And the next words that are in yellow, I think that they all finish in <-ed>").

[6] A *turn* is defined as a unit of conversational exchange where one person speaks, and their interlocutor listens. One turn is generally followed by another person responding by voicing their own thoughts.

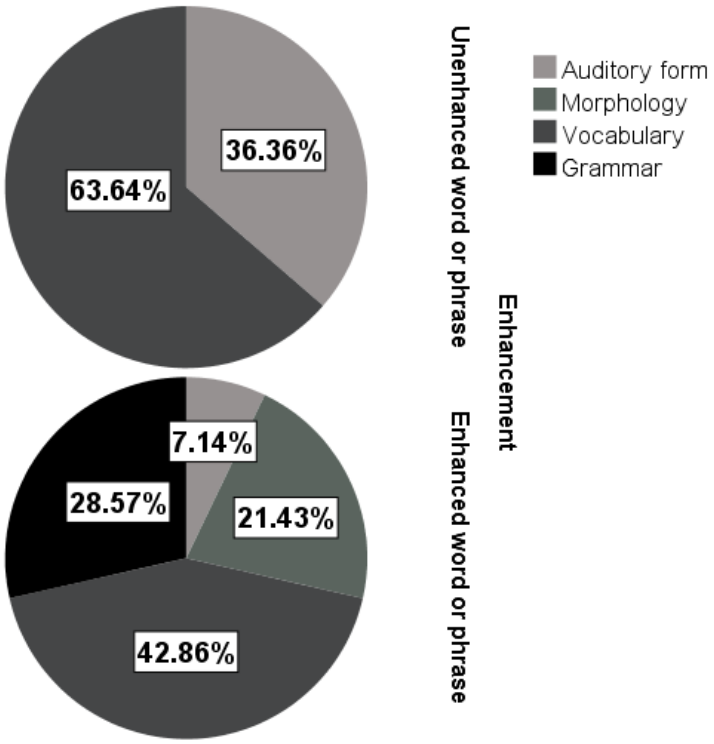


Figure 3.17. Language aspects noticed in fixated words by word enhancement.

In the turns that contained an indication of audiovisual processing (about 30% of the total), participants generally reported directing their attention to the captions due to issues in segmenting speech or recognizing specific auditory forms, as well as inferring the meaning of these words. For example, participant S_02 explained:

> This sentence, I didn't know what he was saying so I tried to read it because I don't know what it means: bent down to pick it up. Well, now I know, but at the moment the video was playing I didn't understand it, so I read it.

The captions also supported the segmentation of fast speech: "I was trying to connect the dots… cause he was speaking a little bit fast and, like, I wanted to take all the data because they are going to ask me later [in the comprehension questions]" (S_06). However, most of the participants (63%) reported actively trying not to read the captions unless they did not understand the dialogue, regardless of the presence of

enhancement. In this quote, participant S_10 reported deliberately avoiding reading the captions:

> "Well, I've read some [words in yellow], and it put "lived", the <-ed> was [underlined]… and then I thought well, they are going to put the past verbs in yellow, so I thought, I'm not going to look at them, because I understand that".

Some participants clarified that they believed that listening and viewing the images without reading was more useful for English practice. The participants who were found to consistently read the captions either attributed this behavior to having limited L2 proficiency or to the physical salience of captions, which automatically attracted their gaze due to their dynamic nature and distinct color ("I don't even want to, but if I have something to read I always think of reading first so... not because I did not understand what they were saying"). The rest of the participants reported focusing on the faces of the characters in an attempt to understand their emotions and improve general comprehension, as this was their usual viewing strategy when watching videos at home. Interestingly, participants' beliefs regarding their perceived reliance on the captions aligned quite closely not only with their level of agreement with the statement "I read the subtitles" in the questionnaire, but also with the online processing data obtained through eye-tracking. Finally, three participants reported increased alertness due to the post-viewing comprehension questions and language test, without explicitly referring to a focus on pronunciation.

To sum up, most of the participants (80%) mentioned at least once that they were watching for general comprehension, and the majority specifically explained that their viewing strategy involved primarily listening to the dialogues and avoiding reading, in favor of attending to other visual elements on screen. Unenhanced captions

typically supported speech processing whenever issues arose in segmenting speech and inferring the meaning of words and expressions. Audio-synchronized textual enhancement was often reported to attract the viewers' attention due to its physical salience, but the participants almost always associated the enhancement of target words with an instructional focus on their semantic or grammatical properties. As a result, despite having some awareness of the upcoming speaking tests, they did not report paying explicit attention to the *pronunciation* of past <-ed> endings.

3.2.6. Discussion

This study investigated English learners' processing of L2 captioned video containing audio-synchronized textual enhancement through the collection of offline and online data from pronunciation tests, eye-tracking, and stimulated recall interviews. To begin with, the participants' production of regular past <-ed> endings and previous knowledge of the past <-ed> pronunciation rule was assessed to determine whether a focus on the target feature would be beneficial for our Spanish L1 intermediate learners of English. Accuracy scores reached ceiling in a read aloud task, but not in the less controlled context of a prompted narrative task, in line with the hypothesis that accurate use of the target feature in spontaneous production can be challenging even for learners at higher proficiency levels (Kruk & Pawlak, 2021; Solt et al., 2003). Although most of the participants were able to identify some elements of the past <-ed> pronunciation rule, they did not seem to have reached yet the type of "awareness at the level of understanding" that would promote explicit restructuring of their internal L2 knowledge system (Leow, 2015).

Regarding research question one, our hypothesis was partially supported, as learners' responses to the questionnaire and stimulated recall interviews indicated that they

noticed the audio-synchronized enhancement and correctly identified the shared grammatical properties among the target words. However, the analysis of eye-tracking data did not highlight clear effects of audio-synchronized enhancement on the allocation of attention to enhanced target words, but rather pointed at a predominant role of individual factors. In other words, although there was a pattern of fixating enhanced target words for longer, the amount of visual attention paid to each word may have largely depended on its intrinsic features and on each participant's viewing preferences (Kam et al., 2020; Wisniewska, 2021). While, in contrast with the results of study 1, audio-synchronized enhancement appeared to be detrimental for audiovisual synchrony, the findings of this study were less reliable due to the very few data points available in the analysis of fixation distance.

Research question two asked what linguistic aspects learners attend to while watching videos with unenhanced captions and with audio-synchronized textual enhancement. The participants' comments pointed at a mismatch in the focus of attention when viewing video under these two conditions, highlighting the different outcomes of internally generated salience and of the external salience generated through input enhancement (Sharwood Smith, 1991). In the absence of enhancement, captions were deliberately used by participants to support speech processing by facilitating speech segmentation and the mapping of auditory word forms onto written forms. However, the presence of audio-synchronized textual enhancement primarily led to noticing of the grammatical properties of the target words, and this focus on grammar was restated in the questionnaire responses. The participants' failure to identify the intended focus on pronunciation is unsurprising, giving the incidental nature of the study and the absence of an explicit focus of instruction (Leow, 2001; Leow & Martin,

2017). On the one hand, a morphophonemic feature like past <-ed> is suitable for an intervention featuring textual enhancement thanks to the underlying phoneme-grapheme correspondence, which facilitates the mapping of targetlike phonological representations onto pre-existing orthographic representations. On the other hand, L2 learners' processing of input enhancement may have been primed by previous encounters with past <-ed> which likely revolved around its grammatical properties. The type of incidental processing stimulated by simultaneous processing of meaning and form, which would have required automatic processing of at least one of the two aspects, probably caused learners to fall back on their previous knowledge of the target words and features (Han et al., 2008). However, the noticing of grammatical aspects should not be perceived as an obstacle in the acquisition of past <-ed> pronunciation, but rather as the basis for the deeper processing of different aspects of the target words, including their auditory form.

The stimulated recall methodology allowed for a nuanced analysis of the effects of modality and proficiency in the acquisition of speech from textually enhanced video. Among learners with an upper-intermediate level of English, who understood the spoken dialogue with relative ease, there was a pattern of processing auditory input for meaning first and making a relatively conscious effort to avoid captions until a comprehension issue arose. As these learners were confident about their knowledge of regular past formation and use, the visual salience of enhanced and unenhanced captions seemed to have limited effects on their noticing of linguistic form. Contrary to our hypotheses, for these learners the most effective instances of modality integration may actually have occurred when the enhancement was internally guided by their own input processing needs (Mitterer & McQueen, 2009). However, learners

with lower English proficiency openly acknowledged regularly reading captions because they represent an invaluable support in the bottom-up processing of auditory information. Although their level of general comprehension was good, following the spoken dialogue without captions may have been challenging due to limitations in vocabulary knowledge and to slower and less automatic processing of auditory information (Montero Perez et al., 2013; Winke et al., 2013). For these less proficient learners, audio-synchronized textual enhancement may have been more effective in promoting a focus on linguistic form, even if this form was grammatical rather than phonological, through the disruption of automatic reading patterns and the redirection of attentional resources (Stenton, 2012).

Finally, hypothesis 3b was supported, as participants' perceptions of audio-synchronized textual enhancement were mixed, with half of the sample indicating that they found it useful and the other half distracting. It is possible that some participants believed that the enhancement was distracting because they assumed that it was aimed at highlighting grammatical aspects of English regular past, which were already familiar to them. If the enhancement had been associated with the intended focus on the pronunciation of past <-ed> endings, which was at least partially unfamiliar to them, more positive perceptions of enhancement may have been reported.

3.2.7. Conclusion and limitations

To sum up, the results of study 2 align with the results of study 1 in indicating that audio-synchronized textual enhancement may direct learners' attention to selected target words during exposure to L2 captioned video. In this study, the enhancement seemed more effective at the lower-intermediate proficiency level, as learners were already used to processing large quantity of written input during the viewing.

However, possibly due to the morphophonemic nature of the target feature, watching the clips with audio-synchronized enhancement but without being explicit instructed to pay attention to pronunciation did not promote a focus on the target words' pronunciation.

This study presents some limitations, starting from the very small number of participants that could be recruited in the school context and the limited time allotted for the study. Possibly as a result of the small sample, the comparison of the participants' eye-tracking data under the enhanced and unenhanced condition was largely inconclusive. Another limitation may have been represented by undetected differences between the two sets of target words, which had to be directly compared due to the adoption of a within group design. In addition, since the prompted narrative task could only be administered at pre-test and the participants achieved ceiling performance in the read aloud even at pre-test, it was impossible to gauge whether there was an improvement in learners' pronunciation of the target feature. However, this was considered a minor limitation, as we did not expect to detect significant gains in these relatively explicit tests from incidental exposure to one video clip without any additional activities. In fact, the narrative task was mainly used in combination with the X-Lex vocabulary test to compare the proficiency of these participants with that of the participants in study 3. Unfortunately, the elicited imitation task was too long and complex to set up in the classroom, therefore the proficiency of the participants in study 2 could not be directly compared to that of the participants in study 1. Finally, a major limitation is represented by the explicit prompts given by the interviewer. However, when using think aloud protocols, participants often fail to address the discrete items that are the focus of the study (Leow et al., 2014). In this study, more

explicit prompts had to be given to learners with a lower level of English and to learners who, possibly due to anxiety, were hesitant to speak unprompted and would have otherwise failed to provide any stimulated recall data. Moreover, the few cases in which the prompt was so explicit that it could guide the learners' answer were excluded, and the relatively high percentage of "I don't know" instances suggests that the learners responded truthfully.

Recommendations for future research include recruiting a larger number of participants to compare the effects of enhanced and unenhanced video adopting a between group design. When using stimulated recall to investigate learners' processing of textual enhancement in captions, the use of explicit prompts should be more carefully avoided. Finally, in a longer study in which learners' attention is more explicitly drawn to pronunciation during the viewing, different results regarding the effectiveness of audio-synchronized textual enhancement may be observed. To address this aspect, study 3 involved repeated exposure to enhanced video in combination with activities designed to foster a more explicit focus on form.

3.3. Teaching pronunciation through L2 video with audio-synchronized textual enhancement and audiovisual activities.[7]

3.3.1. Study aims

In study 1, we found evidence that audio-synchronized textual enhancement can promote a focus on L2 phonological form, as it directs learners' attention to specific words in captions in synchrony with their auditory onset. We subsequently investigated whether the findings from our study involving first-year university students also applied to 10th grade high school students, our target population for a classroom intervention. Although the effects of audio-synchronized enhancement on phonological acquisition were unsubstantiated based on the available tests, learners with a lower-intermediate to upper-intermediate proficiency level noticed the enhanced target words and successfully processed their grammatical and lexical features. In study 3, our aim was to incorporate audio-synchronized enhancement into a pronunciation-focused intervention, in order to maximize learning opportunities and carry out an ecologically valid assessment of its effectiveness in a classroom setting. To this end, we designed a number of activities involving the manipulation of the audiovisual elements in the video clips selected for the study. By engaging learners in collaborative activities that required them to listen carefully to the spoken dialogue as many times as needed and engage in meaningful pronunciation practice, we aimed at promoting the sequential processing of form after meaning, facilitating phonological acquisition.

[7] An article based on this work has been previously published as: Galimberti, V., Mora, J. C., and Gilabert, R. (2023). Teaching EFL pronunciation with audio-synchronised textual enhancement and audiovisual activities: Examining questionnaire data. In A. Henderson & A. Kirkova-Naskova (Eds.), *Proceedings of the 7th international conference on English Pronunciation: Issues and Practices.* University of Grenoble.

3.3.2. Research questions

In the context of a pronunciation-focused classroom intervention involving multiple exposures to video containing audio-synchronized textual enhancement and audiovisual activities:

RQ1: Do learners achieve significant L2 pronunciation accuracy gains?

RQ2: Do the learners' proficiency, listening comprehension skills and phonemic coding ability mediate pronunciation accuracy gains?

RQ3: What are the learners' perceptions of:

a) videos containing audio-synchronized textual enhancement?
b) pronunciation-focused audiovisual activities?

Our hypotheses are that:

HP1: We expect to observe significant pronunciation gains in the spoken production of target words encountered in the intervention. Due to the substantial amount of practice involved in the activities, we expect the pronunciation gains to generalize to unfamiliar items. The gains will be larger for the group watching the video clips with audio-synchronized enhancement, as directing the learners' focus to the target feature during the first viewing should enhance their processing of this feature in the subsequent activities.

HP2: Based on research showing the levelling effects of explicit instruction and audio-synchronized enhancement, we do not expect more proficient learners to obtain larger gains. However, we do expect learners with more advanced listening comprehension

skills and phonemic coding ability to process auditory input more effectively and achieve larger gains.

HP3: We expect learners' perceptions of textual enhancement and of the audiovisual activities to be positive.

3.3.3. Pronunciation target

This study had the same linguistic target as study 2, i.e., the pronunciation of the English regular past tense ending. In particular, at the end of this intervention learners were expected to develop a certain automaticity in the accurate pronunciation of past <-ed> endings. They were also expected to attain understanding of the rule by developing awareness of the different allomorphs, or at least of the detrimental effects of the omission of the <-ed> ending when necessary, and of the erroneous addition of an epenthetic vowel.

3.3.4. Pilot study

The pronunciation tests, activities and questionnaire were piloted in a preliminary study involving three classes of L1 Spanish/Catalan learners of English (N = 70, age 15), who attended the same high school as the participants in study 2 and 3. The pilot study took place over a total of six hours of English, two per class, and involved two components. A small number of students from each class (n = 4 per hour, N = 24) were taken to a quiet room where they carried out the activities in pairs, did the pronunciation tests, and responded to the questionnaire individually. At the same time, we asked their teachers to implement part of the intervention by projecting the video clips on the whiteboard and monitoring the students' completion of each activity as they worked in pairs.

Based on the analysis of the questionnaire and on the recordings of the activities completed under the supervision of the researchers, participants showed good autonomy after a very short training, and valued positively the activities, especially because they gave them the opportunity to learn English through exposure to fun materials and peer collaboration. Participants' performance in the pronunciation tests confirmed that 10th grade L1 Spanish students found the accurate production of English past <-ed> endings challenging. Notably, the accuracy attained for stimuli involving non-salient contexts where the <–ed> ending was pronounced /d/ or /t/ and formed part of a consonant cluster (e.g., *We hoped to find you here*) was lower compared to the accuracy attained in salient contexts where <-ed> was pronounced /id/ and followed by a vowel (*They expected a warm welcome*). The most common mispronunciations involved the omission of the <-ed> ending when it was required and the unnecessary addition of an epenthetic vowel, in line with previous studies on past <-ed> perception (Strachan & Trofimovich, 2019; Solt et al, 2003). In the narrative tasks, the considerable variability in the amount of regular past verbs produced by each participant, as well as in the percentage of <-ed> endings accurately realized, highlighted the need to include both stories in order to collect a sufficiently large sample of verbs. This pilot also led to the decision to exclude from the tests a small number of target words which represented ambiguous models for learners due to the <-ed> ending not being clearly audible in the spoken dialogues of the video clips.

The intervention originally included an oral summary task based on written prompts referring to actions carried out in the past, which was designed to help participants summarize the content of each clip while practicing past <-ed> pronunciation. This

task was removed after the pilot because the students had trouble understanding the instructions and, despite further explanations, they ended up answering the questions in the prompts instead of retelling the content of the clip. In order to optimize time availability while nevertheless assessing comprehension of the video clips, the oral summary was replaced with a written questionnaire and individual responses were collected through a Google Drive form. The teachers reported practical issues involving the poor quality of the speakers, which limited the effectiveness of viewing the video in a whole class setting, and other technical issues which prevented some students from completing the activities in each session within the 50 minutes allocated in the classroom. For this reason, in the intervention, participants completed the entire session on their laptops, and attended a training session before the start of the intervention in which they downloaded the intervention materials and learned how to use the programs needed to do the activities.

To sum up, the results of the pilot confirmed the usefulness of targeting the pronunciation of regular past tense <-ed> endings. The analysis of the performance of a small sample from the target population helped us identify the most common mispronunciations and select the most relevant test stimuli. We also formulated clearer instructions for the activities and identified the need to support the participants by regularly monitoring their understanding of the instructions throughout the intervention. Finally, the teachers' feedback informed practical decisions regarding the best classroom configuration for the implementation of the intervention.

3.3.5. Methodology

This study took place during the semester following the pilot phase and involved six teaching sessions and three testing times distributed over 16 weeks (Figure 3.18). The

pre- and post-tests aimed at assessing not only learners' accurate use of the target pronunciation feature in controlled and spontaneous speech, but also the moderating effect on pronunciation learning of individual factors relating to the learners' proficiency and aptitude. The teaching phase involved the implementation of various activities (explained in detail in the *Materials* subsection) with two intact classes. One group of learners was exposed to video clips with audio-synchronized enhancement, whereas the other watched unenhanced videos. In addition, a control group was pre- and post-tested to measure any practice-related effects in the absence of focused teaching through exposure to the intervention video clips and pronunciation learning activities. Learners in the three groups responded to a final questionnaire containing questions on their language background, English learning experience, knowledge of the past <-ed> pronunciation rule and, for the intervention groups only, their perceptions of the videos and activities. Besides assessing learners' perceptions of the intervention, the analysis aimed at establishing whether this teaching approach led to significant gains in pronunciation accuracy and to the acquisition of the pronunciation rule. Finally, correlation analyses were conducted to investigate whether the learning gains obtained were moderated by individual differences.

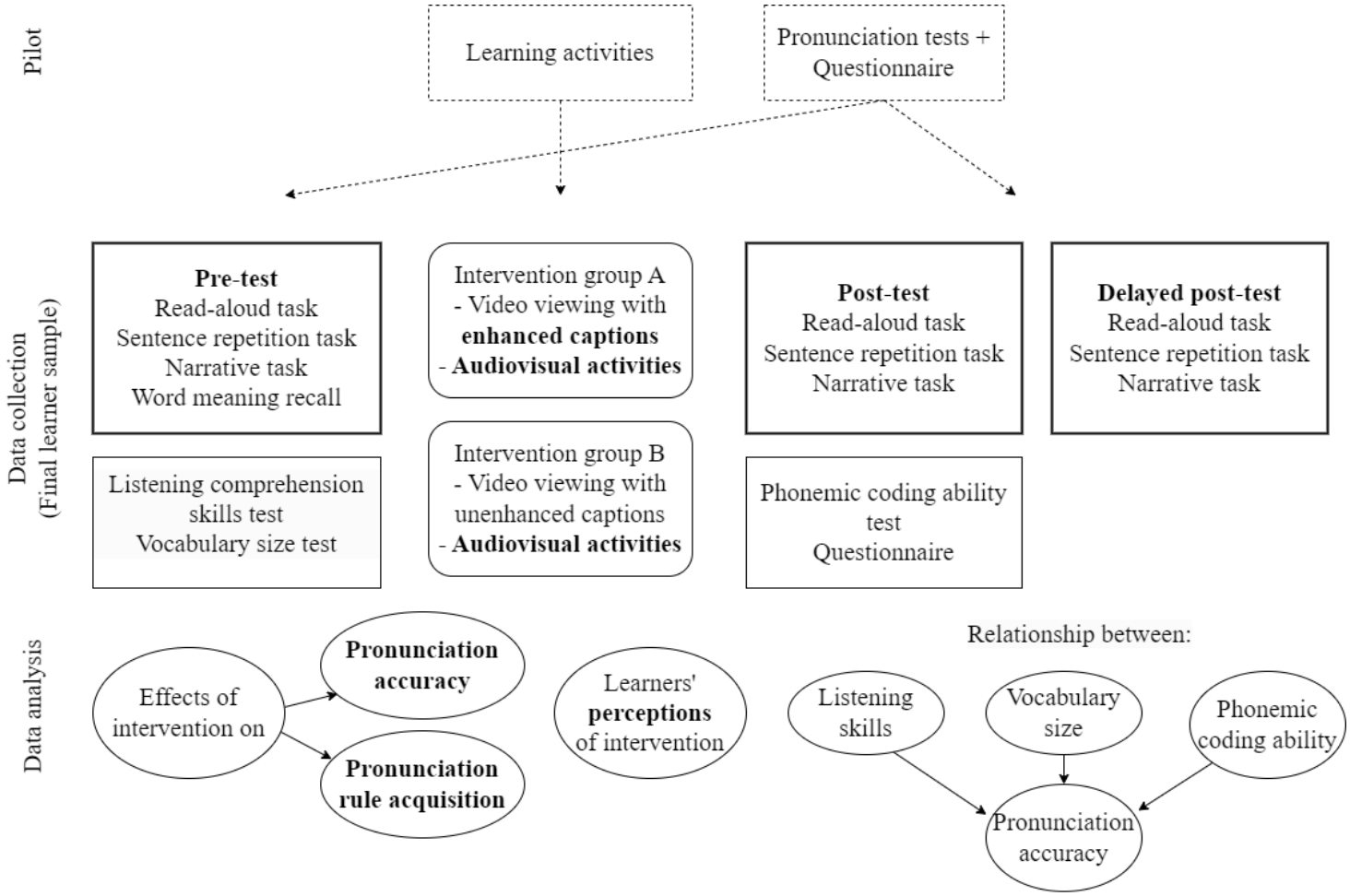


Figure 3.18. Study 3 methodology overview.

Participants

The intervention was implemented with three intact classes of L1 Spanish/Catalan EFL learners (N = 78, age 15), 53 of which (female = 32) also participated in the tests and completed the questionnaire. To ensure the comparability of participant groups in terms of English language proficiency, we tested their vocabulary size, phonetic coding ability and listening comprehension skills with the X_Lex (Meara & Milton, 2003), Llama_E (Meara, 2005) and Oxford Placement Test (OPT) listening section, respectively. ANOVAs found no difference between groups on the results of the X_Lex (F(2, 50) = .52, p = .60), Llama_E (F(2, 47) = .01, p = 1.00) and OPT (F(2, 50) = .30, p = .74). Based on the results of these tests and on the textbook used in class, the average proficiency level was estimated to be lower- to upper-intermediate, and the profile of these participants and those in study 2 appeared to match closely.

To delineate the language learning profile of our participants more clearly, we investigated their exposure to the second language outside the classroom. Most participants reported taking English classes outside the regular school time, with considerable variability in the duration of the classes (m = 3.4 years, SD = 3.8). Around half of them had visited an English-speaking country, and 39 out of 44 participants reported regularly watching English-language TV shows and videos with L2 captions. The three groups were not significantly different in terms of amount of participation in extracurricular classes (F(2, 50) = .90, p = .41), time spent in an English-speaking country (F(2, 50) = .09, p = .92), and time spent watching TV every week (p = .95, Fisher's exact test). For further details regarding the demographic characteristics of the participants, please refer to Table 3.34. Written consent was obtained from the parents of each student in the final sample, in the context of a formal

partnership between the school and the University of Barcelona. A unique identifier was generated for each participant using a combination of alphanumeric characters, e.g., participant 2 in intervention group A was assigned the code A_02.

Table 3.34. Participants' demographics by group.

	Intervention group A (*n* = 18)			Intervention group B (*n* = 17)			Control group C (*n* = 18)		
	M	*SD*	95% CI	*M*	*SD*	95% CI	*M*	*SD*	95% CI
Age at testing	15.00	.00	[15.00, 15.00]	15.00	.35	[14.82, 15.18]	15.00	.00	[15.00, 15.00]
Vocabulary size	2705.56	519.87	[2447.03, 2964.08]	2614.71	538.16	[2338.01, 2891.4]	2819.44	714.58	[2464.09, 3174.8]
Phonetic coding ability	76.67	17.82	[67.8, 85.53]	76.47	19.02	[66.69, 86.25]	76.00	19.57	[65.16, 86.84]
Listening comprehension skills	72.22	4.41	[70.03, 74.42]	73.18	3.59	[71.33, 75.02]	73.44	6.33	[70.3, 76.59]
Extracurricular classes (years)	3.74	3.93	[1.78, 5.69]	4.06	4.45	[1.78, 6.35]	2.44	2.96	[.97, 3.91]
Time spent in an English-speaking country (months)	1.49	3.236	[-0.12, 3.10]	1.44	3.325	[-0.27, 3.15]	1.10	2.656	[-0.22, 2.42]

Materials

Clips

The video clips consisted of five excerpts from the first three episodes of season one of *The Good Place*, with a total duration of 22’ 42”. The shortness of the individual clips was in line with the audiovisual framework, in which a maximum length of 5 minutes is recommended (Zabalbeascoa et al., 2012). Shorter videos are less likely to overload the learners’ short-term memory while they carry out activities that involve holding in mind information about the content of the video while also focusing on linguistic form (Campbell, 2016; Sánchez-Requena, 2017). The clips mainly featured dialogues between two or three characters in quiet situations, and the characters spoke General American English, except one minor character with a British accent (RP). As the overall coverage of the script was 95% at K2 and 98% at K5, the content was deemed appropriate for our participants, who had an overall intermediate proficiency level and an average vocabulary size of 2.5k to 3k words (Rodgers, 2013). Two versions of the clips were created depending on whether the captions were enhanced (for the experimental group) or unenhanced (for the comparison group), except clip 1, as the captions were manipulated differently to match the aim of the perception activity (see description of the activities below). Both unenhanced captions and captions containing target words enhanced 500 ms before auditory onset were in the same format and font as in study 2. Detailed information for each clip can be found in Table 3.33. Participants’ comprehension was tested at the end of each session with five multiple choice and five true/false questions targeting global and local aspects of the video clip content (Rodgers, 2013).

Table 3.35. Video clip characteristics.

	Clip 1	Clip 2	Clip 3	Clip 4	Clip 5	**Total**
Length	4’ 17”	4’ 05”	4’ 24”	5’ 16”	4’ 40”	**22’ 42”**
Number of words	534	571	700	798	760	**3.363**
Coverage 95%	K2	K3	K4	K3	K2	**K2**
Coverage 98%	K5	K6	K6	K6	K6	**K5**
Textual manipulation	Target words missing from captions	Textual enhancement	Textual enhancement	Textual enhancement	Textual enhancement	
Number of enhanced target words	6	5	5	8	4 (1 new + 3 enhanced previously)	**28**

Audiovisual activities

Each session included an audiovisual activity with a focus on phonological form, which created opportunities for learners to practice perception and/or production of the target words encountered during the first viewing of the clip (see Files 4-8 in the Supplementary Materials). Participants were given specific instructions to either: 1) fill in missing words, including the target words, in captions; 2) order and label segments of the clip that include the target words; 3) identify muted target words in shorter excerpts without the support of captions, and repeat the entire sentence; 4) provide voiceover for a muted clip using captions; 5) order uncaptioned segments containing the target words and provide voiceover for the resulting sequence. This section explains the dynamics and underlying rationale of the five audiovisual activities.

In session 1, participants could self-test their perception of the target feature by completing the captions of clip 1 with the target words and the preceding or following words. Successful performance in this audiovisual activity required accurate perception and correct segmentation of L2 speech. It was different from a traditional "dictation" activity in that it involved visual cues as well as auditory ones, the input resembled real-life speech, and learners used a subtitling software (Aegisub) which allowed them to immediately view the video with the captions they created.

In session 2, participants needed to pay close attention to L2 speech, although in the context of a meaning-focused comprehension activity. After watching five short excerpts from clip 2 (5-8 seconds) containing the five target words of the session, they worked in pairs to order the excerpts by copying a brief annotation for each clip in a table. Then, they indicated whether the verbs of each clip were in the present or past tense. The aim was to practice accurate perception of each verb ending in a context where overall meaning comprehension was more essential than in the activity of session 1.

The activity in session 3 featured four uncaptioned clips from clip 3 (7-15 seconds) in which five target words were muted. Working in pairs, the learners identified the words with the help of a hint consisting of a few letters for each word and said out loud the complete soundtrack for each clip. In this revoicing activity, each learner could practice the exact imitation of short sequences of spoken words containing the target feature, thanks to the availability of some of the auditory and written input.

In session 4, the learners practiced saying out loud the dialogue of a muted excerpt from the clip (1'18") with the help of the captions. This activity was different from a traditional "read-aloud" activity due to the availability of the video, but also because

a bigger effort was required to speak fluently and in synchrony with the characters' lip movements. The aim was to practice the automatization of accurate production of the target feature in the presence of textual support.

The activity in session 5 involved six unenhanced clips in random order (6-14 seconds), only four of which had been extracted from the video in session 5. The learners selected the clips, then revoiced each clip in the correct order. This bimodal (audio and moving image without captions) listen-and-repeat activity without textual support encouraged a focus on the target forms and their accurate production in a meaningful context.

Awareness raising activity

At the end of each session, an awareness raising activity provided an explicit focus on one element of the past –ed pronunciation rule through the categorization and analysis of the target verbs and another 19 verbs not included in the clips or tests. The participants read or listened to a list of verbs and were asked, for example, to underline the letter preceding the <–ed> ending of some verbs and indicate if the corresponding sound was voiced or voiceless; group the verbs in the list based on the pronunciation of the <-ed> ending; or determine whether the vowel represented by letter <e> in the <–ed> ending was pronounced or remained silent. The purpose of these activities was to draw explicit attention to key aspects of regular past tense <-ed> pronunciation, including the presence of different allomorphs, the difference between voiced and voiceless consonants, and the effects of the phonetic context preceding and following the <-ed> ending (Strachan & Trofimovich, 2019).

Target words

The pronunciation tests featured 61 verbs that the participants had either encountered in the intervention clips (*familiar* items) or never encountered in the clips or activities (*novel* items), with each familiar and novel verb being tested once only (Table 3.36). It must be noticed that novel does not mean "never seen", since both types of items were chosen among higher frequency bands (mostly K1 and K2) to investigate changes in existing mental representations of regularly inflected verbs (Clahsen et al., 2010; Ullman, 2005). Familiar words appeared on average twice in the clips (range = 1-6), excluding the verb "want" which appeared a total of 16 times, but only 3 of these in <-ed> form. Familiar items consisted of 7 verbs that only appeared in base form and were thus tested, and 25 verbs that were tested either in their base or <–ed> form, as they appeared in the clips in the past simple tense (15 items), past participle in a passive construction (6), present perfect (2) or as predicative adjectives (2). Attributive adjectives were not selected as targets due to exceptions in pronunciation rules (e.g., *beloved father* pronounced /bɪˈlʌvɪd ˈfɑːðə/ although the /d/ allomorph is used in *He loved his father*). Low frequency words and words not clearly audible in the soundtrack of the clips were not selected. A comparable number of test items was selected for each of the three past <–ed> allomorphs (/d/, /t/, /ɪd/). To approximate the variability of real-life use, there was some variety in the preceding phonological context (e.g., the last phoneme of a root verb with a past /t/ ending could be [p], [s], [k] or [ks]).

Table 3.36. Linguistic and presentation properties of the target words by test (RA = Read-aloud; DSR = Sentence repetition task).

Word	Test	Enhanced[a]	Familiar[b]	Orth length	Phonological length[c]	Occurrences in clips	Lexical frequency[d]
Asked	RA	E	F	5	4	3	216
Call	RA	E	F	4	3	3	861
Cause	RA	E	F	5	3	2	310
Died	RA	E	F	4	4	3	157
Dropped	RA	E	F	7	5	1	49
Raised	RA	E	F	6	5	2	26
Rescued	RA	E	F	7	7	1	5
Slipped	RA	E	F	7	5	1	17
Torture	RA	E	F	7	6	1	16
Wanted	RA	E	F	6	6	16	502
Access	RA	U	F	6	5	2	32
Calculate	RA	U	F	9	10	1	2
Need	RA	U	F	4	3	4	1295
Walk	RA	U	F	4	3	2	216
Confess	RA	U	N	7	6	0	16
Danced	RA	U	N	6	5	0	9
Entered	RA	U	N	7	6	0	15
Fixed	RA	U	N	5	5	0	32
Invented	RA	U	N	8	8	0	16
Land	RA	U	N	4	4	0	88
Paused	RA	U	N	6	4	0	1
Produce	RA	U	N	7	7	0	5
Sparked	RA	U	N	7	6	0	1
Study	RA	U	N	5	5	0	49
Suggested	RA	U	N	9	9	0	10
Arrive	DSR	E	F	6	5	2	19
Clean	DSR	E	F	5	4	4	121
Collect	DSR	E	F	7	6	2	20
Created	DSR	E	F	7	8	1	23
Decorated	DSR	E	F	9	10	1	3
Destroyed	DSR	E	F	9	8	1	31
Ended	DSR	E	F	5	5	2	30
Finished	DSR	E	F	8	6	1	84
Happen	DSR	E	F	6	5	5	254
Kicked	DSR	E	F	6	4	2	31
Lived	DSR	E	F	5	4	6	66
Move	DSR	E	F	4	3	1	418
Rolled	DSR	E	F	6	5	1	8

Started	DSR	E	F	7	7	6	188
Traced	DSR	E	F	6	6	1	4
Like	DSR	U	F	4	4	3	3999
Sound	DSR	U	F	5	5	1	143
Stop	DSR	U	F	4	4	1	707
Accept	DSR	U	N	6	6	0	53
Brush	DSR	U	N	5	4	0	14
Change	DSR	U	N	6	5	0	240
Consider	DSR	U	N	8	8	0	52
Continued	DSR	U	N	9	9	0	7
Cross	DSR	U	N	5	4	0	55
Dressed	DSR	U	N	7	5	0	47
Expected	DSR	U	N	8	9	0	104
Hoped	DSR	U	N	5	5	0	321
Married	DSR	U	N	7	5	0	238
Offer	DSR	U	N	5	4	0	75
Opened	DSR	U	N	6	6	0	34
Provided	DSR	U	N	8	9	0	10
Pushed	DSR	U	N	6	4	0	20
Search	DSR	U	N	6	4	0	48
Separated	DSR	U	N	9	8	0	11
Support	DSR	U	N	7	6	0	51
Wished	DSR	U	N	6	4	0	8

[a] E = Enhanced in clips; U = Unenhanced or unapplicable (for words not present in the clips). [b] F = Familiar, encountered in the clips; N = Novel. [c] IPA notation system for American English. [d] Frequency per million words in the $SUBTLEX_{US}$ database (Brysbaert & New, 2009).

Word meaning recall test

To assess previous knowledge of the target words, participants were asked to indicate the meaning of each English word in Spanish or Catalan. The translations were considered acceptable if they appeared in an online Spanish or Catalan dictionary.

Pronunciation tests

a. Read-aloud task

In this task, which was identical to the read aloud task in study 2, the participants read aloud 28 present and past tense verbs that appeared briefly on screen. The internal

consistency reliability of the task was satisfactory, as Cronbach's alpha calculated from the results of the pilot study was .80.

b. Delayed sentence repetition task

The delayed sentence repetition task was designed and piloted specifically for this study. In a typical delayed sentence repetition (DSR) task, the participant listens to several sentences (one at a time) and, after some seconds, repeats each sentence verbatim. As this task involves storing in short-term memory a considerable amount of semantic and syntactic elements, successful repetition of the sentence stimuli depends on efficient processing of different aspects of the second language (Spada et al., 2015). DSR tasks have been used to test problematic L2 phonological and morpho-phonological features, as repeating an utterance after a sufficiently long time-lag involves decoding it and then encoding it again using one's own internal linguistic resources (Spada et al., 2015; Trofimovich et al., 2009). In this study, a DSR task was deemed to be more informative than a perception-based task, as accurate perception does not guarantee accurate production, but the inability to perceive a language feature invariably leads to difficulties in producing it (Trofimovich et al., 2009).

In this study, the DSR task assessed the participants' procedural knowledge of the target pronunciation feature by requiring a higher degree of automaticity, compared to the read-aloud task, in the retrieval and articulation of accurate phonological forms. Sentence stimuli incorporating both base forms and past <–ed> forms were included to establish a performance baseline and increase learners' focus on accurately identifying and conveying verb tense (Strachan & Trofimovich, 2019). Further, lexical cues to verb tense were removed from the input to emphasize its perceptual/acoustic properties, as the presence of time adverbials would have helped

learners understand the sentence without paying attention to verb tense morphology (Bell et al., 2015; Strachan & Trofimovich, 2019).

The 36 sentence stimuli were between 6 and 12 syllables long (m = 8.97, SD = 1.81), and each sentence contained only one verb in second position, coming after a pronoun or short subject phrase (Table 3.37). The subject was never the third person pronoun, so that the absence of the third person morpheme <-s> would not represent a clue as to the verb tense. As the coverage was 95% at K2 (2000 most frequent English words) and 98% at K3, English learners with an intermediate level were expected to be able to repeat the majority of the sentences. To include phonological contexts of varying perceptual difficulty, half of the words following the target verbs in each subset (divided by familiarity, form and <-ed> allomorph) started with a vowel and half started with a consonant or approximant sound (Strachan & Trofimovich, 2019). To avoid assimilation, words starting with [t], [p], [k] and [d] were not used after the [t] ending, and words starting with [d], [g], [b] and [t] were not used after the [d] ending.

Table 3.37. Sentence stimuli used in the delayed sentence repetition task.

Practice items	
We traveled on a sailing boat	
All climbers face their fear of heights	
I have no cash in my pocket	
Test items	
The fires destroyed an ancient building	His stories sound a bit unrealistic
I lived next to a football stadium	The flowers continued to grow
Black clouds rolled up the valley	Some people married against their parents' will
Both trains arrive at platform six	You opened the bottle for me
I clean with natural products	We consider our friends part of the family

Strange things happen during the night	Most universities offer English courses
We move our hands without thinking	Some guests dressed up in suits
You finished all the chocolate	They hoped to find you here
The kids kicked the ball against the wall	We pushed every door open
We traced a call from a foreign number	Many actors wished to get the lead role
Philosophers like to explore the human mind	I cross the bridge to go to school
Two buses stop outside the hotel	We brush our teeth with coconut oil
The candles created a pleasant atmosphere	Seabirds search among the waves for fish
We decorated the Christmas tree with lights	They expected a warm welcome
The shows ended quite late	Service dogs provided comfort to veterans
The lovers started all over again	We separated gold from sand
I collect stones at the beach	We accept the job offer
The players change two letters in each word	The teachers support all students equally

As the predominant variety in the video clips of the intervention was American English, the test stimuli were recorded in a soundproof booth by two native speakers of AmE. Each sentence was recorded twice with a falling intonation. Then, the second repetition of each sentence was extracted using Praat and the audio files were low-pass filtered (60 Hz) and normalized for mean amplitude using the *filter* function in GSU Tools 1.9 (Owren, 2008). The DSR task was delivered to the participants individually via DMDX (Forster & Forster, 2003), through a set of open headphones. Participants saw a logo in the shape of a cross on the screen, then heard the sentence and, after 3 seconds, a different logo appeared to prompt them to speak. They had 6 seconds to repeat the sentence verbatim or at least try to repeat as many words as they could remember from the sentence. The pause between listening and speaking was expected to discourage rote repetition and promote a focus on meaning. The whole test lasted about 8.5 minutes. Cronbach's alpha after piloting the DSR test was .80.

c. Prompted narrative tasks

The same narrative tasks based on comic strips that were used in study 2 were also used in study 3. Since oral summaries are common tasks both in the FL classroom and in real life, all participants were expected to be familiar with the dynamics of this task even before taking the pre-test. As a result, the intervention groups were not expected to gain any additional advantage beyond those obtained from the intervention.

Proficiency tests

a. Vocabulary size test

The X_Lex written vocabulary test (Meara & Milton, 2003), which tests participants' knowledge of the 5000 most frequently used English words, was used as an indicator of vocabulary size. In addition, receptive vocabulary size assessed through the X_Lex test has been found to explain L2 proficiency to a large extent and can therefore be considered a valid indicator of overall proficiency (Miralpeix & Muñoz, 2018). During this test, participants are presented with a series of words and required to indicate their familiarity with each word. If participants falsely claim knowledge of any pseudo words resembling real words, they are strongly penalized.

b. Listening skills test

The listening section of the Oxford Placement Test (OPT) was used as a test of general listening comprehension and sound discrimination. In this test, the participant is presented with 100 written sentences in which a word or expression is missing. While listening to a native speaker say each sentence in full, the participant indicates which of two very similar words or expressions is the correct option to complete the sentence. The test was administered through a set of closed headphones.

c. Aptitude test

An aspect of language aptitude relevant to pronunciation learning is phonemic coding ability, or the ability to discriminate and code foreign sounds, as learners with limited phonemic coding ability may progress more slowly than others with a similar profile (Celce-Murcia et al., 2010). The Llama suite of aptitude tests (Meara, 2005; Meara & Rogers, 2019) is widely used in SLA, and the Llama_E test has been validated as a test of phonemic coding ability (Rogers et al., 2017, Saito, 2017). Llama_E requires participants to work out how a new writing system works, by making connections between language sounds and grapheme combinations.

Questionnaire

The first part of the questionnaire contained the same questions as in study 2 on language background, perceptions of intervention materials and awareness of the target feature (Appendix B.3). In the second part, the participants were asked about their perceptions of AV activities, through a combination of questions adapted from Sokoli (2018) and newly introduced items. The items that were introduced in study 3 to expand upon the questionnaire used in study 2 are presented in Appendix D. In order to assess social interaction, a crucial component of active learning and task engagement (Zabalbeascoa et al., 2012), a question on peer collaboration was included. Due to the impracticality of collecting eye-tracking data during whole-class activities, participants were asked to report whether they read the captions while watching the video clips. This subjective measures of attention allocation was considered useful since, in study 2, learners' self-perceived reliance on captions aligned with the objective metrics obtained from eye-tracking recordings. Sokoli's (2018) questions on the participants' sense of learning were slightly modified to

explore whether the learners primarily directed their attention towards grammar or pronunciation during the activities. This adjustment aimed to account for the potential impact of the intervention on the participants' awareness of the grammatical function and phonological form of the verbs. As a measure of reported noticing, participants in group A were asked the same questions as those in study 2 regarding the presence and usefulness of enhanced target words. A reduced version of the questionnaire that did not contain questions about participants' perceptions of the intervention was created for participants in the control group.

Procedure

Classroom intervention

The intervention spanned six weeks, with each intervention group receiving fifty minutes of instruction per week. The research study comprised three groups, corresponding to three classes: (1) Intervention group A, who watched the video clips with enhanced target words and engaged in pronunciation-focused audiovisual activities and awareness-raising tasks; (2) Intervention group B, who watched the video clips without the target words enhancement (original captions) and then performed the same audiovisual activities and awareness-raising tasks as group A; (3) The control group, who only completed pre and post-tests and continued with their regular EFL classes, while teachers avoided providing any planned or reactive focus on past tense <-ed> pronunciation. The control group served as a baseline for assessing past tense pronunciation accuracy gains and rule knowledge among learners from the same population as the intervention groups but who had not received focused instruction (Thomson & Derwing, 2015).

The researcher took on the role of teacher for the intervention, to ensure consistency in the execution of the activities and in the provision of feedback. In addition, the same two teachers who participated in the pilot study were present in class and occasionally helped out with organizational aspects. After completing a mock session during the first week, each of the following five sessions started by watching a video with enhanced and unenhanced target words for groups A and B, respectively (Table 3.38). Then, students in both groups worked on an AV activity in pairs, while the teacher-researcher circulated around the classroom to provide technical support. Following the activity, two or three pairs reported to the whole class while other students offered feedback on their answers until the correct answer was provided. Subsequently, each student individually answered ten comprehension questions and completed an awareness-raising activity. The class concluded with the teacher prompting a few students to share their responses to the final activity with the whole class and providing feedback.

Table 3.38. Session schedule.

Time	**Event**	**Participation modality**
0-5 min	Setting up computers/headphones	Individual
5-10	Watching video clip	Individual
10-25	Audiovisual activity	Groups of two or three students
25-35	Feedback on AV activity	Whole class
35-40	Comprehension questionnaire	Individual
40-45	Language awareness activity	Individual
45-50	Feedback on language awareness activity	Whole class

Testing

The intervention groups underwent testing for a total of three weeks, with each participant being tested for 30 minutes to 1 hour each week, whereas the control group underwent testing for two weeks excluding the delayed post-test. In total, teaching and testing took around 40 hours over the course of four months. The school administration designated a quiet room for testing, and to adhere to COVID-19 prevention measures, the researchers were responsible for picking up the students at the start of each English class. In the testing room, eight desks were arranged at a minimum distance of two meters from each other. One computer was set up on each desk, and four Marantz PMD-661 recorders with the corresponding mics and stands were placed near the computers at the four corners of the room. The read-aloud and delayed sentence repetition tasks, which required specific computers running DMDX, were recorded using the Marantz recorders on the assigned computers. The narrative tasks were recorded using Tascam DR-05X recorders, which also afforded professional recording quality. The auditory perception tests and written tests had a standard duration, and participants were given around 20 minutes to complete the questionnaire.

To optimize the allocated testing time, the author and another graduate student from the University of Barcelona concurrently managed eight students. Upon entering the testing room, each student took their designated seat and received task instructions from a researcher. The distribution of tasks across each testing session is presented in Table 3.39. To ensure that only four students would be speaking during each half hour, the two halves of the task sequences were counterbalanced within each group of test-takers. Consequently, during approximately half of the session, four students engaged

in speech production tasks, while the remaining four students worked on tasks that did not necessitate speech production. Subsequently, the students swapped desks and tasks.

Table 3.39. Testing sessions and tasks.

Task type	Pre-test week October 2021	Post-test week December 2021	Delayed post-test week January 2022
Production	Read-aloud task	Read-aloud task	Read-aloud task
Production	Delayed sentence repetition	Delayed sentence repetition	Delayed sentence repetition
Production	Prompted narrative task	Prompted narrative task	Prompted narrative task
Perception	OPT	Llama_E	
Written test	X_Lex	Questionnaire	
Written test	Word meaning recall		

Analysis

The pronunciation data was analyzed by the author (and partially by an interrater, as reported below) by assigning a score of 1 to each accurate pronunciation of an <-ed> ending, and a score of 0 to incorrectly pronounced endings. As in study 2, accurate pronunciation fundamentally involved distinguishing the /d/ - /t/ allomorphs from the /id/ allomorph and vice versa, pronouncing the ending when necessary and omitting it when not necessary. In addition, for the delayed sentence repetition task only, the content and linguistic form of the repeated sentences was analyzed following Ortega et al.'s (2002) rubric. The maximum score of 4 was assigned if the original stimulus was repeated with the exact same words in the same order, regardless of slight hesitations and self-repairs. Grammatical and ungrammatical changes, including changes in tense (present/past), led to a score of 3 if the exact meaning was preserved

(e.g., “Two buses stop out of the hotel”). Meaning changes led to a score of 2; this included changes in number (singular/plural), except in those cases where verb agreement was respected. For example, if the subject in the original stimulus was plural and the participant omitted the plural <-s> marking but they also omitted the third person singular <-s> when saying the verb, as if the subject was plural, a score of 3 was assigned (e.g., “The teacher support all students equally”). If only two or three meaningful words of the stimulus were repeated, such as the subject and the verb, a score of 1 was assigned (e.g., “Server dogs provide”). 0 was assigned if the participant repeated only one or two disconnected words, made no effort, or produced an unintelligible sentence. The author listened to each of the 9973 items across the three tests and scored them for <-ed> pronunciation accuracy and, for the DSR task only, general linguistic form. Then, 5% of the data was randomly selected and scored by an expert rater (a fellow applied linguistics graduate student) to obtain interrater reliability. For both the DSR and WR task, an agreement of 94% was found between the scores of the two raters, and for the prompted narrative task, the percentage of agreement was 92%.

The analysis of pronunciation data initially involved the preliminary analysis of word meaning recall data and the descriptive analysis of pronunciation data about the proportion of base form items correctly identified and pronounced. Then, several logistic mixed models were built on the subset of items in the past tense, using the *glmer* function of the *lme4* package and calculating pairwise comparisons and effect sizes using the same approach as in study 1. Three models were run using the control group as baseline, to test the effects of the intervention on the participants’ <-ed> ending pronunciation accuracy in the three pronunciation tests. The models included

Group, *Time* (pre- vs post-test) and the interaction of *Group* * *Time* as fixed effects, and the intercepts of participant and item as random effects. In addition, *Familiarity* was also included as a fixed effect, to control whether possible pronunciation gains generalized to items not encountered in the intervention. Then, one model per pronunciation test was run including the accuracy data collected at pre-, post- and delayed post-test for the intervention groups A and B only (exposed to video with enhanced captions and unenhanced captions, respectively), as the control group had not participated in the delayed post-test. Finally, Spearman correlations were run between the results of the X_Lex, OPT and Llama_E test, and accuracy gains, to highlight any possible relationships between the participants' vocabulary size (a proxy for proficiency), listening skills and aptitude, and their pronunciation gains.

The questionnaire data was analyzed quantitatively, as Yes/No questions were transformed into binary variables (0/1), while the five-point Likert items measuring learner perceptions were transformed into categorical variables with five levels ranging from 1, "totally disagree", to 5, "totally agree". The assessment of responses regarding the past tense <-ed> rule also resulted in a categorical variable with four levels, ranging from 0, indicating no response, to 3, representing a completely correct response. Responses that were partially correct, which were assigned a value of 1, included some relevant elements but missed other important ones, e.g., "it is pronounced like a t" (A_05). Answers were considered mostly correct and assigned a value of 2 if they mentioned the existence of three allomorphs and/or the presence or absence of a vowel sound depending on the context, e.g., "in walked <e> makes no sound, in provided it sounds like /ed/ because the word ends in <e> (sic), other times it sounds like /t/ and others it makes no sound" (B_46). A response was considered

completely correct if it mentioned the three allomorphs and correctly identified some examples of each: “There are verbs that in the past are pronounced as if they ended with /t/ (for example walked), others with /id/ (waited) or with /d/ (turned)” (A_02). The rating of the responses was conducted by the author. Due to the small sample size, the questionnaire data is either presented as count data or in terms of the differences identified between the groups by Fisher’s exact tests with Monte Carlo method.

3.3.6. Results

Video comprehension

Although not all participants’ responses to the comprehension questionnaire were correctly submitted, the data obtained could be considered representative of the entire sample (60%, 83%, 80%, 86%, 83% of responses recorded in session 1 to 5, respectively). Participants’ responses to the comprehension questions were 90.45% correct on average (*SD* = 13.64), indicating that overall participants understood the clips and primarily focused on meaning during the viewing.

Pronunciation accuracy: preliminary analysis

The analysis of the meaning recall test results showed that the participants knew the meaning of most of the target words at pre-test (*m* = 85.15, *SD* = 22.44, 95% CIs [79.41, 90.90]). The descriptive data for *base form* pronunciation accuracy is reported in Table 3.40 by time and group, except for the narrative task, which did not include verbs in the past tense. The three groups achieved ceiling scores in the reading task, but not in the delayed sentence repetition task, possibly due to the higher intrinsic difficulty of the sentence repetition task and the crucial role of auditory perception

and working memory for the successful repetition of longer and more complex stimuli.

Table 3.40. Descriptive accuracy[a] data for base form verbs, aggregated by group and testing time.

					95% Confidence Intervals	
	Time	*N*	*M*	*SD*	Lower	Upper
Word reading task						
Intervention group A	1	180	0.98	0.15	0.96	1.00
(Enhanced video)	2	180	0.99	0.07	0.98	1.01
	3	180	1.00	0.00	1.00	1.00
Intervention group B	1	170	0.99	0.08	0.98	1.01
(Unenhanced video)	2	170	0.98	0.13	0.96	1.00
	3	170	1.00	0.00	1.00	1.00
Control group C	1	180	1.00	0.00	1.00	1.00
	2	180	0.99	0.07	0.98	1.01
Sentence repetition task						
Intervention group A	1	306	0.74	0.44	0.69	0.79
(Enhanced video)	2	306	0.86	0.35	0.82	0.90
	3	306	0.84	0.37	0.79	0.88
Intervention group B	1	289	0.82	0.39	0.77	0.86
(Unenhanced video)	2	289	0.85	0.36	0.81	0.89
	3	289	0.88	0.33	0.84	0.91
Control group C	1	306	0.75	0.44	0.70	0.79
	2	306	0.81	0.40	0.76	0.85

[a] Proportion of accurate responses expressed as decimals, where 1 represents 100% accuracy.

Pronunciation accuracy: analysis of past tense production

From the inspection of the accuracy data on the pronunciation of *past tense* verbs (average accuracy achieved by all the participants), no ceiling effect emerged for any of the tasks (Figure 3.19). Descriptively, the participants seemed more accurate in the word reading task, which involved more controlled production, and progressively less accurate in the sentence repetition and narrative tasks, which required greater automaticity and speed of access. It could be argued that the progressively higher scores obtained by all groups in the sentence repetition tasks, which peaked after the

end of the intervention, were related to the participants' increasing familiarity with the task and with the sentence stimuli, and the consequent reduction in working memory load (Table 3.41). However, the sentence scores remained constant, with the exception that more sentences were repeated exactly as the original (score 4/4) at post-test and delayed post-test (Figure 3.20). Combined with the low percentage of low scores (0s and 1s), this finding suggests that the participants did not simply learn by heart a number of sentences that they could not repeat at pre-test. On the contrary, at post-test they were able to repeat the sentences correctly because the conceptually correct responses produced at pre-test, which contained linguistic error related to several aspects including the pronunciation of past <-ed>, were replaced with error-free responses.

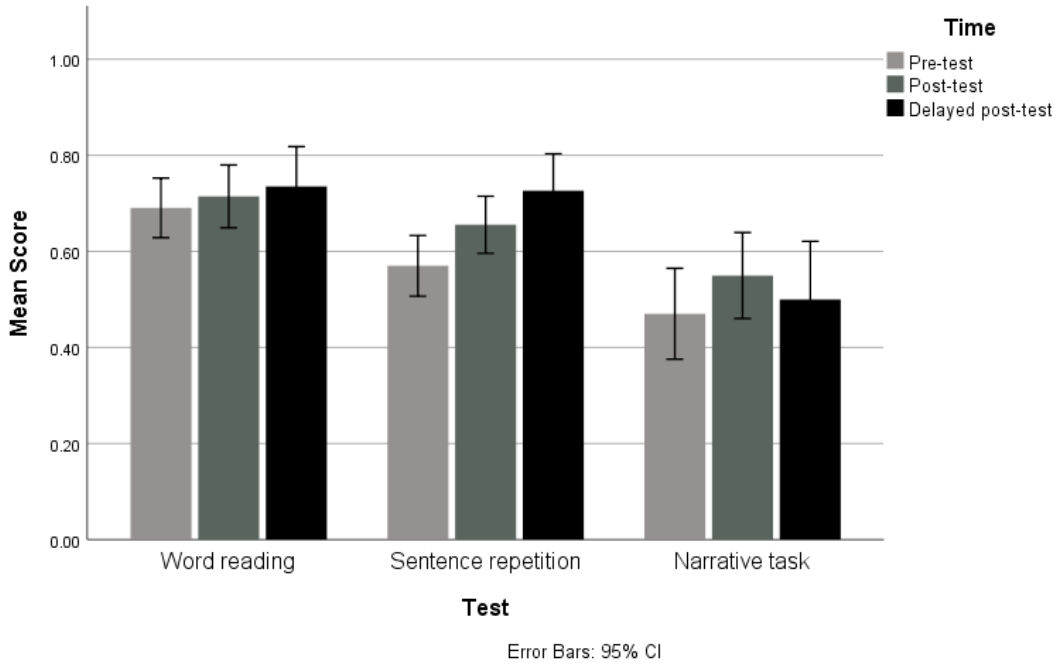


Figure 3.19. Mean accuracy in the pronunciation of past <-ed> endings by task and testing time.

Table 3.41. Descriptive data for verbs in past tense form, aggregated by group and testing time.

					95% Confidence Intervals	
	Time	*N*	*M*	*SD*	Lower	Upper
Word reading task						
Intervention group A	1	270	0.67	0.47	0.62	0.73
(Enhanced video)	2	270	0.74	0.44	0.68	0.79
	3	270	0.76	0.43	0.70	0.81
Intervention group B	1	255	0.69	0.46	0.63	0.75
(Unenhanced video)	2	255	0.71	0.45	0.66	0.77
	3	255	0.71	0.45	0.66	0.77
Control group C	1	270	0.71	0.46	0.65	0.76
	2	270	0.69	0.46	0.64	0.75
Sentence repetition task						
Intervention group A	1	342	0.59	0.49	0.54	0.64
(Enhanced video)	2	342	0.64	0.48	0.59	0.69
	3	342	0.70	0.46	0.66	0.75
Intervention group B	1	323	0.57	0.50	0.51	0.62
(Unenhanced video)	2	323	0.72	0.45	0.67	0.77
	3	323	0.75	0.43	0.70	0.80
Control group C	1	342	0.56	0.50	0.50	0.61
	2	342	0.61	0.49	0.56	0.67
Narrative task						
Intervention group A	1	196	0.53	0.50	0.46	0.60
(Enhanced video)	2	193	0.62	0.49	0.55	0.69
	3	200	0.53	0.50	0.46	0.59
Intervention group B	1	152	0.38	0.49	0.30	0.46
(Unenhanced video)	2	154	0.58	0.49	0.51	0.66
	3	144	0.53	0.50	0.45	0.61
Control group C	1	149	0.50	0.50	0.42	0.58
	2	184	0.56	0.50	0.49	0.63

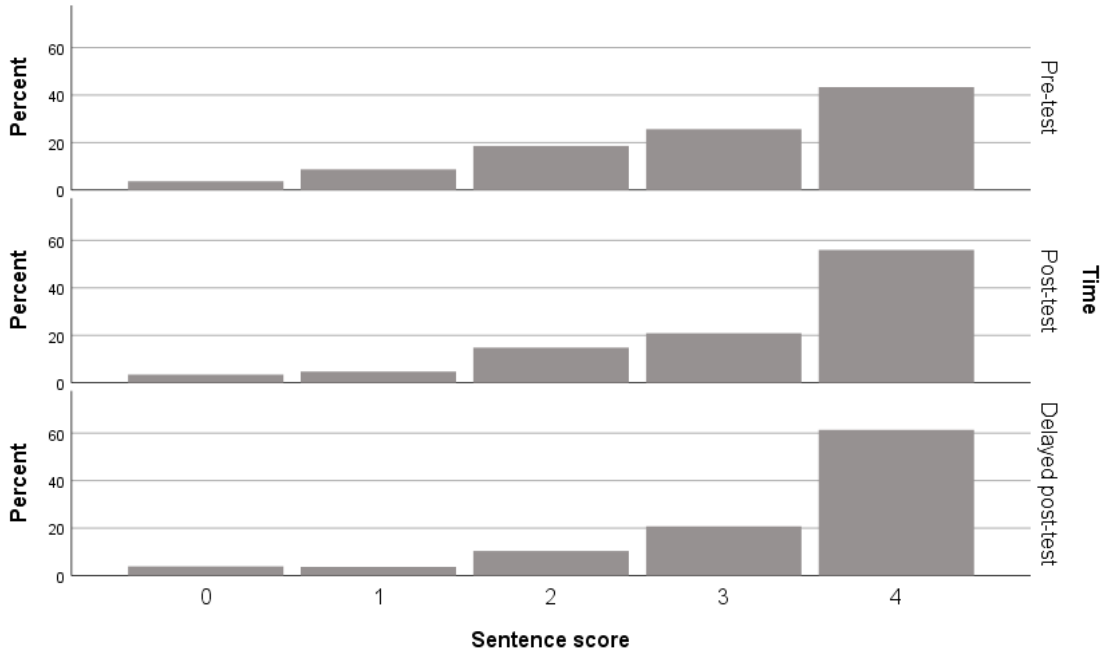


Figure 3.20. Content and linguistic form scores achieved in the sentence repetition task by testing time.

We first report the results of the inferential analysis of the pre- and post-test data obtained from all participants in the three groups. For clarity, we remind the reader that these models included the variable *Group*, with three levels (Intervention group A - enhanced video; intervention group B - unenhanced video; control group), *Time* (two levels), the interaction of *Group* and *Time*, and *Familiarity*, which referred to whether each word constituting a test stimulus occurred in the script of the video clips and therefore in the learning activities. Each level within *Group* was entered in the model as a separate variable and accordingly reported in the text, because *Group* is a nominal variable, i.e., its levels represent discrete and non-ordered categories that refer to arbitrarily assigned categories in reality (contrary to height, for example, as heights can be ordered from smaller to bigger on a continuous scale). Therefore, entering only one variable for *Group* would have resulted in the model calculating the probability that pronunciation accuracy increased linearly as a function of group,

testing the specific hypothesis that group 1 is associated with a smaller coefficient than group 2, which in turn is associated with a smaller coefficient than group 3 (or vice versa).

The analysis of word reading scores was inconclusive, as the amount of variance explained by the fixed effects was very low (*R2m* = 0.004) and none of the variables included in the model was found to have an effect on the participants' accuracy (Table 3.42 and 3.43). On the other hand, the model built on the sentence repetition data found a significant effect of the interaction of *Group B* * *Time* based on asymptotic confidence intervals (Table 3.44), and the pairwise comparisons indicated that the only group that significantly improved at post-test was intervention group B (Table 3.45). However, when bootstrapped confidence intervals were considered, *Group A*, *Time* and *Familiarity* also had a significant effect besides the interaction of *Group B* * *Time*. Unfortunately, the low marginal R-squared value combined with the discrepancies between asymptotic and bootstrapped confidence intervals weaken the robustness of these findings. The model built on the narrative task data showed that, once again, the interaction of *Group B* and *Time* had a significant effect (Table 3.46). However, the bootstrapped CIs contained zero for this interaction, and did not contain zero for the effect of *Group B*. Despite the very low score of group B at pre-test compared to group A and C, pairwise comparisons did not find any significant differences at pre- or post-test between the groups, but only a significant improvement from pre- to post-test for group B (Table 3.47). The variance explained by the fixed effects of the model, despite being higher than in the previous models, remained quite low (*R2m* = 0.04).

Table 3.42. Fixed coefficients for the logistic regression examining pronunciation accuracy in the word reading task at pre- and post-test.

Word reading task							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	SE	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	1.31	0.49	2.67	.01**	0.00	0.47	0.35	2.27	0.85	1.57
Group A	-0.18	0.54	-0.34	.74			-1.25	0.88	-0.64	0.28
Group B	-0.02	0.55	-0.04	.97			-1.11	1.06	-0.53	0.48
Time	-0.09	0.21	-0.44	.66			-0.51	0.33	-0.56	0.37
Familiarity	0.09	0.44	0.22	.83			-0.76	0.95	-0.27	0.44
Group A: Time	0.52	0.31	1.68	.09			-0.09	1.12	-0.15	1.17
Group B: Time	0.26	0.31	0.82	.41			-0.36	0.87	-0.42	0.90

** $p < .01$

Table 3.43. Results of pairwise contrasts involving pre- and post-test accuracy scores in the word reading task.

Group	Testing time	Contrast estimate	SE	*df*	*z*	*p*	95% Confidence Intervals	
							Lower	Upper
Group A	1 - 2	-0.42	0.22	Inf	-1.92	.83	-1.07	0.23
Group B	1 - 2	-0.16	0.23	Inf	-0.71	1.00	-0.84	0.51
Group C	1 - 2	0.09	0.21	Inf	0.44	1.00	-0.53	0.72
A-B	1	-0.16	0.55	Inf	-0.29	1.00	-1.78	1.47
C-A	1	0.18	0.54	Inf	0.34	1.00	-1.41	1.78
C-B	1	0.02	0.55	Inf	0.04	1.00	-1.60	1.65
A-B	2	0.10	0.56	Inf	0.18	1.00	-1.54	1.74
C-A	2	-0.33	0.55	Inf	-0.61	1.00	-1.94	1.27
C-B	2	-0.23	0.55	Inf	-0.42	1.00	-1.86	1.39

Table 3.44. Fixed coefficients for the logistic regression examining pronunciation accuracy in the delayed sentence repetition task at pre- and post-test.

Sentence repetition task							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	SE	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	0.43	0.45	0.94	.35	0.02	0.43	-0.46	1.32	-0.91	-0.19
Group A	0.20	0.42	0.46	.64			-0.63	1.02	0.11	1.03
Group B	0.09	0.43	0.21	.83			-0.75	0.93	-0.30	0.71
Time	0.34	0.18	1.84	.07			-0.02	0.69	0.30	1.23
Familiarity	-0.19	0.48	-0.41	.68			-1.13	0.74	-0.85	-0.14
Group A: Time	-0.04	0.26	-0.16	.87			-0.55	0.47	-1.27	0.05
Group B: Time	0.62	0.27	2.29	.02*			0.09	1.15	0.30	1.63

* $p < .05$

Table 3.45. Results of pairwise contrasts involving pre- and post-test accuracy scores in the delayed sentence repetition task.

							95% Confidence Intervals	
	Testing time	Contrast estimate	SE	*df*	*z*	*p*	Lower	Upper
Group A	1 - 2	-0.29	0.19	Inf	-1.59	1.00	-0.84	0.25
Group B	1 - 2	-0.96	0.20	Inf	-4.78	< .001***	-1.54	-0.37
Group C	1 - 2	-0.34	0.18	Inf	-1.84	.98	-0.87	0.20
A-B	1	0.10	0.43	Inf	0.24	1.00	-1.16	1.37
C-A	1	-0.20	0.42	Inf	-0.46	1.00	-1.44	1.05
C-B	1	-0.09	0.43	Inf	-0.21	1.00	-1.35	1.17
A-B	2	-0.56	0.44	Inf	-1.28	1.00	-1.84	0.72
C-A	2	-0.15	0.43	Inf	-0.36	1.00	-1.40	1.09
C-B	2	-0.71	0.44	Inf	-1.63	1.00	-1.99	0.57

*** $p < .001$

Table 3.46. Fixed coefficients for the logistic regression examining pronunciation accuracy in the prompted narrative task at pre- and post-test.

Narrative task							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	SE	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	0.08	0.46	0.18	.86	0.04	0.50	-0.83	0.99	-0.34	0.58
Group A	-0.10	0.63	-0.15	.88			-1.33	1.14	-0.72	0.47
Group B	-1.08	0.68	-1.60	.11			-2.41	0.25	-1.61	-0.29
Time	0.26	0.26	0.99	.32			-0.25	0.77	-0.36	0.80
Group A: Time	0.33	0.36	0.92	.36			-0.37	1.02	-0.49	1.16
Group B: Time	0.82	0.41	1.99	.05*			0.01	1.64	-0.20	1.63

* $p < .05$

Table 3.47. Results of pairwise contrasts involving pre- and post-test accuracy scores in the prompted narrative task.

	Testing time	Contrast estimate	SE	*df*	*z*	*p*	95% Confidence Intervals	
							Lower	Upper
Group A	1 - 2	-0.58	0.24	Inf	-2.39	.25	-1.30	0.13
Group B	1 - 2	-1.08	0.32	Inf	-3.34	.01**	-2.03	-0.13
Group C	1 - 2	-0.26	0.26	Inf	-0.99	1.00	-1.02	0.51
A-B	1	0.99	0.66	Inf	1.49	1.00	-0.96	2.94
C-A	1	0.10	0.63	Inf	0.15	1.00	-1.76	1.95
C-B	1	1.08	0.68	Inf	1.60	1.00	-0.91	3.08
A-B	2	0.49	0.66	Inf	0.74	1.00	-1.45	2.43
C-A	2	-0.23	0.63	Inf	-0.37	1.00	-2.07	1.61
C-B	2	0.26	0.67	Inf	0.39	1.00	-1.70	2.22

** $p < .01$

The analysis of pre-, post- and delayed post-test data showed interesting patterns in how durable the gains were for the intervention groups. These models included the variable *Group*, with two levels (Intervention group A and intervention group B only), *Time* (three levels), the interaction of *Group* and *Time*, and word *Familiarity*. Each level within *Time* was entered in the model as a separate variable because, although time is a continuous and ordered variable, we could not assume a monotonic pattern between *Time* (predictor) and *Accuracy* (dependent variable). In other words, we did not expect pronunciation accuracy to improve linearly from pre- to delayed post-test, but rather to improve at post-test and plateau or potentially decline at delayed post-test, due to a decrease in focused practice after the end of the intervention. In this case, a logistic regression run with one variable (*Time*) representing all three times may have failed to detect a potential relationship between time and accuracy which was, in fact, significant but curvilinear.

Although both *Post-test* and *Delayed post-test* had a significant effect in the word reading model based on asymptotic CIs, the effect of the *Post-test* was not confirmed through bootstrapping and the *R2m* value was very low (Table 3.48). No significant differences were found between each group's performance at any testing time through pairwise comparisons (Table 3.49).

Table 3.48. Fixed coefficients for the logistic regression examining pronunciation accuracy in the word reading task at pre-, post- and delayed post-test.

Word reading task							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	SE	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	1.13	0.53	2.11	.03*	0.01	0.51	0.08	2.17	0.65	1.41
Group B	0.17	0.58	0.29	.77			-0.97	1.31	-0.34	0.64
Time 2	0.44	0.22	1.94	.05*			0.00	0.88	-0.06	0.90
Time 3	0.57	0.23	2.53	.01**			0.13	1.02	0.05	1.04
Familiarity	0.17	0.49	0.34	.74			-0.80	1.13	-0.19	0.53
Group B: Time 2	-0.26	0.33	-0.80	.42			-0.90	0.38	-0.87	0.39
Group B: Time 3	-0.40	0.33	-1.22	.22			-1.04	0.24	-1.09	0.28

* $p < .05$, ** $p < .01$

Table 3.49. Results of pairwise contrasts between pre-, post- and delayed post-test accuracy scores in the word reading task.

		Contrast					95% CIs	
	Time	estimate	SE	*df*	*z*	*p*	Lower	Upper
Group A	1 - 2	-0.44	0.22	Inf	-1.94	.78	-1.10	0.22
Group A	1 - 3	-0.57	0.23	Inf	-2.53	.17	-1.24	0.09
Group A	2 - 3	-0.14	0.23	Inf	-0.59	1.00	-0.82	0.54
Group B	1 - 2	-0.17	0.24	Inf	-0.73	1.00	-0.87	0.52
Group B	1 - 3	-0.17	0.24	Inf	-0.73	1.00	-0.87	0.52
Group B	2 - 3	0.00	0.24	Inf	0.00	1.00	-0.70	0.70

The model built on the sentence repetition task indicated that the *Delayed post-test* had a significant effect, which was confirmed through bootstrapping (Table 3.50). However, the significant effect found for the interaction between *Group B* and the *Post-test* based on asymptotic CIs was not confirmed through bootstrapping. The pairwise comparisons showed that participants' performance improved from pre- to

post-test for group B, and from pre-test to delayed post-test for group A and B (Table 3.51).

Table 3.50. Fixed coefficients for the logistic regression examining pronunciation accuracy in the delayed sentence repetition task at pre-, post- and delayed post-test.

Sentence repetition task							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	SE	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	0.56	0.47	1.18	.24	0.03	0.47	-0.37	1.49	1.70	2.32
Group B	-0.11	0.49	-0.23	.82			-1.06	0.84	-0.60	0.16
Time 2	0.30	0.19	1.61	.11			-0.06	0.67	-0.76	0.00
Time 3	0.74	0.19	3.82	<.001***			0.36	1.11	-1.34	-0.57
Familiarity	-0.05	0.46	-0.11	.91			-0.96	0.86	0.00	0.50
Group B: Time 2	0.68	0.28	2.46	.01**			0.14	1.22	-0.24	0.88
Group B: Time 3	0.47	0.28	1.67	.09			-0.08	1.02	-0.54	0.58

** $p < .01$, *** $p < .001$

Table 3.51. Results of pairwise contrasts between pre-, post- and delayed post-test accuracy scores in the delayed sentence repetition task.

							95% Confidence Intervals	
	Testing time	Contrast estimate	SE	*df*	*z*	*p*	Lower	Upper
Group A	1 - 2	-0.30	0.19	Inf	-1.61	1.00	-0.85	0.25
Group A	1 - 3	-0.74	0.19	Inf	-3.82	.002**	-1.30	-0.17
Group A	2 - 3	-0.43	0.19	Inf	-2.24	.37	-1.00	0.13
Group B	1 - 2	-0.98	0.20	Inf	-4.85	<.001***	-1.58	-0.39
Group B	1 - 3	-1.21	0.21	Inf	-5.83	<.001***	-1.82	-0.60
Group B	2 - 3	-0.23	0.21	Inf	-1.07	1.00	-0.85	0.39

** p <.01 *** $p < .001$

The analysis of narrative task data provided further support to the finding that, for group B, the pronunciation of <-ed> endings improved after the intervention in spontaneous production. The model found a significant effect of *Time* based on both asymptotic and bootstrapped CIs, and of *Group B* and *Time 3*, although the latter was not confirmed through bootstrapping (Table 3.52). The results of pairwise comparisons showed that only group B improved significantly between pre-test and

post-test and maintained these gains at delayed post-test (Table 3.53). However, once again, the bootstrapped CIs did not contain zero for the effect of Group B, indicating that the performance of this group may have been different from that of group A regardless of the intervention.

Table 3.52. Fixed coefficients for the logistic regression examining pronunciation accuracy in the prompted narrative task at pre-, post- and delayed post-test.

Narrative task							95% Confidence Intervals			
							Asymptotic		Bootstrapped	
	B	SE	*z*	*p*	*R2m*	*R2c*	Lower	Upper	Lower	Upper
Intercept	0.02	0.48	0.05	.96	0.03	0.55	-0.92	0.97	-0.40	0.39
Group B	-0.93	0.71	-1.30	.19			-2.32	0.47	-1.49	-0.10
Time 2	0.58	0.25	2.37	.02*			0.10	1.07	0.04	1.08
Time 3	0.06	0.24	0.25	.80			-0.41	0.54	-0.39	0.60
Group B: Time 2	0.49	0.41	1.21	.23			-0.31	1.30	-0.51	1.31
Group B: Time 3	0.94	0.41	2.29	.02*			0.14	1.75	-0.11	1.67

* $p < .05$

Table 3.53. Results of pairwise contrasts between pre-, post- and delayed post-test accuracy scores in the prompted narrative task.

							95% Confidence Intervals	
	Testing time	Contrast estimate	SE	*df*	*z*	*p*	Lower	Upper
Group A	1 - 2	-0.58	0.25	Inf	-2.37	.27	-1.31	0.14
Group A	1 - 3	-0.06	0.24	Inf	-0.26	1.00	-0.77	0.65
Group A	2 - 3	0.52	0.25	Inf	2.13	.49	-0.20	1.24
Group B	1 - 2	-1.08	0.33	Inf	-3.29	.02*	-2.04	-0.12
Group B	1 - 3	-1.01	0.33	Inf	-3.01	.04*	-1.98	-0.03
Group B	2 - 3	0.07	0.32	Inf	0.23	1.00	-0.88	1.02

* $p < .05$

The final phase of the analysis aimed at ensuring that the results were not biased by individual participants achieving very large gains from pre- to post-test (see Figure 3.21 for an overview and Appendix E for the specific gains achieved by each participant). The most extreme values were observed in participants' performance in the narrative task, possibly due to the high variability in the number of target items

produced by each participant, as well as the variability intrinsic in spontaneous speech (Kormos, 2006). Only two participants in group B exhibited exceptional positive gains, and the largest negative gains experienced by some participants were nevertheless more modest and distributed across all groups, so it is unlikely that these extreme values had a substantial impact on the aggregated results.

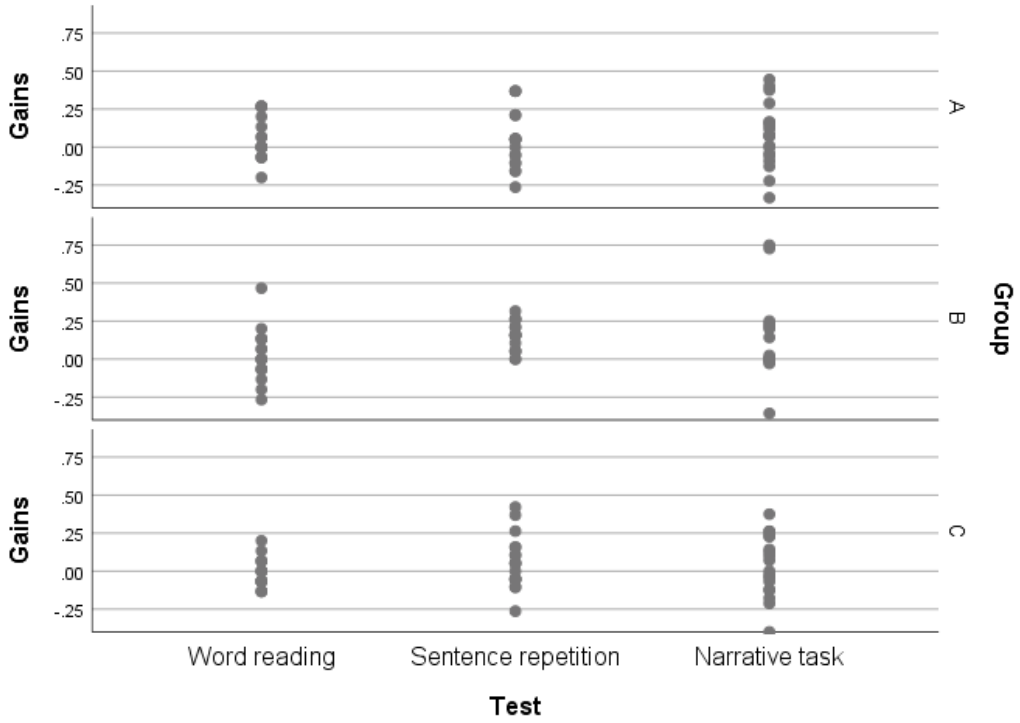


Figure 3.21. Past <-ed> pronunciation gains from pre- to post-test by group and task.

Summary of pronunciation accuracy results

The analysis of the pronunciation test results revealed a complex picture regarding the gains that may have occurred as a result of the intervention. Therefore, this summary will highlight the most important results that emerged from the analysis of pre- and post-test data for the three groups and were confirmed in the subsequent analysis of pre-, post- and delayed post-test data for intervention groups A and B. The models built on the word reading data failed to confirm or contradict the hypothesis that the intervention would make participants achieve significant gains in past <-ed> pronunciation accuracy. The analysis of the delayed sentence repetition task found evidence that only group B had improved significantly at post-test and maintained these gains at delayed post-test. However, group A had also improved at delayed post-test compared to pre-test. The analysis of the narrative task also showed that only group B improved significantly between pre-test and post-test and maintained these gains at delayed post-test. Familiarity did not generally have a significant effect, suggesting that any gains obtained were generalizable to items never encountered during the intervention. As is often the case with social science research modelling, the effect sizes were quite small (Ozili, 2022). The smaller effect size observed in comparison to study 1 could be attributed to the greater number of variables influencing language development and assessment in the classroom, compared to a more controlled environment like a research laboratory.

Effects of proficiency, listening comprehension skills and aptitude

The analysis of pre- to post-test gains found no significant correlations between accuracy gains in any of the tasks and the X_Lex, OPT or Llama E scores. The only correlation that approached significance was a negative correlation between the

participants' accuracy gains in the narrative task and their X_Lex scores ($r_s(53) = -.24$, $p = .08$). The analysis of pre- to *delayed* post-test gains also found no significant correlations. Therefore, it appears that the accuracy gains observed as a result of the intervention were largely independent from individual factors such as the participants' proficiency and auditory skills.

Questionnaire responses

To begin with, we investigated whether the learners noticed the target L2 pronunciation feature and were able to describe the underlying rule. Among the participants who watched the videos with audio-synchronized textual enhancement ($n = 18$), all of them reported noticing the enhanced words in the captions, and 16 found the enhancement to be useful. When asked about the reason behind the enhancement, 14 participants correctly identified the target words to be regular past verbs or verbs ending in <-ed>. Only one participant also specified that the words had been enhanced because of the pronunciation of past <-ed>, and three mentioned that the enhancement was related to pronunciation without providing further details. A considerable percentage of the participants did not attempt to describe the rule for pronouncing regular past <-ed> endings (50%, 12%, and 22% of group A, B, and C, respectively). Among the responses given by the participants in each group, many were incorrect (17%, 71%, and 56%). Out of the twelve acceptable answers, two in group A and one in group B were considered mostly correct (11% and 6%, respectively), and one answer in group A and one in group B were deemed completely correct. Statistical analysis using Fisher's exact tests with the Monte Carlo method did not reveal any significant differences in the responses among the three groups (two-tailed $p = .66$).

Turning to the analysis of participants' perceptions of the intervention, the questionnaire responses offered valuable insights on the effectiveness and appropriateness of the intervention. All participants reported understanding the videos, with approximately 80% in each group finding them enjoyable (Table 3.54). Two-thirds of the participants in group A, who saw the video with enhanced captions, reported reading the captions most of the time, but only half of group B reported doing so. Around 70% of the participants in both intervention groups reported learning some English pronunciation from the video clips. While 65% felt that they had also learned some grammar or vocabulary from the videos in group B, only 50% of group A agreed with this statement. Regarding the audiovisual activities, the majority of the participants understood the instructions, and approximately two-thirds of them used the clues provided to complete the activities (Table 3.55). Eighty percent of the participants in group A and sixty percent of the participants in group B found the activities to be fun. While only one third of the participants in group A indicated that the activities were challenging, nearly two-thirds of the participants in group B found them challenging. Ninety percent of participants in group A and seventy percent in group B indicated that both partners had contributed equally to the completion of the activities. Nearly 70% of the participants in each group reported learning some pronunciation from the activities, but only half reported learning some grammar and vocabulary.

To sum up, the analysis of questionnaire data showed that, while participants in group A reported noticing the enhanced target words, this did not translate into a statistical advantage when comparing the quality of their past <-ed> rule description. Very few participants in each group attempted to answer or provided a correct answer,

suggesting that, overall, awareness of the elements influencing past <-ed> pronunciation was quite low. Although most participants found the intervention materials enjoyable and at the appropriate level of difficulty, participants' perceptions regarding the language learning effectiveness of the video clips and activities were less enthusiastically positive. Finally, some differences emerged between group A and B, most notably regarding their reported reliance on captions while viewing the videos and regarding their experience with the audiovisual activities. To begin with, participants in group A, who watched the clips with enhanced captions, may have read the captions more consistently than participants in group B. Moreover, a larger proportion of participants in group A reported that they successfully collaborated with their peers and thought the activities were fun, suggesting that they not only found the content and assignments proposed enjoyable, but also managed to effectively perform them as intended. On the other hand, the majority of participants in group B reported finding the activities challenging, suggesting that they may have invested more effort into their completion. Notably, despite the perceived challenges and a lower degree of enjoyment compared to group A, most group B participants still reported equally contributing to the successful execution of the audiovisual activities and learning some English pronunciation.

Table 3.54. Responses to statements about the enhanced videos, ranging from 1 (totally disagree) to 5 (totally agree).

	Intervention group A (Enhancement)			Intervention group B (No enhancement)		
	M	SD	95% CI	M	SD	95% CI
I understood the videos	4.83	0.38	[4.64; 5.02]	4.53	0.62	[4.21; 4.85]
The videos were fun	4.22	1.17	[3.64; 4.80]	4.12	0.93	[3.64; 4.59]
I read the captions	3.72	1.02	[3.22; 4.23]	3.53	1.18	[2.92; 4.14]
I learned some English pronunciation from the videos	3.72	0.57	[3.44; 4.01]	3.82	1.07	[3.27; 4.38]
I learned some English grammar or vocabulary from the videos	3.50	0.71	[3.15; 3.85]	3.65	0.99	[3.13; 4.16]

Table 3.55. Responses (1-5) to statements about the audiovisual activities.

	Intervention group A (Enhancement)			Intervention group B (No enhancement)		
	M	SD	95% CI	M	SD	95% CI
I understood the instructions	4.39	0.78	[4.00; 4.78]	4.29	0.85	[3.86; 4.73]
We used the clues to do the activities	3.72	0.89	[3.28; 4.17]	3.76	0.90	[3.30; 4.23]
The activities were fun	3.89	1.23	[3.28; 4.50]	3.53	1.01	[3.01; 4.05]
The activities were challenging	3.00	1.19	[2.41; 3.59]	3.41	1.12	[2.84; 3.99]
My partner and I contributed equally to the activities	4.39	1.24	[3.77; 5.01]	4.06	1.20	[3.44; 4.67]
I learned some English pronunciation from the activities	3.89	0.68	[3.55; 4.23]	4.06	0.97	[3.56; 4.56]
I learned some English grammar or vocabulary from the activities	3.50	0.71	[3.15; 3.85]	3.65	1.11	[3.07; 4.22]

3.3.7. Discussion

In this study, three classes of EFL high school students participated in an intervention consisting of watching video clips containing audio-synchronized textual enhancement and doing activities that involved the audiovisual manipulation of the clips. Our first research question investigated the effects of the intervention on L2 pronunciation accuracy. To assess the effect of audio-synchronized enhancement, we

included a control group that did not participate in the intervention and a group that did the same activities as Intervention Group A but watched the video clips without enhancement. Our first hypothesis was only partially confirmed, as group A achieved modest (non-significant) pronunciation gains at post-test and only improved significantly at delayed post-test in the sentence repetition task. This finding was surprising, as the learners exposed to audio-synchronized enhancement of the target words were expected to achieve higher gains than the other groups. Since their previous knowledge of the target feature and their intermediate L2 proficiency level made them good candidates for the intervention (Ellis, 2016; Han et al., 2008), more prolonged exposure or more explicit instruction may be necessary to detect the effects of a semi-incidental technique like input enhancement (Leow, 2009). It is possible that since non-targetlike representations of past <-ed> verbs had become fossilized in the learners' interlanguage, more repetitions and exposure to target models would be needed to update them (Darcy & Holliday, 2019; Selinker, 1972). It is also possible that the learners exposed to audio-synchronized enhancement did not reap the benefits of multimodal exposure because they relied too much on captions for comprehension and failed to integrate the input in the auditory and written modality (Kruger et al., 2013; Kruger & Doherty, 2016). Finally, the learners' beliefs about language learning may have interfered with their progress, as Reed (2012) observed in a similar classroom intervention involving explicit feedback on the pronunciation of past <-ed> allomorphs. In that case, the learners openly resisted pronunciation instruction, claiming that they had already studied the regular past tense, even though a pronunciation survey revealed uncertainties regarding the pronunciation rule. However, to provide a fair interpretation of this study's findings, it must be pointed out that the participants were high school students with different interests and aims

compared to university students taking specialist linguistics courses. Nevertheless, their enthusiastic participation undoubtedly allowed them to gain from the intervention, even when the results did not align with our specific language learning expectations.

The group that was exposed to captioned video without enhancement, on the other hand, improved significantly at post-test in the two tasks that assessed accuracy in (semi-)spontaneous production and maintained these gains at delayed post-test. Most importantly, the pronunciation gains extended to words that had not been encountered in the intervention, showing that the participants in group B successfully applied the newly acquired knowledge to novel contexts in tasks requiring real-time processing of speech (Kruk & Pawlak, 2021; Solt et al., 2003). An interesting consideration from a methodological point of view was that different tasks elicited production at different levels of accuracy, and the results of different tests at three testing times gave us valuable insights on the participants' gains and their relative stability (Saito & Plonsky, 2019). For example, based on the word reading task alone, it would have been impossible to detect the significant gains obtained by some of the participants in spontaneous production, when the automatization of the accurate pronunciation of the target feature was required (Pellicer-Sánchez, 2015; Tavakoli, 2019). Similarly, the participants did not generally achieve full awareness of the past <-ed> pronunciation rule, indicating that their declarative knowledge of this aspect remained limited. By including tasks involving the planning, encoding and monitoring of longer chunks of speech, we obtained a reliable estimate of the participants' past <-ed> pronunciation accuracy in real-life speech production, gaining valuable insights into their proceduralization of this linguistic aspect (Zhang & Yuan, 2020).

Regarding research question two, our hypothesis was partially supported, as the lack of a significant correlation between accuracy gains and proficiency suggests that the activities proved equally beneficial for learners across a range of proficiency levels, from lower to upper intermediate. Based on previous literature, we had also hypothesized that regular past production would be facilitated by higher phonological ability (Celce-Murcia et al., 2010; Strachan & Trofimovich, 2019). In particular, phonemic coding measured through Llama_E has demonstrated a moderate association with phonological/morphological accuracy and fluency, key factors in acquiring advanced L2 oral ability (Saito, 2017). However, although the comparison of marginal and conditional effect sizes in the models pointed at a substantial effect of individual factors, no significant correlations were found between accuracy gains and the participants' auditory skills.

Finally, our third research question explored learners' perceptions of the video clips and activities used in the intervention. Participants' responses were overall positive, confirming our hypothesis, based on previous literature, that the materials were appropriate for their proficiency level and that the learners would enjoy participating in the intervention (Sánchez-Requena, 2017; Sokoli, 2018; Zhang, 2016). The questionnaire data provided interesting insights into some of the factors that may have affected our participants' classroom experience and the resulting pronunciation gains. Although largely speculative, group A and B may have adopted a different attitude towards the intervention. On one hand, group A may have perceived the activities as an opportunity to deviate from their typical English classes and may have focused on enjoying the novelty of collaborative video-based activities. On the other hand, group B may have found the activities more difficult, due, for example, to limited previous

knowledge of the linguistic target or to a more learning-focused attitude. As a result, group B's perception of the pronunciation learning effectiveness of the intervention may have been more favorable, and, in turn, the positive effects may have reflected in their test performance.

3.3.8. Conclusion and limitations

To sum up, this study has found evidence that video-based collaborative activities that involve the manipulation and integration of auditory and written input can support the acquisition of L2 pronunciation within a pronunciation-focused classroom intervention. Learners with a lower-intermediate to upper-intermediate level of proficiency and varying degrees of auditory aptitude may benefit from the implementation of audiovisual activities. However, the type of audio-synchronized textual enhancement implemented in this study may not have provided additional benefits, possibly due to an excessive focus on written input and the lack of integration with auditory information.

This study has several limitations, starting with the small number of participants included in the final sample, and the relatively short treatment. Unfortunately, the data for this study was collected during the Covid-19 pandemic and the project suffered a number of organizational setbacks. The effect sizes were small, possible due to the small number of participants per group, but also to the relatively limited exposure to the target features and to possible confounds naturally intervening in the collection of data over one semester. As a result, some of our hypotheses were not confirmed and, in the absence of interview data, some speculation was required for the interpretation of the results. The regular survey of learners' extracurricular viewing habits and learning patterns throughout the semester may have shed light on some of the factors

affecting the progress of individual learners. In addition, a separate analysis of learners' accuracy when using the past tense in writing would have provided insights into their level of acquisition of the target feature and the extent to which they consistently used it in obligatory contexts. However, when errors in the pronunciation of inflectional morphemes also occur in a read-aloud task, errors in the suppliance of past <-ed> endings in the other speaking tasks cannot be solely attributed to a lack of grammar knowledge (Reed, 2012). Finally, although the intervention proved engaging and beneficial for students, its standalone nature may have greatly limited its potential impact on learners' linguistic development. The closer integration of this form-focused intervention within the curriculum may have provided learners with further opportunities to notice the target feature in context and to practice the newly acquired knowledge while carrying out different tasks.

To address these limitations, future research should examine the effects of audio-synchronized enhancement and audiovisual activities with a larger participant sample. The number of sessions will depend on the specific educational context and target feature, but it is recommended to consider a minimum of one training session and five learning sessions. Learners' perceptions of the intervention may be surveyed after each session, to identify the most useful activities and explore the factors affecting the development of L2 pronunciation skills. Replicating this study with participants of different ages and backgrounds may open up new possibilities regarding the teaching and learning of several linguistic aspects with audiovisual activities.

The pedagogical recommendations for implementing audio-synchronized enhancement include selecting video clips that are meaningful and relevant for the learners and integrating them into the teaching curriculum. This integration may be

achieved by linking the exposure to L2 video with larger projects or series of tasks (including audiovisual activities) that progressively scaffold learners towards real-life language use. Before implementing audiovisual activities, it is necessary to conduct a comprehensive training session, during which learners are provided with a detailed explanation of what they will be doing and of the reasons behind it. Teachers should be prepared for the increased levels of noise that may arise during the implementation of production activities in larger classes. However, all students should be encouraged to speak, and the teacher should constantly check on the pairs to make sure they are carrying out the activity as intended. Providing regular whole class feedback on the activities as well as individual advice ensures that everyone in the classroom progresses at a similar pace and benefits equally from the session. Finally, some technical recommendations involve double checking in advance the specification of the students' computers. If the activities are administered through PowerPoint, it is important to be aware that not all computers have USB ports. One possible solution is to provide links to an online copy of the materials, alongside bringing USB pen drives with the necessary files and programs. The online copy can be stored in the cloud or published on a website and should be only available for visualization by the students, without the possibility to edit. To make both solutions possible, in this study all the activities which required students to write down words and phrases were designed so that they could be completed orally or on a piece of paper while visualizing the PowerPoint presentation. Another technical issue that may arise is that older computers and those running outdated operating systems may encounter difficulties in decoding the videos embedded in PowerPoint presentations. In this case, students should download codec packs such as K-Lite, which enable an OS to play previously unsupported audio and video formats. At times, it may be necessary for

two students to work on the same computer, and while this situation may require a brief wait as their partner listens to a video before taking the headset, it is generally well tolerated.

CHAPTER 4. GENERAL DISCUSSION

This doctoral dissertation addressed a research gap in the field of second language acquisition from TV series by examining the effects of an input enhancement technique aimed at increasing audiovisual synchrony on L2 pronunciation learning. In this general discussion, the aim is to provide a more comprehensive elaboration of the discussions included within each individual study, highlighting how they collectively contribute to fulfilling the objectives proposed in the introduction. Study 1 and 2 explored the pronunciation learning potential of an innovative input enhancement technique, audio-synchronized textual enhancement, by comparing learners' processing of video clips with enhanced and unenhanced captions. The research design of these two studies focused on internal validity and followed the non-conflated input enhancement research strand, which typically examines the impact of enhancement in the absence of explicit instruction (Leow, 2009). To investigate the potential benefits of audio-synchronized textual enhancement in an ecologically valid context, study 3 integrated it within a form-focused instructional approach. An intervention was conducted in a secondary school classroom, where exposure to audio-synchronized textual enhancement was combined with other video-based activities aimed at promoting a focus on L2 pronunciation. In this chapter, we present a comprehensive overview of the key findings derived from the three studies conducted and of their implications.

The main finding of study 1, conducted with first-year university students, was that the participants became significantly faster at rejecting mispronunciations after being exposed to L2 captioned video with audio-synchronized textual enhancement and unsynchronized enhancement. As higher speed in rejecting mispronounced forms has

been associated with greater automaticity in lexical access, the significantly faster rejection of mispronunciations was interpreted as evidence that the visual enhancement of target words facilitated the updating of the existing representations stored in the learners' mental lexicon (Cook et al., 2016; Pellicer-Sánchez, 2015). This finding aligns with the hypothesis that textual enhancement increases the visual salience of the target words, promoting noticing of the incoming input and the conversion of input into intake (Schmidt, 1990; Sharwood Smith, 1993). In this study, the participants showed both alertness and an orientation towards auditorily presented information that facilitated the detection of this information and its selection for further processing (Leow, 2015; Tomlin & Villa, 1994). The successful mapping of phonological information onto orthographic word forms was possible thanks to the efficient integration of auditory and written input in the viewers' working memory (Mayer, 2005). Despite the large amount of information available in the multimodal input, the focus on form promoted through textual enhancement did not seem to induce cognitive overload (Kruger & Steyn, 2014; Sweller, 2005). On the contrary, the higher automaticity observed in the lexical decision task indicated that input enhancement facilitated the comparison between the L1-biased representations generated through subvocal articulation and the target-like auditory forms spoken by the characters in the video (Stenton, 2013). However, in the absence of production data and long-term retention data, it is impossible to exclude a strong episodic memory component in test performance (Baddeley, 2000). In this case, the enhancement may have led to the unsystematic accumulation of data deriving from minimal processing, rather than to the internalization and restructuring of L2 knowledge regarding the target phonological forms (Leow, 2015; Robinson, 2003).

The second main finding of study 1 relates to the higher degree of audiovisual synchrony observed during exposure to audio-synchronized textual enhancement compared to unsynchronized enhancement. Although both synchronized and unsynchronized enhancement led to improved lexical decision performance, synchronizing the enhancement with the word's auditory onset also reduced the time lag between auditory and visual processing, guiding the viewers' gaze to the target word in synchrony with its auditory onset. The synchronized viewing modality appeared to maximize the benefits provided by captions, as the presentation of text in short segments of one or two lines, even if aligned with extended stretches of auditory input, facilitates the efficient integration of information in different modalities (Kruger et al., 2013; Kruger & Doherty, 2016). On the other hand, enhancing target words from caption appearance may have disrupted automatic reading patterns, with detrimental effects on content comprehension. Redirecting learners' attention to a word too early in relation to its auditory onset may have interrupted the natural left-to-right reading flow, causing their gaze to return from the enhanced word to the beginning of the sentence after the character had already started speaking. This attention shift may have interfered with their inclination to prioritize meaning over form during input processing (VanPatten, 1996, 2004), although due to the limited number of comprehension questions and to the absence of interview data in study 1, definitive conclusions regarding the balance between meaning and form processing cannot be drawn. Although learners may not have been able to fully verbalize and report on their real-time processing of audiovisual input, further data from stimulated recall or retrospective think-aloud protocols may have provided valuable insights into their depth of processing of meaning and linguistic form under different enhancement conditions.

In contrast to previous research findings (cf. Gerbier et al., 2018), audio-synchronized textual enhancement did not promote closer audiovisual synchrony compared to *unenhanced* captions, but only compared to unsynchronized enhancement. However, this discrepancy could be attributed to methodological differences, as Gerbier et al. (2018) highlighted each word in the text and analyzed regressive saccades rather than fixation distance, defined as the time-lag between the viewer's first fixation on a word and the word's auditory onset. Moreover, the characteristics of the participants may have influenced the results, since Gerbier et al. investigated children who were native speakers of French, whereas our studies involved 10th grade students and first-year university students who were native speakers of Spanish and Catalan. The average fixation distance of our participants was in line with that of the first-year university students in Wisniewska and Mora's study (2018), as they tended to read quite fast, often reaching the orthographic representation of words before the auditory onset. However, when a small sample of 10th grade students were exposed to the same videos in study 2, only the less proficient participants consistently read the captions, and when they encountered the target words, it was typically after the auditory onset, irrespective of the presence of audio-synchronized enhancement.

Based on these consideration, the investigation of the moderating effect of reading speed was expected to shed light on the patterns of audiovisual synchrony and attention allocation observed in study 1. In contrast with Wisniewska and Mora (2018), we found a significant effect of L2 proficiency on learners' viewing behavior. The significant correlation found between proficiency and fixation distance in the unenhanced clips indicated that more proficient learners read captions fast, generally fixating on target words in advance of their auditory onset (in line with Wisniewska

and Mora, 2018), whereas less proficient learners read the target words after auditory onset, possibly due to spending more time on each word. The analysis of enhanced target words showed that the visual enhancement mitigated the effect of proficiency on the duration of the time-lag between first fixation and auditory onset, resulting in similar audiovisual synchrony for learners with different levels of proficiency (as in Gerbier et al., 2018). Nevertheless, the higher the learner's proficiency, the shorter time they spent fixating on enhanced target words. On the contrary, less proficient learners, who naturally tend to use captions to reduce cognitive load and support bottom-up speech processing (Kruger & Steyn, 2014; Montero Perez et al., 2013; Winke et al., 2013; Yang & Chang, 2014), needed more time to process the enhanced target words. Finally, despite the fairly advanced level of English proficiency demonstrated by the university students, the pre-test indicated that phono-lexical representations based on L1 decoding of orthographic input were largely considered acceptable (Bassetti & Atkinson, 2015; Erdener & Burnham, 2005). However, more proficient learners did not exhibit greater gains in accuracy or response time, suggesting that the "rich get richer" effect commonly observed in the acquisition of novel vocabulary from exposure to L2 captioned video did not apply to the updating of pre-existing phonological representations (Gesa, 2019; Pattemore & Muñoz, 2020; Pujadas & Muñoz, 2019; Rodgers, 2013).

In study 2, the stimulated recall interviews conducted with a group of 15-year-old high school students revealed valuable insights into their processing of captions containing audio-synchronized enhancement. While all participants attended to the enhanced target words and successfully identified their shared grammatical properties, they struggled to describe the underlying pronunciation rule (a finding that was replicated

in study 3 with a larger sample). On one hand, the observed emphasis on grammar was not as desirable as a focus on phonology but not necessarily detrimental, as accurate pronunciation of morphophonological variants is inherently connected to the learner's knowledge of their grammatical function (Levis, 2018). On the other hand, it is often the case that language learners who notice specific formal aspects in the input can recognize them in subsequent testing without, however, reaching awareness at the level of understanding (Leow, 2001; Leow & Martin, 2017). It is important to note that our participants also exhibited good accuracy in the read aloud task and performed above chance levels in the narrative task. Therefore, previous knowledge of the auditory form of the target words may have been sufficient to successfully read aloud these words and discuss their formal characteristics during the interviews, but not to describe the target pronunciation rule or fully automatize this knowledge in spontaneous production (Darcy, 2018; Pellicer-Sánchez, 2015; Tavakoli, 2019). In addition, the conventional emphasis on grammar and vocabulary learning in foreign language classrooms may have exacerbated difficulties in discussing phonological aspects, as learners often lack training in articulating phonological rules or patterns. It appears that a semi-incidental intervention with input enhancement led to noticing with a low level of awareness, but a clearer focus on auditory input and pronunciation learning could have facilitated deeper processing of the pronunciation target (Leow, 2015). In addition, while the internalization of a linguistic rule can support learning, further opportunities for practice as well as teacher feedback should be provided if the goal is the automatization of accurate perception and production (Celce-Murcia et al., 2010; Han et al., 2008). In summary, although exposure to audio-synchronized enhancement directed learners' attention to the target words during a comprehension-

focused activity, the lack of focus on phonological form prevented them from noticing the target pronunciation feature.

The interviews revealed that, under the unenhanced condition, most learners intentionally used captions to facilitate bottom-up comprehension of fast-paced spoken dialogue. In line with the benefits found in previous studies, captions were seen as a valuable resource for segmenting speech into words, recognizing unfamiliar words and phrases, and mapping auditory word forms onto orthographic representations (Charles, 2017; Mitterer & McQueen, 2009). This finding aligns with the well-established tendency to process input for meaning first before attending to linguistic form (VanPatten, 2004). It also suggests that understanding the input posed a challenging yet manageable task, and that learners naturally allocated attentional resources towards linguistic aspects not specifically addressed by the form-focused intervention (Leow, 2009). In fact, it is possible that learners' internal salience driven by the primary goal of understanding the dialogue conflicted with the external salience of the visually enhanced target words (Sharwood Smith, 1993; Chun et al., 2011). It remains an empirical question whether repeated exposure to the same videos containing audio-synchronized enhancement would allow learners to sequentially allocate attention to aspects deemed crucial for comprehension and then attend to the enhanced features targeted by the intervention (Han et al., 2008). The divergence between learners' focus and the intended targets represents a challenge to the assessment of speech learning outcomes from exposure to captioned video and may partially explain the mixed findings of the present dissertation and of previous studies with multimodal input (e.g., Wisniewska & Mora, 2020). It is nevertheless encouraging that learners reported effectively distributing attentional resources

among different processes (including listening, reading and integrating information in different modalities) without experiencing cognitive overload (Mayer, 2009; Rodgers & Webb, 2011; Sweller, 2005). Finally, the analysis of the impact of exposure to multimodal input is further complicated by learners' resistance, particularly at higher proficiency levels, towards reading captions (Kruger et al., 2015; Vanderplank, 2019). In some cases, when learners believe that exclusively paying attention to the audio and moving image is enough for comprehension and may even allow them to improve their listening comprehension skills, reading captions may be legitimately perceived as unnecessary. In other cases, when captions are ignored in a self-imposed effort to practice unaided comprehension, interventions involving guided practice on effective multimodal processing may gradually foster an appreciation of their value for language learning (Chacón, 2012; Vanderplank, 2019). It is tempting to speculate that the selection of words containing a wider variety of target pronunciation features (as in study 1) may have resulted in higher levels of perceived usefulness and higher attention to pronunciation. However, the decision to focus on a single pronunciation feature in study 2 and 3 was theoretically motivated and aimed to increase the transferability of the findings.

The findings discussed in relation to the first two studies, which examined textual enhancement in isolation to minimize potential confounding factors, provided support for further exploring audio-synchronized enhancement within a classroom intervention featuring various video-based activities. However, in study 3 the intervention group watching the videos with audio-synchronized textual enhancement did not exhibit any advantage in terms of pronunciation gains. This finding has several possible explanations, starting from the semi-incidental nature of interventions

featuring textual enhancement and the generally low gains found in comparison with interventions involving production activities and corrective feedback (Han et al., 2008; Leow, 2009; Pellicer-Sánchez & Boers, 2019). In addition, the benefits previously found for this type of enhancement may have been overridden by the interference of a primary focus on reading rather than listening, leading to difficulties in integrating information from different sources (Kruger, 2016). As the participants in the input enhancement intervention group generally reported reading captions more consistently than those in the no-enhancement intervention group, their orientation towards written input may have facilitated further processing of orthographic word forms at the expense of auditory forms (Tomlin & Villa, 1994). This explanation would be in line with previous findings that accurate pronunciation of *novel* spoken words is more easily retained after exposure to auditory and pictorial input than written and pictorial or multimodal input, even though the availability of text alongside auditory and pictorial input helps establish stronger form-meaning connections than single modality exposure (Uchihara et al., 2022).

Following Mitterer and McQueen (2009), it could be assumed that when the intervention fails to direct learners' attention to the comparison between the auditory input and the phonological forms generated through reading, the inaccurate phonological representations activated during caption processing may enter the phonological loop automatically, impairing auditory processing. Therefore, for the intervention group watching the clips with audio-synchronized enhancement, the enhancement may not have promoted a focus on auditory input by interrupting automatic reading behavior (cf. Stenton, 2013). Instead, it might have inadvertently reinforced the automatic generation and processing of phonological representations

based on L1 symbol-sound correspondences (Woore, 2018). Reliance on visual information can be quite strong for native speakers of languages with phonologically transparent orthographies, such as Spanish and Italian, and inhibiting L1-based decoding may be challenging regardless of L2 proficiency (Bassetti & Atkinson, 2015; Showalter, 2019). In fact, the interference of L1-based decoding is one of the factors that leads to the retention of L1-biased phono-lexical representations even at advanced stages of L2 language development (Bassetti & Atkinson, 2015; Erdener & Burnham, 2005). Finally, a survey of the learners' perceptions suggested that the enhancement group may have been less engaged or motivated by the intervention, although the study did not delve into motivational factors.

Encouraging results were obtained for the group that watched the videos without textual enhancement, as their pronunciation gains were significant and extended to words not encountered in the intervention. These findings suggest that carrying out the video-based activities in the classroom helped learners automatize accurate pronunciation and achieve timely and effortless recall and encoding of the target feature (Kruk & Pawlak, 2021; Solt et al., 2003). The learners' generally incorrect or imprecise descriptions of the target pronunciation rule indicate that the noticing and intake of phonological knowledge occurred at a low level of awareness (Leow, 2015). However, the successful retention of the pronunciation gains in the delayed post-test suggests that extended perception and production practice effectively fostered the implicit long-term restructuring of the learners' developing L2 system (De Jong, 2005; Leow, 2015). In addition, the generalization of the accuracy gains to novel contexts suggests that, contrary to the expectations for late L2 learners at this proficiency level, past <-ed> verbs may have been processed based on their

morphological structure rather than as individual lexical entries (cf. Bassetti & Atkinson, 2015; Clahsen et al., 2010; Ullman, 2005). The gains observed in both the narrative task and sentence repetition task indicate that the learners internalized the new knowledge, resulting in its accurate use in spontaneous and semi-spontaneous language production, respectively. In addition, improved accuracy in the sentence repetition task suggests that bottom-up *perception* of this challenging pronunciation feature also improved through focused listening practice (Strachan & Trofimovich, 2019). The intervention proved beneficial for learners across a range of proficiency and aptitude levels, in line with research showing that, although beginner learners are more likely to mispronounce <-ed> endings, advanced learners also encounter difficulties and may benefit from perception and production practice (Bell et al., 2015; Solt et al., 2003; Strachan & Trofimovich, 2019).

Lastly, the intervention received positive feedback from the participants, who expressed high levels of satisfaction with the videos and activities used. Learners' perceived usefulness of the intervention, in combination with the significant gains obtained by one of the two intervention groups, provided further evidence of the learning potential of audiovisual activities (Chiu, 2012; Navarrete, 2013; Sokoli, 2018). Although intralingual dubbing and captioning had been implemented as teaching tools in previous research, the creation of activities with a proactive focus on form and centered around a target pronunciation feature is a novel contribution of this dissertation. Through repeated exposure to short speech segments containing the target feature, learners were encouraged to pay close attention to pronunciation within meaningful comprehension activities (Lima, 2015b). Moreover, practicing the production of increasingly longer stretches of speech containing this feature, where

the support of captions was gradually eliminated, facilitated the automatization of the newly acquired knowledge and its application in novel contexts (Zhang and Yuan, 2020). Previous research featuring audiovisual activities had also found positive teacher perceptions and overwhelmingly positive learner perceptions (Alonso-Perez & Sánchez-Requena, 2018). However, most previous studies were pedagogically oriented and adopted an ecological perspective exclusively focused on learners' and teachers' perceptions of the activities and on perceived linguistic gains, in the absence of quantitative data on learners' linguistic performance (e.g., Chiu, 2012; Martinsen et al., 2017; Sokoli, 2018; Zhang, 2016). Study 3 in this dissertation, however, adopted a pre-, post-, and delayed post-test design, included a control group, and controlled for individual factors such as proficiency and aptitude. This approach aimed to address not only aspects of external validity, such as ecological validity, related to the generalizability of findings to other educational contexts, but also aspects of internal validity that allow to establish a reliable cause-and-effect relationship between the treatment and learning outcomes.

To sum up, this dissertation has found evidence that audio-synchronized textual enhancement can enhance noticing of L2 pronunciation in multimodal input. However, the high individual variability in learners' allocation of attentional resources and learning outcomes suggests that the effectiveness of this technique may vary depending on factors such as the learner's proficiency level and cognitive maturity. Older and more proficient learners may benefit from an intervention that exclusively features textual enhancement, thanks to their ability to manage the integration of input from auditory and written sources, using captions to support auditory processing. Younger and less proficient learners may still benefit from audio-synchronized textual

enhancement but may require more guidance. In particular, the target items need to be carefully chosen, the learners should be explicitly orientated towards the processing of auditory input, and the viewing may have to be repeated multiple times to allow for the sequential allocation of attention to meaning and form. When implementing audio-synchronized textual enhancement in a classroom intervention involving other video-based activities, the effects of textual enhancement may naturally be less noticeable than those of pronunciation instruction. While audio-synchronized enhancement can be used to supplement traditional pronunciation teaching methods, providing opportunities for production practice and teacher feedback remains crucial. Finally, learners' interest in the video content and eagerness to work with their peers can result in high levels of engagement and satisfaction with the intervention, which in turn can boost their receptiveness towards language learning. The next chapter will provide a conclusion to this dissertation by addressing its limitations and identifying potential areas for further research.

CHAPTER 5. CONCLUSION

5.1. Summary of main findings

This dissertation has contributed to a better understanding of the potential benefits of textual enhancement in captioned video for improving L2 pronunciation. The findings suggest that this type of multimodal input can have a positive impact on pronunciation, but also highlight some important considerations for its effective use. These are the main findings in relation to the aims outlined in the introduction:

1) The investigation of the effectiveness of audio-synchronized textual enhancement in pronunciation teaching and learning yielded mixed findings. In one study, this technique successfully directed learners' attention to the target words in captions and consequently to the corresponding auditory forms, leading to faster rejection of mispronunciations in a lexical decision task. Improved performance in this task suggests that audio-synchronized enhancement successfully triggered noticing of auditory word forms, facilitating the conversion of input into intake. On the other hand, the pronunciation learning gains observed when integrating audio-synchronized enhancement within a classroom intervention were not significant. Therefore, longer exposure and/or more explicit techniques may be required to trigger further processing of the target linguistic feature, leading to its internalization and accurate use in production. In particular, insights from verbal recall suggested that an explicit focus on pronunciation may be needed to redirect learners' attention from the target words' semantic and grammatical features to their phonology.

2) The teaching intervention containing video-based activities led to significant pronunciation gains for the group watching unenhanced videos but, surprisingly, not for the group who watched the video clips with audio-synchronized textual

enhancement. The analysis of learners' perceptions of the intervention hinted at an excessive reliance on captions for comprehension, which in this case may have impaired their ability to integrate auditory and written information. However, it is also possible that other factors not captured by the study may have affected the results of the group who watched the video clips with caption enhancement.

3) Regarding the role of individual factors, when investigating audio-synchronized enhancement in isolation, we found that it increased audiovisual synchrony for all learners, regardless of whether they were faster readers with a higher L2 proficiency level or slower readers with a lower proficiency level. In addition, while the duration of learners' fixations on the unenhanced target words was unrelated to their proficiency and reading speed, learners at lower proficiency levels fixated for longer on enhanced words. However, proficiency and reading speed did not correlate with gains in accuracy and speed of auditory form recognition. Similarly, in the classroom study there was no correlation between pronunciation accuracy gains and proficiency, listening comprehension skills or phonemic coding ability, suggesting that the classroom intervention was equally beneficial for learners with different profiles.

5.2. Pedagogical implications

The findings of this dissertation have several pedagogical implications for the use of captioned video in pronunciation teaching and learning. First, it should be assumed that when learners' attention is not directed to pronunciation (through audiovisual manipulation or other forms of instruction), they will focus on meaning rather than linguistic form, reading captions and/or listening to spoken language depending on how comfortable they are with each modality. This may not guarantee the smooth integration of the two modalities, nor the development of L2 pronunciation through

unfocused viewing. In fact, learners may spend a significant amount of time subvocalizing the written text in captions, preventing the processing of auditory information in the dialogues. Another issue is that, depending on learners' proficiency and reading speed, they may read words in captions too early or too late in relation to the words' auditory onset, missing out on opportunities for modality integration. Audio-synchronized textual enhancement may help to ensure that learners' attention is focused on the target sounds at the right time, enhancing the pronunciation learning potential of captioned video.

Second, enhanced input is processed differently by each learner depending on factors like the difficulty of the input, the learners' internal focus, and their cognitive abilities. As a result, exposure to video containing enhanced words may not directly translate into more accurate pronunciation of these words, but the new data accumulated in the learners' internal system as a by-product of a meaningful activity (watching TV) is likely to merge with previous linguistic knowledge over time, leading to interlanguage restructuring. Providing repeated exposure to the same audiovisual content within an intervention involving extended viewing of captioned video may help learners notice target phonological features, resolve uncertainties regarding their pronunciation and, ultimately, automatize targetlike use of these features in everyday language use. Integrating enhanced or unenhanced video into activities that also involve a component of production practice can help consolidate perception gains and facilitate the transfer to production. Recommendations regarding the design of these activities, discussed in detail in chapter 3, involve the provision of detailed instructions to learners, regularly monitoring learners' execution of the activity, and preparing in

advance for any foreseeable technical difficulty that may prevent successful implementation.

Overall, these recommendations can improve the way pronunciation is taught in classrooms by providing teachers with new insights on how captioned video can help learners improve their pronunciation skills. However, it is important to note that research on the effects of captioned video on pronunciation is still in its early stages, and that the studies in this dissertation have a number of limitations. The remaining part of the section will discuss the limitations of this dissertation and possible directions for future research.

5.3. Limitations

The differences in learner populations among the three studies (university students in study 1 and high school students in studies 2 and 3) represent one of the main limitations of this dissertation. The generalizability of the findings of study 1 to the other two studies was limited due to the possible confounding variables introduced by different learner profiles and motivations. For example, university students may have been driven by a keen interest in learning about the English language, whereas high school students may have been more interested in getting good grades, and this could have influenced their performance on the tasks. However, it was necessary to conduct a preliminary study (study 1) to decide which synchronization to use in the longitudinal intervention (study 3). Recruiting a large number of participants in schools to collect eye-tracking data proved too difficult, especially as the Covid-19 pandemic broke right before the start of data collection and strict prevention measures were implemented in schools. To address this limitation, the participants in the three studies were matched on several variables (L1, EFL learning contexts, range of

proficiency levels, hometown). As a result, the two groups of students were expected to share the same types of issues learning English pronunciation. On the one hand, they had studied English in school for about 10 years and were supposed to have reached an intermediate level of proficiency, which allowed them to understand TV programs quite comfortably and focus on specific linguistic features if encouraged to do so. On the other hand, their interlanguage was still developing, meaning they could still improve their pronunciation by updating their phonological representations and correcting any mispronunciations. In light of this, the three studies were reported in this dissertation because each presents methodological innovations and contributes to our understanding of the learning potential of captioned video and of the effects of audio-synchronized enhancement.

The small number of participants and the type of sampling used (convenience sampling) limits the generalizability of findings to the larger population. In particular, the lack of statistical power in study 2, possibly linked to the small sample size, may have prevented the detection of significant differences between the exposure conditions. To address the issue of the large amount of individual variation intrinsic in caption reading and avoid drawing incorrect conclusions regarding the effects of enhancement, random factors representing *participant* and *item* were included in all the statistical models. Due to the short duration of the treatment in study 1, it was difficult to assess any changes in attention allocation to the enhanced words which may have occurred over time once learners got used to the novelty of the enhancement technique. It must also be pointed out that, despite taking extra steps to hide the eye-tracker and make participants feel at ease, eye-tracking studies may not be entirely

representative of how learners watch TV series at home or in a non-educational context.

In general, the pronunciation tests were matched to the specific aims of one study and may have failed to provide a comprehensive picture of learners' knowledge of the target items. The lexical decision task in study 1 focused on the efficiency of lexical access as an indicator of the stability of the target phonolexical representations in the learners' mental lexicon, without providing information regarding their ability to pronounce the target words. On the contrary, the sentence repetition task in study 3 required both accurate perception and production, and it was therefore impossible to determine whether mispronunciations were due to issues in perception, i.e., whether the learner may have been able to produce an item that they could not perceive. By using a sentence repetition task, we implicitly assumed that accuracy in perception shapes accuracy in production, and if a learner can pronounce a word, they can perceive it. In addition, a written production task may have helped us assess the learners' use of past <-ed> irrespective of their pronunciation. However, in line with previous research, we assumed that inconsistent use of past <-ed> was due to "mapping problems" or the inefficient mapping of the past abstract syntactic feature to its surface morphological realizations due to phonological issues (Solt et al., 2003).

The analysis of learners' perceptions of the intervention was either absent, in study 1, or limited, as in study 2 and 3. The questionnaire did not contain detailed questions about which type of activities the learners deemed more beneficial and about their preferred modality (reading or listening) during the viewing. A more fine-grained analysis of learners' perceptions and feeling of learning throughout the intervention, with diaries or blog entries, may also have provided invaluable insights on the

effectiveness of the video enhancement and audiovisual activities. In addition, since the intervention was conducted by the author, she was unable to collect classroom observation data that may have assisted the interpretation of learners' responses to the questionnaire and, perhaps, shed light on the reasons behind their pronunciation gains or lack thereof.

Finally, another general limitation of this dissertation was a narrow focus on the bimodal aspect of multimodal input (processing of captions and auditory input), with a very limited exploration of learners' processing of the moving image (faces, gestures) in synchrony with the audio. However, it is well attested that viewers tend to fixate on the speaker's face in search of visual cues, and that these cues can enhance speech intelligibility and support L2 perception training (Hardison, 2007). In study 2, we did find that learners often looked at faces and mouths, but they reported that this behavior was either automatic or aimed at reading the characters' expressions and emotions. However, it is possible that they at least partially processed the movements of the character's articulatory organs and associated them to specific sounds while being unaware of it. Similarly, gestures may provide an incidental source of visual enhancement that improve the comprehension and learning of speech (Mathias & von Kriegstein, 2023). The investigation of the effects of the moving image on learners' processing of L2 speech in captioned video was beyond the scope of this dissertation, but it represents a promising area for further research.

5.4. Future research

Future research should both address the limitations of this dissertation and expand on its findings. To begin with, recruiting a bigger sample of learners, divided into experimental groups of similar size, would maximize statistical power and allow to

draw more robust conclusions regarding the effects of each experimental condition. The findings need to be contextualized within the specific educational setting and population targeted in the study, to avoid overgeneralization. Longitudinal research on learners' processing of captioned video and noticing of pronunciation features may help interpret any potential gains in pronunciation accuracy. Oral diaries can provide insights into learners' awareness of their pronunciation gains while viewing captioned video, while also documenting spontaneous use of the foreign language (Yibokou & Jingand, 2023). It is also important to be aware that interventions involving primarily meaning-based activities, such as watching TV, may not produce immediate results for all learners. Using a variety of data collection methodologies and tests tapping into different stages of phonological processing may contribute to obtaining a clearer picture of learners' progress.

The effectiveness of audio-synchronized enhancement may be moderated by a number of individual factors that we have not included in these studies, such as age and motivation. Interestingly, it has even been hypothesized that the impact of orthographic form on language learning may be less significant for current learners than for previous generations due to the exponential increase in exposure to auditory input such as songs, movies and other online resources (Bassetti et al., 2018). Future studies should explore to what extent individual and contextual factors can impact learners' audiovisual processing of captioned video. Relatedly, learners' internal focus during exposure to audio-synchronized enhancement and captioned video in general should be further investigated. It is important to determine whether learners spontaneously focus on pronunciation, whether their internal focus can be redirected to pronunciation, and whether this shift in focus can positively affect their learning

outcomes. In light of learners' tendency to allocate attention to meaning and form sequentially, a first viewing dedicated to comprehension followed by a second viewing dedicated to a focus on form may help them process information more efficiently. However, a study that involves two viewings of a video clip may sacrifice ecological validity, as learners may not be used to watching videos twice. Finally, future research should investigate the effects of the availability of visual cues, including gestures and the articulatory cues that may become visible when a character is speaking directly in front of the camera.

To conclude, while this dissertation has offered valuable insights into the use of captioned video for learning L2 pronunciation, the road to tapping into the full potential of this extraordinary resource is still long. Hopefully, just like with any good TV program, this is just one of the first episodes of a long series, one that will be written by researchers and teachers jointly.

REFERENCES

Aliaga-García, C. (2017). *The effect of auditory and articulatory phonetic training on the perception and production of L2 vowels by Catalan-Spanish learners of English.* [Doctoral dissertation, University of Barcelona]. TDX. http://hdl.handle.net/10803/471451

Almeida, P. A., & Costa, P. D. (2014). Foreign Language Acquisition: The Role of Subtitling. *Procedia - Social and Behavioral Sciences*, *141*, 1234–1238. https://doi.org/10.1016/j.sbspro.2014.05.212

Alonso-Perez, R., & Sánchez-Requena, A. (2018). Teaching foreign languages through audiovisual translation resources: Teachers' perspectives. *Applied Language Learning*, *28*.

Alsadoon, R., & Heift, T. (2015). Textual input enhancement for vowel blindness: A study with Arabic ESL learners. *Modern Language Journal*, *99*(1), 57–79. https://doi.org/10.1111/modl.12188

Araujo, L., & Costa, P. (2013). *The European survey on language competences: School-internal and external factors in language learning.* JRC Scientific and Technical Report (EU 26078 -2013). Luxemburg: Publications Office of the European Union.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human memory: A proposed system and its control processes. In K. W. Spence & J. T. Spence, *The psychology of learning and motivation: II.* Academic Press. https://doi.org/10.1016/S0079-7421(08)60422-3

Baayen, H. (2010). Analyzing reaction times. *International Journal of Psychological Research*, *3*. https://doi.org/10.21500/20112084.807

Baddeley, A. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences, 4*(11), 417-423. doi:10.1016/S1364-6613(00)01538-2

Baddeley, A. D. (2007). *Working memory, thought, and action*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198528012.001.0001

Baddeley, A. D., & Hitch, G. (1974). Working memory. *Psychology of Learning and Motivation*, 8, 47-89. http://dx.doi.org/10.1016/S0079-7421(08)60452-1

Bassetti, B., & Atkinson, N. (2015). Effects of orthographic forms on pronunciation in experienced instructed second language learners. *Applied Psycholinguistics, 36,* 67–91. doi: 10.1017/S0142716414000435

Bassetti, B., Hayes-Harb, R., & Escudero, P. (2015). Second language phonology at the interface between acoustic and orthographic input. *Applied Psycholinguistics, 36*(1), 1–6. doi:10.1017/S0142716414000393

Bassetti, B., Sokolović-Perović, M., Mairano, P., & Cerni, T. (2018). Orthography-induced length contrasts in the second language phonological systems of L2 speakers of English: Evidence from minimal pairs. *Language and Speech, 61(4),* 577–597. https://doi.org/10.1177/0023830918780141

Bell, P., Trofimovich, P., & Collins, L. (2015). Kick the ball or kicked the ball? Perception of the past morpheme <–ed> by second language learners. *Canadian Modern Language Review*, 71, 26-51. DOI: 10.3138/cmlr.2075.

Benitez-Correa, C., Cabrera-Solano, P., Solano, L., & Espinoza-Celi, V. (2020). Improving past tense pronunciation of regular verbs through the use of Audacity: A case study of EFL undergraduate students in Ecuador. *Teaching English with Technology, 20(1),* 3-20.

Best, C. T., & Tyler, M. D. (2007). Nonnative and second-language speech perception: Commonalities and complementarities. In M. J. Munro & O.-S. Bohn (Eds.), *Second language speech learning: The role of language experience in speech perception and production* (pp. 13–24). John Benjamins. https://doi.org/10.1075/lllt.17.07bes

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. http://doi.org/10.1017/S0142716402004022

Birulés-Muntané, J., & Soto-Faraco, S. (2016). Watching subtitled films can help learning foreign languages. *PLoS One,* 11(6), 1–10. https://doi.org/10.1371/journal.pone.0158409

Bisson, M.-J., van Heuven, W. J. B., Conklin, K., & Tunney, R. J. (2014). Processing of native and foreign language subtitles in films: An eye tracking study. *Applied Psycholinguistics, 35(2),* 1–20. https://doi.org/10.1017/S0142716412000434

Broersma, M., & Cutler, A. (2008). Phantom word activation in L2. *System*, 36, 22-34. DOI: 10.1016/j.system.2007.11.003

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, 41(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Burton, M., Jongman, A., & Sereno, J. (1996). Phonological and orthographic priming effects in auditory and visual word recognition. In W. Ham & L. Lavoie (Eds.), *Working papers of the Cornell Phonetics Laboratory* (Vol. 11, pp. 17–41). Ithaca, NY: Cornell University, Department of Linguistics. Cambridge University Press. https://doi.org/10.1017/CBO9781139547369.005

Campbell, A. (2016). *Subtitling for mission accomplishment: An experimental study of the effect of subtitling as a task on listening comprehension for learners of military English for specific purposes* [Unpublished master's thesis, Pablo de Olavide University]. http://hdl.handle.net/10433/3051

Carless, D. (2012). TBLT in EFL settings: Looking back and moving forward. In A. Shehadeh & C. Coombe (Eds.), *Task-based language teaching in foreign language contexts. Research and implementation* (pp. 345–358). Amsterdam, The Netherlands: John Benjamins. https://doi.org/10.1075/tblt.4.20car

Celce-Murcia, M., Brinton, D. M., Goodwin, J. M., & Griner, B. (2010). *Teaching pronunciation: A reference for teachers of English to speakers of other languages* (2nd ed.). Cambridge: Cambridge University Press.

Chacón, C. T. (2012). Task-based language teaching through film-oriented activities in a teacher education program in Venezuela. In: A., Shehadeh, & C. A. Coombe (Eds.). *Task-based language teaching in foreign language contexts: Research and implementation.* Amsterdam: Benjamins. https://doi.org/10.1075/tblt.4.15cha

Charles, T. J. (2017). *The role of captioned video in developing speech segmentation for learners of English as a second language* (Publication No 731589) [PhD Thesis, University of York]. EThOS. https://etheses.whiterose.ac.uk/19156/

Charles, T. J., & Trenkic, D. (2015). The effect of bi-modal input presentation on second language listening: The focus on speech segmentation. In Y. Gambier,

A. Caimi, & C. Mariotti (Eds.), *Subtitles and language learning* (pp. 173-197). Bern: Peter Lang.

Charoy, J., & Samuel, A. G. (2019). The effect of orthography on the recognition of pronunciation variants. *Journal of Experimental Psychology: Learning, Memory, and Cognition*. Advance online publication. http://dx.doi.org/10.1037/xlm0000781

Chiu, Y. (2012). Can film dubbing projects facilitate EFL learners' acquisition of English pronunciation? *British Journal of Educational Technology*, 43(1), pp. 24-27. https://doi.org/10.1111/j.1467-8535.2011.01252.x

Chun, M. M., Golomb, J. D., & Turk-Browne, N. B. (2011). A taxonomy of external and internal attention. *Annual Review of Psychology*, 62, 73–101. https://doi.org/10.1146/annurev.psych.093008.100427

Chung, Y., & Révész, A. (2021). Investigating the effect of textual enhancement in post-reading tasks on grammatical development by child language learners. *Language Teaching Research*, 25(3). https://doi.org/10.1177/13621688211005068

Cintrón-Valentín, M., & García-Amaya, L. (2021). Investigating textual enhancement and captions in L2 grammar and vocabulary: An experimental study. *Studies in Second Language Acquisition*, *43*(5), 1068-1093. doi:10.1017/S0272263120000492

Cintrón-Valentín, M., García-Amaya, L., & Ellis, N. C. (2019). Captioning and grammar learning in the L2 Spanish classroom. *The Language Learning Journal*, *47*(4), 439–459. https://doi.org/10.1080/09571736.2019.1615978

Clahsen, H., Felser, C., Neubauer, K., Sato, M., & Silva, R. (2010). Morphological structure in native and nonnative language processing. *Language Learning*, *60*(1), 21–43. https://doi.org/10.1111/j.1467-9922.2009.00550.x

Cobb, T. *Compleat Web VP v.2.1* [computer program] Accessed March 2019 at https://www.lextutor.ca/vp/comp/

Cook, S. V., Pandža, N. B., Lancaster, A. K., & Gor, K. (2016). Fuzzy nonnative phonolexical representations lead to fuzzy form-to-meaning mappings. *Frontiers in Psychology*, *7(sep)*, 1–17. https://doi.org/10.3389/fpsyg.2016.01345

Corder, S. (1967). The significance of learners' errors. *International Review of Applied linguistics*, 5, 161–170.

Cutler, A. (2000). Listening to a second language through the ears of a first. *Interpreting, 5*(1), 1–23. https://doi.org/10.1075/intp.5.1.02cut

D'Ydewalle, G. (2002). Foreign-language acquisition by watching subtitled television programs. *Journal of Foreign Language Education and Research,* 12, 59–77.

D'Ydewalle, G., & Van de Poel, M. (1999). Incidental foreign-language acquisition by children watching subtitled television programs. *Journal of Psycholinguistic Research,* 28(3), 227–244.

Danan, M., (2010). Dubbing projects for the language learner: A framework for integrating audiovisual translation into task-based instruction. *Computer Assisted Language Learning*, 23:5, 441-456. https://doi.org/10.1080/09588221.2010.522528

Darcy, I. & Holliday, J. J. (2019). Teaching an old work new tricks: Phonological updates in the L2 mental lexicon. In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference* (pp. 10-26). Ames: Iowa State University.

Darcy, I. (2018). Powerful and effective pronunciation instruction: How can we achieve it? *CATESOL Journal,* 30, 13-45.

Darcy, I., Daidone, D., & Kojima, C. (2013). Asymmetric lexical access and fuzzy lexical representations in second language learners. *The Mental Lexicon*, *8*(3), 372–420. https://doi.org/10.1075/ml.8.3.06dar

Darcy, I., Ewert, D, & Lidster, R. (2012). Bringing pronunciation instruction back into the classroom: An ESL teachers' pronunciation "toolbox". In. J. Levis & K. LeVelle (Eds.). *Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference,* Sept. 2011. (pp. 93-108). Ames, IA: Iowa State University.

Darcy, I., Rocca, B., & Hancock, Z. (2021). A window into the classroom: How teachers integrate pronunciation instruction. *RELC Journal, 52(1),* 110–127. https://doi.org/10.1177/0033688220964269

De Jong, N. (2005). Can second language grammar be learned through listening? An experimental study. *Studies in Second Language Acquisition, 27(2),* 205–234. http://www.jstor.org/stable/44486822

DeKeyser, R. M. (2007). Skill acquisition theory. In B. VanPatten & J. Williams (Eds.), *Theories in Second Language Acquisition: An introduction* (pp. 97-112). Mahwah, NJ: Erlbaum.

DeKeyser, R. M. (2010). Practice for second language learning: Don't throw out the baby with the bathwater. *International Journal of English Studies,* 10. https://doi.org/10.6018/ijes.10.1.114021

Derwing, T. M., & Munro, M. J. (2005). Second language accent and pronunciation teaching: A research-based approach. *TESOL Quarterly*, *39*(3), 379–397. https://doi.org/10.2307/3588486

Derwing, T. M., Diepenbroek, L., & Foote, J. (2012). How well do general-skills ESL textbooks address pronunciation? *TESL Canada Journal*, *30*, 22–44. https://doi.org/10.18806/tesl.v30i1.1124

Dickerson, W. B. (2015). Using orthography to teach pronunciation. In M. Reed and J.M. Levis (Eds.) *The Handbook of English Pronunciation* (pp. 488–504). John Wiley & Sons, Ltd. https://doi.org/10.1002/9781118346952.ch27

Dizon, G. (2018). Netflix and L2 learning: A case study. *The EuroCALL Review,* 26 (2), 30-40. https://doi.org/10.4995/eurocall.2018.9080

Edge, B. A. (1991). The production of word-final obstruents in English by L1 speakers of Japanese and Cantonese. *Studies in Second Language Acquisition, 13*, 377–393.

Eiter, B., & Inhoff, A. (2010). Visual word recognition during reading is followed by subvocal articulation. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *36*, 457–470. https://doi.org/10.1037/a0018278

Ellis, N. C. (2017). Salience in usage-based SLA. In S. M. Gass, P. Spinner, & J. Behney (Eds.), Salience in second language acquisition (pp. 21–40). New York, NY: Routledge.

Ellis, R. (2001). Introduction: Investigating form-focused instruction. *Language Learning*, 51, 1-46. https://doi.org/10.1111/j.1467-1770.2001.tb00013.x

Ellis, R. (2012). *Language teaching research and language pedagogy* (pp. 271-306). Malden, MA: John Wiley and Sons. http://dx.doi.org/10.1002/9781118271643

Ellis, R. (2016). Focus on form: A critical review. *Language Teaching Research*, 20(3), 405–428. https://doi.org/10.1177/1362168816628627

Ellis, R. (2020). Teacher-preparation for task-based language teaching. In C. Lambert & R. Oliver (Eds.), *Using Tasks in Second Language Teaching: Practice in Diverse Contexts* (pp. 99–120). Multilingual Matters. https://doi.org/doi:10.21832/9781788929455-008

Erdener, V. D., & Burnham, D. K. (2005). The role of audiovisual speech and orthographic information in nonnative speech production. *Language Learning*, *55*(2), 191–228. https://doi.org/10.1111/j.0023-8333.2005.00303.x

Ernestus, M., Kouwenhoven, H., & van Mulken, M. (2017). The direct and indirect effects of the phonotactic constraints in the listener's native language on the comprehension of reduced and unreduced word pronunciation variants in a foreign language. *Journal of Phonetics*, *62*, 50–64. https://doi.org/10.1016/j.wocn.2017.02.003

Escudero, P. (2007). Second-language phonology: the role of perception. In: M.C. Pennington (Ed.), *Phonology in Context.* Palgrave Macmillan, London. https://doi.org/10.1057/9780230625396_5

Flege, J. E. (1995). Second language speech learning: Theory, findings, problems. In W. Strange (Ed.), *Speech perception and linguistic experience: issues in cross-language research* (pp. 233–277). York Press.

Flege, J. E., & Bohn, O. (2021). The revised speech learning model (SLM-r). In R. Wayland (Ed.), *Second Language Speech Learning: Theoretical and Empirical Progress* (pp. 3-83). Cambridge: Cambridge University Press. doi:10.1017/9781108886901.002

Foote, J., & Mcdonough, K. (2017). Using shadowing with mobile technology to improve L2 pronunciation. *Journal of Second Language Pronunciation*, *3*, 34–56. https://doi.org/10.1075/jslp.3.1.02foo

Forster, K.I., & Forster, J.C. (2003). DMDX: A Windows display program with millisecond accuracy. *Behavior Research Methods, Instruments, & Computers*, 35, 116–124. https://doi.org/10.3758/BF03195503

Frost, R. (1998). Toward a strong phonological theory of visual word recognition: True issues and false trails. *Psychological Bulletin*, *123*(1), 71–99. https://doi.org/10.1037/0033-2909.123.1.71

Galimberti, V., Miralpeix, I. (2018). Multimodal input for Italian beginner learners of English: A study on comprehension and vocabulary learning from undubbed TV

series. In: Coonan, C. M, Bier, A., & Ballarin, E. (Eds.) *La didattica delle lingue nel nuovo millennio: Le sfide dell'internazionalizzazione*. Studi e Ricerche 13 (pp. 615-626). doi: 10.30687/978-88-6969-227-7/036

Gass, S. M., & Mackey, A. (2000). *Stimulated recall methodology in second language research*. Lawrence Erlbaum Associates Publishers.

Gesa, F. (2019). *L1 / L2 subtitled TV series and EFL learning: A study on vocabulary acquisition and content comprehension at different proficiency levels.* [Doctoral dissertation, University of Barcelona]. TDX. http://hdl.handle.net/10803/668505

Ghia, E. (2012). *Subtitling matters. New perspectives on subtitling and foreign language learning* (pp. 7-48). Oxford: Peter Lang.

Gilabert, R. (2007). Effects of manipulating task complexity on self-repairs during L2 oral production. *International Review of Applied Linguistics in Language Teaching (IRAL), 45*(3), 215-240. https://doi.org/10.1515/iral.2007.010

Godfroid, A. (2019). *Eye-tracking in second language acquisition and bilingualism: A research synthesis and methodological guide.* https://doi.org/10.4324/9781315775616.

Godfroid, A., & Kim, K. (2021). The contributions of implicit-statistical learning aptitude to implicit second-language knowledge. *Studies in Second Language Acquisition, 43*(3), 606-634. doi:10.1017/S0272263121000085

Godfroid, A., Housen, A., & Boers, F. (2010). A procedure for testing the noticing hypothesis in the context of vocabulary acquisition. In M. Putz and L. Sicola (Eds.), *Cognitive processing in second language acquisition: Inside the learner's mind* (pp. 169–197). Amsterdam: John Benjamins. DOI: 10.1075/celcr.13.14god

Gor, K., Cook, S., Bordag, D., Chrabaszcz, A., & Opitz, A. (2021). Fuzzy lexical representations in adult second language speakers. *Frontiers in Psychology, 12*. https://doi.org/10.3389/fpsyg.2021.732030

Hall, J. & Stewart, M. E. (2010). Basic psychology. In E. C. Johnstone, D. C. Owens, S. M. Lawrie, A. M. McIntosh & M. Sharpe (Eds.), *Companion to Psychiatric Studies* (pp. 95-108). Churchill Livingstone. doi:10.1016/B978-0-7020-3137-3.00005-X

Hamada, Y. (2016). Shadowing: Who benefits and how? Uncovering a booming EFL teaching technique for listening comprehension. *Language Teaching Research*, *20*(1), 35–52. https://doi.org/10.1177/1362168815597504

Han, Z., Park, E. S., & Combs, C. (2008). Textual enhancement of input: Issues and possibilities. *Applied Linguistics*, *29*(4), 597–618. https://doi.org/10.1093/applin/amn010

Hardison, D. M. (2007). The visual element in phonological perception and learning. In M. C. Pennington (Ed.), *Phonology in Context* (pp. 135–158). Palgrave Macmillan UK. https://doi.org/10.1057/9780230625396_6

Harrington, D. M. (2006). The lexical decision task as a measure of L2 lexical proficiency. *EUROSLA Yearbook,* 6, 147-168, https://doi-org.sire.ub.edu/10.1075/eurosla.6.10har

Haslam, M., & Zetterholm, E. (2019). The role of consonant clusters in English as a lingua franca intelligibility. In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, Ames, IA, September 2018 (pp. 276-284). Ames, IA: Iowa State University.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Front Neurosci,* 8, 150. doi: 10.3389/fnins.2014.00150.

Henderson, A., Frost, D., Tergujeff, E., Kautzsch, A., Murphy, D., Kirkova-Naskova, A., Waniek-Klimczak, E., Levey, D., Cunnigham, U., & Curnick, L. (2012). The English pronunciation teaching in Europe survey: Selected results. *Research in Language*, *10*(1), 5–27. https://doi.org/10.2478/v10015-011-0047-4

Hu, M. & Nation, I.S.P. (2000) Vocabulary density and reading comprehension. *Reading in a Foreign Language 13* (1), 403-430.

Hutchinson, A. E., & Dmitrieva, O. (2022). Exposure to speech via foreign film and its effects on non-native vowel production and perception. *Journal of Phonetics*, *95*, 101189. https://doi.org/10.1016/j.wocn.2022.101189

Incalcaterra McLoughlin, L., & Lertola, J. (2015). Captioning and revoicing of clips in foreign language learning - using ClipFlair for teaching Italian in online learning environments. In C. Ramsey-Portolano (Ed.), *The future of Italian teaching* (pp. 55-69). Newcastle upon Tyne: Cambridge Scholars Publishing.

Isaacs, T. (2018). Shifting sands in second language pronunciation teaching and assessment research and practice. *Language Assessment Quarterly*, *15*(3), 273–293. https://doi.org/10.1080/15434303.2018.1472264

Jarosz, A. (2019). *English pronunciation in L2 instruction: The case of secondary school learners.* Springer: Cham, Switzerland.

Jelani, N.A., & Boers, F. (2018). Examining incidental vocabulary acquisition from captioned video: Does test modality matter? *ITL – International Journal of Applied Linguistics, 169*, 169-190. https://doi.org/10.1075/itl.00011.jel

Jenkins, J. (2002). A sociolinguistically based, empirically researched pronunciation syllabus for English as an international language. *Applied Linguistics*, *23*(1), 83–103+156. https://doi.org/10.1093/applin/23.1.83

Jung, J., Stainer, M. J., & Tran, M. H. (2022). The impact of textual enhancement and frequency manipulation on incidental learning of collocations from reading. *Language Teaching Research, 0(0).* https://doi.org/10.1177/13621688221129994

Kaderoglu, K., & Romeu, F. (2021). Captions in L2 learning from language teachers' perspective: What do teachers believe and do? *CALL-EJ, 23,* 86–106.

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

Kim, Y., & Tracy-Ventura, N. (2011). Chapter 11. Task complexity, language anxiety, and the development of the simple past. In Robinson (Ed.) *Second Language Task Complexity* (pp. 287–306). John Benjamins. https://www.jbe-platform.com/content/books/9789027290274-tblt.2.18ch11

Kormos, J. (2006). *Speech production and second language acquisition*. Lawrence Erlbaum Associates Publishers.

Kostromitina, M., & Plonsky, L. (2021). Elicited imitation tasks as a measure of L2 proficiency: A meta-analysis. *Studies in Second Language Acquisition*, 1-26. doi:10.1017/S0272263121000395

Krashen, S. D. (1982). *Principles and practice in second language acquisition*. Oxford: Pergamon.

Krashen, S. D., & Terrell, T. D. (1983). *The natural approach: Language acquisition in the classroom.* Oxford: Pergamon Press.

Kruger, J. L., & Doherty, S. (2016). Measuring cognitive load in the presence of educational video: Towards a multimodal methodology. *Australasian Journal of Educational Technology*, *32*(6). https://doi.org/10.14742/ajet.3084

Kruger, J. L., & Steyn, F. (2014). Subtitles and eye tracking: Reading and performance. *Reading Research Quarterly*, *49*(1), 105-120. https://doi.org/10.1002/rrq.59

Kruger, J. L., Hefer-Jordaan, E., & Matthew, G. (2013). Measuring the impact of subtitles on cognitive load: Eye tracking and dynamic audiovisual texts. In *ACM International Conference Proceeding Series* (p. 62-66). https://doi.org/10.1145/2509315.2509331

Kruger, J. L., Szarkowska, A., & Krejtz, I. (2015). Subtitles on the moving image: An overview of eye tracking studies. *Refractory: A Journal of Entertainment Media*, *25*.

Kruk, M., & Pawlak, M., (2021). Using internet resources in the development of english pronunciation: The case of the past tense -ed ending. *Computer Assisted Language Learning*, DOI: 10.1080/09588221.2021.1907416

Lachaud, C. M., & Renaud, O. (2011). A tutorial for analyzing human reaction times: How to filter data, manage missing values, and choose a statistical model. *Applied Psycholinguistics*, *32*(2), 389–416. https://doi.org/10.1017/S0142716410000457

Lee, M., & Jung, J. (2021). Effects of textual enhancement and task manipulation on L2 learners' attentional processes and grammatical knowledge development: A mixed methods study. *Language Teaching Research*, 0(0). https://doi.org/10.1177/13621688211034640

Lee, M., & Révész, A. (2020). Promoting grammatical development through captions and textual enhancement in multimodal input-based tasks. *Studies in Second Language Acquisition*, 1-27. https://doi.org/10.1017/S0272263120000108

Leow, R. (1997). The effects of input enhancement and text length on adult L2 readers' comprehension and intake in second language acquisition, *Applied Language Learning*, 82, 151–82.

Leow, R. (2001). Do learners notice enhanced forms while interacting with the L2 input? An online and offline study of the role of written input enhancement in L2 reading. *Hispania, 84*, 496–509.

Leow, R. (2009). Input enhancement and L2 grammatical development: What the research reveals. In: J. Watzinger-Tharp & S. Katz (Eds.), *Conceptions of L2 grammar: Theoretical approaches and their application in the L2 classroom* (pp. 16–34). Heinle.

Leow, R. P. (2015). *Explicit learning in the L2 classroom*. New York: Routledge. https://doi.org/10.4324/9781315887074

Leow, R. P., & Martin, A. (2017). Enhancing the input to promote salience of the L2: A critical overview. In S. M. Gass, P. Spinner, & J. Behney (Eds.), Salience in SLA (pp. 167–186). New York, NY: Routledge. DOI: 10.4324/9781315399027-9

Leow, R. P., Grey, S., Marijuan, S., & Moorman, C. (2014). Concurrent data elicitation procedures, processes, and the early stages of L2 learning: A critical overview. *Second Language Research, 30*(2), 111–127. https://doi.org/10.1177/0267658313511979

Leow, R. P., Donate, A., & Gutierrez, H. (2019). Textual enhancement, type of linguistic item, and L2 development: A depth of processing perspective. In R. P. Leow (Ed.), *The Routledge handbook of second language research in classroom learning* (pp. 317–330). New York, NY: Routledge. https://doi.org/10.4324/9781315165080

Levelt, W. (1989). *Speaking: From intention to articulation*. Cambridge, MA: MIT Press.

Levis, J. (2018). *Intelligibility, oral communication, and the teaching of pronunciation*. Cambridge: Cambridge University Press. doi:10.1017/9781108241564

Lima, E. F. (2015a). *Development and evaluation of online pronunciation instruction for international teaching assistants' comprehensibility* (Publication No. 14561). [Doctoral dissertation, Iowa State University]. https://lib.dr.iastate.edu/etd/14561

Lima, E. F. (2015b). Feel the rhythm! Fun and effective pronunciation practice using *Audacity* and sitcom scenes (Teaching Tip). In J. Levis, R. Mohamed, M. Qian

& Z. Zhou (Eds). *Proceedings of the 6th Pronunciation in Second Language Learning and Teaching Conference*, Santa Barbara, CA (pp. 274-281). Ames, IA: Iowa State University.

Lima, E. F., & Zawadzki, Z. (2019). Teaching Tip: Improving speaker intelligibility: Using sitcoms and engaging activities to develop learners' perception and production of word stress in English. In J. Levis, C. Nagle, & E. Todey (Eds.), *Proceedings of the 10th Pronunciation in Second Language Learning and Teaching Conference*, Ames, IA, September 2018 (pp. 371-381). Ames, IA: Iowa State University.

Lindgren, E., & Muñoz, C. (2013). The influence of exposure, parents, and linguistic distance on young European learners' foreign language comprehension. *International Journal of Multilingualism*, *10*(1), 105–129. https://doi.org/10.1080/14790718.2012.679275

Llompart, M., & Reinisch, E. (2018). Robustness of phonolexical representations relates to phonetic flexibility for difficult second language sound contrasts. *Bilingualism: Language and Cognition, 1–16*. DOI: 10.1017/S1366728918000925

Long, M. (1991). Focus on form: A design feature in language teaching methodology. In: K. de Bot, R. Ginsberg, & C. Kramsch (Eds.), *Foreign language research in cross-cultural perspective* (pp. 39-52). Amsterdam: John Benjamins. http://dx.doi.org/10.1075/sibil.2.07lon

Markham, P., Peter, L., & McCarthy, T. (2001). The effects of native language vs. target language captions on foreign language students' DVD video comprehension. *Foreign Language Annals, 34(5),* 439–445. https://doi.org/10.1111/j.1944-9720.2001.tb02083.x

Marsden, E., Mackey A., & Plonsky, L. (2016). The IRIS Repository: Advancing research practice and methodology. In A. Mackey & E. Marsden (Eds.), *Advancing methodology and practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 1-21). New York: Routledge.

Marslen-Wilson, W. D., & Welsh, A. (1978). Processing interactions and lexical access during word recognition in continuous speech. *Cognitive Psychology*, *10*(1), 29–63. https://doi.org/10.1016/0010-0285(78)90018-X

Martinsen, R., Montgomery, C. and Willardson, V. (2017). The effectiveness of video-based shadowing and tracking pronunciation exercises for foreign language learners. *Foreign Language Annals*, 50: 661-680. https://doi.org/10.1111/flan.12306

Mathias, B., & von Kriegstein, K. (2023). Enriched learning: behavior, brain, and computation. *Trends in cognitive sciences*, *27*(1), 81–97. https://doi.org/10.1016/j.tics.2022.10.007

Mayer, R. E. (2005). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (pp. 31–48). https://doi.org/10.1017/CBO9780511816819.004

Mayer, R. E. (2009). *Multimedia learning* (second edition). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511811678

Mayer, R. E. (2014). Cognitive theory of multimedia learning. In R. E. Mayer (Ed.), *The Cambridge handbook of multimedia learning* (second edition) (pp. 43-71). Cambridge.

Mayer, R. E., Lee, H., & Peebles, A. (2014). Multimedia learning in a second language: A cognitive load perspective. *Applied Cognitive Psychology, 28(5),* 653–660. https://doi.org/10.1002/acp.3050

Meara, P. (2005). *LLAMA Language Aptitude Tests, V.2*. Swansea: Lognostics.

Meara, P., & Milton, J. (2003). *X_Lex, V.2.03, The Swansea Levels Test*. Newbury: Express.

Meara, P., & Rogers, V. (2019). *The LLAMA Tests V.3.* Cardiff: Lognostics. 2019.

Milton, J. (2010). The development of vocabulary breadth across the CEFR levels. In I. Vedder, I. Bartning, & M. Martin (Eds.), *Communicative proficiency and linguistic development: Intersections between SLA and language testing research* (pp. 211-232). Second Language Acquisition and Testing in Europe Monograph Series 1. Rome: EUROSLA.

Miralpeix, I. & Muñoz, C. (2018). Receptive vocabulary size and its relationship to EFL language skills. *International Review of Applied Linguistics in Language Teaching,* 56(1), 1-24. https://doi.org/10.1515/iral-2017-0016

Mitterer, H., & McQueen, J. M. (2009). Foreign subtitles help but native-language subtitles harm foreign speech perception. *PLoS One*, 4(11), 4–8. https://doi.org/10.1371/journal.pone.0007785

Montero Perez, M. (2020). Multimodal input in SLA research. *Studies in Second Language Acquisition, 42*(3), 653-663. doi:10.1017/S0272263120000145

Montero Perez, M. (2022). Second or foreign language learning through watching audiovisual input and the role of on-screen text. *Language Teaching, 55*(2), 163-192. doi:10.1017/S0261444821000501

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

Montero Perez, M., Peters, E., Clarebout, G., & Desmet, P. (2014). Effects of captioning on video comprehension and incidental vocabulary learning. *Language Learning & Technology*, *18*, 118–141.

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. https://doi.org/10.1016/j.system.2013.07.013

Mora, J. C. (2007). Methodological issues in assessing L2 perceptual phonological competence. *Proceedings of the PTLC 2007 Phonetics Teaching and Learning Conference* (pp. 1-5). London: Dept. of Phonetics and Linguistics, University College London.

Mora, J. C., & Levkina, M. (2017). Task-based pronunciation teaching and research: Key issues and future directions. *Studies in Second Language Acquisition, 39,* 381–399. doi:10.1017/S0272263117000183

Muñoz, C. (2008). Symmetries and asymmetries of age effects in naturalistic and instructed L2 learning. *Applied Linguistics*, *29*(4), 578–596. https://doi.org/10.1093/applin/amm056

Muñoz, C. (2017). The role of age and proficiency in subtitle reading. An eye-tracking study. *System, 67*, 77–86. https://doi.org/10.1016/j.system.2017.04.015

Murphy, J. M. (2014). Intelligible, comprehensible, non-native models in ESL/EFL pronunciation teaching. *System, 42*, 258-269. https://doi.org/10.1016/j.system.2013.12.007

Murphy, J. M., & Baker, A. A. (2015). History of ESL pronunciation teaching. In M. Reed & J. M. Levis (Eds.), *The Handbook of English Pronunciation* (pp. 36-65). United Kingdom: Wiley-Blackwell. https://doi.org/10.1002/9781118346952.ch3

Nation, I. S. P. (2001). *Learning vocabulary in another language* (pp. 23-59, 344-379). Cambridge: Cambridge University Press.

Nation, I. S. P. (2007). The four strands. *Innovation in Language Learning and Teaching, 1*(1), 2-13. doi: 10.2167/illt039.0

Navarrete, M. (2013). El doblaje como herramienta en el aula de español y desde el entorno ClipFlair. *MarcoELE, 16*, 75-87.

Ortega, L., Iwashita, N., Norris, J. M., & Rabie, S. (2002, October). An investigation of elicited imitation tasks in crosslinguistic SLA research. Presentation given at *Second Language Research Forum*, Toronto.

Owren, M. (2008). GSU PRAAT tools: Scripts for modifying and analyzing sounds using Praat acoustics software. *Behavior Research Methods, 40,* 822–829. https://doi.org/10.3758/BRM.40.3.822

Ozili, P. K. (2022). The acceptable R-square in empirical modelling for social science research. *Social Research Methodology and Publishing Results*. http://dx.doi.org/10.2139/ssrn.4128165

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

Pattemore, A., & Muñoz, C. (2020). Learning L2 constructions from captioned audiovisual exposure: The effect of learner-related factors. *System, 93,* 102303. https://doi.org/10.1016/j.system.2020.102303

Pattemore, A., & Muñoz, C. (2022). Captions and learnability factors in learning grammar from audio-visual input. *The JALT CALL Journal*, 18(1), 83–109. https://doi.org/10.29140/jaltcall.v18n1.564

Pattemore, A., Suárez, M. D. M., & Muñoz, C. (2020). Exploring L2 TV mode preferences and perceptions of learning. In K.-M. Frederiksen, S. Larsen, L. Bradley & S. Thouësny (Eds), *CALL for widening participation: short papers from EUROCALL 2020* (pp. 272-278). Research-publishing.net. https://doi.org/10.14705/rpnet.2020.48.1200

Pellicer-Sánchez, A. (2015). Developing automaticity and speed of lexical access: The effects of incidental and explicit teaching approaches. *Journal of Spanish Language Teaching*, *2*(2), 126–139. https://doi.org/10.1080/23247797.2015.1104029

Pellicer-Sánchez, A., & Boers, F. (2019). Pedagogical approaches to the teaching and learning of formulaic language. In Siyanova-Chanturia, A., Pellicer-Sánchez, A. (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 153–173). New York: Routledge. DOI: 10.4324/9781315206615-9

Pujadas, G. & Muñoz, C. (2017, April 19–21). *Learning through subtitles. Learners' preferences and task perception*. [Paper presentation] 2017 International Conference on Task-Based Language Teaching, Barcelona, Spain.

Pujadas, G., & Muñoz, C. (2019). Extensive viewing of captioned and subtitled TV series: a study of L2 vocabulary learning by adolescents. *The Language Learning Journal, 47*(4), 479-496. doi: 10.1080/09571736.2019.1616806

Pujadas, G., & Muñoz, C. (2023). Measuring the visual in audiovisual input: The effects of imagery in vocabulary learning through TV viewing. *ITL - International Journal of Applied Linguistics*. https://doi.org/10.1075/itl.22019.puj

Ragni, V. (2018). Didactic subtitling in the Foreign Language (FL) classroom. improving language skills through task-based practice and Form-Focused Instruction (FFI). In: L. Incalcaterra Mcloughlin, J. Lertola & N. Talaván (Eds.), *Audiovisual Translation in Applied Linguistics: Educational Perspectives* (pp. 9-29). John Benjamins.

Ramus, F., Peperkamp, S., Christophe, A., Jacquemot, C., Kouider, S., & Dupoux, E. (2010). A psycholinguistic perspective on the acquisition of phonology. In C. Fougeron, B. Kühnert, M. d'Imperio, & N. Vallée (Eds.), *Laboratory phonology 10: Variation, phonetic detail and phonological representation* (pp. 311–340). Mouton de Gruyter. https://doi.org/10.1515/9783110224917.3.311

Reed, M. (2012). The effect of metacognitive feedback on second language morphophonology. In. J. Levis & K. LeVelle (Eds.). *Proceedings of the 3rd Pronunciation in Second Language Learning and Teaching Conference,* Sept. 2011. (pp. 168-177). Ames, IA: Iowa State University.

Révész, A., & Brunfaut, T. (2013). Text characteristics of task input and difficulty in second language listening comprehension. *Studies in Second Language Acquisition, 35*(1), 31-65. doi:10.1017/S0272263112000678

Révész, A., Michel, M., & Lee, M. (2019). Exploring second language writers' pausing and revision behaviors: A mixed-methods study. *Studies in Second Language Acquisition, 41*(3), 605-631. doi:10.1017/S027226311900024X

Robinson, P. (1995). Attention, memory and the "noticing" hypothesis. *Language Learning*, 45, 283–331.

Robinson, P. (2003). Attention and memory during SLA. In C. Doughty and M. Long (Eds.), *Handbook of second language acquisition (pp. 631–662).* Oxford: Blackwell.

Robinson, P., Mackey, A., Gass, S., & Schmidt, R. (2012). Attention and awareness in second language acquisition. In S. Gass & A. Mackey (Eds.), *The Routledge Handbook of Second Language Acquisition*, (pp. 247-267). New York: Routledge.

Rodgers, M. (2013). *English language learning through viewing television: An investigation of comprehension, incidental vocabulary acquisition, lexical coverage, attitudes, and captions*. [Doctoral dissertation, Victoria University of Wellington]. http://hdl.handle.net/10063/2870

Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related TV programs. *TESOL Quarterly, 45*(4), 689-717. doi: 10.5054/tq.2011.268062

Rogers, V., Meara, P., Barnett-Legh, T., Curry, C., & Davie, E. (2017). Examining the LLAMA aptitude tests. *Journal of the European Second Language Association, 1*(1), 49-60. doi:10.22599/jesla.24

Saito, K. & Lyster, R. (2012), Effects of form-focused instruction and corrective feedback on L2 pronunciation development of /ɹ/ by Japanese learners of English. *Language Learning, 62*: 595-633. https://doi.org/10.1111/j.1467-9922.2011.00639.x

Saito, K. (2017). Effects of sound, vocabulary, and grammar learning aptitude on adult second language speech attainment in foreign language classrooms. *Language Learning, 67*(3), 665–693. https://doi.org/10.1111/lang.12244

Saito, K., & Hanzawa, K. (2016). Developing second language oral ability in foreign language classrooms: The role of the length and focus of instruction and individual differences. *Applied Psycholinguistics, 37*(4), 813-840. doi:10.1017/S0142716415000259

Saito, K., & Plonsky, L. (2019). Effects of second language pronunciation teaching revisited: a proposed measurement framework and meta-analysis. *Language Learning*, 69(3), 652–708. https://doi.org/10.1111/lang.12345

Sánchez-Requena, A. (2016). Audiovisual translation in teaching foreign languages: Contributions of dubbing to develop fluency and pronunciation in spontaneous conversation. *Porta Linguarum,* 26: 9-21.

Sánchez-Requena, A. (2017). *Audiovisual translation in foreign language education: the use of intralingual dubbing to improve speed, intonation and pronunciation in spontaneous speech* [Doctoral dissertation, Manchester Metropolitan University]. http://e-space.mmu.ac.uk/620483/

Sánchez-Requena, A. (2018). Intralingual dubbing as a tool for developing speaking skills. In: L. Incalcaterra Mcloughlin, J. Lertola & N. Talaván (Eds.), *Audiovisual Translation in Applied Linguistics: Educational Perspectives* (pp. 101-128). John Benjamins.

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 129–158.

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 and teaching* (pp. 1–64). Honolulu, HI: University of Hawai'i Press.

Schmidt, R. (2001). Attention. In P. Robinson (Ed.), *Cognition and Second Language Instruction*. New York: Cambridge University Press, 3–32.

Schmidt, R. and Frota, S. (1986). Developing basic conversational ability in a second language: A case study of an adult learner of Portuguese. In R. Day (Ed.), *Talking to learn: Conversation in second language learning* (pp. 237–322). Rowley, MA: Newbury House.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics, 10,* 209–230. http://dx.doi.org/10.1515/iral.1972.10.1-4.209

Sharwood Smith, M. (1981). Consciousness-raising and the second language learner. *Applied Linguistics*, *2*(2), 159-168, https://doi.org/10.1093/applin/II.2.159

Sharwood Smith, M. (1991). Speaking to many minds: On the relevance of different types of language information for the L2 learner. *Second Language Research*, 7, 118–132.

Sharwood Smith, M. (1993). Input enhancement in instructed SLA. *Studies in Second Language Acquisition*, *15*(2), 165–179. https://doi.org/10.1017/S0272263100011943

Sheppard, B. E., Elliott, N. C., & Baese-Berk, M. M. (2017). Comprehensibility and intelligibility of international student speech: Comparing perceptions of university EAP instructors and content faculty. *Journal of English for Academic Purposes*, *26*, 42–51. https://doi.org/10.1016/j.jeap.2017.01.006

Shook, D. (1994). FL/L2 reading, grammatical information, and the input-to-intake phenomenon, *Applied Language Learning*, 52, 57–93.

Showalter, C. E. (2018). Impact of Cyrillic on native English speakers' phono-lexical acquisition of Russian. *Language and Speech*, 61, 565–576. doi: 10.1177/0023830918761489

Showalter, C. E. (2019). Russian phonolexical acquisition and orthographic input: Naïve learners, experienced learners, and interventions. *Studies in Second Language Acquisition*, 1-23. https://doi.org/10.1017/S0272263119000585

Simard, D. (2009). Differential effects of textual enhancement formats on intake. *System*, vol. 37, 1, 124-135. https://doi.org/10.1016/j.system.2008.06.005

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency and lexis. *Applied Linguistics, 30*(4), 510-532. https://doi.org/10.1093/applin/amp047

Sokoli, S. (2018). Exploring the possibilities of interactive audiovisual activities for language learning. *Translation and Translanguaging in Multilingual Contexts*, 4(1), 77–100.

Solt, S., Pugach, Y., Klein, E., Adams, K., Stoyneshka, I., & Rose, T. (2003). L2 perception and production of the English regular past: Evidence of phonological effects. *Proceedings of the Annual Boston University Conference on Language Development.* 28.

Spada, N. (1997). Form-focussed instruction and second language acquisition: A review of classroom and laboratory research. *Language Teaching, 30*(2), 73-87. doi:10.1017/S0261444800012799

Spada, N., Shiu, J. L.-J., & Tomita, Y. (2015). Validating an elicited imitation task as a measure of implicit knowledge: Comparisons with other validation studies. *Language Learning*, *65*(3), 723–751. https://doi.org/10.1111/lang.12129

Stenton, A. (2012). Can simultaneous reading and listening improve speech perception and production? An examination of recent feedback on the SWANS authoring system. *Procedia - Social and Behavioral Sciences*, 34, 219–225. https://doi.org/10.1016/j.sbspro.2012.02.044

Stenton, A. (2013). The role of the syllable in foreign language learning: Improving oral production through dual-coded, sound-synchronised, typographic annotations. *Language Learning in Higher Education*, 2(1). https://doi.org/10.1515/cercles-2012-0009

Strachan, L., & Trofimovich, P. (2019). Now you hear it, now you don't: Perception of English regular past <–ed> in naturalistic input. *Canadian Modern Language Review*, 75(1), 84–104. https://doi.org/10.3138/cmlr.2017-0082

Suárez, M. M., & Gesa, F. (2019). Learning vocabulary with the support of sustained exposure to captioned video: Do proficiency and aptitude make a difference? *The Language Learning Journal*, 47(4), 497–517. https://doi.org/10.1080/09571736.2019.1617768

Swain, M. (1985). "Communicative competence: Some roles of comprehensible input and comprehensible output in its development". In Susan M. Gass and Carolyn G. Madden, (Eds.) *Input in second language acquisition*, 235–253. Rowley, MA: Newbury House.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*, 257–285. doi: 10.1207/s15516709cog1202_4

Sweller, J. (2005). The redundancy principle in multimedia learning. In R. E. Mayer (Ed.), The *Cambridge handbook of multimedia learning* (pp. 159–167). Cambridge University Press. doi:10.1017/CBO9780511816819.011

Tavakoli, P. (2019). Automaticity, fluency and second language task performance. In: Wen, Z. E. and Ahmadian, M. J. (Eds.) *Researching L2 Task Performance and Pedagogy*. John Benjamins, Amsterdam, pp. 39-52. https://doi.org/10.1075/tblt.13.03tav

Thomson, R. I., & Derwing, T. M. (2015). The effectiveness of L2 pronunciation instruction: A narrative review. *Applied Linguistics, 36*(3), 326–344. https://doi.org/10.1093/applin/amu076

Tomlin, R., & Villa, V. (1994). Attention in cognitive science and second language acquisition. *Studies in Second Language Acquisition,* 15, 183–203.

Trofimovich, P., Lightbown, P., Halter, R., & Song, H. (2009). Comprehension-based practice: The development of L2 pronunciation in a listening and reading program. *Studies in Second Language Acquisition, 31*(4), 609-639. doi:10.1017/S0272263109990040

Uchihara, T., Webb, S., Saito, K., & Trofimovich, P. (2022). Does mode of input affect how second language learners create form-meaning connections and pronounce second language words? *The Modern Language Journal, 106*(2), 351–370. https://doi.org/10.1111/modl.12775

Ullman, M. (2005). A cognitive neuroscience perspective on second language acquisition: The declarative/procedural model. In C. Sanz (Ed.), *Mind and context in adult second language acquisition: Methods, theory and practice* (pp. 141–178). Washington, DC: Georgetown University Press.

Vafaee, P., & Suzuki, Y. (2020). The relative significance of syntactic knowledge and vocabulary knowledge in second language listening ability. *Studies in Second Language Acquisition, 42*(2), 383-410. doi:10.1017/S0272263119000676

Van Zeeland, H. (2017). Christopher Brumfit dissertation Award Winner 2014 – Hilde van Zeeland: Four studies on vocabulary knowledge in and from listening: Findings and implications for future research. *Language Teaching, 50*(1), 143-150. doi:10.1017/S0261444816000318

Vanderplank, R. (2010). Déjà vu? A decade of research on language laboratories, television and video in language learning. *Language Teaching 43(1)* 1–37. doi:10.1017/S0261444809990267

Vanderplank, R. (2015). 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 and C. Mariotti (Eds.) *Subtitles and Language Learning* (pp. 19-40). Oxford: Peter Lang.

Vanderplank, R. (2016). *Captioned media in foreign language learning and teaching: Subtitles for the deaf and hard-of-hearing as tools for language learning.* London: Palgrave Macmillan. https://doi.org/10.1057/978-1-137-50045-8

Vanderplank, R. (2019). 'Gist watching can only take you so far': Attitudes, strategies and changes in behaviour in watching films with captions, *The Language Learning Journal, 47:4*, 407-423, DOI: 10.1080/09571736.2019.1610033

VanPatten, B. (1996). *Input processing and grammar instruction: Theory and research*. Norwood, NJ: Ablex.

VanPatten, B. (2004). Input processing in SLA. In B. VanPatten (Ed.), *Processing instruction: Theory, research, and commentary* (pp. 5–31). Mahwah, NJ: Lawrence Erlbaum.

Vaquero, L., Rodríguez-Fornells, A., & Reiterer, S. M. (2017). The left, the better: White-matter brain integrity predicts foreign language imitation ability. *Cerebral cortex*, 27(8), 3906–3917. https://doi.org/10.1093/cercor/bhw199

Webb, S., & Rodgers, M. P. H. (2009). Vocabulary demands of television programs. *Language Learning, 59*(2), 335-366. doi: 10.1111/j.1467-9922.2009.00509.x

Williams, J. N., & Paciorek, A. (2016). Indirect tests of implicit linguistic knowledge. In: *Advancing Methodology and Practice: The IRIS Repository of Instruments for Research into Second Languages* (pp. 23-42). Taylor and Francis. https://doi-org.sire.ub.edu/10.4324/9780203489666

Winke, P. M. (2013). The effects of input enhancement on grammar learning and comprehension. *Studies in Second Language Acquisition*, *35*(2), 323–352. https://doi.org/10.1017/S0272263112000903

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. https://doi.org/10.1111/j.1540-4781.2013.01432.x

Wisniewska, N. (2021). *Pronunciation learning through captioned videos: Gains in L2 speech perception and production.* [Doctoral dissertation, University of Barcelona]. DSpace. http://hdl.handle.net/2445/176592

Wisniewska, N., & Mora, J. C. (2018). Pronunciation learning through captioned videos. In Levis, J. (Ed.), *Proceedings of the 9th Annual Pronunciation in Second Language Learning and Teaching Conference*, (pp. 204-215). Ames: Iowa State University.

Wisniewska, N., & Mora, J. C. (2020). Can captioned video benefit second language pronunciation? *Studies in Second Language Acquisition, 42*(3), 599-624. doi:10.1017/S0272263120000029

Woore, R. (2018). Learners' pronunciations of familiar and unfamiliar French words: What can they tell us about phonological decoding in an L2? *Language Learning Journal, 46*(4), 456–469. https://doi.org/10.1080/09571736.2016.1161062

Yang, J., & Chang, P. (2014). Captions and reduced forms instruction: The impact on EFL students' listening comprehension. *ReCALL, 26*(1), 44-61. doi:10.1017/S0958344013000219

Yibokou, K. S. (2023). Influence of television series on pronunciation. In: Toffoli, D., Sockett, G. & Kusyk, M. (2023). *Language Learning and Leisure: Informal Language Learning in the Digital Age*. Berlin, Boston: De Gruyter Mouton (preview). https://doi.org/10.1515/9783110752441

Yibokou, K. S., & Jingand, A. (2023, July 17–21). *Examining L2 learners' pronunciation using a semi-naturalistic data collection method: What do oral diaries have to offer?* [Paper presentation]. 20th Anniversary Congress of the Association Internationale de Linguistique Appliquée (AILA 2023). École normale supérieure de Lyon.

Zabalbeascoa, P., Sokoli, S., & Torres, O. (2012). *Clipflair: Foreign language learning through interactive revoicing and captioning of clips*. Accessed January 12, 2021. http://clipflair.net/wp-content/uploads/2014/06/D2.1ConceptualFramework.pdf

Zhang, R., & Yuan, Z. (2020). Examining the effects of explicit pronunciation instruction on the development of L2 pronunciation. *Studies in Second Language Acquisition, 42*(4), 905-918. doi:10.1017/S0272263120000121

Zhang, S. (2016). Mobile English learning: An empirical study on an APP, English Fun Dubbing. *International Journal of Emerging Technologies in Learning,* 11, 4. 10.3991/ijet.v11i12.6314.

Zhang, S., & Zhang, X. (2022). The relationship between vocabulary knowledge and L2 reading/listening comprehension: A meta-analysis. *Language Teaching Research, 26(4),* 696–725. https://doi.org/10.1177/1362168820913998

APPENDICES

Appendix A.1: Reading speed test (study 1)

When astronauts first arrive at the space station, they're in awe of the views. It is the sight of our planet that takes the breath away. On board, you can get a panoramic view of Earth. But for the really exceptional views, you need to step outside for a spacewalk. One astronaut describes the experience: 'Sometimes you feel that you are on this big flying building and it's going round the world, but most commonly you feel that someone is rolling this huge ball-shaped map beneath you. You have no feeling of motion.'

Appendix A.2: Comprehension questions for each video clip in study 1 (correct responses underlined)

Clip 1) Where does Eleanor find herself when she opens her eyes?

a) Outside her therapist's office

b) In the afterlife

c) In a big hotel

Clip 2) Who made the most accurate prediction of the afterlife?

a) Doug Forcett

b) Hindus

c) Eleanor

Clip 3) Who goes to the Good Place?

a) Every U.S. president

b) A few outstanding people

c) Most artists

Clip 4) Why is Chidi upset? Because…

a) He cannot study the universe

b) Eleanor tells him she does not love him

c) Eleanor has ended up in the Good Place by mistake

Appendix B.1: Comprehension questions for each video clip in study 2 (correct responses underlined)

Clip 1

1) Where does Eleanor find herself when she opens her eyes?

a) At the doctor's office.

b) In the afterlife.

c) In a big hotel.

2) What happened to the bottle that she was holding?

a) She drank it.

b) Someone stole it from her.

c) The bottle fell on the floor.

3) What crashed into Eleanor outside the supermarket?

a) A bottle of Margarita.

b) A column of shopping carts.

c) The 9 am bus.

1) Eleanor wants to know why she is not on Earth anymore. T / F

2) Michael tells an embarrassing story to Eleanor. T / F

3) Eleanor would have preferred going to the Bad Place. T / F

Clip 2

1) What does the total value of a person's life depend on?

a) Their actions on Earth.

b) The size of their heart.

c) Their bank account.

2) Who can go to the Good Place?

a) Poets and painters.

b) Everyone except the U.S. presidents.

c) Only very generous people.

3) Why is Eleanor given a little cottage? Because…

a) She is humble.

b) She will live alone.

c) Everyone has a small house.

4) Why is Chidi happy? Because he…

a) Was in love with Eleanor on Earth.

b) Will learn everything about the universe.

c) Speaks English fluently in the Good Place.

5) What big mistake does the protagonist mention?

a) Her name is not Eleanor.

b) She does not have any memories.

c) She is not supposed to be in the Good Place.

Very few people can enter the Good Place.	T / F
Eleanor thinks she deserved going to the Good Place.	T / F
Chidi has travelled a lot in his life.	T / F
Eleanor was born and raised in Arizona.	T / F
Chidi is angry because he cannot say bad words.	T / F

Appendix B.2: Prompted narrative task 1 (study 2 and 3)

What happened to Charlie Brown?

Tell the story in the past. Begin with: *"YESTERDAY, Lucy **was** in the garden, when…"*

Look at these verbs before starting. You may use some or all of them:

- Play
- Wait
- Walk
- Look
- Follow
- Roll
- Pass by
- Try
- Sprint (correr rapido)
- Kick

Characters: Lucy (girl), Charlie Brown (boy)

Appendix B.3: Questionnaire used in study 2

Por favor, contesta todas las preguntas antes de pasar a la siguiente sección

Las preguntas marcadas con * son obligatorias

Información personal

1. Código de participante *
2. Sexo *
   - O Hombre
   - O Mujer
   - O Otras opciones
3. Edad *

Uso de la lengua inglesa

4. Idioma nativo (marca todo lo que corresponda) *
   - O Castellano
   - O Catalán
   - O Otro:
5. ¿Tienes algún certificado de inglés? *

   Si SÍ, indica su nombre y nivel, si lo sabes. Si NO, escribe "no".
6. ¿Has realizado cursos de inglés / clases particulares fuera de la escuela? *
   - O Sí
   - O No
7. ¿Cuántos meses de clases extra has tomado? *

   Si no has tomado ninguna clase, escribe 0.
8. ¿Has estado alguna vez en algún país de habla inglesa? *
   - O Sí
   - O No
9. ¿Qué país era? *
   - O Nunca he vivido en el extranjero
   - O England, Scotland, Wales or Ireland
   - O USA
   - O Canada
   - O Australia
   - O India

- Other:

10. ¿Cuánto tiempo has vivido en un país de habla inglesa? Utiliza semanas o meses *

Exposición a series TV

11. ¿Miras la televisión (series, películas o programas) en INGLÉS fuera del aula? *
    - No miro la televisión en INGLÉS
    - Miro entre 30 minutos y 1 hora cada semana
    - Miro entre 1 y 3 horas cada semana
    - Más de 3 horas cada semana
12. ¿Utilizas subtítulos? *
    - No miro la televisión en inglés
    - No uso subtítulos
    - Miro la televisión en inglés con subtítulos en INGLÉS la mayoría de las veces
    - Miro la televisión en inglés con subtítulos en CASTELLANO O CATALÁN la mayoría de las veces
13. ¿Has visto algún episodio de la serie de televisión *The Good Place* (la serie del estudio deinvestigación) FUERA DEL AULA? *
    - Sí, el primer 1 o 2
    - Sí, más de 2
    - No
    - Other:
14. Por favor, describe la regla para pronunciar la "-ed" final de los verbos pasados *

    Incluye ejemplos si puedes. Si no lo sabes y no quieres adivinar, escribe "no".

El estudio

15. Piensa en los vídeos que has visto *

Indica hasta qué punto estás de acuerdo con cada afirmación

	Totalmente en desacuerdo	Un poco en desacuerdo	Indeciso	Un poco de acuerdo	Totalmente de acuerdo
Entendí los vídeos	O	O	O	O	O
Los vídeos fueron divertidos	O	O	O	O	O
He leído los subtítulos	O	O	O	O	O
Aprendí algo de pronunciación en inglés gracias a los vídeos	O	O	O	O	O
Aprendí algo de gramática o vocabulario en inglés gracias a los vídeos	O	O	O	O	O

16. ¿Había algunas letras resaltadas en amarillo en los vídeos? *

- O Sí
- O No
- O No sé

17. SI SÍ: ¿Qué tenían en común las letras resaltadas?

Puedes adivinar si no lo sabes.

18. SI SÍ: ¿Fue útil el resaltado para aprender o fue una distracción?

¿Algún otro comentario?

Appendix B.4: Stimulated recall protocol used in study 2

Switch on Marantz recorder.

State participant number.

"Now the video clips will be played again for you. Circles and lines will appear in the areas your eyes were focusing on and moving over during the viewing. Please tell me to pause the video whenever you remember *why* you were paying attention to something on the screen. I want to know what you thought about at the time of the viewing. Sometimes I will also stop the video to ask you questions.

Remember that I can see *where* you were watching, but I don't know *why*. What I'd like you to do is to tell me what you were thinking during the viewing, not now that you are seeing the video again. Please feel free to point at the screen (*point at screen*) if it helps you explain. Don't worry if you can't remember, but do not invent an answer.

Do you have any questions?"

(Answer possible questions)

"Let's begin. Please say *stop* or *pause* whenever you remember why you were looking at something on the screen."

(*Start the video)*

(Pause the video after a few seconds to check understanding of instructions, then whenever s/he tells you to do so.)

(Pause clip when some of this happens)

- Unenhanced/enhanced target words
- Very long fixations
- Re-reading (regressions)
- (Occasionally, other areas to avoid giving away aim.)

(As appropriate, ask one of the following questions)

- Why were you watching this area?
- What made you watch this area for a long time?
- What made you go back to this area?

Appendix C.1: Complete transcript of stimulated recall interviews in study 2 (622 turns)

INT: Interviewer (author); example of participant coding: S_01 is participant 1.

Event	Turn	Transcript	Analysis
1	1	INT: Ok so here you were watching the subtitles, the face	
	2	**S01**: Yes, here I was watching ehm the person who were talking.	
2	3	INT: Here you finally go back to the subtitles... Why do you think you were skipping until now, and now you went back to watching the captions?	
	4	S01: I read the subtitles because I didn't understand what he was saying and I wanted to have a general idea of what he was saying and [unintelligible]	RECALL
	5	INT: So were you able to follow the dialogue just by watching the faces of the participants?	
	6	S01: Sometimes yes, but there were other times that no.	
	7	INT: So you went back to the subtitles once and then you went back to the face and then you went back again.	
	8	S01: I think because I understand the words that he says here, but here and here not.	
	9	INT: Ok so you understood the first line of the subtitle while they were speaking, but then there was something else, so you went back to the subtitles to see what that actually was, right?	
	10	INT: Why, because you couldn't hear the words or...?	
	11	S01: To... To feel like... To know what he was saying.	RECALL
	12	INT: So to make sure that the word was the word that you were listening?	
	13	S01: Yes.	
	14	INT: You wanted to see if it was the same you were reading.	
3	15	INT: All this dialogue, was this easy for you? Because you never read the subtitles.	
	16	S01: Yes.	
	17	INT: These are like, short sentences...	
	18	S01: Yes, yes. These are short and with understandable words.	
4	19	INT: And then here she said ["Did I have a purse?"] and you were watching the subtitles ... Can you remember why?	

	20	S01: I think because I was watching the decoration of the room and then she talks, so I was despistada [distracted] and I did not know what she said so I read this to make sure.	RECALL
	21	INT: Ok, so you brought back the focus on the dialogue by reading the subtitles.	
	22	S01: [no audible answer]	
	23	INT: Do you know what purse means?	
	24	S01: No	
	25	INT: Ok, so you are trying to read... Even when you do not know a word, you would still read the subtitles to try and understand more. Right?	
	26	S01: [no audible answer]	
5	27	INT: You were not reading the subtitles, right? Not that much. You were trying to watch Michael's face. Why do you think you fixated on this point particularly?	
	28	S01: Because he was talking and he was saying important things because he was saying why they are in the good place.	RECALL
	29	INT: And could you understand what he was saying?	
	30	S01: Ehm... Yes.	
	31	INT: These are easy words for you, so you wouldn't read the subtitles?	
	32	S01: Ehm... yes. I think that I understand it all.	
6	33	INT: Ok so you read "rippled out over time"	
	34	S01: Yes but...	
	35	INT: Do you know what that means?	
	36	S01: I don't know why.	RECALL
	37	INT: You have no idea. So let's say you were watching for meaning, mostly, right? You were trying to understand and then you would read the subtitles if you didn't understand.	
	38	S01: [no audible answer]	
	39	INT: Right, I see that you are fixating on this word, why did you do it?	
	40	S01: Because it was in yellow.	
	41	INT: Because it was in yellow. Ok. So would you have read it or would you not have read it if it wasn't in yellow?	
	42	S01: No, I wouldn't have read it... And the next words that are in yellow, I think that they all finish in -ed. And that's because. I think. [she means "that's why"]	RECALL

	43	INT: Right, ok. So even when you read a subtitle like this, what were you focusing on, mainly? Were you still trying to listen, were you mostly reading? What were you doing, can you remember?	
	44	S01: I was listening and... I was listening.	
	45	INT: Right, so you were trying to listen mostly. So this popped up in yellow and caught your attention, but that's it.	
	46	S01: Yes	
7	47	INT: Ok, so this time you completely ignored [the highlighted word]. You were just watching the person here. Sometimes you would read these captions that are coming up in color?	
	48	S01: Yes, but I... I didn't finish reading them.	
	49	INT: Yes, sure because here there are a lot, but in the beginning there were fewer so you were [reading them] but you are not reading this subtitle. Do you think you are doing this consciously or is it just the way you watch videos normally?	
	50	S01: Normally I watch the person that is talking and when I don't understand what they say, sometimes I put the video more slowly and I put the subtitle.	
8	51	INT: Why do you think you were watching so much [the center of the screen]?	
	52	**S02**: Because at the beginning, I don't normally read the subtitles, so I was watching my first impression of the faces of the characters	RECALL
	53	INT: Right, can you think about why you were not reading the subtitles but you were watching the faces?	
	54	S02: Well, because I understood what they were saying.	RECALL
	55	INT: Right, so if you understand what they say, you don't read the captions, or try not to.	
	56	S02: When I understand I normally don't read it.	
	57	INT: Is this your usual behavior, or are you doing it just now? How do you behave at home, in the same way?	
	58	S02: Yes, it's my usual behavior.	
9	59	INT: Here you were reading the captions. Why do you think that was?	
	60	S02: This sentence, I didn't know what he was saying so I tried to read it because I don't know what it means, "bent down to pick it up". Well, now I know, but at the moment the video was playing I didn't understand it, so I read it.	RECALL

	61	INT: And did reading help you?	
	62	S02: Yes	
	63	INT: Oh ok, so now you have more of an idea what it means, right?	
	64	S02: Yes.	
10	65	INT: Here you were on the subtitle area, then you watched his face, and then you moved back. Why do you think that is?	
	66	S02: Well, because "a long column of shopping carts" at the beginning I didn't know what he was trying to say. But then I understood…	
	67	INT: Which part was unclear?	
	68	S02: "A long column of shopping" ok, but "shopping carts"… The word I did not understand it. Now I do, because it's like when you go to the supermarket, those [unintelligible] where everything is inside, but at the beginning I was not understanding why she died, so I was trying to focus a bit on the story and not... on the faces and everything	
	69	INT: That's great, ok, thanks. So when you heard him saying "a long column of shopping carts", was it hard for you to distinguish the words, or did you understand the single words?	
	70	S02: Yes, the words yes. I didn't know the meaning.	RECALL
	71	INT: Right, but you didn't know what that means.	
11	72	INT: So here… what were you watching?	
	73	S02: When he said that this was not like the earth, that it was the afterlife, I started think "wow, this is really shocking". So I tried to see how it was the idea of the afterlife, so I started watching [around] and I thought that it was really like the earth, it looked like earth. So I started watching the decorations of the place.	RECALL
12	73	INT: Can we just watch this again? Try to see what happens here. So she's talking and then, you're reading the subtitle. What happened?	
	73	S02: Well because I was trying to read what is under that [unintelligible], but then when she said "I don't have a purse", at first I did not understand the word, but then I remembered what it meant. So, after that I started looking everywhere. And like... I was not paying attention to what she was saying, and when she said that she was moving so quickly, I thought "maybe it's something important" and I have not paid attention, so I read the subtitle.	RECALL
	73	INT: Right, and the subtitle had disappeared.	
	73	S02: Yes.	

	73	INT: So… it was too late.	
13	73	INT: What are you doing here?	
	74	S02: I was watching what it was written on the screen, like "buy trashy magazine" and "eat a sandwich" because I think that… I thought that they were so strange examples of what are good things and bad things, at the beginning.	RECALL
	75	INT: So it's like he's giving a presentation and you have some captions of the presentation so you're watching them, ok… then there's too much [hundreds of short captions appear on the screen at this point of the video]	
	76	S02: Yes, and I started looking everywhere.	
	77	INT: So you can't read all of that…	
	78	S02: At some point there's like "commit genocide" and I was like "wait what?"	
14	79	INT: Tell me what happens here.	
	80	S02: Well, I have seen that the subtitle had a yellow word and I was surprised. At the beginning of the recording it said that this could happen, but it had never happened before so my eyes focused on the word in that moment.	RECALL
	81	INT: Ok, and did that change the way that you were reading or listening to the video?	
	82	S02: No, not a lot.	
	83	INT: You just watched the word once, that's it.	
	84	S02: Yes.	
	85	INT: But that made you read the rest of the subtitle too, right?	
	86	S02: Yes, a bit, cause I read fast so I read all of it in a moment.	
15	87	INT: Here you were watching the captions?	
	88	S02: Yes, the part "in a broken down boat" I didn't understand what he was trying to explain, so I was reading so that maybe I could understand it better.	RECALL
	89	INT: So again… when you listen, can you actually make out the single words or not?	
	90	S02: Yes.	
	91	INT: When you listen to something like "in a broken down boat", can you say that that's a "broken - down - boat"?	
	92	S02: Yes.	
	93	INT: …but you don't know what it means so you try to find the meaning.	

16	94	INT: Can you explain what happens here?	
	95	S02: Yes, I'm like comparing a bit the houses and I'm thinking that the personality of the people can affect the Good Place, so I thought that it was really interesting that there could be such a big house with one house that is so small.	RECALL
	96	INT: So you can see your fixations are on the big house, then on the small one, and then you go back to the big one.	
	97	S02: Yes, because the big house is so big and shocking.	
17	98	INT: There you go with another yellow word… but you ignored it completely. Was it even in your peripheral vision [I explained this term to her] or you just didn't see it?	
	99	S02: Well, yes I saw that there was something in yellow but I thought that she said "decorated", so… That was one word that was actually in the other exercise that we've done before, so I thought "well maybe, that's why the yellow words are in yellow", cause they were like... vocabulary that we had in the other test?	RECALL
	100	INT: So you noticed that it was in the test, but for you there was no problem with that word so you just ignored it.	
	101	S02: No, because I already know the definition and everything, so…	
18	102	INT: What were you watching here?	
	103	S02: When I watched the first recording, there were some [comprehension] questions, so I thought "this is a question that will be asked later: did she live in Arizona, did she move somewhere else?" So I started focusing on the words, so that I could remember them better.	RECALL
19	104	INT: [Mixed English-Spanish] Do you always read subtitles at home when you watch videos, or is this uncommon?	
	105	**S04**: I read the subtitles at home because I'm not really good at English.	
	106	INT: Do you always read every single word, or…?	
	107	S04: Sometimes I look at the images because I already know some words.	
20	108	INT: In general, did you understand what they were talking about?	
	109	S04: More or less.	
	110	INT: More or less. Ok, so here for example, did you fixate these words because you did not understand it or... Why?	
	111	S04: Because I did not understand them.	RECALL
	112	INT: Right, so you were trying to read everything.	
	113	S04: [no audible answer]	

	114	INT: What happens with listening? Can you hear what they say?	
	115	S04: It's hard for me.	RECALL
	116	INT: What is hard for you?	
	117	S04: Understanding what they say in English.	
	118	INT: Is it hard to understand each word…? Because when one speaks, they say many things at the same time, right? Is it hard to understand where each word belongs in the sentence, and does the subtitle help with that?	
	119	S04: Yes.	
21	120	INT: Ok, here they are speaking very fast. Could you read everything or was it too fast?	
	121	S04: Not everything.	
	122	INT: Right. And what you managed to read, did it help you? I mean, does reading help you understand the overall meaning?	
	123	S04: Yes.	RECALL
22	124	INT: Here there were no subtitles, did you understand what they said?	
	125	S04: Yes, she is asking if she's in heaven or hell.	
	126	INT: Perfect. So here when there were no subtitles, you looked at the faces.	
	127	INT: Ok, since you always look at the subtitles when there are subtitles, I'm going to ask you a question. Does looking at the faces of the speakers help you understand as well?	
	128	S04: Sometimes.	
	129	INT: But it's not as effective, right? Because you don't do it as much.	
	130	S04: Mhm [unclear if affirmative or negative].	
23	131	INT: Here you understood what was happening right?	
	132	S04: Yes.	
	133	INT: You looked at the big house, then at the small one.	
	134	S04: Yes.	
	135	INT: So you were actually understanding what was going on, you were comparing the two houses, meaning like "so some people have this house and she has this one", ok.	
24	136	INT: So what happened here?	
	137	S04: He talks that the house are more decorate for she.	RECALL
	138	INT: So he's talking about the decoration of the house, and also this word pops up in yellow, and you were watching the word before, you were	

		reading the subtitle, but then your eye goes to the house, and then you go back to the words. Is this because it popped up in yellow?	
	139	S04: Yes	
	140	INT: Ok, why were you watching it? Why do you think it was important? Why do you think I put it in yellow?	
	141	S04: [In Spanish] Why it was in yellow?	
	142	INT: Why do you think I put it in yellow?	
	143	S04: Because, I think, the word describes the house, I mean, because it is decorated.	RECALL
	144	INT: Ok, so you think that I'd like you to focus on the meaning of some words, and then I put them in yellow?	
	145	S04: Yes.	
25	146	INT: So here you have basically read the entire subtitle, but when the yellow word pops up, you read it and then you go back and read it again. Why did you do this?	
	147	S04: Normally, I want to read things twice to understand them better.	RECALL
	148	INT: So if you don't understand something, sometimes you go back and read it again.	
26	149	INT: What were you fixating on here, what can you say?	
	150	**S05**: First of all, I saw his face because, I don't know, I always do. And then I went straight to her face too, just to know how they were, and then to the words I guess, because I... always do. I don't know, I don't even want to, but if I have something to read I always think of reading first so... not because I did not understand what they were saying.	RECALL
	151	INT: Ok, so it's kind of automatic but you understood what [they said].	
	152	S05: Yes, yes. Exactly.	
27	153	INT: So yes, you were watching [all the subtitles]	
	154	S05: Yes, that's what I always do.	
	155	INT: Does all of this make sense to you? Do you think there was a specific reason why you were reading the subtitles at some point?	
	156	S05: No, I don't know why… I always need to have subtitles because, I don't know, [unintelligible] necessary.	RECALL
	157	INT: Ok, so here for you [the dialogue] was easy.	
	158	S05: Yes. I…	
	159	INT: You could just listen and [understand].	
	160	S05: Yes.	

28	161	INT: What about this one, did you understand everything here?	
	162	S05: Yes, it's that, when they talk in this video, for example, I really understood almost everything that they said but, for example, some words I can't hear them properly so I read them to know if I can write them.	RECALL
	163	INT: So you would say that sometimes it's hard to distinguish a specific word in the stream…	
	164	S05: Yes, yes.	
	165	INT: Ok, so you use subtitles for that.	
	166	S05: Yes.	
29	167	INT: So you were reading, and then you were going back to reread something, why?	
	168	S05: Yes, because here in this situation sometimes when they explain something that happened in the past, like here with the shopping cart, I don't process what they are saying or talking about, so I try to imagine the situation, trying… well, reading again what they were saying, because... I don't know, I wasn't focusing on what they were saying so I try to reread it.	RECALL
	169	INT: Right, so do you think you understood the words while you listened?	
	170	S05: I think I understood... all the words.	
	171	INT: Ok, but you couldn't imagine…	
	172	S05: No, I couldn't imagine, I was… I was…	
	173	INT: ...like you couldn't really connect it to a concept.	
	174	S05: Yes, exactly.	
30	175	INT: That's the same, no?	
	176	S05: Yes	
	177	INT: Look. What are you rereading, this?	
	178	S05: This, "rolled" and "plowed" [incorrect pronunciation]. Sometimes there's a word that I don't usually use when I'm talking, so I was trying to imagine what they were talking about.	RECALL
	179	INT: Ok, so you were trying to find out the meaning, or you were trying to understand the context?	
	180	S05: Both, I guess. Maybe I didn't remember what that was… what the meaning was.	RECALL

31	181	INT: What was happening here? You see, you read [the first line of the subtitles], then you go back to the face, and then you read the second one.	
	182	S05: I don't really know, I always see the expressions of people who are talking, because… it's something I always do, but the words... I understood all of them, so I don't know why I was rereading them. I don't really know.	RECALL
	183	INT: So, based on what you were saying before, in a way, you were attracted by the subtitles…	
	184	S05: Yes.	RECALL
	185	INT: ...because they were, you know, appearing…	
	186	S05: Yes.	
	187	INT: ...but you also wanted to look at his face.	
	188	S05: Exactly.	
	189	INT: So you're trying to do both quickly, and that's the result?	
	190	S05: Yes, that's what I always do.	
32	191	INT: Let's see what you were looking at: First [his face], then you read the caption quickly, then you go back to his face, then you go back to the caption and then you start watching this? Why do you think all of this happened, can you give me some insights?	
	192	S05: I don't know, I guess because of their colors… because they kinda glow, and they attract my… my…	RECALL
	193	INT: Your attention.	
	194	S05: Yes, exactly.	
	195	S05: And also, when some subtitles appear, I just have the instinct to read it.	
33	196	INT: Right, but despite this, when I put the yellow word there, you didn't watch it, at all.	
	197	S05: No, I saw it but…	
	198	INT: You ignored it completely.	
	199	S05: Yes, I don't know why.	RECALL
34	200	INT: Ah ok, yes. Here you did notice [the yellow word]. Because you were already watching the subtitle, right?	
	201	S05: Mhm [affirmative].	
	202	INT: But what happened here? So you read everything and then you went back to this word and then to the woman. What happened?	

	203	S05: Because I… didn't really read them properly. I just… I was hearing the words and I knew that they were right, of course [I guess she means that she understood the auditory forms of words correctly because they were easy], so I just went back to "lived" because I didn't read this word properly. So I didn't find it special but I just saw it was yellow... The other one I don't know why I didn't see it.	RECALL
	204	INT: Cool, it's interesting. So to sum up, your attention is mainly focused on listening, and so while you are listening you are able to just skim through the subtitles quickly…	
	205	S05: Yes.	
	206	INT: ...but if you see something weird, like a yellow word, then you will focus on that word.	
	207	S05: Exactly.	
	208	INT: But also in the listening or just in the subtitles?	
	209	S05: Both I guess.	RECALL
	210	INT: Ok, so it also creates a connection or something.	
	211	S05: Mhm [uncertain].	
35	212	INT: So, can you see what you did here? Look at the dots. What happened?	
	213	S05: Well, I guess I was comparing both houses of course, but I don't know why. So a bit how this house was, and then the other too.	RECALL
	214	INT: Yes, it's interesting so you were basically connecting what you were hearing and reading to the image.	
	215	S05: Yes, exactly.	
36	216	INT: There you go, so is this the same as before?	
	217	S05: The same, exactly the same. It's not because I couldn't hear this time, I just read them normally, and then…	
	218	INT: Quickly?	
	219	S05: Yes, I don't read word by word, I just hear, I see if it's right and then I go back again to see what word that is.	RECALL
37	220	INT: What are you focusing on in this scene?	
	221	**S06**: Well, I already saw the series, so that's why at the beginning I was looking like... that I already saw it, and I tried not to look at the subtitles to try to understand English, so I was trying to look at their mouths to, like, see if it could help me. I tried to look at what was going on, actually.	RECALL
	222	INT: That's great.	

223 S06: ...Actually at the scene, like the details that I didn't see when I saw it the first time.

224 INT: This is interesting, I'm going to ask you some follow up questions on what you just said. Does watching the faces help you understand?

225 S06: Well, yes... I usually watch TV series in English, cause it gets on my nerves when they are speaking Spanish, right? But I see in their mouths that they are vocalizing in another language…

226 INT: So they are dubbed.

227 S06: So they are dubbed, right? That's why… And I want to learn English, so it's kind of helping me to see how they are saying something, if they are happy or…

228 INT: ...And their emotions. Ok, sounds good.

229 S06: Yes.

230 INT: So this is the way that you would normally watch a video, when…

231 S06: Yes.

232 INT: You said, mostly because you have watched it already, so this is one thing, but also because this is your normal behavior, right?

233 S06: Yes, yes.

38 234 INT: Ok, so basically you have seen [from the eye-tracking visualization] that you never read the subtitles, right?

235 S06: Sometimes, yes.

236 INT: But I see here that… at some point you go back, no, [I meant] sometimes you go to the subtitles, for example here, right? Can you remember why that happened, here specifically?

237 S06: Well, I don't know, because like… I was a bit nervous, I actually have to say. RECALL

238 INT: Right.

239 S06: And you know, like, "where should I look?" so… and also, the subtitles sometimes, when I watch TV series, they get my attention and I sometimes look at them. That's basically why. RECALL

240 INT: Right. So there was no specific thing in some of these lines…

241 S06: No.

39 242 INT: So here for example, what were you focusing on?

243 S06: Well, I don't know, I'm not really sure, but I think I was… because I remember that this man [she's pointing at a picture on the wall in Michael's office] was a really good man in the series, right? So it's just a RECALL

		detail that I remembered that when I first watched it I didn't see it, so when I rewatch a series I try to focus on...	
	244	INT: on the details.	
40	245	INT: So here when you see this word come up and it becomes yellow...	
	246	S06: Yes, it looked [unintelligible speech] and it was brighter, and…	
	247	INT: …you looked at it.	
	248	S06: …I saw that before, before it said like "there are going to be some words in yellow"…	RECALL
	249	INT: Right.	
	250	S06: …so I saw something in yellow and I thought "I… probably have to see it" because it's like important…	
	251	INT: Ok, so you were thinking more about what you should do for the task than what actually helped you.	
	252	S06: Yes.	
	253	INT: Ok, why do you think this specific word was highlighted?	
	254	S06: Cause it was in the past and we did all the tasks [she means the pronunciation tests] in the past, I don't know how [unintelligible]	RECALL
41	255	INT: Finally you're reading some subtitles. Why do you think this one specifically?	
	256	S06: I don't remember but probably because some words I didn't understand and it went too fast for me, maybe.	RECALL
42	257	INT: [Here] he says "and you're going to spend eternity together", right? And Eleanor here [turns to look at] these guys, and you specifically focus [your gaze] on two guys… that could be her soulmate. You didn't see anything else.	
	258	S06: Yes, I don't know why. Maybe because I knew he was going to be a character in the series…	RECALL
43	259	INT: Why do you think you were watching [the subtitles] and then you were going back here and then you were reading it again? You see the [fixations] go up and then down, why?	
	260	S06: I don't know, probably might be because I didn't understand that phrase…	RECALL
	261	INT: "Went to town?"	
	262	S06: Well, no, I understand what it means… I don't know. Maybe I was just trying to focus on what… because I didn't hear [unintelligible] because I was very nervous so I wanted to know what he was saying.	

	263	INT: Right, because I mean "go to town" is a difficult expression in English, so I would expect that if you don't know the meaning then you would try to understand what it means…	
	264	S06: Yes, yes.	
	265	INT: but also… I'm thinking aloud, which of two options: maybe you couldn't really connecting what you were listening to what you read, and that's why you went back, was it that way?	
	266	S06: I'm not sure, maybe I was trying to connect the dots… cause he was speaking a little bit fast and, like, I wanted to take all the data because they are going to ask me later [in the comprehension questions].	RECALL
	267	INT: Right, so do you think there could be a problem of segmenting whatever you hear into words, like going-to-town or do you think you were trying to understand what the whole sentence means?	
	268	S06: I think I was trying… Because when I hear someone speaking English I usually understand him, like, I don't need repetition, so I was trying to put all together.	
	269	INT: Right.	
	270	S06: I was trying to… as I said before, connect all the dots and create a phrase to know what he was saying, because I heard like "Disney" and "town" but I didn't know what… cause I was really nervous… so I heard "town", "Disney" and "went" and, like, not in that order maybe, but I wanted to know the phrase.	RECALL
	271	INT: Right, so you heard each one of these words… you heard them?	
	272	S06: Yes.	
	273	INT: …ok, but you wanted to connect them.	
44	274	INT: I can see that you are watching [Eleanor's] eyes and her mouth, basically, does this help you? Why do you watch her mouth, does this help you when you're listening?	
	275	**S07**: I don't really know, I mean, maybe because she vocalizes, but…	RECALL
	276	INT: Ok, so you're not very clear on that, let's move on.	
	277	S07: I suppose that it's because when you are listening to someone, you usually watch their eyes…	
45	278	INT: So again you're watching their eyes… so you're trying to read their emotions or something?	
	279	S07: Yes!	
	280	S07: Then I also liked the hair, like the curl in her hair...	

	281	INT: So you were watching like…	
	282	S07: This thing [points at curls in Eleanor's hair]. This is like… perfect.	
	283	INT: Oh, ok. So for some reason, at some point you were focused on her hair, ok.	
46	284	INT: Ok, so, you were never watching the captions, but at some point you do and I'm going to ask you, why did you read this one in particular?	
	285	S07: Because it was a noun or a phrase that was new to me, and I wanted to know what he was saying.	RECALL
	286	INT: Sure. Is it because you couldn't hear the single words, or is it because you knew the words but you didn't know what they meant, or...?	
	287	S07: Well, I knew that it was like the noun of something that he was saying, so for example like a pen is "Bic".	
	288	INT: So it was a brand.	
	289	S07: Yes, so I wanted to know exactly what the brand was like.	
47	290	INT: And then you are actually reading this, is there any specific reason?	
	291	S07: Yes, I think I didn't listen to it well.	
	292	INT: Ok, so you couldn't hear the single words, so you wanted to know what they were, or...?	
	293	S07: Yes, well, cause I heard them but then I wanted to know that I was right, I wanted to check and also to see if I… if I got them correctly or not.	RECALL
	294	INT: Right, so you heard "long column of shopping carts", which was just an entire sentence, and you wanted to see whether…	
	295	S07: Yes, but I… I had my doubts in "shopping carts", I think so.	
48	296	INT: So you read "rolled out of control" and then "and plowed right", and then you went back a little bit, does this have to do with…	
	297	S07: Yes, I'm not used to these words, so…	
	298	INT: Ok, so you don't know what it means, or you don't know how it sounds?	
	299	S07: Both, but then from the context I could…	RECALL
	300	INT: You could understand anyway.	
	301	S07: Yes.	
49	302	INT: Why were you watching here, specifically?	
	303	S07: Cause I wanted to see the decorations and also, these two are big things, and also the paintings had some… I don't know, it's like... similar to their faces.	

	304	INT: Right, cause there's the painting of Doug Forcett here, so you're watching another face, but this is just a random painting, so you're saying that because they're big, they attract your attention more than anything else.	
	305	S07: Yes.	
50	306	INT: So you read the whole thing, you get until "good", then you go back to "simply put", do you remember why?	
	307	S07: Maybe cause I… cause I finished reading and then I… I don't know.	RECALL
	308	INT: Do you know what "simply put" means? Cause it's not a very common expression, so maybe you were going back…	
	309	S07: Yes, well, I think so. It's like… I don't know how to say it. Can I say it in Spanish?	
	310	INT: Yes.	
	311	S07: [translates the caption into Spanish]	
51	312	INT: So you're reading here, then you're watching symbols, why do you think you're watching these things? The words, and the symbols, and then Michael again?	
	313	S07: Cause they're something new, before it was only him speaking, then they put some symbols that are moving…	RECALL
	314	INT: So they pop up…	
	315	S07: Yes, it attracts your attention.	
52	316	INT: Now you went back to the subtitle, because you saw this in yellow, but was there a specific reason why you read it, or just because it was in yellow?	
	317	S07: Yes, I think because this [she meant one of the words popping up on the screen, outside the caption area, during Michael's presentation] was very big and then this [another word] was also very big and this [the yellow word in the captions] was close, and it was like… different, like yellow and underlined.	RECALL
	318	INT: Right, so when you read it, what did you think? Did you think that I put it there for a reason? What reason do you think I had?	
	319	S07: Oh, I though that it was something because of the… of the subtitles, like nothing specific.	RECALL
	320	INT: Right, ok, ok.	
	321	S07: Oh, and maybe because also of the -ed, right? It was in the past… but I don't know.	RECALL

53	322	INT: I see here that you read the caption, "they went to town", no, you don't. You don't read the caption, you watch it quickly, then you go back to Michael, and then you go back here as if you're looking for something, but the caption is not here anymore. What happened, did you try to read it or not?	
	323	S07: No, I think I was waiting for the next subtitles, because this part was a bit complicated, because there were a lot of not [unintelligible speech] people, like Walt Disney, Picasso…	
	324	INT: So… the names.	
	325	S07: Yes, the names. And I thought it was like… phrases that were not so easy to understand… just listening to them, and I was waiting for the next subtitles.	RECALL
	326	INT: Right, ok.	
	327	S07: Cause I did understand this, without having to read it, and then maybe the next phrase or the next thing he was going to say could be more difficult, so…	
	328	INT: Ok, so the overall difficulty increased, and you were trying to... anticipate what he would say next.	
	329	S07: Yes.	
	330	INT: Did you understand "went to town", though? Could you actually hear "went to town" in this scene?	
	331	S07: Ehm… yes…	
	332	INT: Cause this is very complicated, no?	
	333	S07: Well, yes, but I didn't think of it a lot…	RECALL
	334	INT: Ok, so at the time you were not thinking about this.	
	335	S07: Mhm, No.	
	336	INT: Ok, so… [reading captions] "spend eternity"…	
	337	S07: Ah... Cause I didn't understand broken… "broken down boat"	RECALL
54	338	INT: So, I think again, you didn't know the context, you didn't know who "Florence Nightingale" is, I guess, so you were looking for whatever he was saying, but you missed [the caption]	
	339	S07: Yes.	
55	340	INT: So I see you're kind of reading this subtitle a little bit, do you remember why?	
	341	S07: Ehm… yes… [reads caption under her breath] no, I don't.	RECALL
	342	INT: No, ok. But "taking people off death row", is difficult no?	

	343	S07: Yes, this was… difficult and new.	RECALL
	344	INT: That's it. But you were trying to understand the meaning, right?	
	345	S07: Mhm [affirmative].	
	346	INT: You could listen to the single words, when they said them?	
	347	S07: No, I think no. I understood that she was a lawyer, but then the next words I didn't quite… hear them or couldn't listen to them.	RECALL
	348	INT: ...So you went back to the subtitle.	
	349	S07: Yes.	
56	350	INT: So what were you doing here? ...You were watching...	
	351	S07: He was talking about the decorations, so I was…	RECALL
	352	INT: The decoration, ok. And then this yellow word popped up, and what did you focus on, just... you read it and moved on, that's it?	
	353	S07: Yes.	
57	354	INT: So you were... reading here?	
	355	**S08**: Yes.	
	356	INT: Ok, why do you think you were reading that?	
	357	S08: To... know what he says, because it's interesting. I don't know.	RECALL
	358	INT: Right. To help you understand?	
	359	S08: Yes.	
58	360	INT: So where are you watching, can you tell me why?	
	361	S08: Always when I see films in English, always I see the subtitle to learn the vocabulary, and it's because… I don't know.	RECALL
	362	INT: So usually, even at home, you would read all the subtitles or you try to read all of them.	
	363	S08: Yes, yes.	
59	364	INT: So you are reading this sentence, you read "long column of shopping carts" and by the time you arrive here, then you go back to "long column", you read it again.	
	365	S08: Yes.	
	366	INT: Why?	
	367	S08: I don't know, but I think because… I don't understand or something like that. I don't remember well, but I think that.	RECALL
60	368	INT: Look at the dots here. See how big [this one is], why, because you were fixating a lot, you were watching [this point]. Why?	
	369	S08: Eh… of the verbs, because I not so good and if it's something… it's some verbs and this… grammatical verbs and this, I focusing more than	RECALL

in the film [she meant the action on-screen]. I don't know but I always do that.

370 INT: Ah ok, so you always do it.

371 S08: Yes.

372 INT: You're not doing it just because you're watching here…

373 S08: I don't know.

374 INT: Because I give you the test…

375 S08: I don't know, but I always read these… RECALL

376 INT: So I think what you're trying to say, correct me if I'm wrong, is that verbs carry a lot of meaning, and you want to know the meaning, so you read the verbs more carefully.

377 S08: Yes.

61 378 INT: Here you were trying to read the captions but you were also trying to look at these symbols. Why do you think that happened?

379 S08: I think because, so... pictures, well... many things… and I think that I want to look all the things.

380 INT: Right so, you can say something in Spanish if you don't know how to say specific words, so…

381 S08: Ok.

382 INT: So you can say that again. What happened?

383 S08: Ok, I think I want to try to look at the subtitles, and the picture that this man says, and I see all… [struggles]

384 INT: [We switch to Spanish and I sum up my previous remarks] What were you focusing on?

385 S08: Both things [subtitles and images].

386 INT: At the same time?

387 S08: Yes.

388 INT: And could you understand it anyway, despite having two things to focus on?

389 S08: Well, yes… yes.

390 INT: Ok. Were you listening as well, or were you only reading the subtitles?

391 S08: Two things. I don't remember well, but yes. I think I listened but I was paying attention to the subtitles. RECALL

392 INT: So are the subtitles the first thing you focus on?

393 S08: Yes.

	394	INT: You are reading first…	
	395	S08: Yes.	
	396	INT: …and then you also pay some attention to the listening.	
	397	S08: Yes, yes.	
62	398	INT: So when she says house, you look at the house?	
	399	S08: Yes.	
	400	INT: Right? Cause if you see a word then you see what it means, you try to look for that in the image?	
	401	S08: Yes.	
63	402	INT: Basically, you were watching the faces, so you were listening mainly, I think?	
	403	S08: Yes.	
	404	INT: And here, you see this yellow thing appear here, and you see that your eyes move down, why?	
	405	S08: Because I see the girl and I think that in the [comprehension] questions, questionnaire, will be these verbs and I think that, if I see more… I see well, I will remember to put in the questionnaire.	RECALL
	406	INT: Right, so you were trying to answer the questionnaire well…	
	407	S08: Yes.	
	408	INT: …so you think those yellow words… were there to help you.	
	409	S08: Yes.	
64	410	INT: What were you looking at?	
	411	**S09**: I was reading these words and looking at the sofa when she stands up.	
	412	INT: Why do you think you're reading this?	
	413	S09: Because it's big.	
	414	INT: Right, cause it's something big and green on the screen.	
65	415	INT: Alright, I see you were watching a lot... their faces? Right?	
	416	S09: Yes, to see the expressions.	RECALL
	417	INT: Right, so you were seeing the expressions, and… you don't need to read anything?	
	418	S09: No.	
	419	INT: You just listened?	
	420	S09: Yes.	
	421	INT: Can you understand even just by listening?	
	422	S09: [inaudible]	

	423	INT: Ok.	
66	424	INT: Why do you think you fixated a lot on this ["plowed"]?	
	425	S09: Because… I didn't know what that ["plowed"] was.	RECALL
	426	INT: Ok, so you didn't know what the verb "plowed" means, right? So you were watching this.	
	427	S09: [inaudible]	
67	428	INT: Why were you watching the yogurt place?	
	429	S09: It was different and ehm… I looked at it.	
	430	INT: Right, so it stood out and it caught your attention.	
	431	S09: Yes.	
68	432	INT: So when Michael is presenting you look at "good vs bad", because these words come up? So, why do you think that happens?	
	433	S09: Because I was looking at the things in the screen, in the presentation.	
	434	INT: As if someone was giving a presentation, you were looking at the screen.	
	435	S09: Yes.	
69	436	INT: What happened when the yellow word "lived" came on screen?	
	437	S09: I saw the yellow color and I looked at it, cause it was… ehm, like... more… different.	RECALL
	438	INT: …salient? The word "salient" means it stands out, compared to the other ones.	
	439	S09: Yes, it stands out.	
	440	INT: Ok. Did you actually read this subtitle?	
	441	S09: Mhm… yes. Well, some words but…	
	442	INT: …not very carefully? Even despite seeing this, you didn't think the subtitle was for some reason more important?	
	443	S09: No.	
	444	INT: Right. So you just wondered why it was popping up. Why do you think that is?	
	445	S09: Because ehm... it says that "you've lived", and in here [on the screen] it puts the things you've done in your life.	
	446	INT: Right, so you think it's because of the meaning of this word that I highlighted it.	
	447	S09: Yes.	
70	448	INT: Right, so when it says "adorable little cottage"…	
	449	S09: Yes, I was looking at the windows.	RECALL

	450	INT: ...then you look at the house.	
	451	S09: …house, yes.	
71	452	INT: You were reading the subtitle, this one, "the interior has been decorated" a little bit, you looked at the subtitle a little bit and then when it became yellow, you looked somewhere else. Why do you think that is?	
	453	S09: Ehm, because it said that it was decorated and I looked at the decorations in the house.	RECALL
	454	INT: Right, ok, so whenever something… whenever a word catches your attention... and you understand the meaning, you look at whatever that word means…	
	455	S09: Yes.	
	456	INT: …in the video.	
72	457	INT: Why do you think I specifically highlighted these three letters over here?	
	458	S09: Because it's a past tense and it means that it has already been decorated and now, they are not decorating it now.	RECALL
	459	INT: Ok, so you think I was interested in knowing whether you know how to use the past tense? I mean…	
	460	S09: Ehm…	
	461	INT: You know the… you know, the difference between present and past…	
	462	S09: Yes.	
73	463	INT: So, every time you read a subtitle and then something highlights in yellow, even if you're watching somewhere else, you go back to the yellow one.	
	464	S09: Yes… to prove that what I was looking at, is the meaning of the word.	
	465	INT: Right, ok.	
	466	INT: What happens, though, to your listening? Are you still focused on listening?	
	467	S09: Yes.	
	468	INT: What do you think?	
	469	S09: Yes, I think so.	
	470	INT: Right, so you're mostly focused on listening, but you can read at the same time?	
	471	S09: Yes.	

	472	INT: That doesn't cause you any problem? You know...	
	473	S09: No.	
	474	INT: You're also looking at the image…	
	475	S09: Yes.	
	476	INT: But you're understanding everything. You never, kind of, lose…	
	477	S09: No, no.	
	478	INT: …control over this.	
74	479	INT: Why were you specifically watching here?	
	480	S09: Because of the name of the place.	
	481	INT: Because you don't know what it means or…?	
	482	S09: Yes, I know what it is, but, it's just… I don't know.	RECALL
	483	INT: Is it a bit weird that it's written this way but it sounds like Phoenix?	
	484	S09: No. No, it's not strange, but… ehm… because of the big letters, I look at it.	
	485	INT: Mhm, so you see a capital letter and you think, this is important, ok.	
	486	S09: Yes.	
75	487	INT: So you're mostly focusing on their faces… why is that?	
	488	**S10**: Because I'm most of the time watching… the guy that is talk, so that I maybe can describe their emotions and what did they think in that moment, or… what are they… well, yes, thinking about at the time that the other tells the other, what's happening. To describe their emotions.	RECALL
76	489	INT: So, could you understand everything without ever reading the subtitles?	
	490	S10: Yes.	
	491	INT: You were just listening?	
	492	S10: Yes.	
77	493	INT: Why do you think you were watching these specific areas at this point?	
	494	S10: Well, it is… kinda like in the beginning, they are colorful and big, like, sentences or signs… so I look at them, to see what the character [Michael] is talking about.	RECALL
	495	INT: Right, cause this is a presentation, right? So you're basically watching as if this was a slide.	
	496	S10: Yes.	
78	497	INT: So you always watch videos like this, you always try to look at the faces of the people speaking or…?	

	498	S10: Yes.	
	499	INT: Is that all you look at? Or things that pop up?	
	500	S10: Yes, it's like, if something cames [sic] into the screen, I usually watch, well, look at it so I can know what is it or if it's important to find out [unintelligible].	
79	501	INT: So here for example, let's look at this in particular. He's telling them that soulmates are real, and they will spend eternity together. Why do you think you're watching these two guys specifically?	
	502	S10: Well, because they are behind the girl, and the girl turned back and looked at them, so I thought they maybe could be her soulmates.	RECALL
	503	INT: Ok, so you were following her course of… of thinking so, maybe they are her soulmates.	
	504	S10: Yes.	
80	505	INT: Where are you watching and why?	
	506	S10: The house. Well, because has… colorful colors, like the other titles… like he was saying "this is your house, and this is because, well, this almost represents you", so I was watching the house to see how they represent the girl.	RECALL
81	507	INT: Do you normally like subtitles?	
	508	S10: No, because if I have subtitles I most of the time look at them, so I don't like them.	RECALL
	509	INT: Because I can see you're not looking at the subtitles.	
	510	S10: Because if I look at the subtitles, then I could miss something in the screen, so I don't like watching the subtitles.	
	511	INT: Right, and also, I think… How easy is this video for you? Can you understand everything?	
	512	S10: Yes, it's quite easy for me.	
	513	INT: So if they were speaking faster, do you think you might want to look at the subtitles or not?	
	514	S10: No.	
	515	INT: You are trying to avoid it, like, consciously.	
	516	S10: Yes.	
82	517	INT: There's a point, where the video system is presented… No, you were actually looking at the captions before that. And maybe just because it appears.	
	518	S10: Yes, it's like… the color…	

	519	INT: The movement.	
	520	INT: So what do you think about the yellow subtitles?	
	521	S10: Well, they are like, normal [unintelligible]	
	522	INT: Did you notice them?	
	523	S10: Well, I've read some, and it put "lived", the -ed was [underlined]… and then I thought well, they are going to put the past verbs in yellow, so I thought I'm not going to look at them, because I understand that.	RECALL
	524	INT: Ok, why do you think I highlighted specifically these yellow words in the past?	
	525	S10: I don't know.	
	526	INT: Ok, but why do you think I highlighted them like this? [*note*: also underlining the final -ed and the preceding letter]	
	527	S10: To know that these are regular verbs, maybe.	
	528	INT: Do you think because I highlighted some of the words and because I tested you on them, you were actually focusing a bit more on the verbs, or did that not have an effect at all?	
	529	S10: No.	
83	530	INT: What do you usually focus on when you listen to something or you watch a film?	
	531	S10: Well, I usually watch the person that is talking and I sometimes look the mouth, cause it's like I see the movement of the mouth, so I sometimes see them, or just the eyes, or the… where they are, so I watch the things that are around them.	
	532	INT: Right, is that because of an English learning effort, or because you would do it even in Spanish? Would you look at the mouth of the people in Spanish?	
	533	S10: No. Maybe sometimes but not too much time.	
	534	INT: Does it actually help you understand or…	
	535	S10: No.	
	536	INT: Ok, you just do it unconsciously.	
	537	S10: Yes.	
84	538	INT: So, for example here you were following the face of the protagonist, and then here you were kind of reading this. Why do you think you're reading these green words?	

	539	**S11**: Mhm, I thought that there was maybe something wrong with the place, because for me it was very weird this sentence... to be on the wall. I think there was something weird going on…	RECALL
	540	INT: Creepy.	
	541	S11: Yes.	
85	542	INT: Ok, so mainly you are watching their faces.	
	543	S11: Yes.	
	544	INT: Is that something you usually do? … Even like, when you're at home, I mean you're not doing this only because you're here, right? Or are you doing this… are you watching their faces mainly, without [reading the captions]…	
	545	S11: When I watch maybe movies… I tend to watch the mouth, but face to face in real life, I tend to watch the eyes.	RECALL
86	546	INT: Why do you think here you are reading the subtitles?	
	547	S11: Mhm… ah, yes because he was saying the name of maybe a beverage, and I wanted to know… and maybe with hearing I couldn't know exactly what beverage it was, because I didn't know which one it was, but I wanted to know exactly how it was spelled and how was it wrote [sic].	RECALL
	548	INT: Right, ok. Do you think it was difficult for you to understand each single word, like, to separate the words while you were listening?	
	549	S11: Mhm, I remember that I wanted to read it to exactly know what it was, but I think that without reading then I would have… maybe the name I would… I wouldn't remember the name after hearing it, but I think that in the moment I would have understood without... having to read.	
87	550	INT: So here also you go down [to the caption area], right? Why do you think that is?	
	551	S11: Maybe I was watching his, like, his clothes… or what he was wearing.	
	552	INT: So you weren't… because here it looks like you're actually reading… no. You're just looking at the clothes.	
	553	S11: Yes, or maybe I was reading "shopping carts".	
	554	INT: Right, because… why were you reading "shopping carts"?	
	555	S11: I don't remember.	RECALL
	556	INT: But you know what it means, right?	

	557	S11: Yes.	
88	558	INT: Why did you go back and forth between the place and this yellow word here?	
	559	S11: Because I… the image of the person began to slowly be smaller and I had two options: continue watching the person, seeing all these numbers and phrases, or read the subtitles. And I saw the yellow word and I wanted to see what it was.	
	560	INT: Ok, so it attracted your attention. And why do you think it was there? Why do you think it was yellow?	
	561	S11: I really don't know.	RECALL
89	562	INT: So here you also focused on this word in particular and you went back to it, you see?	
	563	S11: Yes.	
	564	INT: Why do you think that was, at the time?	
	565	S11: Maybe because the first image of the house was shown to me, and I began watching all the screen, all the decorations of the house and then the word "decorated" appeared, and was in yellow and, I don't know, maybe they connected, in some way.	RECALL
90	566	INT: So you understood everything, even just by listening... you understand...?	
	567	S11: I only read in the moments that you saw, the beverage… and also in the moment of all the data, of all the deaths…	
	568	INT: Yes.	
	569	S11: But only in there. I didn't have to read to understand.	
	570	INT: So you were trying actively not read, or is this something you always do?	
	571	S11: I normally don't read the… I try not to read the subtitles, but sometimes I end up reading them, because it's something I cannot help. But well, in this case I didn't have to read the subtitles to understand.	RECALL
	572	INT: Because it was easy.	
	573	S11: Yes.	
91	574	INT: So, you are mostly looking at their faces, or…?	
	575	**S12**: yes, at their faces and the… like, the…	
	576	INT: …the clothes…	
	577	S12: Yes, the clothes and t-shirts, and… I was thinking that, the boy…	
	578	INT: The man.	

	579	S12: The man, yes… I knew.	
	580	INT: Ok, so you were focusing on the actors now, but… were you actually listening? Like, were you able to listen to what they were saying?	
	581	S12: Yes, yes.	
	582	INT: Ok, and you understood everything?	
	583	S12: Yes.	
92	584	INT: So here you actually move down to the subtitles, why is that?	
	585	S12: Because I thought that at some time, there would be some… some words that I don't know or I listen like very fast or something [she may have meant "not very carefully"].	RECALL
	586	INT: So, because…they were speaking a little bit fast, right?	
	587	S12: No, because… I know everything that it's going on and his face and the character, so… I thought that.	
	588	INT: Ok, so the subtitles were more interesting because you knew about the characters already.	
	589	S12: Yes.	
	590	INT: Right, do you always do this? Do you always look at the faces mostly?	
	591	S12: Yes, but when I… when I see a movie or something, I don't put subtitles, because if I put subtitles I'm like distracting with the subtitles.	RECALL
	592	INT: Right, so the subtitles are actually distracting you from looking at the action, which is what you're interested in.	
	593	S12: Yes, but they're not doing anything, so…	
	594	INT: Because you're listening.	
	595	S12: Yes.	
	596	INT: And you can hear, you can understand everything.	
	597	S12: [no audible answer]	
93	598	INT: So… did you see that? This subtitle comes up, and… [your gaze moves] down to "lived".	
	599	S12: Yes, because after [she meant before the video] it was a phrase that say that there was… yellow…	
	600	INT: Right.	
	601	S12: …yellow words, so I… I wanted to know if they're different or something.	
	602	INT: And did you find out? Did you find out why they were yellow?	
	603	S12: Yes, because they're in the past, like with the -ed.	RECALL

	604	INT: What do you think I'm interested in?	
	605	S12: Ehm… that I focus on the verbs, or I don't know.	
	606	INT: Ok, on the verbs.	
	607	INT: Why do you think specifically the -ed [morpheme] is underlined? [*note*: on top of being yellow, like the rest of the letters in the word]	
	608	S12: Because it's the… the… I don't know how to say this.	
	609	INT: Suffix.	
	610	S12: Yes, like, the thing that you add… to make that it's in the past.	RECALL
	611	INT: …to a regular verb.	
	612	S12: Yes.	
94	613	INT: Do you think reading this distracted you from listening? Or do you think you could actually hear all of this, you could listen to everything?	
	614	S12: I think [I could listen] to everything. Because there was… a lot of numbers and things that I don't have to… to see.	
	615	INT: So you were not specifically watching the image, so you could actually read the subtitles.	
	616	S12: [no audible answer]	
95	617	INT: So here again you are looking at this, right? And then you look again at Eleanor… You were watching the decorations…	
	618	S12: Yes.	
	619	INT: ...then you read "decorated" [target word in yellow] and then you look at Eleanor. What happened here?	
	620	S12: Ehm... so the man was talking about that the house, it was maded [sic] for her, like, all the things that she loved, so I was looking… what she liked. And the decorations and these things.	RECALL
	621	INT: Right, so you were interested in what the meaning of that sentence was, in the [visual] context…	
	622	S12: Yes.	

Appendix C.2: Stimulated recall coding rubric (study 2)

Category	Description	Justification	Levels	Examples
Noticing of word or phrase	A focus on a specific word or phrase is reported	*Noticed intake*, moving to the intake processing stage (Leow, 2015)	-	"The part "in a broken down boat", I didn't understand what he was trying to explain, so I was reading so that maybe I could understand it better."
Language aspect noticed	If a word or phrase was noticed, which aspect specifically		Auditory word form	"I really understood almost everything that they said but, for example, some words I can't hear them properly so I read them to know if I can write them."
			Morphology	"The next words that are in yellow, I think that they all finish in -ed. And that's because [why]. I think."
			Vocabulary	"This sentence, I didn't know what he was saying so I tried to read it because I don't know what it means."
			Grammar	"[The word was highlighted] because it's a past tense and it means that it has already been decorated and now, they are not decorating it now."
General comprehension	The participant reports watching or reading captions to understand the general meaning	No reported attention to linguistic form	-	"I was trying to connect the dots… cause he was speaking a little bit fast and, like, I wanted to take all the data because they are going to ask me later [in the comprehension questions]."
Audiovisual processing	The participant talks about the auditory input in	Simultaneous attention to auditory and visual information	-	"I read the subtitles because I didn't understand what he was saying, and I wanted to have a general idea of what he was saying"

	connection with visual (verbal only) input			
Mentions test	The participant mentions the pre- or post-test	Attention as *orientation* to linguistic form (Tomlin & Villa, 1994)	-	"When I watched the first recording, there were some [comprehension] questions […] So I started focusing on the words, so that I could remember them better."
Captions attract gaze	The appearance of the caption or of the enhancement attracted the participant's gaze (generally against their will)	Low-level processing (*attended* / *detected intake* in Leow, 2015)	-	"Well, I have seen that the subtitle had a yellow word and I was surprised. At the beginning of the recording it said that this could happen, but it had never happened before so my eyes focused on the word in that moment."
Face/mouth attracts gaze	The participant fixates on the faces of the characters	Focus on image for general context, listening for general comprehension (no attention to form)	-	"I'm most of the time watching… the guy that is talk, so that I maybe can describe their emotions and what did they think in that moment, or… what are they… thinking about at the time […] To describe their emotions."
Does not know	The participant does not know why they were doing something back at the time of the viewing	Occasionally responding "I don't know" indicates that the participant responded truthfully	-	"I don't really know, I always see the expressions of people who are talking, because… it's something I always do […] I don't really know."
Explicitness of prompt	Whether the participant's comment follows an explicit prompt by the interviewer or a more general question	To control for a possible confound	Explicit	"[Did you read this caption] to make sure that the word was the word that you were hearing?"
			Non-explicit	"So here… what were you watching?"

Appendix D: Additional items introduced in study 3 to expand upon the questionnaire used in study 2

El presente estudio: las actividades de aprendizaje de idiomas

Piensa en las actividades de comprensión y producción oral, como identificar las palabras pronunciadas por los actores o practicar un diálogo en voz alta.

Indica hasta qué punto estás de acuerdo con cada afirmación.

	Totalmente en desacuerdo	Un poco en desacuerdo	Indeciso	Un poco de acuerdo	Totalmente de acuerdo
Entendí las instrucciones	O	O	O	O	O
Las actividades fueron divertidas	O	O	O	O	O
Las actividades fueron un reto	O	O	O	O	O
Utilizamos las pistas para realizar las actividades	O	O	O	O	O
Mi pareja y yo contribuimos por igual a las actividades	O	O	O	O	O
Aprendí algo de pronunciación en inglés gracias a las actividades	O	O	O	O	O
Aprendí algo de gramática o vocabulario en inglés gracias a las actividades	O	O	O	O	O

Appendix E: Accuracy gains for each participant in study 3, aggregated by task and testing time and presented in ascending order

ID	Test	Gains	Group	Test	Gains
C_73	Narrative task	-0.40	C_80	Delayed sentence repetition	0.05
B_32	Narrative task	-0.36	A_21	Delayed sentence repetition	0.05
A_15	Narrative task	-0.33	A_22	Delayed sentence repetition	0.05
B_53	Read-aloud task	-0.27	B_31	Delayed sentence repetition	0.05
A_13	Delayed sentence repetition	-0.26	A_11	Delayed sentence repetition	0.05
C_77	Delayed sentence repetition	-0.26	B_38	Delayed sentence repetition	0.05
A_4	Narrative task	-0.22	C_59	Delayed sentence repetition	0.05
C_78	Narrative task	-0.21	C_81	Delayed sentence repetition	0.05
A_18	Read-aloud task	-0.20	A_19	Read-aloud task	0.07
B_42	Read-aloud task	-0.20	A_22	Read-aloud task	0.07
C_80	Narrative task	-0.18	B_46	Read-aloud task	0.07
A_23	Delayed sentence repetition	-0.16	B_50	Read-aloud task	0.07
A_20	Delayed sentence repetition	-0.16	C_61	Read-aloud task	0.07
C_81	Read-aloud task	-0.13	C_66	Read-aloud task	0.07
B_47	Read-aloud task	-0.13	C_70	Read-aloud task	0.07
C_55	Read-aloud task	-0.13	A_23	Narrative task	0.07
C_77	Read-aloud task	-0.13	C_68	Narrative task	0.07
A_18	Narrative task	-0.13	A_12	Narrative task	0.08
C_77	Narrative task	-0.13	A_29	Narrative task	0.08
C_65	Narrative task	-0.12	C_76	Narrative task	0.10
A_2	Delayed sentence repetition	-0.11	C_66	Delayed sentence repetition	0.11
A_7	Delayed sentence repetition	-0.11	C_75	Delayed sentence repetition	0.11
C_70	Delayed sentence repetition	-0.11	B_37	Delayed sentence repetition	0.11
C_73	Delayed sentence repetition	-0.11	A_7	Narrative task	0.12
A_20	Narrative task	-0.09	C_61	Narrative task	0.13
C_75	Narrative task	-0.07	B_38	Read-aloud task	0.13
A_4	Read-aloud task	-0.07	C_57	Read-aloud task	0.13
A_11	Read-aloud task	-0.07	B_41	Read-aloud task	0.13
A_23	Read-aloud task	-0.07	A_5	Read-aloud task	0.13
B_31	Read-aloud task	-0.07	B_44	Read-aloud task	0.13
B_35	Read-aloud task	-0.07	A_2	Narrative task	0.14
B_37	Read-aloud task	-0.07	B_38	Narrative task	0.14
C_64	Read-aloud task	-0.07	C_81	Narrative task	0.14
C_65	Read-aloud task	-0.07	B_52	Delayed sentence repetition	0.16
C_68	Read-aloud task	-0.07	B_47	Delayed sentence repetition	0.16
C_75	Read-aloud task	-0.07	C_68	Delayed sentence repetition	0.16
C_78	Read-aloud task	-0.07	B_29	Delayed sentence repetition	0.16
C_80	Read-aloud task	-0.07	B_41	Delayed sentence repetition	0.16
A_27	Narrative task	-0.06	B_50	Delayed sentence repetition	0.16

A_19	Delayed sentence repetition	-0.05	C_64	Delayed sentence repetition	0.16
A_24	Delayed sentence repetition	-0.05	A_3	Narrative task	0.16
C_76	Delayed sentence repetition	-0.05	A_21	Narrative task	0.17
C_78	Delayed sentence repetition	-0.05	B_49	Read-aloud task	0.20
C_61	Delayed sentence repetition	-0.05	B_46	Narrative task	0.20
C_72	Delayed sentence repetition	-0.05	A_27	Read-aloud task	0.20
A_13	Narrative task	-0.04	C_73	Read-aloud task	0.20
C_71	Narrative task	-0.04	A_5	Delayed sentence repetition	0.21
B_49	Narrative task	-0.03	B_32	Delayed sentence repetition	0.21
C_59	Narrative task	-0.03	B_49	Delayed sentence repetition	0.21
B_41	Narrative task	-0.02	A_27	Delayed sentence repetition	0.21
A_3	Read-aloud task	0.00	B_35	Narrative task	0.21
A_7	Read-aloud task	0.00	C_55	Narrative task	0.23
A_12	Read-aloud task	0.00	B_53	Narrative task	0.23
A_20	Read-aloud task	0.00	C_70	Narrative task	0.23
A_21	Read-aloud task	0.00	B_51	Narrative task	0.25
A_24	Read-aloud task	0.00	C_64	Narrative task	0.26
B_29	Read-aloud task	0.00	C_57	Narrative task	0.26
B_32	Read-aloud task	0.00	B_42	Delayed sentence repetition	0.26
B_34	Read-aloud task	0.00	B_44	Delayed sentence repetition	0.26
B_51	Read-aloud task	0.00	B_51	Delayed sentence repetition	0.26
C_59	Read-aloud task	0.00	C_57	Delayed sentence repetition	0.26
C_71	Read-aloud task	0.00	A_15	Read-aloud task	0.27
C_72	Read-aloud task	0.00	A_2	Read-aloud task	0.27
C_76	Read-aloud task	0.00	A_10	Read-aloud task	0.27
A_15	Delayed sentence repetition	0.00	A_13	Read-aloud task	0.27
B_46	Delayed sentence repetition	0.00	A_11	Narrative task	0.29
B_53	Delayed sentence repetition	0.00	B_35	Delayed sentence repetition	0.32
C_71	Delayed sentence repetition	0.00	A_18	Delayed sentence repetition	0.37
A_19	Narrative task	0.00	A_10	Delayed sentence repetition	0.37
B_34	Narrative task	0.00	A_3	Delayed sentence repetition	0.37
B_37	Narrative task	0.00	C_55	Delayed sentence repetition	0.37
B_42	Narrative task	0.00	A_5	Narrative task	0.38
B_47	Narrative task	0.00	C_66	Narrative task	0.38
B_52	Narrative task	0.00	A_22	Narrative task	0.40
C_72	Narrative task	0.00	C_65	Delayed sentence repetition	0.42
A_24	Narrative task	0.01	A_10	Narrative task	0.44
B_44	Narrative task	0.02	B_52	Read-aloud task	0.47
A_4	Delayed sentence repetition	0.05	B_50	Narrative task	0.73
A_12	Delayed sentence repetition	0.05	B_31	Narrative task	0.75
B_34	Delayed sentence repetition	0.05			